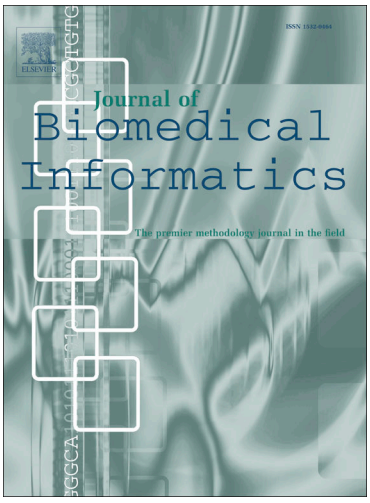


Differential splicing analysis based on isoforms expression with NBSplice

Gabriela Alejandra Merino, Elmer Andrés Fernández

PII: S1532-0464(20)30005-8
DOI: https://doi.org/10.1016/j.jbi.2020.103378
Reference: YJBIN 103378

To appear in: *Journal of Biomedical Informatics*

Received Date: 14 June 2019
Revised Date: 7 December 2019
Accepted Date: 13 January 2020

Please cite this article as: Alejandra Merino, G., Andrés Fernández, E., Differential splicing analysis based on isoforms expression with NBSplice, *Journal of Biomedical Informatics* (2020), doi: https://doi.org/10.1016/j.jbi.2020.103378

Differential splicing analysis based on isoforms expression with NBSplice

Gabriela Alejandra Merino[a,b], Elmer Andrés Fernández[b,c,*]

[a] *Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática (IBB), Universidad Nacional de Entre Ríos, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ruta 11 Km 10.5, E3100XAD, Oro Verde, Argentina*

[b] *Centro de Investigación y Desarrollo en Inmunología y Enfermedades Infecciosas (CIDIE), Universidad Católica de Córdoba, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Av. Armada Argentina 3555, X5016DHK, Córdoba, Argentina*

[c] *Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Av. Vélez Sarsfield 1611, X5016GCA, Córdoba, Argentina*

Abstract

Alternative splicing alterations have been widely related to several human diseases revealing the importance of their study for the success of translational medicine. Differential splicing (DS) occurrence has been mainly analyzed through exon-based approaches over RNA-seq data. Although these strategies allow identifying differentially spliced genes, they ignore the identity of the affected gene isoforms which is crucial to understand the underlying pathological processes behind alternative splicing changes. Moreover, despite several isoform quantification tools for RNA-seq data have been recently developed, DS tools have not taken advantage of them.

Here, the NBSplice R package for differential splicing analysis by means of isoform expression data is presented. It estimates differences on rela-

*Corresponding author.
E-mails: gmerino@ingenieria.uner.edu.ar (G.A.Merino), efernandez@cidie.ucc.edu.ar (E.A.Fernández)

tive expressions of gene transcripts between experimental conditions to infer changes in gene alternative splicing patterns. The developed tool was evaluated using a synthetic RNA-seq dataset with controlled differential splicing. NBSplice accurately predicted DS occurrence, outperforming current methods in terms of accuracy, sensitivity, F-score, and false discovery rate control. The usefulness of our development was demonstrated by the analysis of a real cancer dataset, revealing new differentially spliced genes that could be studied pursuing new colorectal cancer biomarkers discovery.

Keywords: Alternative splicing, RNA-seq, Transcriptomics, Gene isoforms, Cancer.

1. Introduction

Alternative splicing (AS) is a post-transcriptional mechanism of higher eukaryotes responsible for transcriptome complexity and functional diversity. During this process, specific regions of a gene can be included or excluded from messenger RNA (mRNA), leading to different transcript isoforms (mRNA variants) [1]. Although changes in AS patterns occur under normal conditions, they have also been related to several diseases and have been especially associated to cancer progression and metastasis [2, 3]. Thus, the study of the AS dynamic is a key issue for translational cancer medicine in order to understand how genes are regulated, for instance, during cancer development and/or progression.

RNA sequencing (RNA-seq) is the most widely used technique to analyze transcriptome expression dynamics, including AS and its changes [4]. When

analyzing modifications at the transcript level, two kinds of expression alterations can be inquired: differential expression of transcripts, and differential transcript usage (DTU), i.e. differences in the relative expression of a transcript between conditions [5]. In particular, DTU could involve potential functional consequences leading to a substantial biological impact which has been found especially prominent in cancer [6]. In spite of this, the complexity of the quantification process of gene transcripts, highly overlapped, has led to the development of tools for DS analysis based on differential exon usage (DEU) [7, 8]. Although this approach allows the discovery of alternatively spliced genes (ASG), it does not provide the identity of the isoforms of ASG, which is crucial, for instance, to identify new therapeutic targets and/or biomarkers of disease progression [3].

In the last years, several methods for transcript quantification [9] and for DTU [5, 10, 11] have been developed. However, despite the multiple improvements they have achieved, many challenges remain unsolved. As an example, although relative abundances of transcript isoforms are frequently described in terms of the percentage of spliced-in (PSI), DS results are not always given in terms of the difference of PSI between conditions [5]. In addition, DS results provided by some DTU tools are hard to interpret since they have been obtained by indirect tests, instead of evaluating the significance of the ΔPSI, or reporting only fold changes at the gene level [10].

Here, we present NBSplice which is a new method for DS analysis to overcome the issues described above. It allows fitting a negative binomial generalized linear model (GLM, 12) for each gene in order to estimate the mean relative expressions of gene transcripts. Our tool provides methods

for performing hypothesis tests to evaluate both DTU and DS, and for identifying differentially spliced genes (DSGs) between experimental conditions. NBSplice was evaluated and compared against three R packages for DS analysis using a synthetic RNA-seq database with DS control. The utility of our tool was demonstrated by analyzing a real cancer dataset from The Cancer Genome Atlas (TCGA) project where new and previously reported DSGs were identified. NBSplice is freely available on the Bioconductor site.

2. Materials and Methods

2.1. Model

NBSplice is based on GLMs, a widely used technique in transcriptomic data analysis[13]. Expression counts of the i-th transcript from the j-th gene in the k-th sample, y_{ijk}, are assumed as realizations of a negative binomial (NB) distribution with mean μ_{ijk} and dispersion ϕ_j:

$$y_{ijk} \sim NB(\mu_{ij} = p_{ijk}\mu_{jk},\ \phi_j) \tag{1}$$

Particularly, μ_{ijk} is assumed to be the product of the isoform relative expression, p_{ijk}, and the mean of the total counts from the j-th gen, μ_{jk}. The distributional parameters are unknown and need to be estimated. To account for library size differences, isoform and gene counts are considered in counts per million (CPM) scale. The μ_{jk} parameter is computed from the observed gene counts in each sample. Then, a log-linear GLM is fitted to estimate μ_{ijk},

$$ln(\mu_{ijk}) = \boldsymbol{x}_{ijk}^T \boldsymbol{\beta}_{ij} + ln(\mu_{jk}) \tag{2}$$

where $\boldsymbol{x}_{ijk}^T$ is the design matrix and $\boldsymbol{\beta}_{ij}$ is the vector of unknown coefficients. The columns of the desing matrix correspond to experimental factors applied to samples k and affecting the i-th transcript isoform.Thus, $\boldsymbol{\beta}_{ij}$ can be interpreted as the vector of fold changes (in a logarithmic scale) of relative expression of isoform i from the j-th gene for each column of the design matrix. The $ln(\mu_{jk})$ is an offset term representing the logarithmic-mean expression of gene j in sample k.

Considering a single factor imposing different experimental conditions, the proposed model decomposes the linear predictor, $\boldsymbol{x}_{ijk}^T\boldsymbol{\beta}_{ij}$, into four components depicted as:

$$ln(p_{ijk}) = \mu_0 + \alpha_{ij} + \delta_{jr(k)} + \gamma_{ijr(k)} \tag{3}$$

In Eq. 3, μ_0 represents the overall mean isoform relative expression, in logarithmic scale, whereas α_{ij} is the change of the expected relative expression to the i-th isoform from the j-th gene. $\delta_{jr(k)}$ represents the fold change in the mean relative expression of isoforms from gene j under condition r, for sample k. The interaction term, $\gamma_{ijr(k)}$, is the effect of condition r on the relative expression of isoform i from the j-th gene.

