BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 9 2013, pages 1141–1148
doi:10.1093/bioinformatics/btt101

Gene expression

Advance Access publication February 28, 2013

SplicingCompass: differential splicing detection using RNA-Seq data

Moritz Aschoff[1,2,*], Agnes Hotz-Wagenblatt[1], Karl-Heinz Glatting[1], Matthias Fischer[3], Roland Eils[2,4] and Rainer König[2,4,5,6,*]

[1]Bioinformatics ‘HUSAR’, Genomics Proteomics Core Facility, German Cancer Research Center (DKFZ), Im Neuenheimer Feld (INF) 580, [2]Division of Theoretical Bioinformatics, DKFZ, INF 580, 69120 Heidelberg, Germany, [3]Department of Pediatric Oncology and Hematology and Center for Molecular Medicine Cologne (CMMC), University Children’s Hospital, Kerpener Str. 62, D-50924 Cologne, Germany, [4]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Bioquant, University of Heidelberg, INF 267, 69120 Heidelberg, Germany, [5]Center for Sepsis Control and Care, University Hospital Jena, Bachstrasse 18, 07743 Jena, Germany and [6]Leibniz Institute for Natural Products Research and Infection Biology, Hans-Knöll-Institute, Beutenbergstrasse 11a, 07745 Jena, Germany

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Alternative splicing is central for cellular processes and substantially increases transcriptome and proteome diversity. Aberrant splicing events often have pathological consequences and are associated with various diseases and cancer types. The emergence of next-generation RNA sequencing (RNA-seq) provides an exciting new technology to analyse alternative splicing on a large scale. However, algorithms that enable the analysis of alternative splicing from short-read sequencing are not fully established yet and there are still no standard solutions available for a variety of data analysis tasks.

Results: We present a new method and software to predict genes that are differentially spliced between two different conditions using RNA-seq data. Our method uses geometric angles between the high dimensional vectors of exon read counts. With this, differential splicing can be detected even if the splicing events are composed of higher complexity and involve previously unknown splicing patterns. We applied our approach to two case studies including neuroblastoma tumour data with favourable and unfavourable clinical courses. We show the validity of our predictions as well as the applicability of our method in the context of patient clustering. We verified our predictions by several methods including simulated experiments and complementary *in silico* analyses. We found a significant number of exons with specific regulatory splicing factor motifs for predicted genes and a substantial number of publications linking those genes to alternative splicing. Furthermore, we could successfully exploit splicing information to cluster tissues and patients. Finally, we found additional evidence of splicing diversity for many predicted genes in normalized read coverage plots and in reads that span exon–exon junctions.

Availability: SplicingCompass is licensed under the GNU GPL and freely available as a package in the statistical language R at http://www.ichip.de/software/SplicingCompass.html

Contact: m.aschoff@dkfz.de or r.koenig@dkfz.de

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

Received on May 15, 2012; revised on February 1, 2013; accepted on February 22, 2013

1 INTRODUCTION

Alternative splicing is a central cellular process that produces different mRNA isoforms from a single gene. The qualitative and quantitative identification of such isoforms is essential for understanding the different roles of alternatively spliced genes in a cell. In addition, the detection of disease-specific isoforms is an important task because aberrant splicing is known to be responsible for various diseases (reviewed in Kim *et al.,* 2008a) and associated with different cancer types (e.g. Christofk *et al*., 2008; Venables *et al*., 2009).

Isoform structures and abundances can be inferred from short sequence reads generated by RNA sequencing (RNA-seq). In contrast to hybridization-based approaches using, for example, microarrays, such sequencing-based approaches better allow for the detection of previously unknown splicing events and can avoid cross-hybridization and limited dynamic range (reviewed in Wang *et al*., 2009). Computational methods for analysing differential expression and differential splicing from RNA-seq data ‘are only beginning to emerge’ (reviewed in Garber *et al*., 2011). Existing methods detect alternative splicing mainly by considering sequencing reads that map uniquely to single isoforms or by assembling transcripts and estimating the most likely isoform abundance levels according to the given sequencing reads [an overview and discussion of limitations is given in Martin and Wang (2011) and Garber *et al*. (2011)]. Short-read assemblers have been developed for genome and transcriptome assembly [e.g. Velvet (Zerbino and Birney, 2008), Scripture (Guttman *et al*., 2010), Cufflinks (Trapnell *et al*., 2010) and others (reviewed in Martin and Wang, 2011)]. Despite the success of such approaches, assembly artifacts remain a common phenomenon (Birney, 2011), and in the context of transcriptome assembly, many reads cannot be assigned to single isoforms unambiguously. When trying to interpret these reads, different sets of

*To whom correspondence should be addressed.

