Alternative splicing analysis benchmark with DICAST

Amit Fenn [1,2,†], **Olga Tsoy** [2,†], **Tim Faro**[1], **Fanny L.M. Rößler**[1], **Alexander Dietrich**[1], **Johannes Kersting** [1], **Zakaria Louadi**[1,2], **Chit Tong Lio** [1,2], **Uwe Völker** [3,4], **Jan Baumbach** [2,5], **Tim Kacprowski** [6,7,†] **and Markus List** [1,*,†]

[1]Chair of Experimental Bioinformatics, Technical University of Munich, 85354 Freising, Germany, [2]Institute for Computational Systems Biology, University of Hamburg, Notkestrasse 9, 22607 Hamburg, Germany, [3]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Felix-Hausdorff-Straße 8, D-17475 Greifswald, Germany, [4]DZHK (German Centre for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany, [5]Institute of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5000 Odense, Denmark, [6]Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of Technische Universität Braunschweig and Hannover Medical School, Braunschweig, Germany and [7]Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany

ABSTRACT

Alternative splicing is a major contributor to transcriptome and proteome diversity in health and disease. A plethora of tools have been developed for studying alternative splicing in RNA-seq data. Previous benchmarks focused on isoform quantification and mapping. They neglected event detection tools, which arguably provide the most detailed insights into the alternative splicing process. DICAST offers a modular and extensible framework for analysing alternative splicing integrating eleven splice-aware mapping and eight event detection tools. We benchmark all tools extensively on simulated as well as whole blood RNA-seq data. STAR and HISAT2 demonstrated the best balance between performance and run time. The performance of event detection tools varies widely with no tool outperforming all others. DICAST allows researchers to employ a consensus approach to consider the most successful tools jointly for robust event detection. Furthermore, we propose the first reporting standard to unify existing formats and to guide future tool development.

INTRODUCTION

Alternative splicing (AS) affects around 95% of eukaryotic genes with multiple exons (1,2) and gives rise to a large number of isoforms. AS is involved in cellular processes and disease development (see recent reviews on cancer (3), muscle (4) and neuron (5) development). The most popular technology to study AS is short-read RNA sequencing (RNA-Seq). Possibilities for AS analysis from short-read RNA-Seq data comprise splice-aware mapping, *de novo* transcriptome assembly, AS detection and/or quantification on the exon, isoform, or event level as well as differential splicing analysis. Each year, several new tools are published for each of these analysis types.

Existing benchmark studies that could guide users on which tool to use for which analysis have several limitations (6–16). First, the studies use as ground truth either only a small subset of well-studied genes or simulated RNA-Seq data with randomly introduced AS events and limited event types. Second, they focused on AS detection tools that operate on the isoform level (6–11), splice-aware mapping tools (12,13), and differential splicing analysis tools (14–16). However, AS detection tools that operate on the event level are more precise. Finally, benchmark studies are seldomly updated.

To perform a standardised benchmark, we created a modular pipeline called DICAST (Docker Integration for Comparison of AS tools). It uses Docker to handle the installation process and Snakemake to handle AS analysis and the benchmark workflow. Its modularity allows for adding new tools in the future and for quickly updating the benchmark. As ground truth, we used RNA-Seq data sets that were simulated genome-wide and contained a predefined number and distribution of AS events. The simulation with ASimulatoR (17) allowed us to evaluate the challenges and limitations of the tested tools systematically across data sets with varying levels of difficulty.

While our main focus was to benchmark AS event detection tools, we started from splice-aware mapping to provide the best possible input to those tools. Thus, we evaluated 11 splice-aware mapping tools in addition to eight AS

detection tools. We demonstrate the benefits and limitations of AS event detection tools and suggest a putatively optimal strategy for comprehensive AS analysis.

MATERIALS AND METHODS

DICAST

DICAST uses Docker for full reproducibility and to simplify deployment (18,19). Each docker image in DICAST can be used individually or as part of the benchmark/AS analysis pipeline. DICAST orchestrates the pipeline using Snakemake (20). The DICAST source code and documentation are available at https://github.com/CGAT-Group/DICAST and https://dicast.readthedocs.io/en/master/contents.html, respectively.

Benchmark workflow

Supplementary Figure S1 illustrates the workflow for benchmarking AS analysis tools. RNA-Seq data sets were simulated using the R package ASimulatoR (17) with a sequencing depth of 200 million reads (read length 76 bp) based on the human genome hg38 and the Ensembl genome annotation, version 99 (21). We limited the simulation to chromosomes 1–22, X, Y and MT.

We tuned the complexity of the simulated data sets changing AS events distribution, combinations, and sequencing error rate (Table 1). We started from the data set S0 with equal proportions of the four main types of AS events and 0% sequencing error rate. Using this data set, we filtered out tools that demonstrate poor performance or long run time. Next, we added MES, ALE, and AFE to simulate data set S1 and to evaluate the impact of complex AS events on tool performance. For data set S2, we raised the sequencing error rate up to 0.1% and investigated the impact of not-perfect sequencing. Data sets S0–S2 contain only one event per transcript which is an artificial setting helping us to explore performance under simpler and well-controlled conditions. As all AS event types combinations are possible, we evaluated how the AS event types combinations impact the performance: for data set S3 we allowed two types of events to occur within one transcript but with different exons; for data set S4 we allowed two types of events to occur with the same exon. We also aimed to evaluate AS analysis tools in a biologically relevant setting but the real RNA-Seq data sets lack the genome-wide ground truth. We addressed this challenge and simulated a data set S5 with realistic event proportions, which were obtained by analyzing 117 RNA-Seq samples from the SHIP cohort (the Study of Health in Pomerania) (22).