2.2. Differential splicing detection

In NBSPlice, differences on the relative expression of the i-th isoform between two experimental conditions, i.e differential transcript/isoform usage, are evaluated through a linear hypothesis test analyzing the interaction term. If the number of analyzed samples is large, the χ^2 distribution can be assumed for the test statistic. However, the use of the F distribution is strongly recommended given that the number of samples is generally small.

The results of all gene isoforms tests are then combined to evaluate differential splicing at the gene level using the Simes test [14]. It defines a new gene-based statistic associated with a global null hypothesis defined by the family of null hypotheses related to the isoforms of the gene. Both, isoform and gene, p-values are corrected with the Benjamini-Hochberg method [15].

2.3. NBSplice implementation

NBSplice is available as a Bioconductor R package [16], and it is based on several R packages commonly used for GLMs fitting and hypothesis testing. For instance, it uses the glm.nb function of the MASS R package [17] to estimate model coefficients and dispersion parameters by means of the maximum likelihood method. In addition, the car [18] and mppa [19] packages are used for linear hypothesis and Simes tests, respectively.

Since filtering low-expressed isoforms helps to reduce false positive detections [10, 20, 21], NBSplice considers that an isoform is low-expressed if: **i)** It has a mean absolute expression, in at least one condition, lower than a count threshold (cT); or **ii)** Its relative expression, in at least one sample, is lower than a ratio threshold (rT), where both cT and rT are user-defined parameters. The low-expressed isoforms detected by NBSplice are tagged as unreliable and they are ignored during models fitting. Even though, they are considered to compute the total gene counts, avoiding over-estimation of relative expression of the isoforms detected as reliable.

As input, count expression matrices of any transcript quantification tool such as RSEM [22] can be used. Estimated expression counts from novelty free alignment tools such as Kallisto [9] has also been previously considered as input for DTU methods assuming NB distributions [20, 23, 24, 25]. Thus,

Kallisto counts can also be used by NBSplice. Independently of the software used for transcripts quantification, NBSplice will automatically round (if it is necessary) and transform them to CPM before model fitting and DS analysis. The outcome of the developed tool includes the mean relative expressions of gene isoforms estimated for each condition, and the DTU results and DS analyses. As a complement, the tool also incorporates several useful methods for results exploration.

The NBSplice functionality is illustrated here showing the analysis of a small example, extracted from the package's vignette [16]. After loading NBSplice, the isoform expression, the gene isoform relationship and the design matrices are loaded, followed by the specification of the column name of the design matrix to be contrasted. These data are stored in an `IsoDataSet` object that is then analyzed to detect the low-expressed isoforms. Following this, the NB models and hypothesis tests are performed, and the DS results are finally extracted to be explored.

```
# Loading the package
> library(NBSplice)
# Data loading
> data(isoCounts, package ="NBSplice")
> head(isoCounts)
#                 C1R1 C1R2 C1R3 C1R4 C2R1 C2R2 C2R3 C2R4
# ENST00000228345   79    0  106   99   76    0   68   23
# ENST00000358495  567  212  162  715   70  176  255  243
# ENST00000397230   15    0    0   28    0    0   22    3
# ENST00000430095    0    6    0   77    0    0   27   59
# ENST00000461568   11    0    0    0    0    2    0    0
```

```
# ENST00000463750   10   31  126   60   88    0    0   31
> data(geneIso, package ="NBSplice")
> head(geneIso)
#                         gene_id      isoform_id
# ENST00000228345 ENSG00000002016 ENST00000228345
# ENST00000358495 ENSG00000002016 ENST00000358495
# ENST00000397230 ENSG00000002016 ENST00000397230
# ENST00000430095 ENSG00000002016 ENST00000430095
# ENST00000461568 ENSG00000002016 ENST00000461568
# ENST00000463750 ENSG00000002016 ENST00000463750
> data(designMatrix, package ="NBSplice")
> head(designMatrix)
#     sample condition
# C1R1   C1R1    Normal
# C1R2   C1R2    Normal
# C1R3   C1R3    Normal
# C1R4   C1R4    Normal
# C2R1   C2R1     Tumor
# C2R2   C2R2     Tumor
> colName <- "condition"
# Building an IsoDataSet object
> myIDS <- IsoDataSet(isoCounts, designMatrix, colName,
                      geneIso)
# Identifying low-expresed isoforms
> myIDS <- buildLowExpIdx(myIDS)
# Model fitting and differential splicing testing
> myNBSpliceRes <- NBTest(myIDS, colName, test = "F")
```

```
# Extracting the table of those cases with evidence
# of differential splicing
> myDSRes <- GetDSResults(myNBSpliceRes)
# DTU and DS detection are based on the FDR and geneFDR
# columns, respectively
> head(myDSRes, n=3)
#                 iso            gene ratio_Normal ratio_Tumor
# 378 ENST00000206514 ENSG00000092068   0.03696968  0.07501606
# 379 ENST00000316902 ENSG00000092068   0.51890400  0.17702007
# 381 ENST00000339733 ENSG00000092068   0.03601217  0.05323293
#          odd      stat         pval   genePval         FDR
# 378  0.7076040  6.065408 0.0170055687 0.00281695 0.128238715
# 379 -1.0754558 14.340033 0.0003851829 0.00281695 0.009577521
# 381  0.3908204  1.841038 0.1804745117 0.00281695 0.566677648
#        geneFDR
# 378 0.02540778
# 379 0.02540778
# 381 0.02540778
# Extracting the differentially spliced genes' names
> myDSGenes <- GetDSGenes(myNBSpliceRes)
> head(myDSGenes, n=3)
# [1] "ENSG00000092068" "ENSG00000103275"  "ENSG00000103942"
```

2.4. Data simulation and processing

The characterization and performance evaluation of NBSplice was carried out using a synthetic RNA-seq experiment with controlled DS patterns. This database was simulated based on the RSEM simulation tool [22] and

following the same approach as in [4]. A real human prostate cancer RNA-seq dataset (GSE22260; 26) was used as a reference for the simulation in order to mimic a realistic sequencing process. As in 4, eight samples, four for control (C) and four for tumor (T) conditions were simulated. Ten realizations of the control-tumor experiment were performed. Each realization considered around 110,000 isoforms belonging to 16,100 genes, 10% of which were simulated with different levels of DS.

Since DS and differential gene expression can occur simultaneously, two groups of DS events were defined. Moreover, as it was previously demonstrated in [4], the DS detection could be influenced by the mangnitude of the expression change. For this reason, several subgroups characterizing different change levels in those two DS groups were defined (see Table 1). The first group was called **DS**, and it involved genes where the overall gene expression remained unchanged, whereas isoform proportions were modified between T and C conditions. The second group was named **DIEDS**, and it was related to genes where both absolute and relative expression of gene transcripts were modified. As in [4], the subgroups were designed to mainly control the change in the relative expression on the major (M) isoform in both conditions. As an example, the subgroup *DS-0.3-0.5* involves genes with DS without modification in their absolute expression across condition, and with the relative expression of their M isoforms simulated as 0.3 in condition C, and as 0.5 in condition T. On the other hand, the subgroup *DIEDS-2-0.3-0.5* involves genes with changes in both absolute isoforms expression and DS. In particular, the absolute expression of their isoforms in condition T was simulated as twice (2 times) of their absolute expression in condition C. Whereas, the

relative expression of their M isoforms were 0.3 and 0.5 in condition C and T, respectively.

Table 1: **Simulated cases with differential splicing**. **DS** groups imply differential splicing (DS) without change in overall gene expression between tumor (T) and control (C) conditions; **DIEDS** groups refer to changes in both DS and absolute isoform expression. Fold changes of the absolute expression, affecting equally all gene isoforms, and the relative expression of the major gene isoforms in both C and T conditions are listed.