possible assemblies are theoretically compatible with the sequencing data (Garber *et al*., 2011), and this leads to the misassembly of false transcripts and a wrong estimation of specific isoform expression levels. This problem becomes more severe in the presence of complex alternative splicing patterns and depends on factors like the sequencing depth and the length of generated reads. These considerations are also the basis for recent developments that aim at identifying alternative splicing patterns from De-Bruijn graphs without actually assembling complete isoforms (Sacomoto *et al*., 2012) or methods that integrate RNA-seq data with existing gene annotations and ESTs to predict more reliable splice graphs and annotate ambiguous events as unresolved (Rogers *et al*., 2012). A systematic assessment of transcriptome assemblies is difficult because appropriate quality metrics have not been established yet and require a well-defined gold standard that is difficult to find (Martin and Wang, 2011). During the revision of this article, Anders *et al*. (2012) published a tool called ‘DEXSeq’ for the detection of differential exon usage from RNA-seq data. Rather than using an assembly approach and comparing abundance levels of predicted transcripts, DEXSeq avoids the assembly step and calculates *P*-values for every annotated exon (or parts of exons if an exon modification event can be derived from the provided annotation). Similarly, our method directly identifies genes with significantly altered ratios of (potentially unknown) isoforms. The power to detect significant expression differences depends on the number of available reads and therefore on the sequencing depth and length of the region of interest. Regions that are specific to a single isoform (like cassette exons) are often much shorter compared to full transcripts, and the detection of significant differences may be severely hampered by their low number of reads. Besides this, the number of statistical tests that are required to detect previously known and unknown events for all expressed genes in the observed cells raises to such an extent that makes it challenging to avoid high false-positive levels or to reach statistical significance after using multiple testing correction. A critical aspect for all methods is the potentially complex composition of isoforms. Some genes generate many different transcripts in varying abundances that comprise a large variety of different combinations of exons. RNA-seq data gives the opportunity at hand to yield a global insight into this situation. Consequently, we developed a new method (and implemented it into the software tool ‘SplicingCompass’) to predict genes that show differential splicing taking all of their exons into account. To our knowledge, our method is the first approach that detects differential splicing on a gene level by taking into account all exons of a gene at once instead of considering single exons independently. This reduces the number of statistical tests considerably and better accounts for combined effects. We applied our method to high-throughput sequences of the transcriptomes from human brain and liver samples as well as from neuroblastoma patients with low-risk and high-risk tumours and to simulated data. In addition, we present results that show the validity of our method as well as its applicability in the context of clustering tissues and cancer patients.

2 METHODS

2.1 Data

We analysed two RNA-seq datasets and compared high-throughput sequences of the transcriptomes from four human brain and seven liver samples [data published in Au *et al*. (2010); Blekhman *et al*. (2010) and via the ‘Human BodyMap project’ from Illumina] as well as 14 neuroblastoma patients with low-risk and high-risk tumours {i.e. seven favourable [stage 4S according to the International Neuroblastoma Staging System (Brodeur *et al*., 1993)] and seven unfavourable [stage 4] outcome} (Supplementary Table S1). In addition, we used the simulation engine BEERS (Benchmarker for Evaluating the Effectiveness of RNA-Seq Software) (Grant *et al*., 2011) to simulate RNA-seq reads of known transcripts and introduced a variety of simulated differential splicing events (see Supplementary Note 1a and Supplementary Figs S1 and S2 for a detailed description).

2.2 Mapping and pre-processing

TopHat (Trapnell *et al*., 2009) was used for a splice junction sensitive mapping of sequence reads to the reference genome UCSC hg19. We used default parameters but did not allow for multihits. To normalize reads for different sequencing depths, we divided read counts by a library size parameter as described previously (Anders and Huber, 2010). To remove genes that are barely expressed, we calculated RPKM values (‘reads per kilobase of exon model per million mapped reads’) (Mortazavi *et al*., 2008) on an exon level and included genes that had at least one exon in the union transcript that was above a threshold of five RPKM averaged over the samples. Union transcripts were defined as described in the following section. We also excluded genes with only one annotated exon. With this, 127 634 (79 819) of 182 291 exons in 10 987 (6944) of 18 121 annotated genes satisfied the criterion in the neuroblastoma (brain/liver) comparison.

2.3 Splicing angles

For every gene g, we defined a union transcript $ut_g = (e_1, \ldots, e_n)$ by combining all exons e_i from all corresponding isoforms that were annotated in Consensus CDS coding sequences (CCDS) (Pruitt *et al*., 2009). In the following, we represented the expression of that gene as a vector $\bar{v}_g$, called read count vector. Every component of that read count vector corresponded to the number of sequencing reads R_{e_i} that uniquely aligned with a given exon e_i of the union transcript: $\bar{v}_g = (R_{e_1}, \ldots, R_{e_n})$. Hence, $\bar{v}_g$ combined the expression levels of all isoforms of g. Systematic differences in alternative splicing affect the relative proportions between exon-expression levels and thereby determine the orientation of read count vectors. Hence, we used the geometric angles between read count vectors as a means to measure differences in alternative splicing. Furthermore, we added additional dimensions to $\bar{v}_g$ for all observed exon–exon connections based on junction reads found by TopHat (Fig. 1). This generally increased the sensitivity of our method and helped to identify even small differences that arise from exon modification events like alternative $5'/3'$ exons (see Supplementary Note 2 for a more detailed description). If a gene has a constant composition of isoforms, i.e. the ratios between isoforms do not change between conditions, the read count vectors of that gene will be roughly parallel between conditions yielding a low angle between them, even if the overall gene expression levels differ (see Supplementary Fig. S3). Therefore, our method inherently distinguishes between differences in overall gene expression and differences in alternative splicing.

