Bioinformatics, 32(14), 2016, 2111–2119
doi: 10.1093/bioinformatics/btw145
Advance Access Publication Date: 21 March 2016
Original Paper

OXFORD

Gene expression

jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data

Yann Christinat†, Rafał Pawłowski† and Wilhelm Krek*

Institute of Molecular Health Sciences, ETH Zurich, Zurich 8093, Switzerland

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: Ziv Bar-Joseph

Received on March 5, 2015; revised on March 3, 2016; accepted on March 11, 2016

Abstract

Motivation: Alternative splicing represents a prime mechanism of post-transcriptional gene regulation whose misregulation is associated with a broad range of human diseases. Despite the vast availability of transcriptome data from different cell types and diseases, bioinformatics-based surveys of alternative splicing patterns remain a major challenge due to limited availability of analytical tools that combine high accuracy and rapidity.

Results: We describe here a novel junction-centric method, jSplice, that enables *de novo* extraction of alternative splicing events from RNA-sequencing data with high accuracy, reliability and speed. Application to clear cell renal carcinoma (ccRCC) cell lines and 65 ccRCC patients revealed experimentally validatable alternative splicing changes and signatures able to prognosticate ccRCC outcome. In the aggregate, our results propose jSplice as a key analytic tool for the derivation of cell context-dependent alternative splicing patterns from large-scale RNA-sequencing datasets.

Availability and implementation: jSplice is a standalone Python application freely available at http://www.mhs.biol.ethz.ch/research/krek/jsplice.

Contact: wilhelm.krek@biol.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Alternative splicing, a mechanism unknown more than 30 years ago, is now regarded a major contributor to transcriptome and proteome diversity (Modrek and Lee, 2002). It is estimated that, in human, 95% of multi-exon genes undergo alternative splicing (Pan *et al.*, 2008; Wang *et al.*, 2008), underscoring the importance of this post-transcriptional mechanism in the regulation of gene expression. Indeed, many key biological processes and disease states involve profound alternative splicing pattern changes (Singh and Cooper, 2012). Nowadays, technological advances in RNA sequencing (RNA-seq) are yielding data at an enormous scale and allow for genome-wide analysis of alternative splicing patterns from patient data. However, the availability of in silico tools able to reliably interrogate alternative splicing events from RNA-seq data remains limited (Ozsolak and Milos, 2011). For example, transcript-based approaches, such as CuffDiff (Trapnell *et al.*, 2012b), allow for reconstruction of the complete transcriptome but are limited in their accuracy of extracted events (Garber *et al.*, 2011; Rehrauer *et al.*, 2013). Exon-based approaches, such as DEXSeq (Anders *et al.*, 2012), provide more reliable results but due to their focus on a subset of events, yield only partial information. In addition, both of these approaches suffer from scalability issues, hindering the analysis of large-scale datasets. Hence, there is pressing need to expand methodologies for fast and accurate analysis of alternative splicing patterns.

Over the past few years, an increased ability to align split reads to the genome and greater sequencing depths raised the status of junction reads—reads that span exon–exon junctions—to an independent and reliable source of information. To the contrary of other reads, they permit a precise localization of exon boundaries and allow testing whether two exons exist consecutively within a transcript, an essential step for the detection of alternative splicing events. An increasing number of methods, such as DiffSplice (Hu *et al.*, 2013), MISO (Katz *et al.*, 2010), JuncBASE (Brooks *et al.*, 2011), or SplicingCompass *(*Aschoff *et al.*, 2013), are including junction reads in their analysis but only few—e.g. Spanki (Sturgill *et al.*, 2013) and rMATS (Shen *et al.*, 2014)—consider them as the primary source of information.

Here we present jSplice, a junction-centric method to identify differentially expressed alternative splicing events at the exon level. Unlike Spanki and rMATS, which rely on databases of splicing events, jSplice performs a *de novo* reconstruction of all possible alternative splicing events based on junction positions, allowing thus detection of simple and complex alternative splicing events even in poorly annotated genomes. Distinctive to other currently applied methods, jSplice does not assess the statistical significance of the observed changes in expression between two groups of samples. Rather it relies on a set of user-defined thresholds to select and rank alternative splicing changes by the amplitude. The performance of jSplice vis-à-vis other splicing analysis software tools was rigorously assessed on simulated and real datasets and shows a greatly enhanced sensitivity and specificity. Also, application of jSplice to RNA sequencing datasets derived from human clear cell renal cell carcinoma (ccRCC) cell lines and patient tumors identified validatable alternative splicing events. Our work suggests that jSplice is a highly reliable and accurate method with remarkably low execution times, making it greatly suitable for routine large-scale analysis of RNA-seq data.

2 Methods

2.1 The jSplice computational pipeline

jSplice, with the 'j' standing for 'junction', follows a classical alternative splicing analysis pipeline (Fig. 1). It starts with an external read alignment procedure (e.g. STAR (Dobin *et al.*, 2013) or TopHat (Kim *et al.*, 2013)) followed by the detection and scoring of alternative splicing modules (ASMs) to finally generate a ranked list of differentially expressed ASMs. Distinctive to other methods such as DEXSeq (Anders *et al.*, 2012) or MISO (Katz *et al.*, 2010), jSplice detects alternative splicing modules solely based on junction positions. jSplice adopts thus a junction-centric approach as emphasized in Spanki and rMATS but addresses their limitations by relying on a *de novo* reconstruction, as proposed in DiffSplice and JuncBASE, and a novel scoring system. Indeed, current methods rely heavily on genome annotations or databases of splicing events and, unlike jSplice, might miss complex or novel alternative splicing events.

The definition of an ASM in jSplice is inspired by Hu *et al.* (2013) and represents the set of exon-exon junctions of a genomic region where transcripts diverge in more than one isoform. From their definition of an ASM—a region of the splice graph with a single entry point and a single exit point—one can observe that given a junction *i* from an ASM *A*, there always exists a junction *j* in *A* s.t. *j* overlaps position-wise with *i*. By extension, if a junction *k* overlaps with *i* then *k* belongs to *A*. A maximal set of overlapping junctions, which defines an ASM in jSplice, is thus the set of junctions of an ASM as defined by Hu *et al.*, including its nested ASMs. Retained introns contain only one junction and therefore have to be treated as special cases in jSplice that require information on exons. The latter, if available, is added to the ASM if they are fully contained position-wise within it but does not contribute to their identification. Note, that the detection of all ASMs can be done in linear time once elements are sorted.

