Université Claude Bernard Lyon 1

Internship Report

Academic year 2023 – 2024

Comparison of Vast-tools and rMATS, two tools to detect differential alternative splicing events from RNA sequencing

Mehdi Marchand – Master 1, Bioinformatics

mehdi.marchand@etu.univ-lyon1.fr

Supervisors: Hélène Polvèche and Cyril Bourgeois

Ecole normale supérieure de Lyon, Laboratory of Biology and Modeling of the Cell (LBMC), Regulation of Genome Architecture and Dynamics of Splicing (ReGArDS)

Comparison of Vast-tools and rMATS, two tools to detect differential alternative splicing events from RNA sequencing

Mehdi Marchand, William Desaintjean, Hélène Polvèche, Cyril F Bourgeois

Abstract

DDX5 and DDX17 are two proteins involved in RNA metabolism, in particular in alternative splicing regulation. To investigate the impact of these proteins on this mechanism, a transcriptomic analysis was previously performed on human neuroblastoma cells in which the expression of DDX5 and DDX17 was silenced compared to control cells, and differential alternative splicing (DAS) analyses were performed with rMATS. Nevertheless, due to the complexity and redundancy of the rMATS results, the team focused its interest on Vast-tools, an alternative tool for analyzing alternative splicing that interacts with a database (VastDB) containing a fixed set of alternative splicing events. During my internship, I implemented and used Vast-tools to perform DAS analyses and compared my results to those obtained with rMATS. Vast-tools and rMATS detected 1,584 and 5,991 DAS events, respectively. Since exon skipping (ES) is the most frequent alternative splicing pattern and the easiest to analyze, I focused on the results corresponding to this type of event. Several analyses were conducted to estimate the proportion of unique differential ES events detected by rMATS, determine the number of rMATS differential ES that could be identified by Vast-tools and determine the proportion of differential ES events identified by both tools. Finally, the results were compared with experimentally validated DAS events. I concluded that Vast-tools is an interesting toolset for detecting reliable differential alternative splicing events while providing a simpler and more user friendly output than rMATS. Nevertheless, since it uses a limited set of annotated alternative splicing, Vast-tools should be used along with a complementary tool to expand the number of detectable events.

Keywords: Vast-tools, rMATS, differential alternative splicing, RNA sequencing, DDX5, DDX17

Table of contents

Table of figures

Glossary

A3SS = alternative 3' splice site, **A5SS** = alternative 5' splice site, **cDNA** = complementary deoxyribonucleic acid, **DAS event** = differential alternative splicing event, **DAS gene** = differentially alternatively spliced gene, **DDX** = dead box, **EEEJ** = exon-microexon-exon junction, **EEJ** = exon-exon junction, **EIJ** = exon-intron-junction, **ES** = exon skipping, **EX** = alternative exon skipping events, **MES** = microexon skipping, **mRNA** = messenger RNA, **MXE** = mutually exclusive exons, **nt** = nucleotide, **PIR** = percent intron retention, **pre-mRNA** = pre-mature messenger RNA, **PSI** = percent spliced in, **ΔPSI** = difference of percent spliced in, **PSU** = percent splice-site usage, **RI** = retained intron, **rMATS** = replicate multivariate analysis of transcript splicing, **RNA** = ribonucleic acid, **siRNA** = small interference RNA, **RT-PCR** = reverse transcription polymerase chain reaction, **VastDB** = vast data base, **Vast-tools** = vertebrate alternative splicing and transcription

Tools and database used

Bash 5.1.16, BLAT (06/17/2024), Bowtie 1.3.1, hg19 GCF_000001405.13, Nextflow 23.09.2-edge, Python 3.8, R 4.4.0, Slurm Workload Manager 20.11.4, Vast-tools 2.5.1, Vastdb.hsa.23.06.20, Vastdb.hs2.23.06.20

Introduction

This study is part of Cyril Bourgeois's project "Control of gene expression by ribonucleic acid (RNA) helicases DDX5 and DD17" within the "Regulation of Genome Architecture and Dynamics of Splicing (ReGArDS)" team.

Transcription of coding genes is a highly regulated mechanism through which a gene is read by RNA polymerase II, resulting in the synthesis of a pre-mature messenger RNA (pre-mRNA). The pre-mRNA undergoes various processing steps such as 5' capping, splicing and 3' polyadenylation, to form a mature messenger RNA (mRNA) which will be translated into a protein. The splicing process is catalyzed by ribonucleoprotein complexes named spliceosomes and it consists in the excision of pre-mRNA segments, in a constitutive or an alternative manner. On one hand, the constitutive process removes introns and ligates exons together. On the other hand, whole exons or a part of them, as introns, may be removed or conserved to form different combinations of mRNAs. Thus, transcription of one gene can lead to the synthesis of a lot of different mRNAs (also called isoforms) that differ according to their conserved exon content. The majority of protein-coding genes (>95%) in human are subject to this mechanism (1).