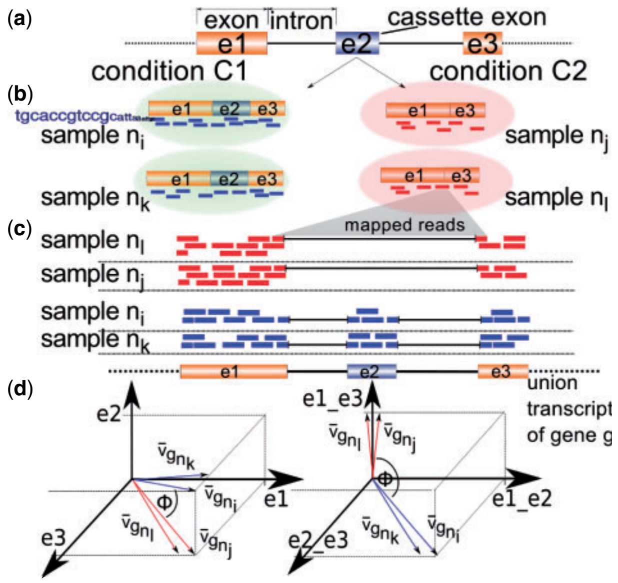

Fig. 1. Defining the angles for detecting differential splicing: Systematic differences in alternative splicing lead to systematic differences in the orientations of read count vectors. (**a**) Alternative splicing generates different mRNA isoforms that comprise different sets of exons. (**b**) The condition-specific composition of isoforms is sequenced in a high throughput manner for a number of biological replicates. (**c**) Sequence reads are mapped against the reference genome and counted for every exon of the union transcript. Reads that span known and novel exon–exon junctions are reported by TopHat. (**d**) Multidimensional read count vectors that incorporate exon-body reads (left) as well as exon-junction reads (right) are used as a combined representation of exon expression. For visualization purposes, the theoretical six-dimensional read count vectors are depicted in two different coordinate systems. Geometric angles between read count vectors bring out splicing diversity between the different conditions

2.4 Detection of differentially spliced candidate genes

To detect differentially spliced genes between two conditions c_1 and c_2 with n samples, we calculated all $\binom{n}{2}$ pairwise angles as given in equation (1)

$$\Phi_{g_{n_i},g_{n_j}} = arccos\left(\frac{\bar{v}_{g_{n_i}} \cdot \bar{v}_{g_{n_j}}}{||\bar{v}_{g_{n_i}}|| \cdot ||\bar{v}_{g_{n_j}}||}\right) \cdot \frac{180}{\pi} \qquad (1)$$

and tested if *splicing angles* within conditions were significantly smaller than *splicing angles* between conditions. $\Phi_{g_{n_i},g_{n_j}}$ is the angle between samples n_i and n_j for gene g. $\bar{v}_{g_{n_i}}$ is the read count vector for gene g and sample n_i. We performed two tests per gene. To reduce false negatives from high dimensional vectors, we restricted the dimensions of the read count vectors in the first test to exons for which a skipping or modification event was generally supported by junction reads and then used their pairwise *splicing angles* to test if these events were condition specific. Because alternative splicing events could not always be recognized from reported junction reads [Transcript variation affecting the first/last exon(s) do not give rise to simple junction read patterns that normally span nested exons and would be missed with the first test alone. In addition, tools that discover splice junction reads still suffer from false negatives (Rogers *et al.*, 2012)], we performed a second test that included all exons and all exon–exon junctions from the union-transcript. We assessed statistical significance with a one-sided *t*-test and used the Benjamini–Hochberg procedure for multiple testing correction (Benjamini and Hochberg, 1995).

2.5 Verification of candidate genes

A major issue in verifying differential splicing predictions is the lack of an established gold standard. Quantitative PCR can be used to experimentally verify a subset of genes but in general the performance of prediction methods is largely assessed by simulations. We used the simulation engine BEERS to simulate a variety of artificial differential splicing events but also pursued additional *in silico* strategies to assess the quality of our biological predictions. This included the following:

- Evaluation of splicing factor motifs, which are essential to specifically enhance or silence the splicing of alternative exons in a tissue-specific manner (sections 2.6 and 3.2).
- Assessment of the number of publications found in an unbiased automatic PubMed search (sections 2.7 and 3.3).
- Clustering of tissues and patients based on splicing information (sections 2.8 and 3.4).
- Assessment of different kinds of evidence of splicing diversity for predicted genes such as reads that span exon–exon junctions and exonic read coverage plots that were normalized for differences in gene expression (sections 2.9 and 3.1).

2.6 Splicing factor motifs

To detect splicing factor motifs, we used the R packages ‘Biostrings’ and ‘BSgenome’ (http://www.bioconductor.org) and scanned for all binding motifs characterized in the database SpliceAid2 (Piva *et al.*, 2012). We used the ‘matchPattern’ function without mismatches or indels and scanned for motifs around exon–intron boundaries (100 bps up- and down-stream) from all exons that were analysed with our method. For every splicing factor, we calculated a confusion matrix comprising the number of exons with and without nearby motifs in predicted versus remaining genes. As a control, an identical analysis was performed with the same number of exons from random genes. We assessed statistical significances with Chi-square tests and corrected for multiple testing with the Benjamini–Hochberg procedure. *P*-values for differential expression of genes that encode for splicing factors were calculated from the RNA-seq data with DESeq (Anders and Huber, 2010).