For each simulated data set, ASimulatoR estimated how many genes have enough exons to create AS events and can, hence, be used for the simulation. For the simulated set S3, this number was the lowest—37 648. We used this number of genes as a parameter for all simulated data sets to preserve the read coverage level.

We used seqtk (available at https://github.com/lh3/seqtk) to downsample fastq files uniformly at random to 10 million reads to benchmark splicing-aware mapping tools. We then mapped these data sets to the human genome using 11

splice-aware mapping tools and calculated the proportion of unmapped reads relative to the number of all simulated reads ('fraction of unmapped reads') and the proportion of correctly mapped reads and junctions relative to the number of all mapped reads ('precision'). Precision and fraction of unmapped reads for splice-aware mapping tools were calculated based on the resulting alignment file and the correct coordinates provided by ASimulatoR.

We chose the mapping tool with the best balance between precision and fraction of unmapped reads values and used it to produce alignments for 8 AS detection tools. We analyzed fastq files with 200 million reads and additionally used seqtk to downsample them to 50 and 100 million reads. We then calculated the proportion of correctly found events relative to the number of events in the simulated set ('recall') and relative to the number of events found by the tool ('precision'). Precision and recall for AS event detection tools were calculated based on the unified output of DICAST and the description of correct events provided by ASimulatoR.

We estimated the variability in performance of splicing-aware mapping tools and AS detection tools by simulating five samples for the most simple data set (S0) and five samples for the biologically-relevant data set (S5).

SHIP cohort

The Study of Health in Pomerania (SHIP) is a longitudinal population-based cohort study located in the area of West Pomerania (Northeast Germany). For RNA-Seq analysis, total RNA was extracted from whole blood with a mean RNA integrity of 8.5. Based on 500ng total RNA per sample, a library was prepared using the TruSeq Stranded mRNA kits (A and B) with 24 barcodes and 117 samples were sequenced on Illumina HiSeq 4000, $2\times$ 101 bp paired-end reads with a sequencing depth 40 million clusters per sample. Written informed consent was obtained from SHIP-TREND study participants, and all protocols were approved by the institutional ethical review committee in adherence with the Declaration of Helsinki. We used these 117 RNA-Seq samples from the SHIP-TREND cohort (22) in this study. They were mapped to the human genome hg38 with STAR using the Ensembl genome annotation, version 99, and analyzed further with MAJIQ (23). Custom scripts in Python 3 were used to obtain the number of AS events and genes with AS.

RESULTS

Tools for benchmark

We collected splice-aware mapping and AS event detection tools released between 2010 and 2021 based on a Pubmed and Google Scholar search with the following inclusion criteria. A tool should be (i) available, (ii) documented, (iii) under an open-source license, (iv) available as stand-alone software (web-services can usually not process a large amount of data), (v) used in project(s) other than described in the tool publication, (vi) able to process widely used file formats such as fasta/fastq, gtf/gff3, and bam/sam. Since we controlled the number of AS events in a custom genome annotation, a tool should be (vii) able to work with those as

Table 1. Characteristics of the simulated data sets. 'Transcripts per gene' indicates the number of alternative transcript variants simulated for a gene; 'Events per transcript' indicates the number of AS events simulated within one transcript; 'Events per exons' indicates the number of AS events which an exon can be simultaneously involved in (e.g., exon skipping and alternative splice site)

Simulated data set	Transcripts per gene	Events per transcript	Events per exon	Event types							Sequencing error rate
				ES	IR	A5	A3	MES	ALE	AFE	
S0	1	1	1	x	x	x	x				0%
S1	1	1	1	x	x	x	x	x	x	x	0%
S2	1	1	1	x	x	x	x	x	x	x	0.1%
S3	>=1	2	1	x	x	x	x	x	x	x	0.1%
S4	>=1	2	>=1	x	x	x	x	x	x	x	0.1%
S5	>=1	2	>=1	x	x	x	x				0.1%

well. The final list of tools contains 11 splice-aware mapping tools and 8 AS detection tools (Table 2).

Alternative splicing analysis and benchmark pipeline

A unified output format for AS event detection. The main challenge for comparing AS event detection tools is the lack of standard output format. For an exon skipping event, AS-GAL reports the coordinates of neighboring exons; ASpli reports the coordinates of a skipped exon itself; MAJIQ reports the coordinates of a neighboring exon and its junctions. We propose a unified format for all AS event types that reports the coordinates of (i) skipped exons for exon skipping, multiple exon skipping, mutually exclusive exons; (ii) a retained intron for intron retention; (iii) an alternative part of an exon for alternative splice sites (Supplementary Figure S2).

Additionally, for each event, the unified output format contains the gene name, chromosome, strand as well as a unique ID.

DICAST: Docker Integrated Comparison of Alternative Splicing Tools. The next necessity for the standardization of the AS analysis and benchmark process is a unified pipeline. DICAST handles (i) the installation process for every tool using Docker; (ii) the execution of tools using Snakemake; (iii) the output format unification; (iv) the comparison of AS events detected by different tools.

The general workflow starts with short-read RNA-Seq data (optionally simulated by ASimulatoR (17), Figure 1) that serves as input for splice-aware mapping tools. The resulting alignment files then serve as input for AS event detection tools. Next, DICAST converts the output files of the AS event detection tools to the unified format described above, compares detected events across the tools, and reports the results of the comparison. A user could run all or separate steps of the workflow with their RNA-Seq reads, alignment and genome annotation files.