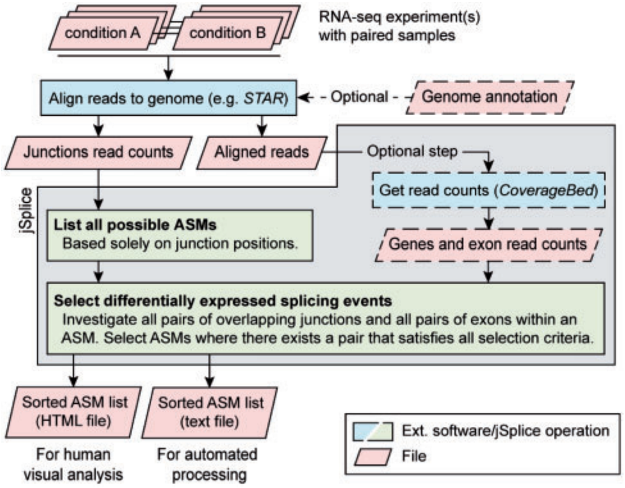
condition A
condition B
RNA-seq experiment(s) with paired samples
Align reads to genome (e.g. STAR)
- Optional -
Genome annotation
Junctions read counts
Aligned reads
Optional step
jSplice
Get read counts (CoverageBed)
List all possible ASMs
Based solely on junction positions.
Genes and exon read counts
Select differentially expressed splicing events
Investigate all pairs of overlapping junctions and all pairs of exons within an ASM. Select ASMs where there exists a pair that satisfies all selection criteria.
Sorted ASM list (HTML file)
Sorted ASM list (text file)
For human visual analysis
For automated processing
Ext. software/jSplice operation
File

Fig. 1. The jSplice pipeline. Schematic outline of jSplice distinguishing data files (light red) and software operations (light blue and light green). jSplice's inputs are aligned RNA-seq reads from paired experiments. The call to CoverageBed is executed within jSplice but the read alignment procedure has to be performed independently, e.g. with STAR or TopHat

The identification of differential expression within an ASM in jSplice differs greatly from current methods. Instead of estimating a read count distribution, which would account for the technical and biological variation inherent to RNA-seq experiments, jSplice only considers the expression fold-change between conditions on every possible pair of overlapping junctions (position-wise), or exons if available. By definition, overlapping elements of an ASM belong to different sets of transcripts. Therefore pairwise comparisons assess whether there exists one or several transcripts whose expression is changed with respect to the others. jSplice tests then, for each ASM, all pairs of overlapping junctions and defines an ASM as differentially expressed if there exists at least one pair that satisfies the five thresholds defined below. Among all differentially expressed pairs, the one with the maximum fold-change value is highlighted in jSplice's output and its average value across replicates is used as the ranking criterion for the ASM. The rationale for using a fold-change ratio is that, due to their flanking DNA sequences, two different junctions will have different read-binding affinities. A fold-change ratio of a single junction over two conditions will eliminate this particular bias and render junction-junction comparisons feasible. Note that the fold-change is still subject to technical and biological variability and thus entails assessment of statistical significance. For the sake of comparison, the popular 'percent spliced in' (PSI) metric (Sturgill *et al.*, 2013) assumes that all junctions have the same read-binding affinity.

As mentioned above, five user-defined thresholds are used to test if a pair of overlapping junctions is differentially expressed:

- Fold-change ratio. We define the relative log2 fold-change, given two overlapping elements *i*, *j* of an ASM and their associated read counts in condition A and B ($c_{i,A}$, $c_{i,B}$, $c_{j,A}$ and $c_{j,B}$), as
- relFC $= \log2\left({}^{c_{i,A}}/_{c_{i,B}}\right) - \log2\left({}^{c_{j,A}}/_{c_{j,B}}\right)$. Infinite values are returned if counts are zeros. The absolute relFC has to be above or equal to the user-defined threshold.

- Fisher exact test. The latter is carried out on the read counts to assess whether there exists, per replicate and junction pair, a non-random difference between the two conditions. By default, a *P*-value threshold of 0.05 is considered.
- Minimum read count. The threshold t_c enforces a lower limit on the read count to compute expression ratios. A valid junction pair *(i,j)* has $c_{i,A} \geq t_c$ and $c_{j,B} \geq t_c$ or vice-versa ($t_c = 20$ by default). In the case in exon pairs, length-normalized counts are used.
- Transcript inclusion percentage: The threshold t_i ensures a focus on major transcripts or isoform switches if set to a high value. A valid junction pair *(i,j)* has $c_{i,A} \geq t_i * \max(c_{k,A} | k \in ASM)$ and $c_{j,B} \geq t_i * \max(c_{k,B} | k \in ASM)$ ($t_i = 10\%$ by default). In the case of exon pairs, length-normalized counts are used.
- Total gene expression: The threshold t_r limits the analysis to well-expressed genes ($t_r = 1$ RPKM by default). A valid gene has a total gene expression higher than t_r in all conditions.

Similar to MISO (Katz *et al.*, 2010), jSplice handles replicates independently. By default all thresholds have to hold in all replicates but in case of large cohorts, one can focus on ASMs that happen in at least *N* patients.

To obtain read counts per exon and genes, jSplice relies on CoverageBed from the BEDtools package (Quinlan and Hall, 2010). Note that jSplice distributes the CoverageBed commands—one per sample file—to the available CPU cores (user-defined). For efficiency reasons, the genome annotation is compared to the junction files before the call to CoverageBed is performed. Exons without any junction matching one of their extremities are discarded and new exons are added by intersecting annotated exons with junctions. The annotation is then used to assign gene names and exons to ASMs.

2.2 Data

Data simulation was performed through BEERS (Grant *et al.*, 2011) with the human RefSeq annotation. Two different setups were applied. First, we generated a 'simple' dataset where we randomly selected 100 multi-transcript genes on chromosome 1 and changed the expression of the lowliest expressed transcript to twice the total gene expression. Simulations were run in triplicates with 10M 100-bp reads. Second, we created a 'complex' dataset by selecting 1000 multi-transcript genes and randomly assigning the expression of one transcript to another value sampled from the original distribution. Additionally, biological noise was added to each replicate in the 'complex' dataset as a Gaussian distribution with a standard deviation of 1 RPKM. Combinations of 20 M, 50 M, 100 M of 48-bp, 100-bp, up to five replicates were used for the simulations. Unless specified otherwise, jSplice was run with a *relFC* threshold of log2(1.5), t_c threshold of 10, and t_i threshold of 10% to capture all types of events.

Real datasets were downloaded from the Gene Expression Omnibus (GEO) database (Edgar *et al.*, 2002) with the following accession numbers: GSE37704 (Trapnell *et al.*, 2012a) and GSE23776 (Griffith *et al.*, 2010). The 189 RT-PCR gel images from Griffith *et al.* were visually categorized as differentially expressed or not by Dr. Rafal Pawlowski, Dr. Peter Mirtschink and Dr. Yann Christinat independently based on the relative band intensity. The category was assigned in case of agreement between at least two examiners. jSplice was run with a *relFC* threshold of $\log_2(2)$, a t_c threshold of 10 (20 for the Griffith dataset as the read length is shorter), and a t_i threshold of 10% to capture all types of events.