Group	*Absolute expression change**	*Relative expression in T-C conditions ***	*Subgroup*
DS	No change	0.3-0.5	DS-0.3-0.5
	No change	0.3-0.7	DS-0.3-0.7
	No change	0.5-0.7	DS-0.5-0.7
	No change	0.5-0.9	DS-0.5-0.9
DIEDS	0.5	0.5-0.7	DIEDS-0.5-0.5-0.7
	0.5	0.5-0.9	DIEDS-0.5-0.5-0.9
	2	0.3-0.5	DIEDS-2-0.3-0.5
	2	0.5-0.3	DIEDS-2-0.5-0.3
	2	0.3-0.7	DIEDS-2-0.3-0.7
	2	0.7-0.3	DIEDS-2-0.7-0.3
	2	0.5-0.7	DIEDS-2-0.5-0.7
	2	0.5-0.9	DIEDS-2-0.5-0.9
	4	0.5-0.7	DIEDS-4-0.5-0.7
	4	0.7-0.5	DIEDS-4-0.7-0.5
	4	0.5-0.9	DIEDS-4-0.5-0.9
	4	0.9-0.5	DIEDS-4-0.9-0.5

*Change in the absolute expression of all gene isoforms in condition T regarding to their expression in condition C; **Relative expression of the major isoform in C-T conditions.

Once synthetic expression matrices were generated, RSEM was used to simulate the RNA-seq reads for each experiment replication. After that, isoform expression quantification for each sample of the ten experiment replications were obtained using Kallisto, a free-alignment quantification tool [9]. Since Kallisto generates estimated transcript counts, with decimal precision,

its resulting counts were then transformed to CPM and rounded before be used by NBSplice. Finally, the isoform expression matrix for each experiment realization was built and processed following the steps described in the NBSplice vignette.

2.5. TCGA cancer dataset

Isoform expression matrix from a subset of subjects of the colorectal adenocarcinoma (COAD) database of the TCGA project was used as a real application example. Particularly, twenty-eight non-matched samples randomly selected, representing tumor and normal cases, were considered. The expression matrix of gene transcripts obtained using Kallisto and generated by [27] was used. Metadata information about the selected samples is listed in Supplementary File S1.

After rounding Kallisto counts and converting it to the CPM scale, the mean-variance relationship of reliable isoforms was explored. The obtained results, shown in Figure 1, revealed that most of the isoform counts do not follow the mean-variance linear relationship expected for non-overdispersed data (i.e Poisson distribution). Thus, CPM rounded Kallisto counts can be assumed as negative-binomially distributed. In addition, goodness of fit tests to NB distributions were performed for those isoforms. Significant results (Bonferroni adjusted p-value < 0.05) were found for less than 20% of the isoforms, supporting the assumption of NB distribution for most of the data.

2.6. NBSplice evaluation

The configuration of the user-defined cT and rT thresholds, required for low-expressed isoforms detection, could impact on the efficiency of DS analy-

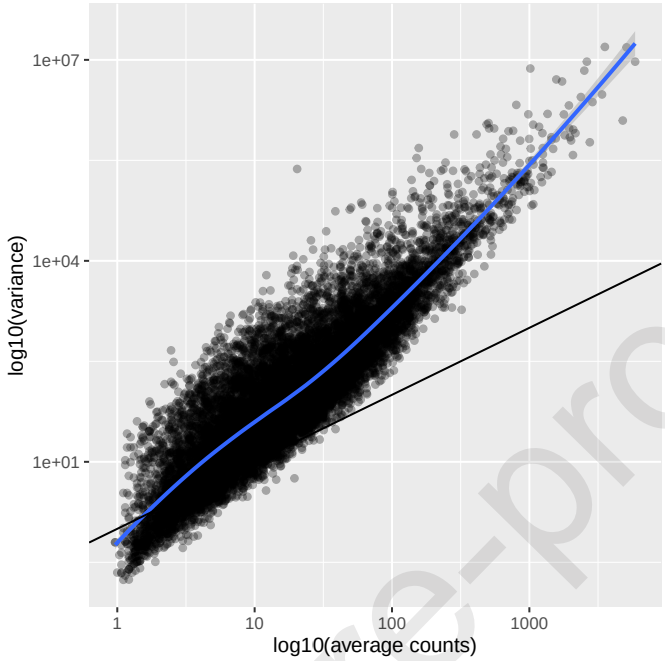


Figure 1: **Mean-variance relationship for Kallisto rounded counts.** Dispersion plot of mean and variance of Kallisto expression counts in the counts per million (CPM) scale and after rounding. Each dot represents an isoform that has been identified as reliable by NBSplice. The mean and variance were calculated across all replicates of each condition. Since NBSplice assumes equal-variance for all isoforms from the same gene, the variance was computed at the gene level. The black line represents the Poisson relation, where variance and mean are equals, and the blue curve shows the smoothed data (in logarithmic space) using local polynomial regression fitting.

sis tools. Thus, in order to choose their optimum values that lead to the best NBSplice performance, the nine configurations of these parameters listed in Table 2 were evaluated. In particular, the first configuration is called without filtering (WF) since it uses both cT and rT equals to zero, whereas the other setups involve a non-zero value for at least one of the two thresholds.

The NBSplice performance on DS detection, through the ten experiment replications, was evaluated based on commonly-used performance measures for classification tasks as in 4. Taking into account the gene status, defined during the simulation step, NBSplice results were used to classify each gene as: True positive (TP), false positive (FP), true negative (TN) or false negative (FN). Then, the number of: Analyzed genes after low-expressed isoforms detection, differentially spliced genes (DSGs) detected by NBSplice, TPs and FPs measures were obtained. After that, accuracy, sensitivity (TPR or recall), precision, false discovery rate (FDR) and F-score were computed. The NBSplice ability to deal with FPs was also assessed by using ten null experiment replications which were obtained by random permutation of two replicates, per condition, of the simulated dataset in order to confound the effect of the experimental condition.

The state-of-art DS tools DEXSeq [28] and DRIMSeq [10], and a recently developed tool for DTU, RATs [11], were selected for comparison purposes. Briefly, DEXSeq considers each gene as a set of features (originally, the gene exons), and assumes a NB distribution for the feature counts. Then, it fits GLMs, one for each feature, including an interaction term to relate the feature counts with the total counts for all the other features of the same gene [28]. Although DEXSeq was originally designed for DEU, it can be used as a DTU tool considering each gene isoform as a gene feature [20, 21]. DRIMSeq assumes a Dirichlet Multinomial model for each gene. In this model total gene counts are considered fixed, and the interest is focused on the proportion of its transcripts, for each sample. Thus, if the proportion of one transcript increases, it must result in a decrease in the proportions of the other gene

Table 2: **NBSplice configurations**. Nine NBSplice setups defined in terms of the values used for its cT and rT thresholds were evaluated. Unreliable isoforms are defined based on these thresholds. Particularly, cT refers to the minimum value admitted for the mean of the absolute isoform expression per condition, and rT refers to the minimum value admitted for the relative expression of isoforms per sample.

Threshold for mean absolute expression per condition (cT)	**Threshold of minimum relative expression per sample (rT)**	**Denomination**
No threshold	No threshold	*WF*
No threshold	0.01	*rT=0.01*
1	No threshold	*cT=1*
1	0.01	*rT=0.01; cT=1*
2	0.01	*rT=0.01; cT=2*
1	0.05	*rT=0.05; cT=1*
2	0.05	*rT=0.05; cT=2*
1	0.1	*rT=0.1; cT=1*
2	0.1	*rT=0.1; cT=2*

transcripts. The variation in proportions seen within condition are considered and modelled by means of a single precision parameter per gene. Thus, genes detected as having statistically significant DTU are those in which the proportion changes observed across condition are large [10]. On the other hand, RATs identifies DTU at both the gene and the transcript levels using the independence G-test for detecting significant differences above a user-defined threshold. Then, it evaluates the reproducibility of its DTU calls by means of bootstrapping the abundance estimations from free alignment tools. Finally, the DTU state reported by RATs is based on the combination of an FDR threshold at both the gene and transcript level, an effect size threshold and a reproducibility threshold [11]. DEXSeq, DRIMSeq, RATs, and NBSplice were used to analyze both, the simulated and the COAD databases.