Benchmark results: splice-aware mapping tools

Splice-aware mapping tools differ in the mapping approach and their use of genome annotations. Most tools use variations of the seed-and-extend algorithm and start with aligning parts of a read (seeds). ContextMap and MapSplice2 first align reads end-to-end and use the seed-and-extend algorithm for reads that could not be aligned in the first step. Some tools can use genome annotations for additional information (e.g., splice sites) (Table 2), while others (e.g. BBmap) do not need it.

We provided a genome annotation where possible and used the best mode for AS analysis (e.g. a 2-pass mode for STAR). We downsampled the input files to 10 million reads uniformly at random to reduce the analysis time. For data sets S0 and S5 we repeated the alignment with five simulated datastets to evaluate the robustness of the performance.

The mapping approach has a limited effect on the performance of the tool. ContextMap and MapSplice2 show comparable performance metrics as the tools with the best performance such as STAR and HISAT2 (Figure 2, Supplementary Figure S3). Tools that use genome annotation generally show better performance. Surprisingly, the complexity of the data sets only marginally affects the performance of the tools. Only DART suffers from a decrease in precision with an increasing sequencing error rate. Splice-aware mapping tools are robust in terms of precision (Figure 2B): the standard deviation is less than 0.02. Concerning the fraction of unmapped reads we observe two categories: BBMap, CRAC, DART, HISAT2 and STAR perform well with a low fraction of unmapped reads and a low standard deviation while others perform poorly in both metrics.

In summary, STAR and HISAT2 showed a balance between precision and fraction of unmapped reads compared to other tools. ContextMap demonstrated the best precision and could be considered for the analysis if the sequencing depth of samples is deep enough and a user can allow the loss of 10–15% of reads. Unfortunately, ContextMap has long run time (see run time subsection below) which might be inconvenient when analyzing a large number of samples. We chose STAR for further analysis, as it is more widely used (28 433 Google Scholar citations by January 2023). We downsampled data sets to 200, 100 and 50 million reads uniformly at random to investigate the effect of sequencing depth on performance.

Alternative splicing detection tools

AS event detection tools detect events from genome annotation (ASpli, SGSeq (annotated transcripts), IRFinder), RNA-Seq alignment (EventPointer, SGSeq (*de novo*), MAJIQ), or both, i.e. by augmenting the genome annotation through alignment (ASGAL, splAdder, Whippet). As most tools (Table 2) do not detect mutually exclusive exons and

Table 2. The list of AS analysis tools chosen for benchmark

Name (version)	Dependencies	Reference	Genome Annotation*
A. Splice-aware mapping tools			
BBMap (38.94)	Java 7+	(24)	-
ContextMap2 (2.7.9) In the current study used with bowtie 2 (25)	Java, BWA, bowtie 1, bowtie2	(26)	+
CRAC (2.5.2)	Perl, htslib	(27)	-
DART (1.4.6)	GCC, GNU make, libboost-all-dev, libbz2-dev, and liblzma-dev	(28)	-
GSNAP (2020-03-12)	GCC,GNU make, Perl	(29)	+
HISAT2 (2.2.1)	GCC, GNU make, MSYS, zlib	(30)	+
MapSplice2 (2.2.1)	GCC 4.3.3+, GNU make, python 2+	(31)	+
minimap2 (2.17)	None (precompiled binaries) or GCC, GNU make, zlib	(32)	-
segemehl (0.3.4)	GCC, GNU make, htslib	(33)	-
STAR (2.7.5)	GCC, GNU make	(34)	+
Subjunc (2.0.0)	None (precompiled binaries) or GCC, GNU make	(35)	+
B. Alternative splicing event detection tools			
Name (version)	Dependencies	Reference	Supported events
ASGAL (1.1.6)	python3.6+, biopython, pysam, gffutils, pandas, cmake, samtools, zlib	(36)	ES, IR, A5, A3
ASpli (1.12.0)	R, BiocManager	(37)	ES, IR, A5 A3
EventPointer (2.4.0)	R, BiocManager	(38)	ES, IR, A5, A3, MEE, MES
IRFinder (1.3.1)	GLIBC 2.14+, GCC 4.9.0+, Perl 5+, STAR 2.4.0+, samtools 1.4+, bedtools 2.4+	(39)	IR
MAJIQ (2.3)	htsilb, python3, python packages	(23)	ES, IR, A5, A3
SGSeq (1.24.0)	R, BiocManager	(40)	ES, IR, A5, A3SS, AFE, ALE, AF, AL, MES (with two skipped exons)
splAdder (2.4.3)	python3, python packages	(41)	ES, IR, A5, A3, MEE, MES
Whippet (0.11.1)	julia	(42)	New junctions could be added from alignment as an option ES, A3SS, A5SS, IR, AFE, ALE, tandem transcription start, tandem alternative polyadenylation, circular back splicing

* '-': does not use; '+' - can use as an option

ES, exon skipping; IR, intron retention; A3, alternative 5′-splice site, A5, alternative 5′-splice site, MES, multiple exon skipping, MEE, mutually exclusive exons, AFE, alternative first exon, ALE, alternative last exon.

multiple exon skipping, we treated such events as exon skipping events.

Figure 3A (upper part) shows results for data set S0. After examining the results, we excluded EventPointer, SGSeq (*de novo*), and splAdder from the further detailed analysis as these tools demonstrate low recall (less than 10% of recovered events). We also excluded ASGAL because of the long run time (a genome-wide analysis of 50 million reads took around 3 days). Supplementary Figure S4 shows results for data sets S1-S4. The ranking of the tools differs only slightly. Whippet and SGSeq_Anno (annotated transcripts) show the best performance. For all tools in data sets S0-S4, recall depends on the sequencing depth, while precision does not.