Level 3 RNA-seq data from 65 ccRCC patients with matching samples (normal and primary tumor tissues) were downloaded from the TCGA web archive (http://tcga-data.nci.nih.gov, accessed on November 4, 2013). As exon read count is readily available, it was included as outlined in the jSplice pipeline. jSplice was run with a *relFC* threshold of $\log_2(2)$, a t_c threshold of 20 (as the data consists of 42bp reads), a t_i threshold of 50% to focus on isoform switch events. A minimum of 5 samples had to match all criteria. Clustering of jSplice results was performed first on the 1225 ASMs (Ward linkage), then on samples separated by their VHL mutation status.

2.3 Read alignment and other software

For all experiments (simulated and real data), reads where aligned with STAR (v2.3.0)—or TopHat (v2.0.11) when specified—and the University of California, Santa Cruz (UCSC) hg19 genome annotation. All software were executed using default parameters. An FDR value of 0.05 was used for DEXSeq (v1.8.0), rMATS (v3.0.8), DiffSplice (v0.1.1) and SplicingCompass (v1.0.1) to enforce high specificity. Results based on exon and junction reads were considered for rMATS. The complete database of splicing events (SE, MXE, AFE, ALE, RI, A3SS and A5SS) was used for MISO (v0.4.9) in all cases except for the Griffith dataset where only cassette exons (SE) were considered. A Bayes factor threshold of 3, to hold in all replicates, was used for MISO. Multicore processing was used whenever available and depended on the cluster load. Results for CuffDiff2 on the Trapnell dataset and ALEXA-Seq on the Griffith dataset were taken from their respective publications (Griffith *et al.*, 2010; Trapnell *et al.*, 2012a).

2.4 Cell culture

RCC4, HEK293T and Lynex cells were cultured in DMEM supplemented with 10% FCS, L-Glutamine and Pen/Strep. RCC4 cells re-expressing pVHL ('VHL') or control cells ('EMPTY') were generated by lenti- (using pLKO1) or retro-viral (pBABE) transduction as previously described (Thoma *et al.*, 2007; Troilo *et al.*, 2014). Pools of RCC4 cells stably expressing pVHL were obtained by culturing in the presence of puromycin (2 μg/ml) for at least one week.

2.5 RNA isolation and sequencing

Total RNA was isolated using miRNeasy kit (Qiagen), including DNase digestion step, according to the manufacturer's protocol. The quality of the RNA samples was verified by measuring absorbance at 280 and 260 nm, assessment of rRNA 28S/18S ratio and RNA Integrity Number (RIN). RNA sequencing was carried out at the Beijing Genomics Institute (BGI) on an Illumina HiSeq 2000 sequencer and resulted in 50 M of 100-bp paired-end reads per sample. Reads were aligned with STAR on hg19 and UCSC annotations. jSplice was run with a *relFC* threshold of $\log_2(2)$, a t_c threshold of 10 (as the data consists of 100 bp reads), and a t_i threshold of 50% to focus on major isoforms only. Data are available under accession number SRP045624 at http://www.ncbi.nlm.nih.gov/sra.

2.6 Immunoblot analysis

Cells were lysed directly on the dish using standard 2× SDS sample buffer. Protein samples were sonicated, resolved on acrylamide gels, transferred to nitrocellulose membranes and visualized using antibodies against pVHL (Hergovich *et al.*, 2003), HIF1 (Novus NB100-479), HIF2 (Novus NB100-122) and GLUT1 (Abcam ab14683).

2.7 Gene and RNA isoform expression analysis by quantitative and semi-quantitative PCR

2–3 μg of total RNA was reverse-transcribed into cDNA using EcoDry Premix Random Hexamers kit (Clontech) according to the manufacturer's instructions. cDNA samples were then analyzed by real time PCR using LightCycler 480 SYBR Green (Roche) and primer pairs designed to recognize either the specific RNA isoforms or the total mRNA. Semi-quantitative PCR was carried out using PrimeSTAR Max DNA Polymerase (Clontech) and primer pairs designed to bind exons flanking the alternatively spliced regions. Subsequently, PCR products were resolved on 2% agarose gels and visualized using a UV lamp. The primer pairs used in this study with the corresponding sequences are listed in supplementary material (Supplementary Table S5).

3 Results

3.1 jSplice is a fast and accurate method to detect differentially expressed alternative splicing events

As no benchmark dataset exists to validate methods for the detection of differential alternative splicing, we first employed simulated data. For that purpose, we used BEERS (Grant *et al.*, 2011), a simulation engine for generating RNA-seq data to create two different scenarios. The 'simple' dataset contains no biological variation and introduces obvious splicing changes (median fold-change of 4.43, 5% percentile at 4.01, IQR of 2.15) while the 'complex' dataset had biological variation and a wider range of splicing changes (median fold-change of 2.80, 5% percentile at 1.09, IQR of 5.42). We then compared the performance of jSplice with several other currently available methods. We included software that have a broad user base (DEXSeq (Anders *et al.*, 2012), MISO (Katz *et al.*, 2010) and rMATS (Shen *et al.*, 2014)), that introduced the notion of ASMs (DiffSplice (Hu *et al.*, 2013)) and that has just recently been released (SplicingCompass (Aschoff *et al.*, 2013)). We define a true positive as a gene where at least one of the originally changed alternative exon was correctly identified by the method.

On the 'simple' dataset, jSplice provides by far the shortest running time when compared to other methods (Fig. 2a). A comparison of jSplice and DEXSeq, which is second to jSplice, shows a 28-fold improvement when accounting for the number of cores. In terms of accuracy, jSplice displays the highest specificity and sensitivity (Fig. 2b). Its *F*-score—a performance measure that combines sensitivity and specificity into one single value—is largely superior to the second best method: DEXSeq (0.77 for jSplice and 0.51 for DEXSeq). DEXSeq compares favorably against rMATS and MISO as it provides a higher sensitivity and similar specificity. DiffSplice identified only three genes with differential alternative splicing and two turned out to be true positives. SplicingCompass, for an unknown reason and despite our best efforts, kept returning an error. Of note, all but 3 true positives identified by rMATS were also identified by jSplice (Chi-square test *P*-value of 2.75×10^{-5}). To the exception of MISO, no bias in genes with low total expression was observed in any method (Fig. 2c).

As jSplice depends on the quality of the read alignment, we tested its performance using TopHat instead of STAR. Regarding the alignment performance, TopHat detected on average fewer true junctions than STAR (18 649 versus 18 820) but demonstrated a better specificity (93% versus 90%). However, the junction read count of STAR showed a better correlation to the true junction read count (Pearson's rho of 0.995 versus 0.961). In terms of jSplice's accuracy, it translated into a slightly lower specificity (89% versus 93%) with TopHat. A comparison with the true alignment, i.e. as defined by the simulation engine, revealed that the read aligner has little influence on jSplice (Supplementary Fig. S1). Lastly, the running time of the two methods differs enormously. While it took only 8 min per file for STAR, TopHat completed the task in 2 h and 34 min, a 20-fold difference.