Alternative splicing is usually classified into different categories according to the event pattern (Fig.1). **Exon skipping** (ES) is the most common pattern (40% of the alternative splicing patterns) (2). It results in the omission of an exon from the mature mRNA. **Microexon skipping** (MES) is a variant of ES that involves a short exon. The **mutually exclusive exons** (MXE) pattern produces isoforms with only one included exon from a cluster of neighboring exons. **Retained intron** (RI), as the name suggests, includes an intron in the mature transcript. **Alternative 5' splice site** (A5SS) and the **alternative 3' splice** (A3SS) refer to alternative splicing sites that are located upstream and downstream, respectively, of the constitutive splicing site within the involved exons, producing isoforms with excluded exon regions. Finally, the alternative first and last exon patterns involve alternative first and terminal exons in exclusive manners.

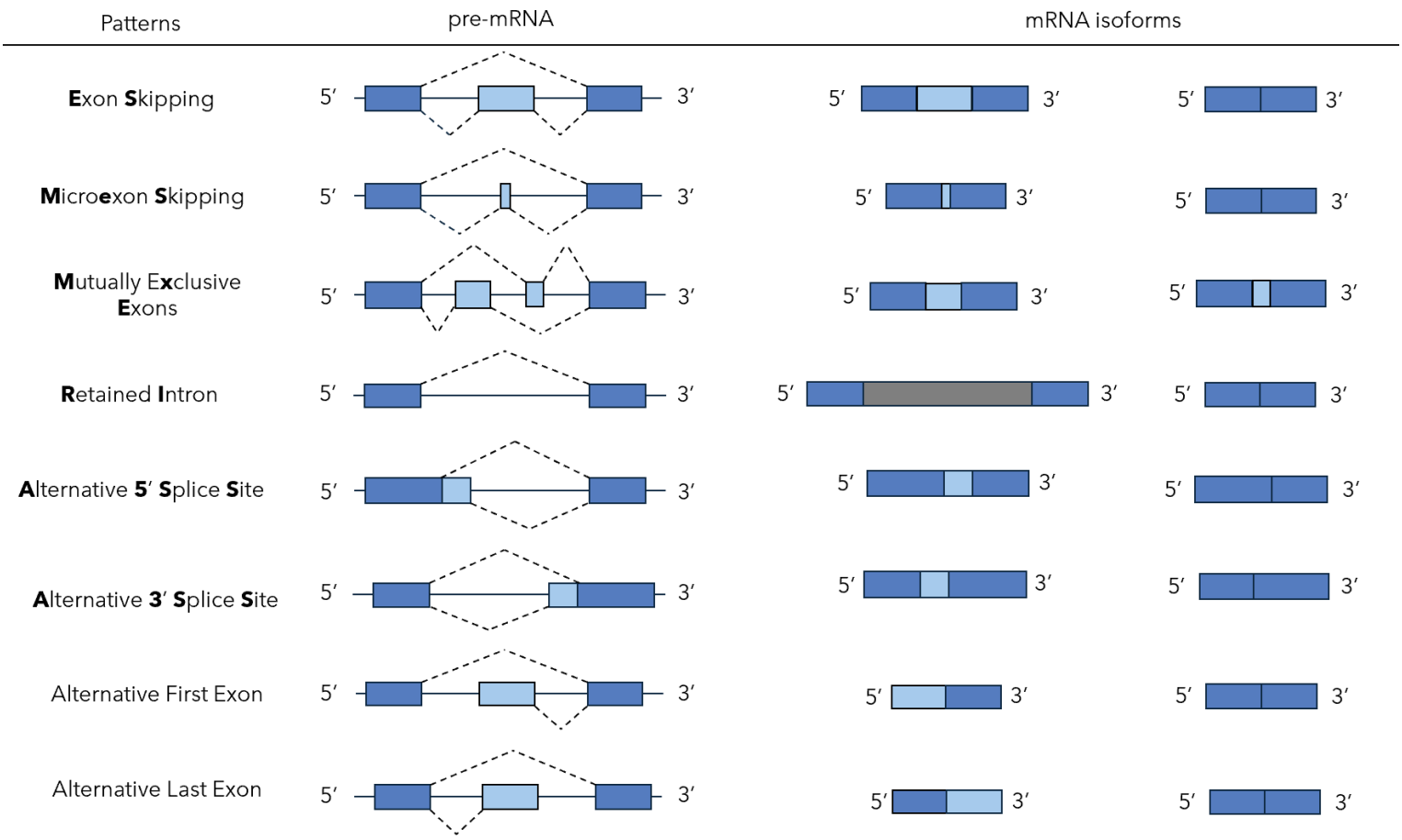


Figure 1 - The different patterns of alternative splicing. Seven alternative splicing patterns are illustrated. The blue boxes and the black lines represent the exons and introns respectively. The different isoforms resulting from these patterns are illustrated on the right side.