Since the use of DS tools requires a previous step for unreliable transcript detection, different strategies were considered. On the one hand, the RATs tool has implemented this step as part of the DTU call process, so its own filtering criteria was used. On the other hand, the optimum cT and rT thresholds, and the filtering strategy proposed by [21] named as *SwimmF* were employed for DEXSeq, DRIMSeq and NBSplice. Thus, the different DS tools using these filtering strategies, through the simulated database, were compared.

DS results for the COAD dataset were contrasted in order to detect both, the overlapping in the lists of DSGs of each tool and the advantages of the NBSplice use. Particularly, NBSplice was used considering its optimum cT and rT thresholds, whereas DRIMSeq and DEXSeq were used combined with the *SwimmF* filtering strategy as suggested [21]. Meanwhile, expression counts were analyzed by RATs using the optimal configuration for its parameters [11]. DSGs only detected by NBSplice were also explored and validated by bibliography search looking for recent publications where those genes have been related to CRC .

The Wilcoxon test (wilcox.test R method), considering a significance threshold of 0.05, was used to compare two samples distributions. More details on the synthetic database generation, scripts used for DS analyses and the comparison strategy can be found in the GitHub repository gamerino/NBSpliceSuppInformation.

3. Results and Discussion

3.1. NBSplice evaluation

3.1.1. Overall results

Overall NBSplice results of DS analysis over the synthetic data with the nine parameters setups (see Table 2) are summarized in Table 3. As it can be expected, as more restrictive are the cT and rT thresholds, fewer genes and isoforms are analyzed by NBSplice. Moreover, the use of at least a single threshold for unreliable isoforms detection (*rt=0.01* or *cT=1* NBSplice configurations) discarded, on average, more than the 70% of transcripts. However, although significant differences (Wilcoxon test p-value < 0.05) were found between the results of the without filtering setup (*WF*) and *rt=0.01*, or *cT=1* configurations, the average number of analyzed genes was barely reduced by 10% when a single threshold was used. Moreover, for these two NBSplice setups, the number of DSG and true positive genes (TPG) were significantly higher (Wilcoxon p-value, one-tail, < 0.05) than the ones observed for the WF configuration.

The highest average values of DUT and TPT were found by the *cT=1* NBSplice setup, whereas the highest DSG and TPG values were found for *rT=0.01*, *rT=0.01;cT=1*, and *rT=0.01;cT=2* configurations. The comparison of these three NBSplice setups revealed that differences below 1% for DSG and TPG, and below 10% for DUT and TPT were found, suggesting that those configurations are equivalent.

As it is depicted by Table 3, the average values of the four listed indicators resulted to be mainly reduced when the NBSplice setup presented a change at the rT parameter than when it was fixed and the cT parameter was changed.

Table 3: **NBSplice results**. The average (± standard deviation) of the number of genes and transcripts analyzed, differentially-spliced genes (DSG), true positive genes (TPG), differentially-used transcripts (DUT) and true positive transcripts (TPT) for the nine NBSplice configurations over the 10 simulated RNA-seq experiments.

Method	Analyzed genes	DSG	TPG	Analyzed transcripts	DUT	TPT
WF	**13284** (± 24)	**942** (± 27)	**876** (± 24)	**99622** (± 921)	**2913** (± 77)	**2745** (± 72)
rT=0.01	**12292** (± 31)	959 (± 27)	**915** (± 24)	29015 (± 85)	2158 (± 53)	2037 (± 55)
cT=1	12412 (± 31)	**967** (± 26)	896 (± 21)	**41393** (± 87)	**3476** (± 66)	**3230** (± 56)
rT=0.01;cT=1	12091 (± 27)	957 (± 27)	913 (± 23)	27812 (± 72)	2179 (± 53)	2050 (± 54)
rT=0.01;cT=2	11547 (± 21)	947 (± 26)	902 (± 22)	25154 (± 80)	2165 (± 51)	2033 (± 48)
rT=0.05;cT=1	11939 (± 25)	833 (± 26)	813 (± 26)	18495 (± 59)	1090 (± 34)	1049 (± 34)
rT=0.05;cT=2	11444 (± 24)	845 (± 28)	823 (± 28)	17591 (± 47)	1111 (± 34)	1066 (± 36)
rT=0.1;cT=1	11511 (± 33)	756 (± 21)	744 (± 23)	14547 (± 71)	844 (± 22)	826 (± 25)
rT=0.1;cT=2	11109 (± 24)	770 (± 24)	758 (± 26)	14040 (± 54)	856 (± 21)	837 (± 23)

For instance, the differences in the average numbers of DSG and TPG found between setups *rT=0.05;cT=1* and *rT=0.1;cT=1* were around 10%, whereas between *rT=0.05;cT=1* and *rT=0.05;cT=r* configurations these differences were lower than 2%. Thus, this suggests that a threshold imposed at the level of the relative expression of isoforms has a greater impact on the DS results than a threshold imposed at the absolute expression level. This could be explained by the fact that NBSplice models are defined in terms of relative expression instead of the absolute expression of the gene transcripts.

3.1.2. Performance results

The boxplot of the performance measures reached by the nine NBsplice setups, considering a nominal FDR value of 0.05, are shown in Figure 2. All configurations achieved an accuracy higher than 0.9 in all the experiment replications (Figure 2**A**), suggesting that most genes (DSG and not DSG) were correctly detected. On the opposite, all NBSplice configurations reached

sensitivity values lower than 0.6, revealing that less than the 60% of the DSG are found by our tool (Figure 2**B**). One possible reason of this result is the higher level of complexity of the human transcriptome, with many more isoforms per gene, which makes more difficult the transcript quantification process.

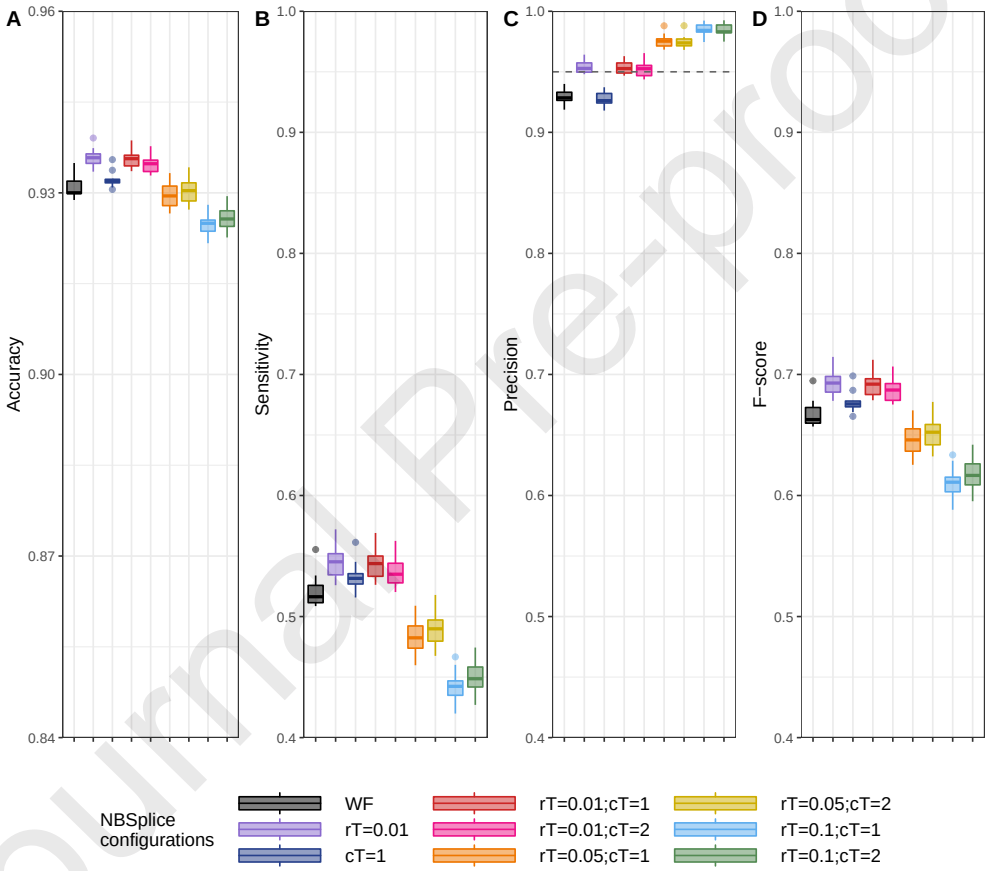


Figure 2: **NBSplice performance.** Boxplots of the performance measures for the nine NBSplice configurations. **A)** Accuracy. **B)** Sensitivity or true positive rate (TPR). **C)** Precision (1-False Discovery Rate, FDR): The dashed gray line depicts the precision value correspond to the nominal FDR (0.05). **D)** F-score.