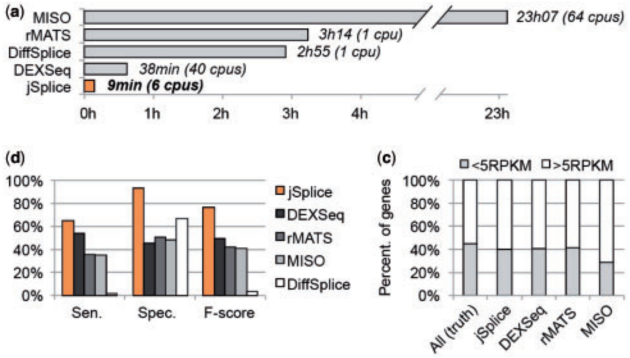

Fig. 2. Performance of jSplice on the 'simple' simulated data. (**a**) Comparison of execution time in hours (h) and minutes (min) of several methods. Multithreading was used whenever available (depending on the software and cluster load) and the number of used cores (cpus) is indicated next to each result. All pipelines were run with STAR whose execution time is not included. (**b**) Accuracy comparison of the indicated methods. (**c**) Percentage of identified genes with low expression (below 5RPKM) for each indicated method. 'All (truth)' represents the 100 genes that were originally modified to have differential alternative splicing

jSplice's method relies on junctions as a primary source of information and then complements this with exons ('optional step' in Fig. 1). In principle, this second step can be skipped resulting in a small drop in sensitivity—65% to 61%—without affecting specificity (93% versus 94%). This indicates that even though exon information is beneficial, junction reads are the main contributors to jSplice's accuracy.

Based on these results, it appears that jSplice demonstrates superiority over MISO, rMATS, SplicingCompass, DiffSplice and DEXSeq in detecting major splicing changes in the absence of biological noise. Importantly, it provides a tremendous speedup when compared to all tested software, enabling thus the analysis of large datasets.

3.2 jSplice is robust to biological noise and parameter changes

Application of jSplice and its two closely related analytical tools (DEXSeq and rMATS) on the 'complex' dataset confirmed its advantages in identifying genomic regions with consequent alternative splicing changes (Fig. 3a). Interestingly, in the presence of biological noise, DEXSeq's performance dropped. Nonetheless, if test criteria were relaxed to focus at genes only, that is irrespective of the identified exons, DEXSeq's sensitivity raises to 26% but remains inferior to rMATS and jSplice (35% and 37%). Quite noticeably, rMATS and jSplice share only two third of their true positives (Fig. 3b). However, jSplice was able to identify many true positives identified by DEXSeq but not by rMATS. This tends to indicate that jSplice is able to capture a wider and different range of splicing events than DEXSeq or rMATS. An analysis of sensitivity with respect to the amplitude of the true change revealed that jSplice's sensitivity increases with the amplitude of the change (Fig. 3c). A similar behavior is observed with rMATS and DEXSeq but rMATS seems to plateau at 50% sensitivity.

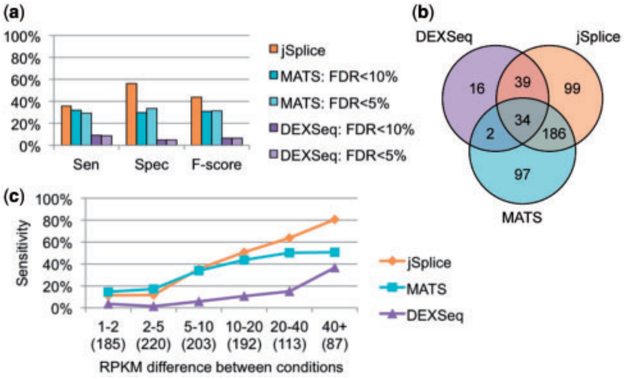

Fig. 3. Performance of jSplice on the 'complex' simulated data (50 M of 100-bp reads, aligned with STAR). (**a**) Accuracy comparison of the indicated methods. (**b**) Overlap of true positive genes identified by the different methods. (**c**) Sensitivity with respect to absolute changes in transcript expression. The number of genes per category is indicated in parentheses

jSplice relies on four user-defined thresholds (relative fold-change, minimum read count, Fisher's exact test *P*-value and inclusion percentage) and consequently their influences on the results were tested. For the simulations, true alignments were used to discard any bias induced by the read aligner. One could observe that although the removal of one or another threshold affects the sensitivity or specificity of jSplice, its *F*-score remained unchanged (Supplementary Fig. S2a). Additionally, smaller coverage (20M reads), lack of replicates, or use of 48-bp reads (Supplementary Fig. S2b, c and d), reduces the amount of available information and impacts the performance of jSplice. This however translates into a lower sensitivity and a higher specificity, affecting, in the end, only moderately the accuracy of jSplice. An increase in the total number of reads (100 M) or the use of more than two replicates did not actually improve jSplice's accuracy (Supplementary Fig. S2b and d). For the sake of comparison, we investigated the use of rMATS' parameter to threshold the amplitude of the splicing change—a parameter similar to our relative fold-change threshold—but did not observe any improvement of its accuracy, rather a trade-off between sensitivity and specificity (Supplementary Fig. S3).

Taken together, the results from the 'simple' and the 'complex' simulations suggest that the flexibility offered by the four thresholds, the presence of biological noise, or the use of different RNA-seq setups, does not influence the high specificity of jSplice. Splicing changes identified by jSplice may therefore by more likely validatable through PCR reactions of biological material.

3.3 Application of jSplice to real data sets confirms its rapidity and accuracy

Simulated data have their limitations and hence we tested jSplice on two publicly available datasets from Griffith *et al.* (2010) and Trapnell *et al.* (2012a). Griffith *et al.* performed RNA-sequencing on fluorouracil-resistant and –nonresistant human colorectal cancer cell lines (50 M of 42-bp reads, no replicates) and assessed 189 cassette exons by RT-PCR, which makes it the most extensively tested dataset for differential alternative splicing to date. Based on those results, we categorized each cassette exon as differentially expressed (DE) or not, resulting in 41 DE events (38 genes) and 77 non-DE events (77 genes) (Supplementary Table S1). Using this dataset, we observed striking differences between different methods (Fig. 4a). rMATS returned only 7 genes whereas jSplice, MISO and ALEXA-Seq yielded 280, 324 and 1724 genes respectively. Of note, all three true positives identified by rMATS were also identified by jSplice, which is consistent with the results on simulated data. Even though jSplice and MISO identified a similar number of events, the agreement between the two methods was weak; 43 common genes. Nonetheless, for a similar number of genes in the RT-PCR validated set, jSplice greatly outperformed MISO. Note that the gene selection for the PCR validation was based on the results from ALEXA-seq, which introduces a bias. DEXSeq, DiffSplice and SplicingCompass require replicates and hence could not be run on the Griffith dataset.