MAJIQ was used to derive the proportion of AS event types and combinations from the SHIP cohort (Supplementary Figure S5). We used the proportions as a parameter for ASimulatoR to simulate the biologically-inspired set S5. Note that only the proportion values, but not the detected AS events were used, so MAJIQ and the other tools re-

mained unaware of exact AS events in S5. We used MAJIQ here, since it does not depend on genome annotation and can extract events directly from alignments. The results for data set S5 (Figure 3A, lower part) demonstrate almost the same ranking of the tools.

Additionally, we investigated the ability of tools to detect events *de novo*. For this purpose, we edited a genome annotation file used as input for AS detection tools. Genome annotation describes the exon composition for all transcripts simulated from a gene: main and alternatives. We truncated annotations and kept only the description of the main transcript for each gene. We denote this data set as S5-tr(uncated).

The tested tools differ significantly in their abilities to detect *de novo* events based on truncated genome annotation. ASpli and SGSeq (annotated transcripts) did not detect any novel events. For other tools, we compared precision and recall for S5 and S5-tr for each AS event type individually (Figure 3B). Whippet did not find any *de novo* intron retention events. The recovery rate of alternative splice sites
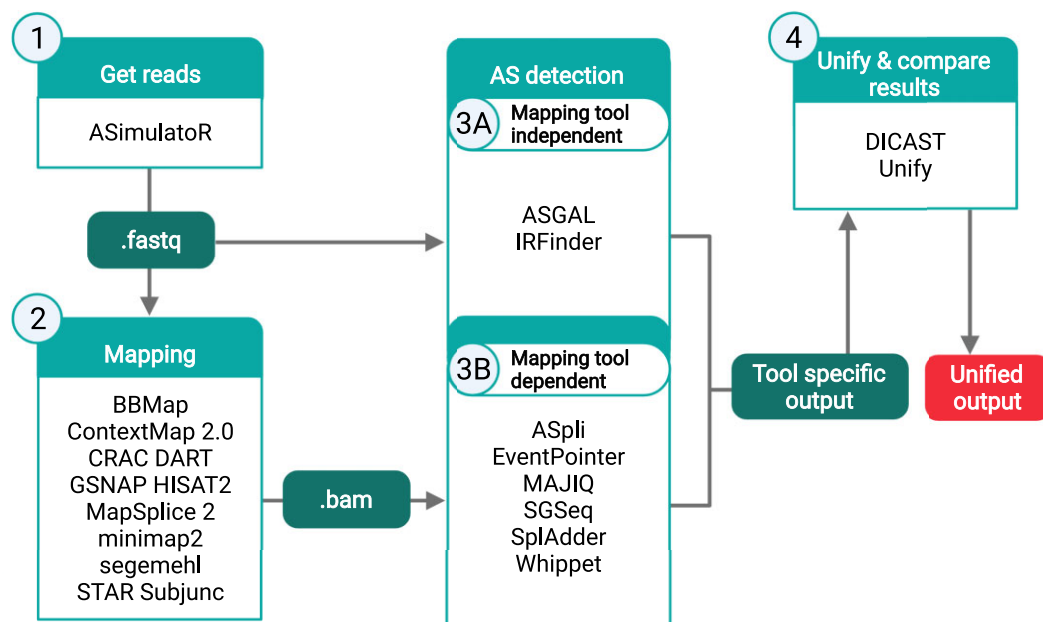
Figure 1. DICAST framework: 1) simulated (ASimulatoR) or user-provided fastq files 2) bam files could be generated by any of 11 supported splice-aware mapping tools; 3) AS events detected by any of 8 AS event detection tools based on files generated in the previous steps; 4) the output files of the AS detection tools are unified by DICAST. Created with BioRender.com.

also dropped for all tools: for MAJIQ., precision decreased from around 0.9 to 0.5; recall decreased from 0.4 to 0.1; for Whippet, precision decreased from around 0.7 to 0.4; recall decreased from 0.8 to 0.2 (evaluated for 50M reads).

For data sets S0 and S5 we repeated the AS detection with the full genome annotation five times to evaluate the robustness of the performance. All evaluated AS detection tools demonstrated robust performance (Figure 3C). On data set S5 with the most complex combinations of AS events, the increasing sequencing depth leads to increasing recall but slightly decreasing precision.

Notably, the results of different tools show comparably little overlap (Figure 4). Most events are detected exclusively by one tool (e.g., Whippet), suggesting that an integrative approach is needed for the comprehensive analysis of AS events from short-read RNA-Seq data: (i) detect known events using tools based on genome annotation (e.g. ASpli, Whippet); (ii) detect intron retention *de novo* using IRFinder; (iii) detect *de novo* events using Whippet and/or MAJIQ.

Run time

Using the data set S0, we estimated the run time of the tools tested on Intel(R) Xeon(R) Gold 6148 Processor with 27.5M Cache, 2.40 GHz (Figure 5). For the mapping tools, we used the downsampled data set with 10 million reads from the benchmark. Most tools perform indexing and mapping within an hour. ContextMap is the slowest tool with a run time of around 11 h. minimap2 and BBmap are the fastest tools with a run time of only several minutes. We also es-

timated the run time of AS detection tools depending on the sequencing depth (Figure 5). The run time of MAJIQ does not depend on the sequencing depth and is in the range of minutes. Most other tools run within several hours, with splAdder taking up to 10 hours for 200 million reads. AS-GAL was only evaluated with 50 million reads, already running longer than 72 h to complete.