The maturation steps of pre-mRNA involve a wide range of factors, among which RNA helicases constitute the largest enzyme family implicated in mRNA metabolism (3). These proteins contribute to RNA structure reorganization and RNA interactions with proteins.

In the context of this study, we are interested in DDX5 and DDX17, two proteins produced by paralogous genes belonging to the DEAD box ATP-dependent RNA helicase family, that are most particularly involved in the regulation of alternative splicing. Indeed, it has been established that these proteins control the inclusion of a large number of exons by modulating the interaction between splicing regulators and the targeted RNA in addition to modifying RNA folding (4–7). This activity is important in the context of myogenic differentiation and epithelial-mesenchymal transition (7), the latter occurring in the context of early development and metastatic processes. Moreover, the team has also shown that DDX17 is implicated in early stages of neuroblastoma (8), a common childhood cancer resulting from the nerve cell degeneration in the sympathetic nervous system.

Previously, genome-wide analyses were performed to investigate the impact of those proteins on alternative splicing (9). The transcriptome of human neuroblastoma cells in

which the expressions of DDX5 and DDX17 were silenced was compared to that of control cells. Silencing was achieved by transfecting small interfering RNAs (siRNA) targeting DDX5 and DDX17 mRNAs. Six samples (three per condition) were sequenced using the Illumina RNA sequencing technology.

Bioinformatical tools to analyze alternative splicing have been developed in parallel with next-generation sequencing technology. Currently, this type of analysis is common and many tools are available, making their selection challenging. Two main approaches are used by the differential alternative splicing tools. On one hand, the differential analyses are conducted through the reconstruction of the transcripts. This method is called isoform-based. On the other hand, the inclusion level of RNA segments involved in the events of interest are measured using different metrics. This approach is called event-based (10).

The previous differential alternative splicing analyses were conducted using the FaRLine (11) pipeline and rMATS tool (8), both using event-based approaches. However, the FaRLine is now outdated (due to compatibility issues with R packages), the rMATS output can be complex to interpret for biologists and can present issues related to splicing event redundancy. Indeed, a considerable proportion of the identified events differ only based on slight variations in genomic coordinates. To find an alternative user-friendly tool for analyzing alternative splicing, I tested Vast-tools (12,13), a "toolset for profiling and comparing alternative splicing events". During my internship, I performed differential alternative splicing analyses with Vast-tools and I compared the results with those previously generated using rMATS. The objective was to ensure the robustness of the results and determine whether Vast-tools is a suitable tool for analyzing alternative splicing according to the context described above.

Materiel and Methods

Cell culture et transfections (pre-internship work)

To silence DDX5 and DDX17, human SH-SY5Y neuroblastoma cells (ECACC) were transfected with 20 nM of siDDX5 and siDDX17. Control neuroblastoma cells were transfected with 20 nM of siGL2 targeting the luciferase RNA. The cells were harvested 48h post-transfection (9).

RNA Sequencing (pre-internship work)

PolyA+ RNA were enriched using TriPure Isolation Reagent (Roche). The complete protocol can be found in the previous article of the laboratory (9).

Reverse-stranded paired-end sequencing (2 x 125bp) was conducted using an Illumina HiSeq 2500 platform (Genewiz). Three replicates per condition (siDDX5/17 vs Control) with an average of 45 million of matched pairs of reads per sample were obtained.

Quality Control

Although quality control was performed in the context of the previous team study, I performed it a second time on pre-trimmed fastq files to fully understand the Vast-tools output according to the baseline data. Fastq data were imported from the Sequence Read Archive (SRA) (14) database (available with the GSE183205 ID).

Illumina Universal adapters were detected at the 3' ends of the reads. Thus, a trimming step was achieved using Cutadapt (15) through the lab trimming script.

Vast database

Vast database (VastDB) for Vertebrate Alternative Splicing and Transcription (12) is "an atlas of alternative splicing profiles in vertebrate cell and tissue types". It offers a collection of insights about alternative splicing events and gene expression **obtained using Vast-tools** for seven species, namely Human, Mouse, Rat, Cow, Chicken, Zebrafishand Fruitfly. The data cover a wide range of developmental stages, tissues and cells and presents insights about inclusion level, sequences involved in the alternative splicing events of interest, splice site strength *etc*.