In terms of precision, most of the tool configurations achieved average values higher than 0.95, being the highest for those involving the most restrictive rT threshold of 0.1 (Figure 2**C**). In particular, the NBSplice setups involving rT=0.01 (*rT=0.01*, *rT=0.01;cT=1*, *rT=0.01;cT=2*) not only controlled the FDR, but also they reached the highest accuracies, sensitivities and F-scores (Figure 2**D**). On the opposite side, although the configurations which involve more restrictive rT thresholds (rT=0.05 and rT=0.1) achieved the highest precisions, they exhibited lower accuracies, sensitivities, and F-scores with the worst performance for the most restrictive setups (*rT=0.1;cT=1* and *rT=0.1;cT=2*).

Figure 3 shows the mean values of sensitivity and FDR reached by the NBSplice setups at three nominal FDR values (significance thresholds), i.e. 0.01, 0.05, and 0.1. Measures values are shown as dots which are filled if the observed FDR was lower than the corresponding nominal FDR. As it can be noted, the more restrictive NBSplice setups, involving rT of 0.05 and 0.1, always controlled their FDR. Meanwhile, the NBSplice configurations without rT filtering (*WF* and *cT=1*) never controlled their FDR. Furthermore, the highest sensitivity was found for the NBSplice configurations using rT=0.1 (*rT=0.01*, *rT=0.01;cT=1* and *rT=0.01;cT=2*).These findings suggest that in order to combine both higher sensitivity and FDR controlling, the optimum configuration of NBSplice parameters is *rT=0.01,cT=1*.

The NBSplice ability to deal with FPs was evaluated only using the best setup, *rt=0.01;cT=1*. Obtained results revealed that in three out of the ten null-realizations, one gene was incorrectly identified as a DSG, indicating a mean percentage of false positive rate (proportion of FP regarding to N)

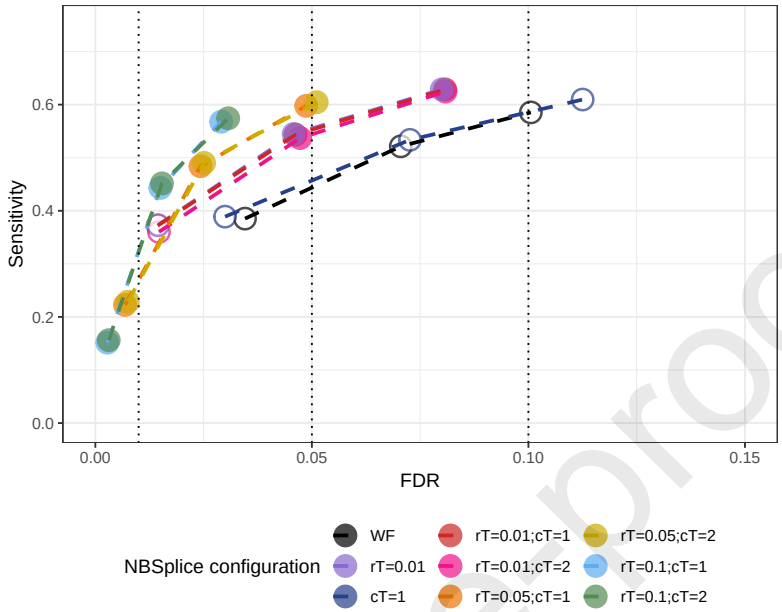


Figure 3: **NBSplice screening for differential splicing.** Average sensitivity (y-axis) over average false discovery rate (FDR, x-axis) achieved by nine evaluated NBSplice setups for three nominal FDR values, i.e 0.01, 0.05, and 0.1. Circles are filled if the observed FDRs were lower than the target nominal FDR, shown as vertical dotted lines.

lower than 0.01 %.

Considering the previous results and the use of a nominal FDR of 0.05 or 0.1, the optimum parameters setup is *rt=0.01;cT=1*. Using this configuration, NBSplice controls its FDR with the highest sensitivity (about to 55%) and F-score (up to 70%). Moreover, it achieves similar results to those obtained with the *rT=0.01* configuration but analyzing many fewer transcripts.

3.1.3. Evaluation on simulated groups and subgroups

Sensitivity over the gene groups and subgroups defined in Table 1 were explored. The results found using the NBSplice optimal configuration (*rT=0.01,cT=1*)

are shown in Figure 4. Panel **A** has the boxplots of the overall measure in the simulated groups, **DS** and **DIEDS**. Averaged scores for those groups were 0.494 and 0.563, respectively. The sensitivity for the **DIEDS** group was significantly higher (Wilcoxon p-value, one-tail, < 0.05) than for the **DS** group, in agreement with the results of [4]. In this work the authors reported that the simultaneous occurrence of differences in both absolute and relative expression of isoforms, simulated here as the **DIEDS** group, contributes to the identification of differential splicing.

Detailed sensitivity boxplots for each simulated subgroups are depicted in Figure 4**B**. The boxplots of TPR achieved in **DS** and **DIEDS** subgroups show a broad separation between subgroups, mainly related to the magnitude of the relative expression change of the most expressed isoform (i.e. the mayor isoform, M). For instance, in subgroups only considering DS, i.e. **DS** group, sensitivities of 78.3% and 73.8% were found for the *DS-0.3-0.7* and *DS-0.5-0.9* subgroups, respectively. Whereas, the observed TPR means were lower than 30% for the *DS-0.3-0.5* and *DS-0.5-0.7* subgroups (26% and 22.8%, respectively). Therefore, the DS detection on the *DS-0.3-0.5* and *DS-0.5-0.7* subgroups is a more difficult task than in the *DS-0.3-0.7* and *DS-0.5-0.9* subgroups. This behaviour could be explained by the fact that in the former **DS** subgroups, the relative expression of the different isoforms of a single gene might be more similar between them than in the case of the later subgroups. Thus, the differences on their relative expressions might not be well detected, suggesting that genes with DS involving large changes in the relative expression of their M isoform are easier to identify than genes which smaller changes in the M isoform proportion.