2.7 PubMed search

We performed an automatic PubMed search and counted for every predicted gene the number of publications with a co-occurrence of the corresponding gene symbol and the term ‘alternative splicing’. A control search was performed with an equal number of random genes that were sampled from the same filtered gene set (i.e. only genes were sampled for the control that were actually tested by SplicingCompass and retained after expression and single-exon filtering). The control set was drawn such that the candidate set and control set were disjunctive and no gene was sampled twice (i.e. the predicted candidate genes were removed from the control set and sampling was done without replacement).

2.8 Clustering

To demonstrate that functional information was captured by SplicingCompass within the different splicing profiles, and not merely by differential gene expression, we performed an alternative splicing-based hierarchical clustering of neuroblastoma patients and brain versus liver samples with the genes identified by SplicingCompass. Analogously to the detection of differential splicing events, we clustered using all pairwise angles between samples. In addition, we then removed all genes that were significant in a gene-expression ranking based on the commonly used RPKM measure (i.e. we removed all genes with a RPKM-based

P-value ≤ 0.1) and repeated the clustering with this reduced set. For clustering, Pearson distance and average linkage was used in all cases.

2.9 Read coverage plots

To visually compare exonic read count profiles in terms of alternative splicing, we normalized the number of reads R_{e_i,n_j} that mapped to all exons e_i from gene g in sample n_j. We normalized for exon lengths and the overall gene expression level of g in a given sample n_j by

$$R^{norm}_{e_i,n_j} = asinh\left(\frac{\frac{R_{e_i,n_j}}{l_{e_i}}}{g^{exp}_{n_j}}\right) \tag{2}$$

In which l_{e_i} was the length of exon e_i, and $g^{exp}_{n_j}$ was the overall gene expression level of gene g in sample n_j. We estimated $g^{exp}_{n_j}$ by the median number of length-normalized reads that mapped to the exons of the union transcript from gene g in the given sample. Equivalently, we normalized and depicted the number of junction reads that TopHat reported between all expressed regions (i.e. known and unknown exons) of gene g.

2.10 Comparison with DEXSeq and Cufflinks

We compared the predictions made by SplicingCompass, DEXSeq (Anders *et al*., 2012) and Cufflinks/Cuffdiff (Trapnell *et al*., 2010) for the described simulated and biological datasets. To compare DEXSeq with SplicingCompass (which assigns *P*-values to whole genes), we considered a gene differentially spliced in DEXSeq if at least one of its exons was differentially used. For Cufflinks/Cuffdiff, we followed the steps described in (Trapnell *et al*., 2012) and considered a gene differentially spliced if any of the transcripts assembled for that gene was differentially expressed according to Cuffdiff (irrespective of the fact if the assembled transcripts correctly matched the simulated ones). Strictly speaking, this definition does not only include transcript variation produced by the splicing machinery but also transcript variation due to multiple promoter and polyadenylation sites.

3 RESULTS

3.1 Differentially spliced genes

We predicted 1595 genes to be differentially spliced between brain and liver samples and 24 genes to be differentially spliced between neuroblastoma stages (P-value $\leq$ 0.1 after Benjamini–Hochberg correction, Supplementary Tables S2 and S3). Supplementary Figure S4 exemplifies one top scoring gene predicted by our method that shows a clear exon exchange event. The event is strongly supported by exon junction reads and independent annotation information from the UCSC alternative splicing track (Karolchik *et al*., 2008). Read counts in the figure are normalized as described in Methods (Section 2.9) and plotted for all exons of the union transcript from that gene. Exon 7 is specifically skipped in liver tissues, whereas exon 8 is specifically skipped in brain tissues (SLC7A2 P-value $= 10^{-32}$). The splicing events are directly supported by splice junctions. According to UCSC, exons 7 and 8 are known to undergo alternative splicing. Another example that revealed a liver-specific exon modification event in the transcripts of the gene MAGI1 is depicted in Supplementary Figures S5a and S5b. MAGI1 showed several condition-specific splicing events that affected different exons. One of these events specifically shortened exon 7 in liver samples. We used Ingenuity (Ingenuity Systems, www.ingenuity.com) and tested for gene set enrichment in the categories 'Diseases and Disorders', 'Molecular and Cellular Functions' and 'Physiological System Development and Function'. We found enrichments for 'Neurological Disease', 'Cellular Growth and Proliferation' and 'Cell Death and Survival', which point at processes important for brain tissues, while 'Protein Synthesis' and 'Nucleic Acid Metabolism' are central functions of the liver, which also plays an important role in 'Small Molecule Biochemistry' (see Supplementary Table S4). This shows that SplicingCompass identified genes that are enriched in processes relevant to the investigated tissues.