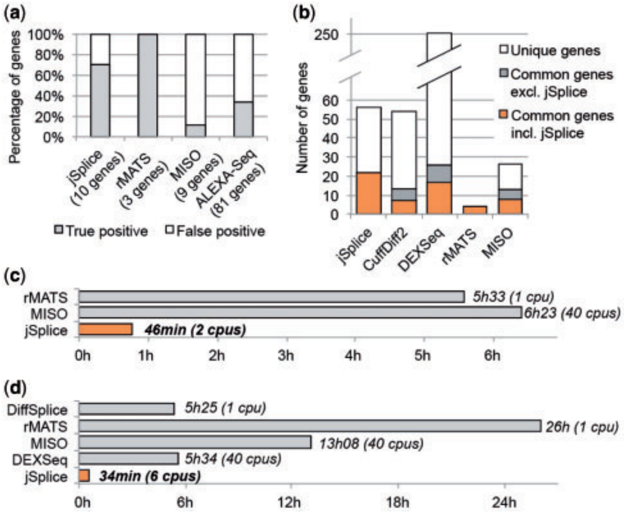

Fig. 4. Performance of jSplice on real data. (**a**) Results of rMATS, MISO, ALEXA-Seq and jSplice on the Griffith dataset. Other methods require replicates and could not be run on this dataset. The number in parenthesis represents the total number of PCR-tested genes identified by each method. For instance, among the differentially spliced genes identified by jSplice, 10 were tested by PCR by Griffith *et al.* and 7 were validated as differentially spliced. (**b**) Result comparison of several methods on the Trapnell dataset. Execution times of several methods on (**c**) the Griffith dataset and (**d**) the Trapnell dataset

Trapnell *et al.* performed RNA-sequencing on a *HOXA1* knockdown experiment in human fibroblasts and RNA-sequencing (20 M of 100-bp reads in triplicates). However, few alternatively spliced genes were validated through PCR experiments. Application of jSplice and other methods to the Trapnell dataset revealed an extensive disagreement in terms of identified genes with differential alternative splicing (Fig. 4b). SplicingCompass failed to complete the task (unknown error) and DiffSplice returned no results at a 10% FDR. DEXSeq predicted the highest number of genes with DE splicing events (250 genes), but only 10% of these were also identified by another method, suggesting low specificity. Consistent with the results on the Griffith dataset, rMATS identified few genes and all were detected by jSplice and at least another method, suggesting low sensitivity but high specificity for rMATS. In general, two thirds of the genes identified by several methods were also identified by jSplice, which hints at a high specificity. In their original publication, Trapnell *et al.* (2012a) selected 4 genes with DE transcripts for PCR validation: *TBX3*, *CDC14B*, *ORC6* and *CDK2*. However, with the exception of DEXSeq, which detected *CDC14B* and jSplice, which identified *TBX3*, no other method reported any of the four genes.

The two above-mentioned datasets represent examples of RNA-seq experiments performed with biological samples and thus can be used to assess jSplice's speed improvement over existing methods. Consistent with the results on the simulated data, jSplice provides a massive speedup over existing methods (Fig. 4c and d). Accounting for the difference in computing cores, it yielded a 65-fold improvement over DEXSeq on the Trapnell dataset. In conclusion, a test on two public real datasets confirmed the results from the simulated data—a tremendous speedup and a superior accuracy—and revealed large discrepancies in results between state-of-the-art methods.

3.4 Experimental examination of VHL tumor suppressor-regulated alternative splicing changes in renal carcinoma cell lines

Tumorigenesis is associated with profound alternative splicing changes affecting virtually all aspects of cancer biology (Oltean and Bates, 2013; Sette *et al.*, 2013) but only a handful of genes have been well characterized with respect to their splicing isoform function in tumor formation, growth and metastasis. Biallelic inactivation of the von Hippel-Lindau (*VHL*) tumor suppressor gene is a signature feature of ccRCC but its potential role in alternative splicing regulation has not been yet investigated. To investigate the matter, we engineered RCC4 renal cell carcinoma cells, which are VHL-deficient, to re-express the VHL gene product pVHL, and carried out transcriptome sequencing in cells expressing VHL (RCC4-VHL) and control cells (RCC4-EMPTY). Western blot analysis of the protein samples confirmed pVHL re-expression, as well as diminished abundance of pVHL's degradation targets HIF1 and HIF2 and repression of GLUT1, a prominent target of HIF regulation (Fig. 5a). Total gene expression levels from the RNA-sequencing results confirmed inhibition of canonical HIF target genes *EGLN3*, *LDHA* and *SLC2A1* in RCC4-VHL cells, whereas *ACTB* remained unchanged (Fig. 5b). Thus, the ±VHL RCC4 cell line pair functionally recapitulates the situation of VHL inactivation in RCC.

To obtain insight into the regulation of alternative splicing by *VHL*, we analyzed the RNA sequencing results with jSplice and identified four genes—*MYO6* (myosin VI), *DNMT3B* (DNA (cytosine-5)-methyltransferase 3 beta), *NEDD4L* (neural precursor cell expressed, developmentally down-regulated 4-like, E3 ubiquitin protein ligase) and *TMCC1* (transmembrane and coiled-coil domain family 1)—whose transcript abundance was consistently altered across three independent pVHL re-introduction experiments (Fig. 5c; genomic coordinates of the identified ASMs are given in Supplementary Table S2). A comparative analysis using SplicingCompass, DiffSplice, DEXSeq, rMATS and MISO on the pVHL re-introduction RNA-seq samples revealed no overlap across methods except for genes identified by jSplice, which were, to the exception of *TMCC1*, all identified by at least another method. Again, a huge reduction in execution time was observed for jSplice when compared to the other methods.