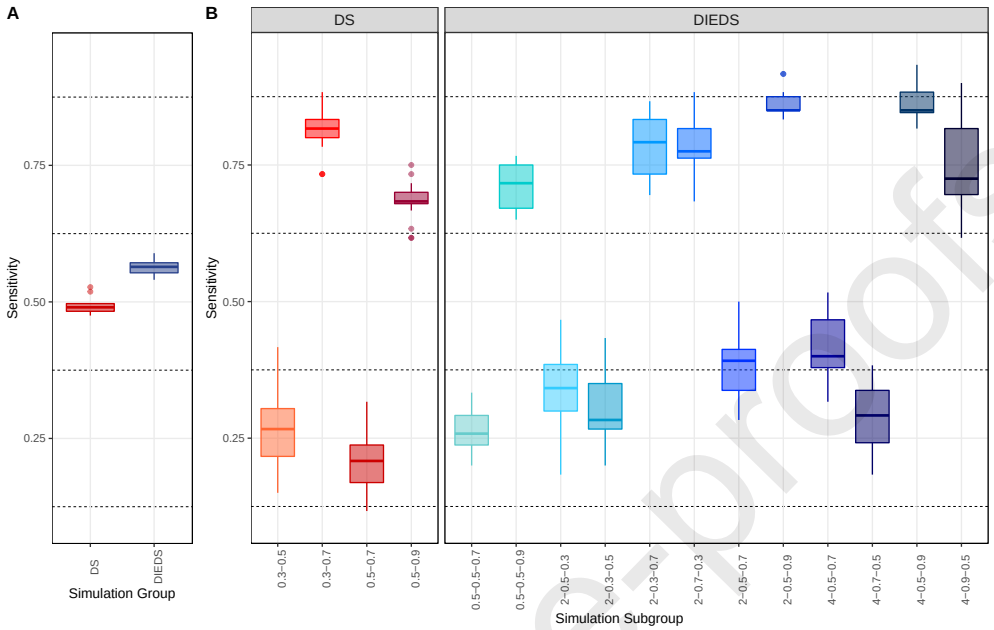


Figure 4: **NBSplice sensitivity evaluation.** Boxplots of the sensitivity (TPR) reached by NBSplice with its parameters set as rT=0.01 and cT=1, along ten experimental replications. **A)** TPR for the simulated groups: **DS**, with differential splicing; **DIEDS**, with differential splicing and differential isoform expression. **B)** TPR over the simulated subgroups described in Table 1 and defined by the change in the relative expression of the major gene isoform and by the fold change in absolute isoform expression between tumor and control conditions.

In the subgroups where DS was combined with changes in the absolute expression of gene isoforms, i.e. **DIEDS** group, higher sensitivities ($> 60\%$) were also found in those involving the largest relative expression changes in M gene isoforms. In particular, the highest TPRs were observed for DIEDS-2-0.5-0.9 and DIEDS-4-0.5-0.9 subgroups, without evidence of significant differences between them (Wilcoxon test p-value > 0.05). In agreement with the trend observed for the **DS** subgroups, in those genes with lower changes

in the relative expression of the M gene isoforms, the DS detection is a difficult task. Furthermore, and as it was previously evaluated in [4], the DS detection could also be affected by the number of transcripts per gene.

Considering pairs of **DIEDS** subgroups with the same fold change but different relative expression of the M isoform, i.e *DIEDS-0.5-0.5-0.7* and *DIEDS-0.5-0.5-0.9* or *DIEDS-2-0.3-0.5* and *DIEDS-2-0.3-0.7*, sensitivities significantly higher (Wilcoxon test p-values > 0.05) were found in these subgroups involving higher changes in the proportion of the M isoforms. Furthermore, the highest variability on TPRs was mainly found for those simulation subgroups where the absolute isoform expression was increased in one direction, whereas the relative expression of the M isoform was increased in the opposite direction. For example, in the *DIEDS-4-0.9-0.5* subgroup. Thus, large changes in the relative expression of the M isoform are critical to correctly detect genes as DSG with NBSplice.

3.2. Comparison against differential splicing tools

The performance measures achieved by NBSplice and three R packages used for DS analysis, DEXSeq, DRIMSeq, and RATs, are shown in Figure 5. Except for RATs, which has been run using its own filtering strategy, each method is represented twice since they have been used considering two different strategies to detect unreliable transcripts; *SwimmF* and the optimal NBSplice setup. The comparison of measures distributions revealed that for DEXSeq, DRIMSeq and NBSplice the best results on the simulated database were reached using the NBSplice filtering strategy (Wilcoxon test p-values < 0.05). Thus, only the results of these DS tools using the *rT=0.01;cT=1* filtering strategy and the RATs results were compared.

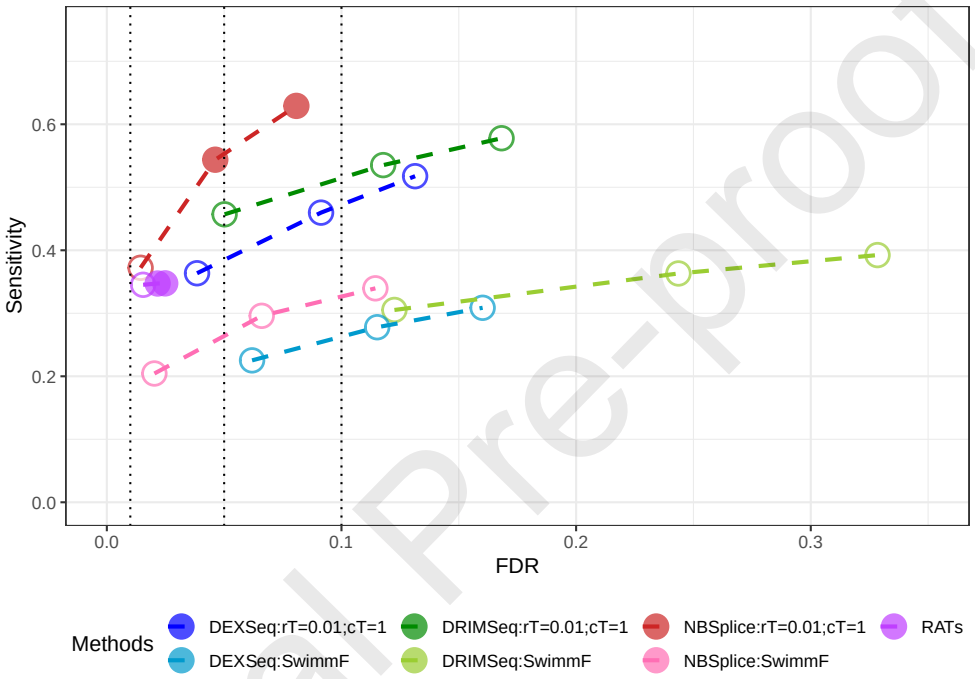


Figure 5: **Comparison of performance measures distributions.** Boxplots of the performance measures achieved by NBSplice, DEXSeq, DRIMSeq and RATs, using two different strategies for filtering unreliable transcripts, *SwimmF* and *rT=0.01;cT=1* are considered. **A)** Accuracy. **B)** Sensitivity or true positive rate (TPR). **C)** Precision (1-False Discovery Rate, FDR): The dashed grey line depicts the precision value corresponding to the nominal FDR (0.05). **D)** F-score.

Considering accuracy, sensitivity and F-score, NBSplice was the best tool for DS detection (Wilcoxon one-tail test p-values < 0.05). Comparing the achieved precision scores, it also overcame both DEXSeq and DRIMSeq R packages. Although the NBSplice precision was significantly lower than the score achieved by RATs, our tool was able to control the imposed FDR (0.05). Significant differences were found between the accuracy of the four methods (Figure 5**A**). However, the lowest value was about 90% suggesting that all of them have been performed similarly with regard to this measure. In particular, the highest average accuracy, achieved by NBSplice, was 0.938. In terms of sensitivity (Figure 5**B**), only DRIMSeq and NBSplice reached average measures higher than 0.5, being the latter the most sensitive tool (Wilcoxon one-tail test p-value, 0.0333). Moreover, its high sensitivity was complemented with high precision (Figure 5**C**). NBSplice also achieved the best balance between these two scores, reaching an average F-score of 0.693. Therefore, these results suggest that NBSplice is appropriate for DS analysis.