Most of the 24 genes that we predicted from the neuroblastoma data were known to undergo alternative splicing in general. Many of the genes have been implicated in processes relevant to tumour progression or neuronal events: UBE2V1 is involved in cell cycle control and can cause transcriptional activation of the human FOS proto-oncogene (Rothofsky and Lin, 1997). PELO plays a role in maintaining genomic stability (Adham *et al*., 2003), and ANKS1A affects the signalling pathway of growth factors (Pandey *et al*., 2002). PCDHAC2 is a member of the protocadherin alpha (Pchdh-α) gene family. It codes for a neural cadherin-like cell adhesion protein, which is present at synaptic junctions (Wu and Maniatis, 1999). Different Pchdh-α mRNAs are generated in a process that involves both differential promoter activation and alternative splicing (Ribich *et al*., 2006). Ribich *et al*. suggested that Pcdh-α proteins might serve as diverse but distinct synaptic tags that play an important role in establishing the 'identities' of individual neurons (i.e. they are critical in determining the highly specific interactions of synapses). PIK3R1 codes for the regulatory subunit alpha of the phosphatidylinositol 3-kinase. It influences tumour cell growth and motility and is a potential therapeutic target in glioblastoma multiforme (Weber *et al*., 2011). Multiple isoforms are annotated for PIK3R1 that differ in the first six to eight exons. Our analyses suggest an increased expression of short isoforms that exclude the first six exons in stage 4S patients compared with stage 4 patients (see Supplementary Figs S6a and S6b). The short isoforms lack the SH3 domain that mediates specific protein–protein interactions (Koyama *et al*., 1993) as well as the Rho-Gap domain that influences different regulators of the cytoskeleton (Peck *et al*., 2002). CRTAC1 codes for the cartilage acidic protein 1 and can be used as a biomarker to discriminate chondrocytes from osteoblasts in humans (Steck *et al*., 2001). A previously unknown isoform of CRTAC1 (CRTAC1-B) has been identified in brain tissue and RT-PCR confirmed that both isoforms (CRTAC1-A and CRTAC1-B) are being expressed in human (Steck *et al*., 2007). The two isoforms differ by alternate last exon usage (i.e. the most 3′ exon), which matches our in-depth analyses of the read coverage derived from the neuroblastoma data (not shown). Only CRTAC1-A contains a functional O-glycosylation site resulting in a glycosylated isoform. Recent studies in mice showed that the unglycosylated isoform CRTAC1-B (also called LOTUS) binds to Nogo receptor-1 (NgR1) and inhibits the binding of Nogo-A (Sato *et al*., 2011). Nogo-A binding destabilizes the actin cytoskeleton of the nerve cell and triggers a growth cone collapse. This induced growth cone collapse leads to a stabilization of the CNS wiring at the expense of plastic rearrangements and regeneration (Schwab, 2010) supporting the hypothesis for functional relevance in neuroblastoma cells, which originated from the primitive neural crest. Recent studies suggested that the carboxyl-terminal region (UA/EC domain) is a functional domain of CRTAC1 responsible for NgR1 binding (Kurihara *et al.*, 2012). In contrast to the 3′ exon of CRTAC1, which contains the O-glycosylation site, the exons coding for the carboxyl-terminal region (UA/EC domain) do not seem to be affected by alternative splicing directly. CRTAC1-A products could not be detected in mouse tissues by rapid amplification of cDNA ends (RACE) experiments (Steck *et al.*, 2007) suggesting that this isoform is not expressed in mouse. To our knowledge, a possible influence of the O-glycosylation motif on the binding to NgR1 has therefore not been investigated yet but seems to be a promising future aspect, in particular in the context of alternative splicing of neuroblastoma tumour cells.

3.2 Splicing factor motifs

During splicing, a protein complex known as the spliceosome assembles on the pre-mRNA to remove introns and join exons. This process is guided by short consensus motifs at the ends of introns called splice sites. In addition, a complex interplay between trans-acting splicing factors and auxiliary cis-acting regulatory elements takes place. This is essential to ensure accurate splicing and to achieve tightly regulated tissue-specific alternative splicing. Cis-acting elements can function as silencers or enhancers and are found in the vicinity of splice sites in introns and exons. In general, alternative splicing is determined by the combined effect of multiple positively and negatively acting elements, and the fate of cassette exons is decided by the presence and arrangement of surrounding motifs as well as the condition-specific ratio and modification status of splicing factor proteins (splicing regulatory principles are reviewed in Matlin *et al.*, 2005). Therefore, we investigated if a significant number of exons with specific types of splicing factor motifs can be observed in genes predicted by SplicingCompass (Supplementary Fig. S7).

For 9 out of 62 splicing factors, we found a significant enrichment or depletion of motif-associated exons (i.e. exons with at least one binding motif of the specific factor in 100 bp up- or down-stream of the exon–intron boundaries) in our predicted genes from brain versus liver tissues. This strongly suggests a special pattern of splicing regulation. For seven of the nine factors, we found an enrichment of motif-associated exons in our gene list. The factors hnRNP K and SRp55 showed a significant depletion. Notably, hnRNP K has been described for its potential to act as a splicing silencer presumably by antagonizing the binding or activity of nearby enhancers (Revil *et al.*, 2009). A depletion of silencer motifs might thus serve as a similar regulatory mechanism as an enrichment of enhancer motifs. As expected, a control analysis with the same number of exons from random genes did not yield any significant results. Many factors with enriched motifs did not show a significant difference in corresponding gene expression values. These factors might impose regulation in a more interrelated manner, and corresponding protein activities may be regulated post-translationally. For instance, Sam68 showed the most significant enrichment of motif-associated exons in our data but did not show differential gene expression. Sam68 has been described as ‘a prototype regulator of alternative splicing whose function depends on protein modification in response to extracellular cues’ (Matter *et al.*, 2002). Matter *et al.* showed in detail that phosphorylation of Sam68 is crucial in regulating the alternative splicing of a specific exon in CD44. For other factors that were differentially expressed but did not show an enrichment of motif-associated exons in the complete set of predicted genes, a distinct pattern of binding motifs could still be observed in single genes. For instance, we found an accumulation of motifs of the tissue-specific splicing factor Fox-1 in the vicinity of differentially spliced exons from SLC7A2 (Supplementary Fig. S4).