To ensure the reliability of alternative splicing events initially added to the database, alternative splicing inclusion levels have been validated at a high rate by RT-PCR (136 of 88,778 events in Human, 761 of 69,151 events in Mouse and 36 of 40,263 events in Chicken) (12).

A special feature of Vast-tools is its limit of detection to a fixed set of alternative splicing events annotated from VastDB (2) in addition to assigning a stable ID to each events identified. Thus, the events IDs are identical in VastDB and Vast-tools, allowing users to directly obtain rich information on events identified through the database.

VastDB libraries

The libraries can be divided into four sub-libraries, namely: the Exon-Exon Junction (EEJ), the Exon-Microexon-Exon Junction (EEEJ), the Exon-Intron Junction (EIJ) and the Intron sub-libraries (12).

The EEJ was built based on cDNAs and Expressed Sequence Tags (ESTs) alignments to genomic sequence, and RNA sequence-based transcripts using Cufflinks (12,16). The EEEJ was constructed from scanned intronic sequences in which sequences of 3 to 15 nucleotides flanked by AG and GT were identified (13). Finally, the EIJ and Intron libraries consist of a comprehensive set of reference sequences (17).

The libraries are stored in VastDB. The vastdb.hsa.23.06.20 library was used (corresponds to the hg19 human genome version).

Vast-tools

The Vast-tools software (v2.5.1) consists of several modules to perform differential alternative splicing or differential expression analyses (align, merge, combine, compare, diff, plot, compare_expr).

It detects and groups the ES, MES and MXE patterns into **alternative exon skipping** (EX)events. Simple EX events involving small exons (3 ≤ exon size ≤ 15 nt) are defined as MESevents. The RI, A5SS and A3SS patterns are also detected by Vast-tools.

The **align** module of Vast-tools accepts fastq or fasta files as input (used from fastq in this study). Firstly, the reads are trimmed into fragments of 50 nt using a sliding window of 25nt in order to increase the number of reads that map on splicing junctions (each read is trimmed into four fragments). For paired-end sequencing, the paired reads are consolidated into one read group, resulting in groups of eight fragments. Secondly, the fragments are mapped using Bowtie (1.3.1) (18), a splice unaware tool, to the human reference genome hg19 (GCF_000001405.13, the same version used for differential alternative analyses with rMATS to optimize the comparison between the results). In this context, only unique mappings with no more than two mismatches are authorized, except for microexon analysis in which no mismatch is authorized (13). Then, the unmapped fragments, presumed to cover transcript splice junctions, are mapped against human splice junction libraries (VastDB libraries).

The fragments mapping to the VastDB libraries and the quantification of alternative splicing event are achieved through different **align** sub-modules depending on the type of alternative splicing events (19). The ES and MXE events are quantified with the **Transcript-based module** that uses the fragments that map to the Exon-Exon Junction (EEJ) libraries, and the **Splice site-based module** that relies on a library containing all theorical possible combinations of exons for each gene (20). In the other hand, the MES events are quantified through the **Microexon module** using EEJ libraries complemented with Exon-Microexon-Exon Junction (EEEJ) libraries (13). The output of these three sub-modules are merged to quantify EX events, through Percent Spliced In (PSI) calculation in a non-redundant manner (Fig. 2). The A5SS and A3SS events are quantified from the **Splice site-based module** output (13) with Percent Splice-site Usage (PSU) calculation. Finally, the INT events are identified with an independent module (17) that directly maps the trimmedreads to Exon-Intron Junctions (EIJ), EEJ and Intron libraries, ignoring the first mapping step described before. Only the fragments that map on EIJ and EEJ are used to measure the intron inclusion level through Percent Intron Retention (PIR). Fragments mapping to the Intron library are used to test the read imbalance between fragments that map to the EIJ