Average values of sensitivity and FDR achieved by the evaluated NBSplice, DEXSeq, DRIMSeq and RATs tools at three nominal FDR values, i.e. 0.01, 0.05, and 0.1, are shown in Figure 6. As it was previously found, the use of the optimal NBSplice filtering strategy leads to better performance results than the *SwimmF* strategy for NBSplice, DEXSeq and DRIMSeq. Moreover, NBSplice (*rT=0.01;cT=1*) and RATs were the only two strategies that controlled their FDR for most of the nominal FDR evaluated (filled dots), excepting the 0.01 threshold. By analyzing the RATs results, no sensitivity improvement was found when the FDR threshold was changed, probably because the other thresholds used for DTU detection (such as the

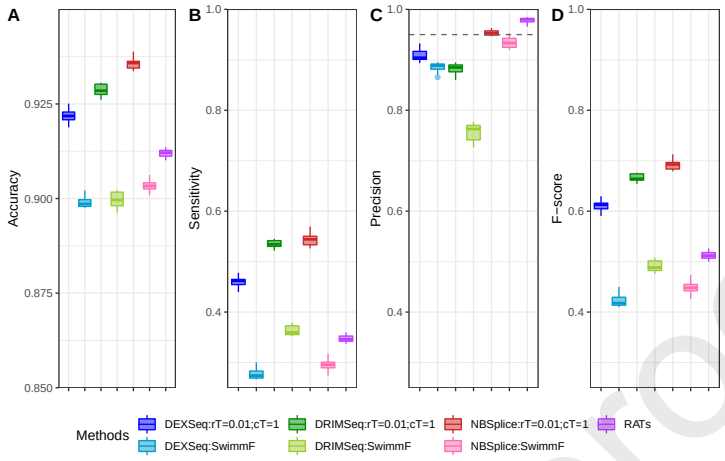


Figure 6: **Sensitivity and FDR evaluation.** Average sensitivity (TPR, y-axis) over average false discovery rate (FDR, x-axis) achieved for NBSplice, DEXSeq, DRIMSeq and RATs, considering three nominal FDR values, i.e 0.01, 0.05, and 0.1. Excepting RATs, which implements its own filtering strategy, the methods were evaluated using two alternative filtering strategies to detect unreliable transcripts, identified as *SwimmF* and *rT=0.01;cT=1*. Filled circles indicate if the achieved average FDR is lower than the target nominal FDR, shown as vertical dotted lines.

reproducibility bootstrap threshold) influence more in the DS identification than the FDR threshold. Thus, NBSplice with its filtering parameters set as $rT = 0.01$ and $cT = 1$ not only is the most sensitive and accurate tool, but also is precise and control its FDR when using 0.05 or 0.1 significance thresholds.

3.3. Application to the TCGA cancer dataset

Isoform expression matrix for the 28 COAD analyzed samples involved 178,581 isoforms. The use of the NBSplice filters (*rT=0.01* and *cT=1*) resulted in 162,533 unreliable isoforms meanwhile 159,679 isoforms were identified as low-expressed using the *SwimmF* method. On the other hand, RATs

identified 93,443 ineligible transcripts. NBSplice detected 569 isoforms with significant differences in their relative expression between normal and cancer patients, and 517 DSGs. Whereas, the number of DSGs found by DRIMSeq, DEXSeq and RATs were 1,715, 1,134, and 251, respectively. Detailed results are available at the Supplementary File S2.

The comparison of the DSG lists obtained with the four R packages is illustrated in the Venn diagram shown in Figure 7. It is worth noticing that the highest number of exclusive DSGs was achieved by DRIMSeq (653), followed by RATs (206). The Venn diagram also revealed that the number of overlapped DSG between DRIMSeq and DEXSeq (670) is greater than the number of DSG commonly identified (12). This could be related to the fact that both tools analyze the same set of reliable transcripts defined by the *SwimmF* strategy as recommended [21].

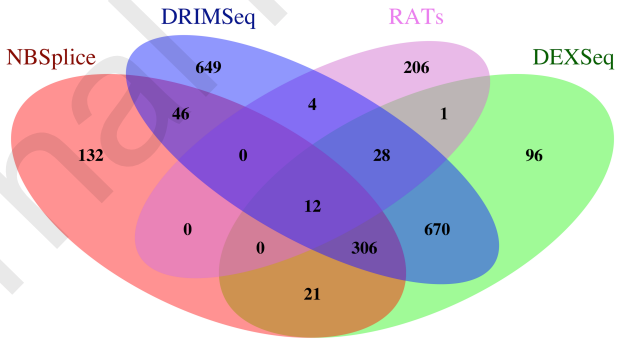


Figure 7: **Differentially spliced genes overlapping.** Venn diagram for the differentially spliced gene lists obtained with NBSplice, DRIMSeq, DEXSeq and RATs over the colorectal cancer data set.

On the other hand, the highest consensus between three different tools

was observed for NBSplice, DEXSeq and DRIMSeq (306 DSGs). In this sense, the 74% of the DSGs detected by NBSplice were also found by at least one of the other tools, being 61.5 the percentage of DSGs also detected by both DRIMSeq and DEXSeq. Among those differentially spliced genes detected by NBSplice, several genes previously related to colorectal cancer and other cancer types were found. For instance, the *FBLN2* gene, recently reported as differentially spliced in CRC [29], has been identified by the four DS tools. Other examples are the *ZG16B* and the *ATXN3* genes, which have been only detected by NBSplice. Particularly, mRNA alterations of *ZG16B* have been associated with poor prognosis in CRC patients [30, 31]. The *ATXN3* gene has been recently identified as differentially regulated in CRC patients by means of expression changes in the microRNA-25. Also, it has been related to the promotion of proliferation and metastasis of CRC suggesting them as potential therapeutic targets for this type of tumor [32]. Relative expressions of significant isoforms from those genes are shown in Figure 8. In particular, the *ZG16B* gene has six annotated transcripts which relative expressions are shown in Figure8**A**. From those, only one, *ZG16B-201*, has been identified by NBSplice as a reliable isoform and analyzed for DTU. Particularly, for this transcript, the estimated relative expressions were 0.612 and 0.897, for Tumor and Normal conditions, respectively. Similarly, for the *ATXN3* gene (Figure 8**B**) only one transcript, *ATXN3-247*, has been detected as reliable and analyzed for DTU. Since the *ATXN3* has 43 annotated transcripts which have been identified as unreliable, they have not been included in the graphics of Figure 8**B** generated by NBSplice methods in which only the DUT is shown.

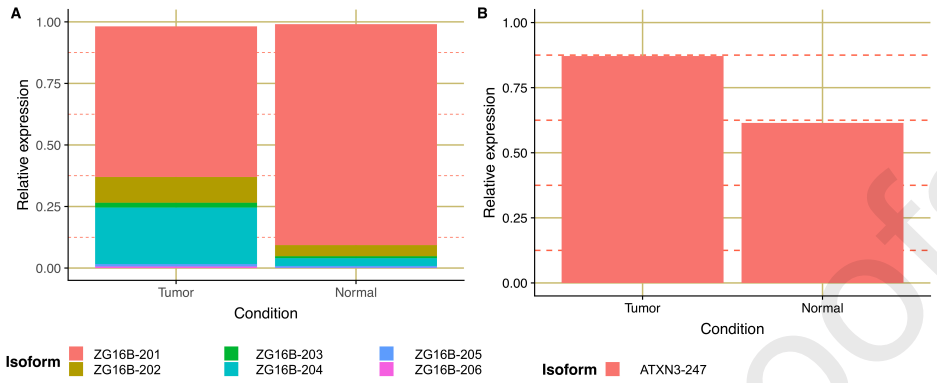


Figure 8: **Relative expression of isoforms differentially used**. Per condition isoform relative expression of significant isoforms from genes **A)** *ZG16B* and **B)** *ATXN3*, obtained with NBSplice.

4. Conclusion

Here we present NBSplice, a novel tool for differential splicing (DS) analysis based on the identification of differential transcript usage (DTU). It uses negative binomial generalized linear models, fitted at the gene level, to estimate changes in the relative expression of gene isoforms. NBSplice is implemented as an R package available at Bioconductor. NBsplice outperfoms current state-of-the-art DS methods in terms of accuracy, sensitivity, F-score, and in the FDR controlling.

Our tool detects more than the 50% of the simulated differentially spliced genes (DSG), being more sensitive to identify genes where the change in the relative expression of their major isoform is greater. The comparison of NBSplice against DEXSeq, DRIMSeq, and a recently developed tool, RATs, revealed that our tool is adequate for DS analysis.

The NBSplice usefulness was demonstrated by analyzing a real colorectal

cancer dataset (COAD) where it found 500 DSG. Approximately, 26% of these genes were only identified using our tool. Some of these findings were explored finding that they have been previously reported in CRC, suggesting the NBSplice ability as a discovery tool.

In conclusion, NBSplice resulted an accurate and precise tool useful for identifying differential spliced genes and their isoform counterparts.