3.3 PubMed search

For 636 out of the 1595 candidate genes from the brain versus liver comparison, our automatic literature search yielded at least one publication with a co-occurrence of the corresponding gene symbol and the term ‘alternative splicing’. This is significantly more than revealed by a control search with the same number of random genes (482 out of 1595, Pearson’s Chi-squared *P*-value: 1.4e-08). For 72 candidate genes, we found more than 10 publications by automatic searching compared with 34 genes in the random set. Supplementary Figure S8 shows the number of publications found for our candidate genes and the control set.

3.4 Clustering results

Splicing differences can serve as powerful biomarkers to discriminate tissues and have a great potential to improve existing stratification methods of cancer patients. In addition, the ability to discriminate patients based on genes that do not show any distinct gene expression variation is supporting evidence for post-transcriptional differences (caused by splicing) in the gene set under consideration. For this reason, our clustering analyses were also a means to test our method.

A hierarchical clustering analysis based on predicted genes yielded two main clusters, which clearly separated all neuroblastoma patients according to their tumour stage (Supplementary Fig. S9) and all brain from liver tissues (Supplementary Fig. S10). Even after removing all genes with significant differences in their overall expression levels (i.e. we removed all genes with an RPKM-based $P\text{-value} \leq 0.1$), we could robustly cluster tissues based on differentially spliced genes (Supplementary Figs S11 and S12). As expected, an RPKM-based clustering with the same reduced gene set did not yield meaningful clusters any more (Supplementary Figs S13 and S14). This indicates that *splicing angles* exploit post-transcriptional splicing information, which is not apparent in gene-level RPKM values. Correspondingly, *P*-values computed by SplicingCompass for differential splicing only showed little correlation to *P*-values computed by DESeq for differential expression [Pearson correlation of 0.27 (0.02) for brain versus liver (neuroblastoma stage 4S versus stage 4) comparisons (see Supplementary Fig. S15)]. This further demonstrates that SplicingCompass does not merely detect differential gene expression but, at the same time, supports findings that indicate some cross-talk between splicing regulation and gene expression regulation (Kornblihtt *et al.*, 2004).

3.5 Comparison with DEXSeq and Cufflinks

We compared our method with the two well-established and well-known methods DEXSeq (Anders *et al.*, 2012) and

Cufflinks/Cuffdiff (Trapnell *et al*., 2010). SplicingCompass and DEXSeq predicted a similar amount of differentially spliced genes for the brain/liver study. In brain versus liver tissues, 1595 genes were predicted by SplicingCompass and 1489 genes by DEXSeq. In neuroblastoma patients, 24 genes were predicted by SplicingCompass and eight genes by DEXSeq. Four of these genes were predicted with both methods, and in the brain versus liver comparison 873 of the genes were predicted with both methods (see Supplementary Fig. S16). Cuffdiff did not report a significant expression difference between the Neuroblastoma stages for any transcript that was assembled by Cufflinks. For the brain/liver study, we ran Cuffdiff for 150 h on 15 processing cores before we stopped the analysis. The estimated running time for Cuffdiff on eight processing cores is around 6 h according to Trapnell *et al*. (2012) and we did not observe such excessive running times with the other datasets. In addition, we performed a comparison using simulated data in which we selected 30 genes and randomly 'knocked out' half of their transcript isoforms (experiment 1) and a second experiment, where we focused on events where multiple exons were affected in an interdependent manner (experiments 2–4, see Methods). From the 30 genes in simulated experiment 1, SplicingCompass, DEXSeq and Cufflinks/Cuffdiff detected 23, 28 and 17 events, respectively (see Supplementary Table S5). From the 10 interdependent events simulated in experiments 2–4, SplicingCompass and DEXSeq detected all events with an isoform switch of 50%, eight versus six events with an isoform switch of 30% and one versus three events with an isoform switch of 10% (see Supplementary Table S6). Transcripts that were assembled by Cufflinks in the regions of the simulated true positive events were not reported as significant by Cuffdiff. See Supplementary Note 1b for a more detailed description. Hence, SplicingCompass and DEXSeq showed a similar performance to detect these simulated differentially spliced genes, whereas Cufflinks/Cuffdiff did not report any of the events simulated in experiments 2–4.