DISCUSSION

We investigated the performance of AS event detection tools in a comprehensive benchmark. We started with splice-aware mapping tools to obtain the best possible input for event detection. Among splice-aware mapping tools, STAR and HISAT2 present the best balance between the precision value, fraction of unmapped reads values, and the run time, which is in concordance with previous findings (12).

Concerning AS event detection, we still find much room for improvement since tools with high recall values (Whippet, SGSeq (annotated transcripts), ASpli) can not detect events *de novo*. Vice versa, tools that can identify events *de novo* demonstrate poor recall values. We suggest using a combination of existing tools for a comprehensive AS analysis on the event level using short-read RNA-Seq data.

While we aimed for a comprehensive, objective, and standardized benchmark, we faced some limitations. First, the simulated data sets might not reach the same level of complexity as real biological data sets. We mitigated this by mimicking the proportions of AS event types as observed in the SHIP cohort. Second, for run time reasons, we refrained from a tool-specific parameter tuning and relied on
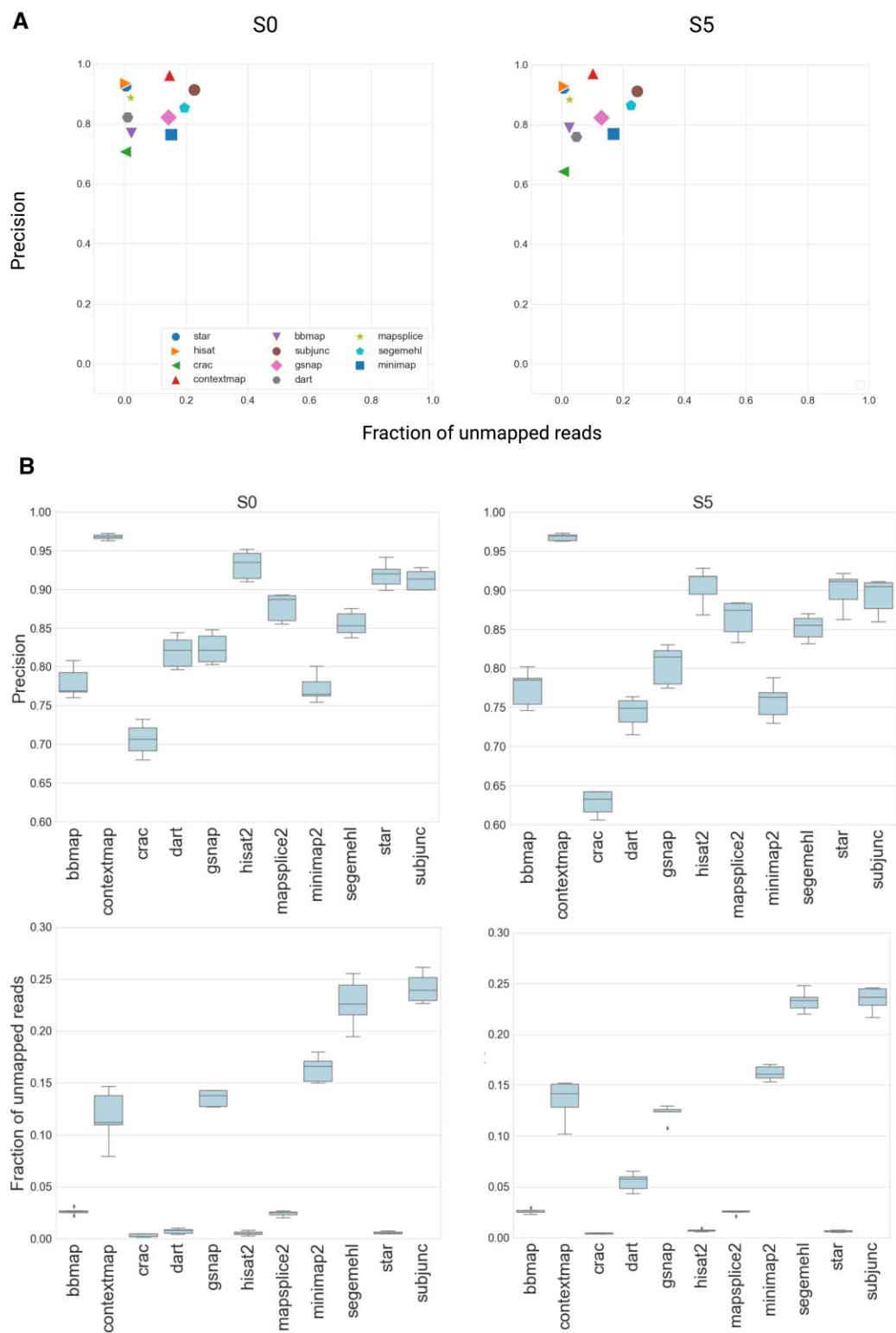
Figure 2. (**A**) The plots for precision and fraction of unmapped reads plots for splice-aware mapping tools for the data sets S0 and S5. The results for S1–S4 are presented in Supplementary Figure S3. The values are taken from the analysis of one out of five simulated datasets as an example. (**B**) The boxplots for precision and fraction of unmapped reads for splicing-aware mapping tools calculated for S0 (repeated 5 times) and S5 (repeated 5 times)
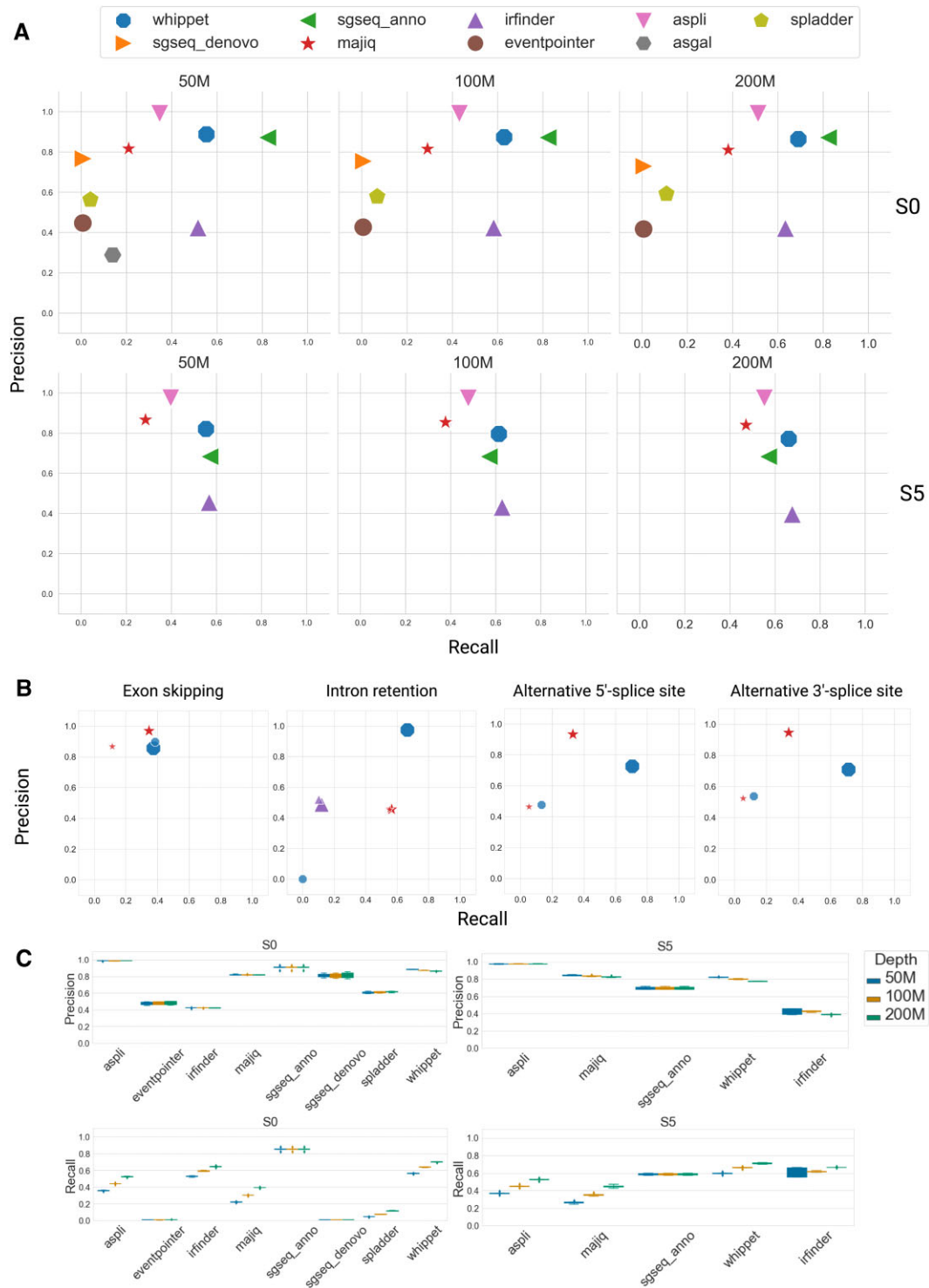
Figure 3. (**A**) Precision/recall plots for AS event detection tools for data sets S1 and S5. The values are taken from the analysis of one out of five simulated datasets as an example. (**B**) Tool performance on S5 (bigger markers) and S5-tr (smaller markers) sets by AS event type for the sequencing depth of 50M reads.(**C**) Precision/recall boxplots for AS event detection tools for the S0 (repeated 5 times) and S5 (repeated 5 times)
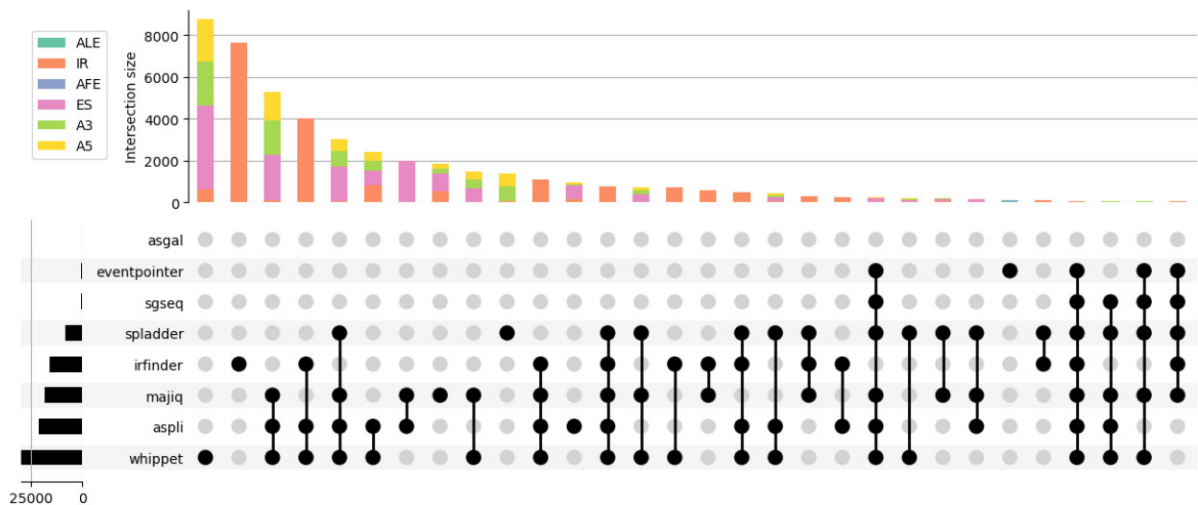
Figure 4. The UpSet plot shows the intersection of results from evaluated AS detection tools.
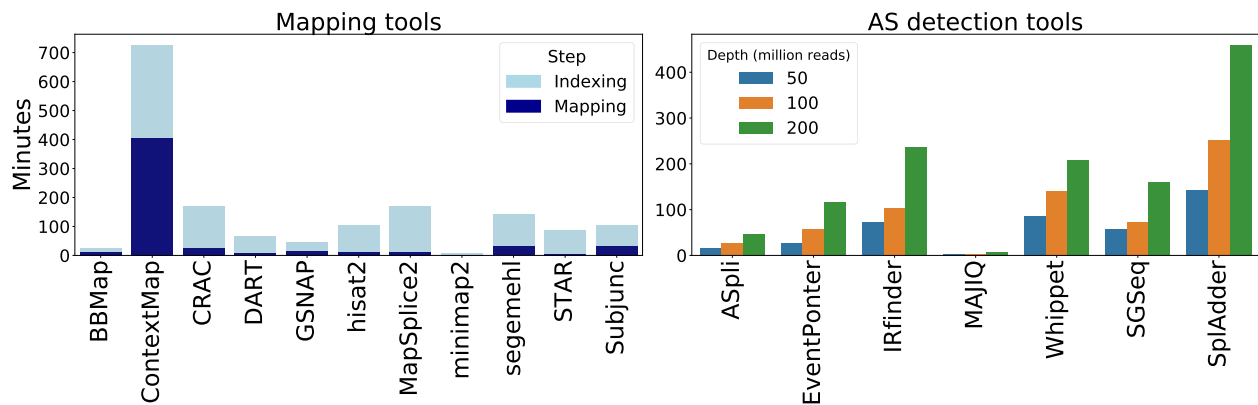
Figure 5. The run time in minutes of the splice-aware mapping tools and AS event detection tools.