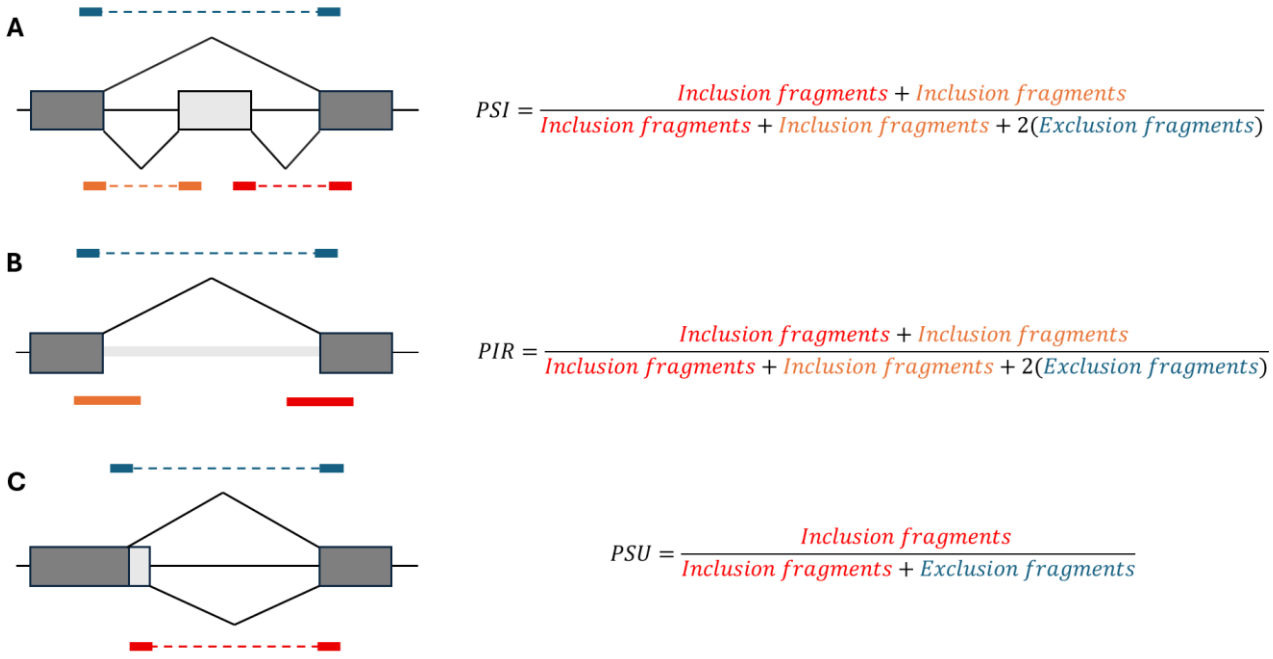


Figure 2 - Formula to quantify the inclusion isoform proportion in function of the alternative splicing event type with Vast-tools. The different alternative splicing events are illustrated on the left side. The boxes represent the exons and the black lines represent the introns. Dark and light grey boxes correspond to constitutively included and alternative fragments, respectively. In orange and red are represented the fragments that map on exon or intron inclusion isoforms (inclusion fragments). Conversely, the fragments that map on exon or intron exclusion isoforms (exclusion fragments) are represented in blue. (A) Vast-tools measures the inclusion percentage for Skipping Exon events using Percent Spliced In (PSI). (B) The Percent Intron Retention (PIR) formula is used to quantify intron inclusion for Retained Intron events. (C) The metric used to quantify inclusion in the context of Alternative 5' or 3' Splice Site events is the Percent Splice-site Usage (PSU).

The quantification of the reads that map on splice junction libraries is performed using one random fragment per read group to prevent multiple counting from the original RNA, resulting in one count for six fragments (13).

Merge is an optional module to fuse replicates from the same condition if the coverage is considered as not deep enough, to prevent bad estimation for poorly expressed genes.

The **combine** module constructs an inclusion table containing Percent Spliced In (PSI), Percent Splice-site Usage (PSU) and Percent Intron Retention (PIR) values depending on the alternative splicing events (Fig 2) with gene and event IDs, in addition to the event coordinates. The table provides information on all the alternative splicing events that can be identified by Vast-tools according to the version of the VastDB library used. Each eventID is unique and contains insights about the alternative splicing event type (example: HsaEX0039231). Moreover, different scores are calculated to assess the reliability of PSI, PSU and PIR values. The main one informs on the coverage of splice junctions, categorizing events into different groups: N/VLOW/LOW/OK/SOK, with N corresponding to a coverage under the minimum threshold (less than 10 or 15 reads according to event

type) and SOK meaning that the alternative splicing event is supported by more than 100 reads.

To simplify this internship report, the term “PSI” will be used for further discussions on inclusion level.

As the name suggests, the **compare** device identifies differential alternative splicing events through a non-statistical approach from the **combine** inclusion table. The rationale behind is that events supported by a large number of reads have more statistical power than those supported by fewer reads. Thus, using a statistical test to identify differential alternative splicing events results in under-estimating the latter. The **compare** module’s strategy is to filter events based on the read coverage score determined by the **combine** module, to obtain reliable results. Then, some criteria are proposed to compare the PSI values between the two conditions, such as minimum ΔPSI (|PSI A – PSI B|), or the distance between the smallest PSI value of the group with the largest PSI average and the greatest PSI value of the group with the lowest PSI average. The latter criterion reflects whether the PSI distributions are allowed to overlap or not. In addition, the module can generate Ensembl gene IDs (https://www.ensembl.org), for example for further Gene Ontology (GO) analyses. GO analyses (21) consist in testing whether a group of differentially alternatively spliced genes is enriched in genes expressed in particular cellular compartments, or in genes involved in particular molecular functions or biological processes. The compare module was not used in the context of this study.