Some of the genes found by SplicingCompass in the brain versus liver comparison that were not detected by DEXSeq involved differential splicing of unannotated exons. E.g. SplicingCompass detected a liver-specific exon in PCYT2 that was not annotated in CCDS when we downloaded the reference but, in the meantime, has been included in the current database (see Supplementary Fig. S17a and S17b). SplicingCompass detected this event because it considers all reported junction reads (even between not-annotated regions), while DEXSeq relies more on the provided annotation and therefore missed the event (see Supplementary Fig. S17c). Providing DEXSeq with the most recent annotation in this case or generally with an annotation that is constructed from the sequencing data itself probably would have also allowed it to detect the event. However, this case demonstrates that SplicingCompass can even identify differential splicing that involves unknown exons if provided with appropriate junction read information. In other cases, we saw strong evidence of differential splicing in genes predicted by SplicingCompass where DEXSeq did not reach statistical significance (see Supplementary Fig. S18 for an example of such an event in PAPSS2 that was not reported by DEXSeq). In turn, DEXSeq reported significant exons in genes that were not significant in SplicingCompass. E.g. for NR1D1 (one of the four genes that were not reported by SplicingCompass in the neuroblastoma comparison), DEXSeq reported a differential usage of exon 1. A relative coverage difference of exon 1 was also visible in the plots produced by SplicingCompass but the gene was not considered to be significant. This might be due to the fact that transcript variation affecting the first and last exons are generally less supported by junction reads, which are used by SplicingCompass if possible. However, a differential usage of the last exon in CRTAC1 that was more distinct between neuroblastoma stages was detected by both methods in good agreement.

While DEXSeq reports *P*-values on an exon level, the plots produced by SplicingCompass were often equally informative and in many cases suggested the same exons to be affected by differential splicing. In some cases, the plots produced by SplicingCompass even allowed for an interpretation of the underlying events that was more consistent with existing knowledge. E.g. in the case of PIK3R1 DEXSeq marked the first exon as differentially used (see Supplementary Fig. S6c), while the plots produced by SplicingCompass allowed to infer three isoforms that differed in the first six to eight exons, an interpretation that is consistent with annotated transcripts in CCDS (see Supplementary Fig. S6a and S6b).

4 DISCUSSION AND CONCLUSION

We presented a new approach to predict genes that undergo differential splicing. SplicingCompass is able to detect complex and previously unknown events and inherently distinguishes between gene expression and splicing differences. Our method allowed us to detect tissue-specific alternative splicing in brain versus liver samples and tumour stage-specific alternative splicing in neuroblastoma patients. We showed that a significant number of exons in our predicted genes are associated with specific splicing factor motifs and that a statistically significant number of publications link those genes to reported alternative splicing. Moreover, we successfully exploited splicing information to cluster samples of different tissues and neuroblastoma tumours. Our clustering results showed that we captured post-transcriptional modifications of genes, which are not apparent in the general gene expression. Risk stratification of neuroblastoma patients is a major challenge that led to the development of expression classifiers based on mRNA levels (Oberthuer *et al*., 2006) as well as miRNA levels (De Preter *et al*., 2011). Our results bear the perspective of improved stratification methods for the diagnosis of cancer patients based on differential splicing. In eukaryotes, intron retention is the rarest type of alternative splicing, whereas exon-skipping is the most prevalent (Kim *et al*., 2007), (Sultan *et al*., 2008). SplicingCompass currently ignores coverage differences of introns, but we plan to incorporate dimensions for observed retained introns in the next version of SplicingCompass. Mutations that cause aberrant splicing are known to be responsible for many diseases, and at least 15% of all disease-causing, single base-pair mutations affect splicing (reviewed in Kim *et al*., 2008b). The particular involvement of alternative splicing in cancer development has been demonstrated by splice variants of oncogenes and tumour suppressor genes that were found specifically in tumours from diverse tissues (reviewed in

Venables, 2004 and Hu and Fu, 2007), making alternative splicing analysis a promising approach in cancer research. The increasing amount of RNA-seq data from cancer patients now paves the way for a broad application of splicing analysis tools in this respect. We identified several candidate genes in neuroblastoma patients with yet unknown or already reported functional relevance in tumour biology but it is clear that the differentially spliced genes together with their predicted effects have to be verified experimentally.

While SplicingCompass can detect differentially spliced genes it does not aim to assemble the isoforms itself or infer isoform expression levels like assembly algorithms such as Cufflinks (Trapnell *et al.*, 2010), Scripture (Guttman *et al.*, 2010) and others. Transcriptome assembly from RNA-seq reads is challenging especially in the presence of complex alternative splicing patterns that often occur in plants and mammalians. While assembly algorithms benefit from longer reads, algorithms that estimate gene-level expression values often recommend shorter reads [e.g. Li *et al.* (2010) suggests that 20 to 25 bases is optimal]. This makes it difficult to choose the most appropriate sequencing protocol when a variety of analyses should be applied to the data. SplicingCompass works with all common read lengths as well as single and paired end data. However, exon junction reads and distance information of read pairs are well supporting the determination of the exact splicing events and isoforms involved. In addition, strand-specific protocols are used more and more to preserve the strand information of reads, which is usually lost during cDNA conversion (Levin *et al.*, 2010). If the strand information for every read is available, reads can be unambiguously assigned to exons of genes annotated on the same strand. Our software comes with a plotting functionality that allows for the visual assessment of predicted genes (see Supplementary Fig. S4). SplicingCompass enables to detect genes affected by previously unknown splicing events. We applied our method to a reference assembly for which unknown events can involve known exons from existing annotations. With appropriate junction read information, SplicingCompass can also detect unknown events involving exons not represented in the annotation. Alternatively, SplicingCompass can easily be combined with algorithms that define exonic regions from the sequencing data itself and might thus reveal additional unknown events involving newly defined exons.