the default parameters, which should ideally lead to optimal results in typical settings. Additionally, we did not want to favor tools with higher flexibility simply due to an increased number of parameters to tune. Finally, we did not explore the performance of the machine and deep learning-based tools (e.g. DARTS (43), IRFinder-S (44)). The simulated AS events were added randomly and do not account for, e.g., regulatory sequences that are used in machine learning approaches. Finally, our benchmark is based on a human genome annotation. The results might differ for organisms with dissimilar AS patterns (e.g. for plants). However, Baruzzo et al. (12)showed that at least the performance of RNA-Seq mapping tools does not differ dramatically between human and *Plasmodium falciparum* genomes.

Many challenges in AS analysis have yet to be addressed, including general difficulties for tool development such as the lack of efficient parallelization and substantial run time. We found that the standard format for aligned reads—bam/sam/cram—differs between tools and might not be compatible with some AS event detection tools. For instance, only few mapping tools add an XS tag that indicates the strand orientation of an intron which is needed by tools such as SGSeq. We gathered the information about tools compatibility here: https://dicast.readthedocs.io/en/master/tools/tools.html. This incompatibility limits the us-

ability of many splice-aware alignment tools for AS analysis. While DICAST introduces a unifying standard for AS event reporting, AS event detection tools utilize inherently different approaches and lead to inconsistent results. To mitigate this, DICAST allows users to execute any combination of tools and facilitates adding newly published tools. By standardizing the output of AS event detection tools, DICAST significantly simplifies downstream analysis. In summary, DICAST offers a unified interface for existing methods and boosts method development by offering an easily extensible framework for benchmarking of existing and novel AS analysis approaches.

DATA AVAILABILITY

Access to the SHIP data for research purposes may be requested at https://www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php. The description of simulated data sets and the corresponding R scripts are available at https://doi.org/10.5281/zenodo.7573144.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