The **diff** module identifies differential alternative splicing events from the **combine** inclusion table through Bayesian inference. For each alternative splicing event and sample, a beta distribution is set as the *prior* distribution with parameters $\alpha = 1$ and $\beta = 1$. It represents the initial knowledge before any data is observed. Then, for each replicate, a likelihood function is determined with regard to the observed data. The *posterior* distribution is defined by combining the prior distribution and the likelihood function using Bayes’ theorem. It represents the belief after considering the observed data. The *posterior* distributions from the same conditions are combined to perform differential alternative splicing analyses through joint distribution comparisons. The differential alternative splicing events were defined under the constraint that the probability that the ΔPSI exceeds a threshold of 0.1 was greater than 0.95, such as $P(|PSI\ A - PSI\ B| > 0.1) > 0.95$.

The **plot** module uses a R plotting script based on the psiplot (12) package.

Finally, the **compare_expr** module works on a similar logic to the **compare** device to identify differentially expressed genes between conditions. It was not used in the context of this study.

Nextflow

The script executing the Vast-tools modules of interest (align, combine and diff) from trimmed fastq files was written based on Nextflow (22). Nextflow is a workflow system

that coordinates tools execution on large datasets in reproducible manner through parallelization and distributed computing. The Nextflow version 23.09.2-edge was used.

PSMN

The Nextflow script was executed on the « Pôle Scientifique de Modélisation Numérique » (PSMN) cluster from ENS-Lyon, through Slurm Workload Manager (20.11.4) (23).

Containerization

To ensure the test reproducibility, I created a Docker image from the Vast-tools (v2.5.1) dockerfile of github (https://github.com/vastgroup/vast-tools). The Docker image was pushed on DockerHub (https://hub.docker.com/r/marchandrm/vast-tools). Then, it was converted into Charliecloud image (24) and pushed to the PSMN. Indeed, the Charliecloud technology is specifically made for high performance computing environments, offering greater security and performance compared to Docker.

rMATS (pre-internship work)

rMATS for replicate Multivariate Analysis of Transcript Splicing (4.1.2) (19) detects and identifies the five standard alternative splicing events, namely : ES, MXE, A5SS, A3SS and RI.

A special feature of rMATS consists in the detection of unannotated alternative splicing events. These events refer to splicing events that are not indexed among the potential alternative splicing events listed in the annotation. They can be divided into two classes; the novel junctions and the novel splice sites. The first one corresponds to unindexed events with indexed splice sites, in contrast to the second that corresponds to unindexed events with unindexed splice sites (Fig.3). In the study, novel events were not studied.

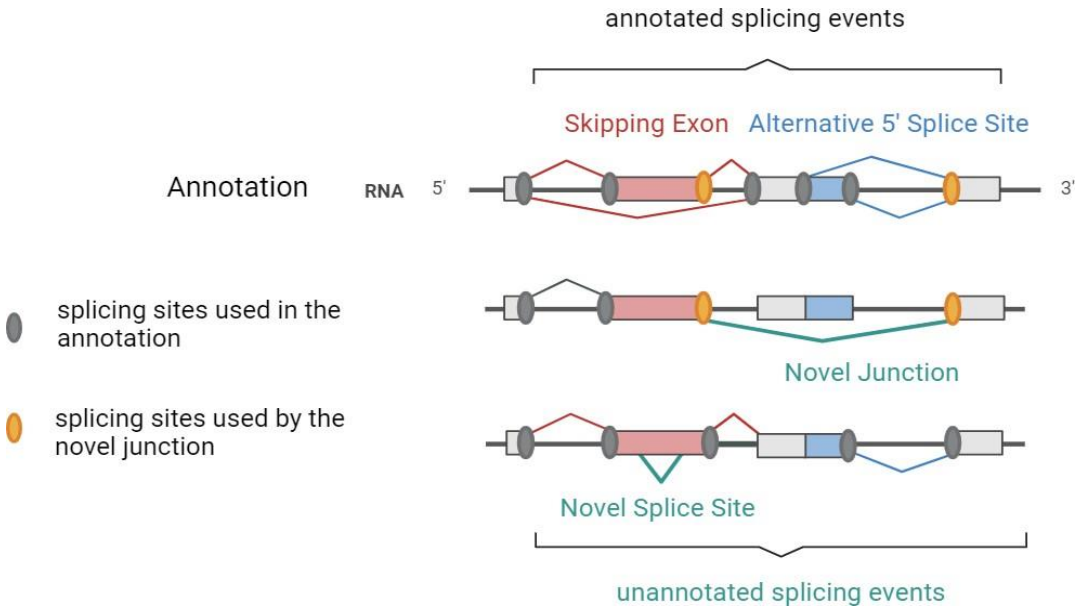
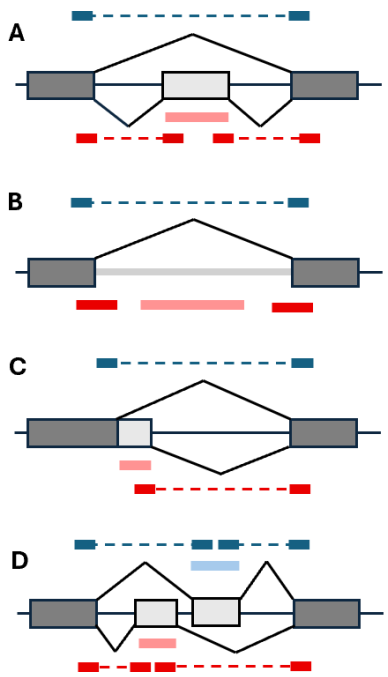