Availability of data and material

NBSplice is provided as an R package and is freely available at Bioconductor repository http://bioconductor.org/packages/NBSplice/.

All the source code used for this article can be found at the URL https://github.com/gamerino/NBSpliceSuppInformation.

Competing interests

The authors have declared no competing interests.

Funding

This work has been supported by the following Argentine institutions: Universidad Católica de Córdoba [grant no. BOD/2016 to E.A.F.], Ministerio de Ciencia, Tecnología e Innovación Productiva [grant no. PPL6/2011 to E.A.F.], Secretaría de Ciencia y Tecnología de la Universidad Nacional de Córdoba [grant no. 30720150101719CB to E.A.F.] and Consejo Nacional de Investigaciones Científicas y Técnicas.

Authors' contributions

GAM and EAF conceived of and designed the method and performed the data analysis. GAM implemented the method and performed all computational experiments. GAM and EAF drafted the manuscript. All authors read and approved the final manuscript.

References

[1] L. Gallego-Paez, M. Bordone, A. Leote, N. Saraiva-Agostinho, M. Ascensao-Ferreira, N. Barbosa-Morais, Alternative splicing: the pledge, the turn, and the prestige, Human genetics 136 (9) (2017) 1015–1042.

[2] C. Ghigna, C. Valacca, G. Biamonti, Alternative splicing and tumor progression, Current genomics 9 (8) (2008) 556–570.

[3] S. Oltean, D. Bates, Hallmarks of alternative splicing in cancer, Oncogene 33 (46) (2014) 5311.

[4] G. A. Merino, A. Conesa, E. A. Fernández, A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies, Briefings in bioinformatics 20 (2) (2017) 471–481.

[5] J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, E. Eyras, SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions, Genome biology 19 (1) (2018) 40.

[6] K. Vitting-Seerup, A. Sandelin, The landscape of isoform switches in human cancers, Molecular Cancer Research 15 (9) (2017) 1206–1220.

[7] R. Liu, A. E. Loraine, J. A. Dickerson, Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems, BMC bioinformatics 15 (1) (2014) 364.

[8] J. Wang, Z. Ye, T. H.-M. Huang, H. Shi, V. Jin, A survey of computational methods in transcriptome-wide alternative splicing analysis, Biomolecular concepts 6 (1) (2015) 59–66.

[9] C. Zhang, B. Zhang, L.-L. Lin, S. Zhao, Evaluation and comparison of computational tools for RNA-seq isoform quantification, BMC genomics 18 (1) (2017) 583.

[10] M. Nowicka, M. D. Robinson, DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics, F1000Research 5 (2016) 1356.

[11] K. Froussios, K. Mourão, G. Simpson, G. Barton, N. Schurch, Relative Abundance of Transcripts (RATs): Identifying differential isoform abundance from RNA-seq, F1000Research 8.

[12] P. McCullagh, N. J, Generalized linear models, Chapman and Hall, 1989.

[13] C. Soneson, M. Delorenzi, A comparison of methods for differential expression analysis of RNA-seq data, BMC bioinformatics 14 (1) (2013) 91.

[14] R. J. Simes, An improved Bonferroni procedure for multiple tests of significance, Biometrika 73 (3) (1986) 751–754.

[15] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal statistical society: series B (Methodological) 57 (1) (1995) 289–300.

[16] G. A. Merino, E. A. Fernandez, NBSplice: Negative Binomial Models to detect Differential Splicing, URL `http://www.bdmg.com.ar`, r package version 1.0.6, 2019.

[17] W. Venables, B. Ripley, Modern applied statistics with S Springer-Verlag, New York .

[18] J. Fox, S. Weisberg, An R companion to applied regression, Sage Publications, 2018.

[19] M. P. Rubin-Delanchy, Package mppa URL `https://CRAN.R-project.org/package=mppa`.

[20] C. Soneson, K. L. Matthes, M. Nowicka, C. W. Law, M. D. Robinson, Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage, Genome biology 17 (1) (2016) 12.

[21] M. I. Love, C. Soneson, R. Patro, Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification, F1000Research 7.

[22] B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC bioinformatics 12 (1) (2011) 323.

[23] K. Van den Berge, C. Soneson, M. D. Robinson, L. Clement, stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage, Genome biology 18 (1) (2017) 151.

[24] L. Yi, H. Pimentel, N. L. Bray, L. Pachter, Gene-level differential analysis at transcript-level resolution, Genome biology 19 (1) (2018) 53.

[25] C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, F1000Research 4.

[26] 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, Proceedings of the National Academy of Sciences 108 (22) (2011) 9172–9177.

[27] P. Tatlow, S. R. Piccolo, A cloud-based workflow to quantify transcript-expression levels in public cancer compendia, Scientific reports 6 (2016) 39259.

[28] S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from RNA-seq data, Genome research 22 (10) (2012) 2008–2017.

[29] J. Liu, H. Li, S. Shen, L. Sun, Y. Yuan, C. Xing, Alternative splicing

events implicated in carcinogenesis and prognosis of colorectal cancer, Journal of Cancer 9 (10) (2018) 1754.

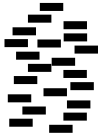
[30] R. Barderas, M. Mendes, S. Torres, R. A. Bartolome, M. Lopez-Lucendo, R. Villar-Vazquez, A. Peláez-García, E. Fuente, F. Bonilla, J. I. Casal, In-depth characterization of the secretome of colorectal cancer metastatic cells identifies key proteins in cell adhesion, migration, and invasion, Molecular & Cellular Proteomics 12 (6) (2013) 1602–1620.

[31] B. W. Kang, S. J. Lee, Y. J. Lee, J. G. Kim, Y. S. Chae, S. K. Sohn, J. H. Moon, Genetic variations in miRNA binding site of TPST1 and ZG16B associated with prognosis for patients with colorectal cancer., Journal of Clinical Oncology 31 (15_suppl) (2013) 3553–3553, doi:\let\ @tempa\bibinfo@X@doi10.1200/jco.2013.31.15_suppl.3553.

[32] D. Li, T. Zhang, J. Lai, J. Zhang, T. Wang, Y. Ling, S. He, Z. Hu, MicroRNA-25/ATXN3 interaction regulates human colon cancer cell growth and migration, Molecular medicine reports 19 (2019) 4213–4221.

- **Isoforms expression analysis enables the detection of differentially spliced genes.**
- **NBSplice successfully detects both differential splicing and transcript usage.**
- **The transcript usage analysis allows the identification of potential biomarkers.**
- **NBSplice identifies alternatively spliced genes with a potential therapeutic role.**

Reads



Mapping

Alignments

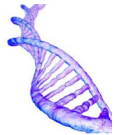
Quantification

Normal Tumor

...	...	...	...	...	...
8	9	...	12	13	...
5	4	...	8	11	...
12	11	...	4	5	...
...	...	...	...	...	...

Isoforms

Gene



Differentially-spliced genes

...

ZG16B

ATXN3

...

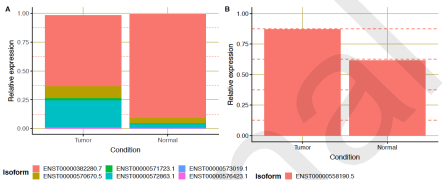
Differentially-used isoforms

...

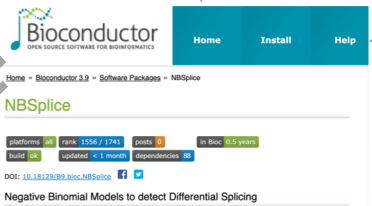
ZG16B-201

ATXN3-247

...



Expression matrix



Isoform I	Gene A
Isoform II	Gene A
...	...
Isoform I	**Gene X**
Isoform II	**Gene X**
Isofom III	**Gene X**
...	...
Isoform I	Gene Z
Isoform II	Gene Z
...	...

Gene-isoform matrix

Authors' contributions

GAM and EAF conceived of and designed the method and performed the data analysis. GAM implemented the method and performed all computational experiments. GAM and EAF drafted the manuscript. All authors read and approved the final manuscript.

The authors have no competing financial conflict of interest to declare.