REFERENCES

1. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
2. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
3. Bonnal,S.C., López-Oreja,I. and Valcárcel,J. (2020) Roles and mechanisms of alternative splicing in cancer—implications for care. *Nat. Rev. Clin. Oncol.*, **17**, 457–474.
4. Nikonova,E., Kao,S.-Y. and Spletter,M.L. (2020) Contributions of alternative splicing to muscle type development and function. In: *Seminars in cell & Developmental Biology*. Elsevier, Vol. **104**, pp. 65–80.
5. Zheng,S. (2020) Alternative splicing programming of axon formation. *Wiley Interdiscipl. Rev.: RNA*, **11**, e1585.
6. Zhang,C., Zhang,B., Lin,L.-L. and Zhao,S. (2017) Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, **18**, 583.
7. Jin,H., Wan,Y.-W. and Liu,Z. (2017) Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics*, **18**, 51–59.
8. Dapas,M., Kandpal,M., Bi,Y. and Davuluri,R.V. (2017) Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. *Brief. Bioinform.*, **18**, 260–269.
9. Hayer,K.E., Pizarro,A., Lahens,N.F., Hogenesch,J.B. and Grant,G.R. (2015) Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, **31**, 3938–3945.
10. Leshkowitz,D., Feldmesser,E., Friedlander,G., Jona,G., Ainbinder,E., Parmet,Y. and Horn-Saban,S. (2016) Using synthetic mouse spike-in transcripts to evaluate RNA-Seq analysis tools. *PloS one*, **11**, e0153782.
11. Chandramohan,R., Wu,P.-Y., Phan,J.H. and Wang,M.D. (2013) Benchmarking RNA-Seq quantification tools. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 647–650.
12. Baruzzo,G., Hayer,K.E., Kim,E.J., Di Camillo,B., FitzGerald,G.A. and Grant,G.R. (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods*, **14**, 135–139.
13. Engström,P.G., Steijger,T., Sipos,B., Grant,G.R., Kahles,A., Rätsch,G., Goldman,N., Hubbard,T.J., Harrow,J., Guigó,R. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.
14. Mehmood,A., Laiho,A., Venäläinen,M.S., McGlinchey,A.J., Wang,N. and Elo,L.L. (2020) Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief. Bioinform.*, **21**, 2052–2065.
15. Liu,R., Loraine,A.E. and Dickerson,J.A. (2014) Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics*, **15**, 364.
16. Merino,G.A., Conesa,A. and Fernández,E.A. (2019) A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. *Brief. Bioinform.*, **20**, 471–481.
17. Manz,Q., Tsoy,O., Fenn,A., Baumbach,J., Völker,U., List,M. and Kacprowski,T. (2021) ASimulatoR: splice-aware RNA-Seq data simulation. *Bioinformatics*, **37**, 3008–3010.
18. Merkel,D. (2014) Docker: lightweight linux containers for consistent development and deployment. *Linux J.*, **2014**, 2.
19. Nüst,D., Sochat,V., Marwick,B., Eglen,S.J., Head,T., Hirst,T. and Evans,B.D. (2020) Ten simple rules for writing Dockerfiles for reproducible data science. *PLoS Comput Biol.*, **16**, e1008316.
20. Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A. *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Research*, **10**, 33.
21. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
22. Völzke,H. (2012) Study of health in pomerania (ship). *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, **55**, 790–794.
23. Vaquero-Garcia,J., Barrera,A., Gazzara,M.R., Gonzalez-Vallinas,J., Lahens,N.F., Hogenesch,J.B., Lynch,K.W. and Barash,Y. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *elife*, **5**, e11752.
24. Bushnell,B. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. *Conference: 9th Annual Genomics of Energy & Environment Meeting, Walnut Creek, CA, March 17-20, 2014*.
25. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
26. Bonfert,T., Kirner,E., Csaba,G., Zimmer,R. and Friedel,C.C. (2015) ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics*, **16**, 122.
27. Philippe,N., Salson,M., Commes,T. and Rivals,E. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.*, **14**, R30.
28. Lin,H.-N. and Hsu,W.-L. (2018) DART: a fast and accurate RNA-seq mapper with a partitioning strategy. *Bioinformatics*, **34**, 190–197.
29. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
30. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotech.*, **37**, 907–915.
31. Wang,K., Singh,D., Zeng,Z., Coleman,S.J., Huang,Y., Savich,G.L., He,X., Mieczkowski,P., Grimm,S.A., Perou,C.M. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

32. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

33. Hoffmann,S., Otto,C., Kurtz,S., Sharma,C.M., Khaitovich,P., Vogel,J., Stadler,P.F. and Hackermüller,J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, **5**, e1000502.

34. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

35. Yang,L., Gordon,K.S. and Wei,S. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, **41**, e108.

36. Denti,L., Rizzi,R., Beretta,S., Della Vedova,G., Previtali,M. and Bonizzoni,P. (2018) ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events. *BMC Bioinformatics*, **19**, 444.

37. Mancini,E., Rabinovich,A., Iserte,J., Yanovsky,M. and Chernomoretz,A. (2021) ASpli: integrative analysis of splicing landscapes through RNA-Seq assays. *Bioinformatics*, **37**, 2609–2616.

38. Romero,J.P., Muniategui,A., De Miguel,F.J., Aramburu,A., Montuenga,L., Pio,R. and Rubio,A. (2016) EventPointer: an effective identification of alternative splicing events using junction arrays. *BMC Genomics*, **17**, 467.

39. Middleton,R., Gao,D., Thomas,A., Singh,B., Au,A., Wong,J.J., Bomane,A., Cosson,B., Eyras,E., Rasko,J.E. *et al.* (2017) IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.*, **18**, 51.

40. Goldstein,L.D., Cao,Y., Pau,G., Lawrence,M., Wu,T.D., Seshagiri,S. and Gentleman,R. (2016) Prediction and quantification of splice events from RNA-seq data. *PloS One*, **11**, e0156132.

41. Kahles,A., Ong,C.S., Zhong,Y. and Rätsch,G. (2016) SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics*, **32**, 1840–1847.

42. Sterne-Weiler,T., Weatheritt,R.J., Best,A.J., Ha,K.C. and Blencowe,B.J. (2018) Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Mol. cell*, **72**, 187–200.

43. Zhang,Z., Pan,Z., Ying,Y., Xie,Z., Adhikari,S., Phillips,J., Carstens,R.P., Black,D.L., Wu,Y. and Xing,Y. (2019) Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods*, **16**, 307–310.

44. Lorenzi,C., Barriere,S., Arnold,K., Luco,R.F., Oldfield,A.J. and Ritchie,W. (2021) IRFinder-S: a comprehensive suite to discover and explore intron retention. *Genome Biol.*, **22**, 307.