Figure 3 - The rMATS unannotated splicing events. The boxes represent the exons, the grey lines represent the introns. The splicing sites are represented in grey and orange. Two potential alternative splicing events are represented in the annotation, namely a skipping exon in red and an alternative 5' splice site in blue. The novel junction and the novel splice site events correspond to events that are not represented in the annotation. Nevertheless, the first oneuses splicing sites reported in the annotation (orange ones) in contrast with the novel splice site which is based onunannotated splicing sites.

rMATS was executed on BAM files processed with Samtools (0.1.19) (25). They were obtained by mapping the trimmed fastq files on the human reference genome hg19 using STAR (2.7.10a) (26), a splice aware alignment tool. rMATS proceeded to read counting as a first step, to quantify exon or intron inclusion level through Percent Spliced In for the five alternative splicing events. Note that the PSI formula of rMATS differs from the Vast-toolsone (Fig. 2,4).

In this context, the junction count (JC) method has been used, thus only the reads thatmap on the splicing junctions or the paired reads that map on two different exons involved in the alternative splicing event of interest were used to calculate the PSI. In contrast to the junction and exon count (JCEC) method where the reads that map only on the inclusion exon are used in addition to those used with the JC method (Fig. 4).



$$PSI = \frac{\frac{isoform\ A\ reads}{effective\ length\ of\ isoform\ A}}{\frac{isoform\ A\ reads}{effective\ length\ of\ isoform\ A} + \frac{isoform\ B\ reads}{effective\ length\ of\ isoform\ B}}$$

Figure 4 - The reads used in percent spliced in (PSI) calculation with rMATS. The different alternative splicing events are illustrated on the left side (A – Skipping Exon, B – Intron Retention, C – Alternative 5' or 3' Splice Site, D – Mutually Exclusive Exons). The boxes represent the exons and the black lines represent the introns. Dark and light grey boxes correspond to constitutively included and alternative exons, respectively. Reads that map on the A or B isoforms are illustrated in red and blue respectively. rMATS measures inclusion percentage with PSI for all events. It can be quantified in two different ways, namely through the junction count (JC) method using only the reads that map on the splicing junctions (reads in darker colors), or through the junction and exon count (JCEC) method using the reads that map on alternative exons or introns (reads in brighter colors) in addition to the junction reads.

Differential alternative splicing events are determined through likelihood-ratio tests based on likelihood functions constructed from the combination of a normal distribution summarizing the PSI among replicates and a binomial distribution accounting the relation between PSI and inclusion or exclusion read counts. The test consists in comparing the maximum likelihood values from two distributions based on the joint likelihood function, the first one is defined using the entire parameter space (all the parameters), while the second one is determined by imposing a constraint on the parameters. If the maximum likelihood values of the two models are not significantly different, thus there is no difference between the conditions. The p-values, according to the test, were calculated under the constraint that the ΔPSI between the conditions exceeds a given threshold (fixed by default at 0.0001). The events with a minimum average of 20 reads were conserved. A False Discovery Rate (FDR) and ΔPSI thresholds of 0.05 and 0.1 were used respectively to define the differential alternative splicing events.

rMATS redundancy and coverage between the tools

The redundancy of ES events identified by rMATS was estimated using the start and end coordinates of alternatively spliced exons. The Vast-tools events that share the exon coordinates of rMATS within a margin of 5 nt were selected, taking into account that the

events of interest involved the same genes. Then, the rMATS results were merged based on the corresponding Vast-tools events, ensuring that the merged events had the same ΔPSI variation signs.