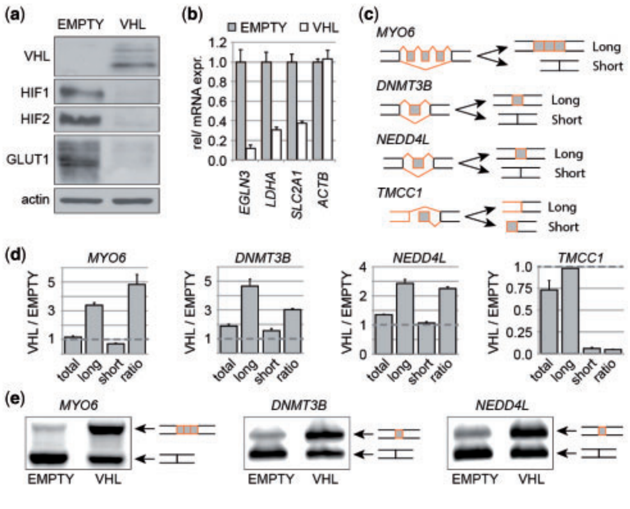

Fig. 5. Relative mRNA abundance of MYO6, DNMT3B, NEDD4L and TMCC1 isoforms as a function of pVHL status. (**a**) Western blot of protein samples isolated from the RCC4-EMPTY and RCC4-VHL cell lines with antibodies against indicated proteins. (**b**) qPCR results on RCC4-EMPTY and RCC4-VHL cell lines and indicated genes ($N=3$ independent experiments, error bars indicate SEM). (**c**) Schematic view of the four alternative splicing events identified by jSplice. (**d**) qPCR analysis of transcript expression of indicated genes. 'Ratio' represents the long/short ratio of the qPCR signal ($N=3$ independent experiments, error bars indicate SEM). (**e**) RT-PCR results. Primers were designed to bind flanking exons of the alternatively spliced regions and to give rise to the PCR products corresponding to the two alternative splicing transcripts differentially expressed in ±pVHL RCC4 cells

In order to verify the results of the software analysis, we performed real time quantitative PCR (qPCR) measurement of specific gene transcripts as well as total gene expression. In all four cases, qPCR results concurred with the in silico analysis (Fig. 5d).

The analysis by qPCR allows quantitative measurements of mRNA species across conditions, but it does not provide direct information about the relative contribution of each mRNA transcript to the total gene expression. Consequently, we carried out a semi-quantitative PCR analysis using a similar setup. Results recapitulate the quantitative isoform measurements by real-time qPCR, which are an isoform switch for *MYO6* and the up-regulation of an alternative transcript for *DNMT3B* and *NEDD4L* (Fig. 5e). Experimental validation confirmed thus the findings of jSplice.

3.5 Application of jSplice identifies tumor-specific alternative splicing events in human kidney cancer

Next, we set out to further identify genes undergoing alternative splicing in ccRCC and assess the applicability of jSplice to large-scale datasets. To this end, we analyzed a previously described genomic data from the TCGA consortium from 65 ccRCC patients (Supplementary Table S3) and identified 1225 ASMs with a relative fold-change superior to 2 in at least 5 patients (Fig. 6a and Supplementary Table S4). In this regard, we note that jSplice analyzes each patient independently and does not attempt to identify splicing events significantly changed across the whole cohort. This allows, in turn, for an independent and subsequent clustering analysis of the identified splicing events in order to identify subgroups within the cohort. Exon read counts were already provided by the TCGA and hence removed the need to perform CoverageBed calls, which in turn allowed the analysis to complete in a mere 45 min on a desktop computer. A similar analysis executed with MISO or rMATS would have taken weeks to complete and required a few terabytes of disk space. The application of DEXSeq or DiffSplice is not possible due to the availability of only one sample per patient.

Of the four splicing events identified in the VHL re-introduction experiment, only *MYO6* was also reported in the jSplice analysis of the TCGA data. *DNMT3B* is too lowly expressed to meet the read count threshold, the short transcript of *NEDD4L* is never a major one, and the short junction of *TMCC1* is not reported. Nonetheless, an expression analysis at the exon level confirmed that the splicing changes identified in RCC4 cell lines upon VHL re-introduction do also occur in patients (Supplementary Fig. S3). Of note, the VHL mutation rate in the 65 patients is of 55% but did not reveal any pattern in alternative splicing (Fig. 6a).

In the TCGA data, jSplice detected a *FGFR2* (fibroblast growth factor receptor 2) mRNA isoform switch (Fig. 6b), which had already been reported (Zhao *et al.*, 2013), and a cassette exon in *SYK* (spleen tyrosine kinase), whose tyrosine kinase protein product is a

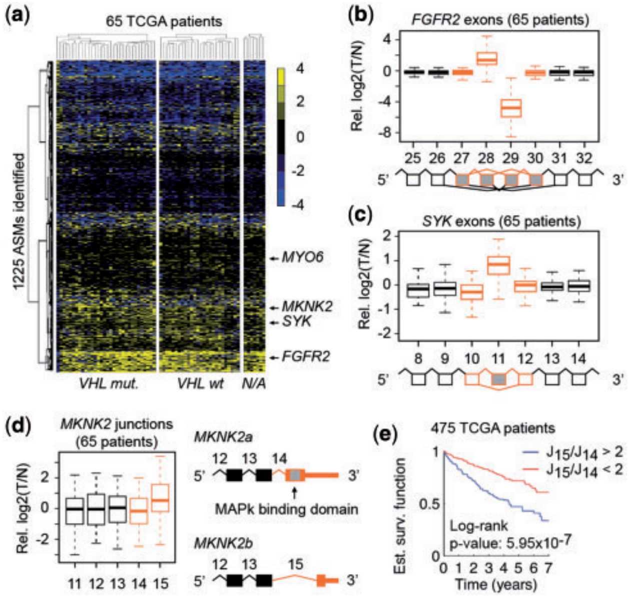

Fig. 6. jSplice analysis of human ccRCC cases from the TCGA archive. (**a**) Biclustering of patient-specific relative fold-change, as returned by jSplice, for each of the 1225 identified ASMs across the 65 TCGA patients with respect to their *VHL* mutation status. (**b**) Expression levels of *FGFR2* exons. Only the first 3 flanking exons are displayed. (**c**) Expression levels of the identified *SYK* cassette exon together with the first 3 flanking exons on each side. (**d**) Expression levels of *MKNK2* junctions. The alternative end exons are not annotated in the TCGA but the corresponding junctions are. Hence analysis was performed on junctions. (**e**) Kaplan-Meier curves of 475 ccRCC patients from TCGA, grouped by their *MKNK2b*/*MKNK2a* ratio. The latter is computed as the ratio of junction read counts. A threshold of 2 corresponds to the median. The red line represents patients above the threshold and the blue line represents patients below the threshold. In (b), (c) and (d) events identified by jSplice are outlined in orange. Splice graphs, with alternative exons in gray, were inferred from UCSC annotations

key regulator of immune signaling (Fig. 6c) (Lowell, 2011; Mocsai *et al.*, 2010). Up-regulation of the long isoform of *SYK* has been previously reported in breast and ovarian cancers (Klinck *et al.*, 2008; Prinos *et al.*, 2011; Wang *et al.*, 2003) but has not been yet associated to ccRCC.

Another interesting example from jSplice's results is the MAP kinase interacting serine/threonine kinase 2 (*MKNK2*). Alternative splicing of *MKNK2* has been recently shown to be a tumor-promoting mechanism present in breast, lung and colon cancer (Maimon *et al.*, 2014). The pro-oncogenic isoform, MKNK2b, lacks the MAPK-binding site but enhances protein synthesis. Based on junction information, we could observe that the *MKNK2b* mRNA transcript has a higher expression in tumor tissue (Fig. 6d). To assess whether its expression with respect to the *MKNK2a* transcript correlates with patient outcome, 475 ccRCC patient samples from the TCGA archive were stratified into two equally sized groups based on their *MKNK2b*/*MKNK2a* ratio. Kaplan-Meier plots showed that patients with high levels of *MKNK2b* display a significantly worse overall survival (log-rank *P*-value: 5.95×10^{-7}, Fig. 6e).