ACKNOWLEDGEMENT

We thank Rolf Kabbe and Karl-Heinz Groß for IT support and Serkan Ölmez for helping with the homepage.

Funding: Joint Germany-Israeli Research Program (ca-139) (DKFZ/MOST) (to M.A. and A.H.W); Nationales Genom-Forschungs-Netz (NGFN+) for the neuroblastoma project, ENGINE (#01GS0898).

Conflict of Interest: none declared.

REFERENCES

Adham,I.M. *et al.* (2003) Disruption of the pelota gene causes early embryonic lethality and defects in cell cycle progression. *Mol. Cell Biol.*, **23**, 1470–1476.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

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

Au,K.F. *et al.* (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Birney,E. (2011) Assemblies: the good, the bad, the ugly. *Nat. Methods*, **8**, 59–60.

Blekhman,R. *et al.* (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, **20**, 180–189.

Brodeur,G.M. *et al.* (1993) Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment. *J. Clin. Oncol.*, **11**, 1466–1477.

Christofk,H.R. *et al.* (2008) The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature*, **452**, 230–3.

De Preter,K. *et al.* (2011) miRNA expression profiling enables risk stratification in archived and fresh neuroblastoma tumor samples. *Clin. Cancer Res.*, **17**, 7684–7692.

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

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

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Hu,A. and Fu,X.D. (2007) Splicing oncogenes. *Nat. Struct. Mol. Biol.*, **14**, 174–175.

Karolchik,D. *et al.* (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.

Kim,E. *et al.* (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.*, **35**, 125–131.

Kim,E. *et al.* (2008a) Alternative splicing and disease. *RNA Biol.*, **5**, 17–19.

Kim,E. *et al.* (2008b) Insights into the connection between cancer and alternative splicing. *Trends Genet.*, **24**, 7–10.

Kornblihtt,A.R. *et al.* (2004) Multiple links between transcription and splicing. *RNA*, **10**, 1489–1498.

Koyama,S. *et al.* (1993) Structure of the PI3K SH3 domain and analysis of the SH3 family. *Cell*, **72**, 945–952.

Kurihara,Y. *et al.* (2012) The carboxyl-terminal region of Crtac1B/LOTUS acts as a functional domain in endogenous antagonism to Nogo receptor-1. *Biochem. Biophys. Res. Commun.*, **418**, 390–395.

Levin,J.Z. *et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, **7**, 709–715.

Li,B. *et al.* (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

Matlin,A.J. *et al.* (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell. Biol.*, **6**, 386–398.

Matter,N. *et al.* (2002) Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature*, **420**, 691–695.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Oberthuer,A. *et al.* (2006) Customized oligonucleotide microarray gene expression based classification of neuroblastoma patients outperforms current clinical risk stratification. *J. Clin. Oncol.*, **24**, 5070–5078.

Pandey,A. *et al.* (2002) Cloning of a novel phosphotyrosine binding domain containing molecule, Odin, involved in signaling by receptor tyrosine kinases. *Oncogene*, **21**, 8029–8036.

Peck,J. *et al.* (2002) Human RhoGAP domain-containing proteins: structure, function and evolutionary relationships. *FEBS Lett.*, **528**, 27–34.

Piva,F. *et al.* (2012) SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum. Mutat.*, **33**, 81–85.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Revil,T. *et al.* (2009) Heterogeneous nuclear ribonucleoprotein K represses the production of pro-apoptotic Bcl-xS splice isoform. *J. Biol. Chem.*, **284**, 21458–21467.

Ribich,S. *et al.* (2006) Identification of long-range regulatory elements in the protocadherin-alpha gene cluster. *Proc. Natl Acad. Sci. USA*, **103**, 19719–19724.

Rogers,M.F. *et al*. (2012) SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol*., **13**, R4.

Rothofsky,M.L. and Lin,S.L. (1997) CROC-1 encodes a protein which mediates transcriptional activation of the human FOS promoter. *Gene*, **195**, 141–149.

Sacomoto,G.A. *et al*. (2012) KISSPLICE: *de-novo* calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, **13** (**Suppl. 6**), S5.

Sato,Y. *et al*. (2011) Cartilage Acidic Protein-1B (LOTUS), an endogenous Nogo receptor antagonist for axon tract formation. *Science*, **333**, 769–773.

Schwab,M.E. (2010) Functions of Nogo proteins and their receptors in the nervous system. *Nat*. *Rev*. *Neurosci*., **11**, 799–811.

Steck,E. *et al*. (2001) Chondrocyte expressed protein-68 (CEP-68), a novel human marker gene for cultured chondrocytes. *Biochem*. *J*., **353** (Pt. 2), 169–174.

Steck,E. *et al*. (2007) Chondrocyte secreted CRTAC1: a glycosylated extracellular matrix molecule of human articular cartilage. *Matrix Biol*., **26**, 30–41.

Sultan,M. *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat*. *Biotechnol*., **28**, 511–515.

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

Venables,J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res*., **64**, 7647–7654.

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

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat*. *Rev*. *Genet*., **10**, 57–63.

Weber,G.L. *et al*. (2011) Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells. *Oncotarget*, **2**, 833–849.

Wu,Q. and Maniatis,T. (1999) A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*, **97**, 779–790.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*., **18**, 821–829.