To conclude, a jSplice analysis of 65 ccRCC mRNA samples identified, in less than an hour, multiple alternative splicing events associated to ccRCC. Alternative splicing of *MYO6*, which was linked to VHL in RCC4 cells, has been confirmed in ccRCC patient data. Furthermore, this analysis revealed several other splicing events that were previously reported in other cancer types but not in clear cell renal cell carcinoma providing a new entry point for treatment for ccRCC.

4 Discussion

Recent studies concur that most if not all of human multi-exon genes yield multiple splice isoforms through alternative splicing. Alternative splicing is also highly context-dependent and promotes an enormous diversification of gene function that dictates tissue differentiation, development and, when misregulated, disease development. With transcriptomes of different tissues and disease states becoming readily available through the application of high-throughput RNA-sequencing, analytical methods to distinguish and quantify mRNA isoforms hold the potential for an unbiased and thorough insight into alternative splicing. However, data analysis is not trivial and, as a result, the development of software for fast, reliable and ultimately routine detection of alternative splicing lags currently behind the sequencing capabilities. The development of jSplice described here provides an important advance as it provides a method of unmatched specificity, sensitivity and rapidity for the analysis of AS events, key elements that currently hinder the comprehensive analysis of large-scale experiments. Thus, jSplice is highly suitable for reliable analysis of alternative splicing events in large-scale transcriptome data. Application of jSplice on RNA-seq datasets of renal carcinoma cells and human ccRCC tissues revealed novel, validatable VHL- and renal cancer-dependent alternative splicing events substantiating the value of jSplice as an analytical tool for genome-wide analysis of AS events.

jSplice relies on a novel junction-based definition of alternative splicing modules that allows for a rapid detection of alternative splicing events, simple or complex, without having to rely on a predefined set of transcripts such as implemented in rMATS or MISO. The concept of ASMs confers generality while keeping complexity at a low level but the associated software, DiffSplice, suffers from high execution time and low accuracy. Another key feature of jSplice is its entirely different method for the assessment of differential expression. That is, to focus on expression fold-change instead of modeling biological and technical variations. Models of read count distribution are necessary to estimate transcript abundance for a given experiment and to assess the statistical significance of the change. However a simple fold-change threshold, as implemented in jSplice, seems, in practice, sufficient to detect large splicing changes in comparisons of experiments. Supporting that argument, Rehrauer *et al.* (2013) recently pointed out that a fold-change measurement is sufficient to identify differential expression of genes. Of note, as the complexity of an ASM increases, so does the number of pairwise comparisons, which increases the possibility of identifying a false positive. In this specific state, jSplice may potentially overreport complex ASM. Variability between replicates is inherent to biological experiments and, in the context of our simulations, affected all tested methods. Nonetheless jSplice aims at ranking splicing events by the amplitude of the change in order to provide a list of validatable genes for experimental biologists. In that respect the larger changes are potentially the most relevant ones and jSplice showed a better accuracy than DEXSeq and rMATS in detecting those. Additionally, we demonstrated that the use of a different read aligner, different parameters, or a different RNA-sequencing setup did not impact jSplice's high specificity.

Application of jSplice to detect alternative splicing events in renal carcinoma cells as a function of pVHL status identified alternative splicing events in *MYO6*, *DNMT3B*, *NEDD4L* and *TMCC1*. Since loss of VHL represents a signature lesion in human ccRCC, it is attractive to consider that one or more of these changes in alternative splicing contribute, in part, to renal carcinogenesis. Myosin VI (*MYO6*), whose alternative splicing event was identified by jSplice in both cell lines experiment and TCGA data, moves toward the minus end of actin filaments and is involved in the transport of vesicles and organelles (Tumbarello *et al.*, 2013). The three exons that we found alternatively spliced in renal carcinoma cells encode a fragment referred to as 'long insert' located in the cargo binding tail of myosin VI (Buss *et al.*, 2001; Tumbarello *et al.*, 2013). Thus, the presence of the 'long insert' may affect motor-cargo interactions. It may also confer preferential expression in polarized cells (such as kidney cells (Buss *et al.*, 2001; Hasson, 2003)). As the long isoform of Myosin VI is more abundant in pVHL-proficient renal carcinoma cells and in healthy tissue compared to tumor, it is conceivable that a myosin VI switch from the long to the short isoform may contribute to the loss of polarization of transformed kidney cells. *DNMT3B* is implicated in chromatin methylation and has been previously shown to be differentially spliced in multiple cancer types (Klinck *et al.*, 2008; Ostler *et al.*, 2007; Saito *et al.*, 2002; Vasanthakumar *et al.*, 2013; Venables *et al.*, 2008; Wang *et al.*, 2007). Also the alternative exon identified in this study has been previously shown to discriminate between pluripotent and differentiated cells (Gopalakrishna-Pillai and Iverson, 2011). Thus, pVHL-dependent changes in *DNMT3B* splicing may be added to the growing list of chromatin modifying enzymes functionally altered in human ccRCC. *NEDD4L* has been shown to regulate TGF-beta signaling (Gao *et al.*, 2009) and its gene expression has been associated with several cancers (Hu *et al.*, 2009; Sakashita *et al.*, 2013; Tanksley *et al.*, 2013). However the reported alternative splicing event has not been explored. *TMCC1* is the least studied of the four genes identified in the jSplice analysis in renal carcinoma cells. In a recent study, it has been shown to anchor in the endoplasmic reticulum (ER) to regulate ER membrane organization and the attachment of ribosomes to the ER (Zhang *et al.*, 2014). The short isoform identified in this study would encode only the C-terminal transmembrane domain and create a loss of function.

It has been previously reported that splicing events are not necessarily specific to one cancer type, but rather that common splicing signatures exist (Klinck *et al.*, 2008; Oltean and Bates, 2013; Sette *et al.*, 2013; Venables *et al.*, 2008, 2009; Zhao *et al.*, 2013). Consistent with this notion, our jSplice analysis of RNA-seq data from human ccRCC obtained from the TCGA archive unveiled potentially generic alternative splicing events in *SYK* (previously shown to occur in breast and ovarian cancer (Klinck *et al.*, 2008; Prinos *et al.*, 2011; Wang *et al.*, 2003)), and *MKNK2* (found in breast, lung and colon cancers (Maimon *et al.*, 2014)). In the case of *MKNK2*, we observed a highly significant correlation between the relative expression of the *MKNK2* isoforms and patient outcome. Given the very encouraging reports on the recent development of orally available small molecule splicing modifiers for the treatment of spinal muscular atrophy (Naryshkin *et al.*, 2014), it will be exciting to explore the possibility of targeting *MKNK2* alternative splicing as a line of therapy in cancer.

AS analysis of large-scale patient data, such as provided by the TCGA, is currently unachievable with other software such as DEXSeq (Anders *et al.*, 2012) or rMATs (Shen *et al.*, 2014). Their long execution time or their inability to use read counts as inputs renders them impractical for large-scale analyses. In that respect, jSplice offers a convenient and fast solution. As a consequence, a pan-cancer analysis of alternative splicing in patient data is now feasible and may reveal potential splicing signatures suitable for diagnostic and therapeutic applications.

Acknowledgement

The authors would also like to thank Dr. Peter Mirtschink for his help in evaluating the Griffith dataset.

Funding

This work was supported by an Long-term EMBO Fellowship (EMBO ALTF 1120-2011) to R.F. and a Swiss National Science Foundation Grant (Project Number 310030B_138669) to W.K.

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Aschoff,M. *et al.* (2013) SplicingCompass: differential splicing detection using RNA-Seq data. *Bioinformatics*, **29**, 24–21.

Brooks,A.N. *et al.* (2011) Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.*, **21**, 193–202.

Buss,F. *et al.* (2001) Myosin VI isoform localized to clathrin-coated vesicles with a role in clathrin-mediated endocytosis. *EMBO J.*, **20**, 3676–3684.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, **29**, 15–21.

Edgar,R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Gao,S. *et al.* (2009) Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. *Mol. Cell*, **36**, 457–468.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Gopalakrishna-Pillai,S. and Iverson,L.E. (2011) A DNMT3B alternatively spliced exon and encoded peptide are novel biomarkers of human pluripotent stem cells. *PLoS One*, **6**, e20663.

Grant,G.R. *et al.* (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.

Griffith,M. *et al.* (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods*, **7**, 843–847.

Hasson,T. (2003) Myosin VI: two distinct roles in endocytosis. *J. Cell. Sci.*, **116**, 3453–3461.

Hergovich,A. *et al.* (2003) Regulation of microtubule stability by the von Hippel-Lindau tumour suppressor protein pVHL. *Nat. Cell Biol.*, **5**, 64–70.

Hu,X.Y. *et al.* (2009) Nedd4L expression is downregulated in prostate cancer compared to benign prostatic hyperplasia. *Eur. J. Surg. Oncol.*, **35**, 527–531.

Hu,Y. *et al.* (2013) DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.*, **41**, e39.

Katz,Y. *et al.* (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Klinck,R. *et al.* (2008) Multiple alternative splicing markers for ovarian cancer. *Cancer Res.*, **68**, 657–663.

Lowell,C.A. (2011) Src-family and Syk kinases in activating and inhibitory pathways in innate immune cells: signaling cross talk. *Cold Spring Harb. Perspect. Biol.*, **3**. doi:10.1101/cshperspect.a002352.

Maimon,A. *et al.* (2014) Mnk2 alternative splicing modulates the p38-MAPK pathway and impacts Ras-induced transformation. *Cell Rep.*, **7**, 501–513.

Mocsai,A. *et al.* (2010) The SYK tyrosine kinase: a crucial player in diverse biological functions. *Nat. Rev. Immunol.*, **10**, 387–402.

Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.

Naryshkin,N.A. *et al.* (2014) SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy. *Science*, **345**, 688–693.

Oltean,S. and Bates,D.O. (2013) Hallmarks of alternative splicing in cancer. *Oncogene. Science*, **33**, 5311–5318.

Ostler,K.R. *et al.* (2007) Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins. *Oncogene*, **26**, 5553–5563.

Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.

Pan,Q. *et al*. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet*., **40**, 1413–1415.

Prinos,P. *et al*. (2011) Alternative splicing of SYK regulates mitosis and cell survival. *Nat. Struct. Mol. Biol*., **18**, 673–679.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rehrauer,H. *et al*. (2013) Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching. *BMC Bioinformatics*, **14**, 370.

Saito,Y. *et al*. (2002) Overexpression of a splice variant of DNA methyltransferase 3b, DNMT3b4, associated with DNA hypomethylation on pericentromeric satellite regions during human hepatocarcinogenesis. *Proc. Natl. Acad. Sci. U. S. A*., **99**, 10060–10065.

Sakashita,H. *et al*. (2013) Identification of the NEDD4L gene as a prognostic marker by integrated microarray analysis of copy number and gene expression profiling in non-small cell lung cancer. *Ann. Surg. Oncol*., **20**, S590–S598.

Sette,C. *et al*. (2013) Alternative splicing: role in cancer development and progression. *Int. J. Cell Biol*., **2013**, 421606.

Shen,S. *et al*. (2014) rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A*., **111**, E5593–E5601.

Singh,R.K. and Cooper,T. (2012) Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med*., **18**, 472–482.

Sturgill,D. *et al*. (2013) Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics*, **14**, 320.

Tanksley,J.P. *et al*. (2013) NEDD4L is downregulated in colorectal cancer and inhibits canonical WNT signaling. *PLoS One*, **8**, e81514.

Thoma,C.R. *et al*. (2007) pVHL and GSK3beta are components of a primary cilium-maintenance signalling network. *Nat. Cell Biol*., **9**, 588–595.

Trapnell,C. *et al*. (2012a) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol*., **31**, 46–53.

Trapnell,C. *et al*. (2012b) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., **7**, 562–578.

Troilo,A. *et al*. (2014) HIF1alpha deubiquitination by USP8 is essential for ciliogenesis in normoxia. *EMBO Rep*., **15**, 77–85.

Tumbarello,D.A. *et al*. (2013) Myosin VI and its cargo adaptors - linking endocytosis and autophagy. *J. Cell. Sci*., **126**, 2561–2570.

Vasanthakumar,A. *et al*. (2013) Dnmt3b is a haploinsufficient tumor suppressor gene in Myc-induced lymphomagenesis. *Blood*, **121**, 2059–2063.

Venables,J.P. *et al*. (2008) Identification of alternative splicing markers for breast cancer. *Cancer Res*., **68**, 9525–9531.

Venables,J.P. *et al*. (2009) Cancer-associated regulation of alternative splicing. *Nat. Struct. Mol. Biol*., **16**, 670–676.

Wang,E.T. *et al*. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,J. *et al*. (2007) Delta DNMT3B variants regulate DNA methylation in a promoter-specific manner. *Cancer Res*., **67**, 10647–10652.

Wang,L. *et al*. (2003) Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer Res*., **63**, 4724–4730.

Zhang,C. *et al*. (2014) Transmembrane and coiled-coil domain family 1 is a novel protein of the endoplasmic reticulum. *PLoS One*, **9**, e85206.

Zhao,Q. *et al*. (2013) Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas. *Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res*., **19**, 2460–2472.