First, the rMATS redundancy was estimated using the events coordinates of the **inclusion table** from the Vast-tools combine module to know the number of alternative exon skipping events that could be identified by Vast-tools. In a second time, the rMATS ES events were merged using the coordinates of the **differential EX events detected by Vast-tools** to identify the number of differential alternative exon skipping events in common between the tools.

RT-PCR validated events comparison with Vast-tools and rMATS events

Previously validated differential ES events through RT-PCR were checked against the Vast-tools and rMATS outputs (9). The coordinates of the exons of interest were obtained by mapping the primers used for amplification on human genome (hg19) using BLAT (27) on the University of California, Santa Cruz (UCSC) Genome Browser (28) website (https://genome.ucsc.edu/). The coordinates were manually compared with those of differential ES events from Vast-tools or VastDB and rMATS. All ES events were identified with a maximum coordinate discrepancy of one nt.

Visualization

The data were visualized using the ggplot2 (3.5.1) (29) and tidyr (1.3.1) (https://tidyr.tidyverse.org) packages.

Script

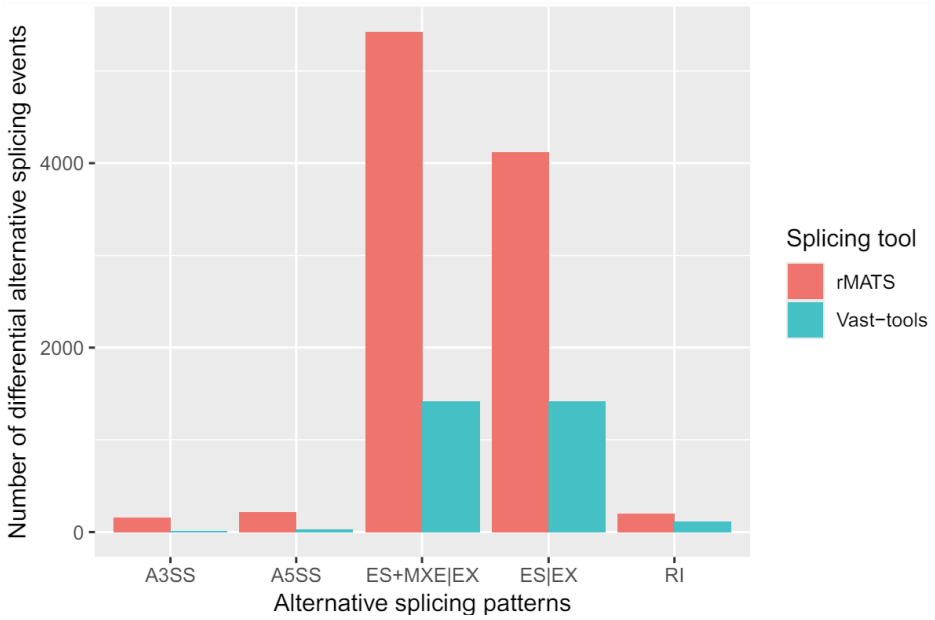
All the scripts used in the context of this study are available at https://github.com/mehdiretif/M1_internship.

Results

Splicing tools

After implementing the Vast-tools in a Nextflow pipeline, differential alternative splicing (DAS) analyses were conducted. Vast-tools detected **1,584 DAS events** between the DDX5/17-depleted and the control conditions, corresponding to **1,070 differentially alternatively spliced (DAS) genes**. On the other hand, rMATS identified **5,991 DAS events** corresponding to **2,661 DAS genes**. The most frequently detected pattern by both tools is the **exon skipping**, that represents around **90%** of the identified alternative splicing events (Vast-tools: EX – 90%, rMATS: ES – 68.8% MXE – 21.7%) (Fig.5). No microexon skipping event was detected with Vast-tools. For both tools, the other types of alternative splicing events, namely RI, A3SS and A5SS, represent less than 10% of the detected events. There are **879 DAS genes common** to both tools.

A



B

	rMATS		Vast-tools		
	DAS events	DAS genes	DAS events	DAS genes	DAS genes in common
A3SS	157	144	12	6	5
A5SS	215	187	33	18	9
RI	197	178	117	108	6
ES	4,120	2,289	1,422	972	831
MXE	1,302	768			

Figure 5 - Number of differential alternative splicing (DAS) events and differentially alternatively spliced (DAS) genes detected with rMATS and Vast-tools. The five standard alternative splicing patterns identified by rMATS are represented, namely the Alternative 3' Splice Site (A3SS), the Alternative 5' Splice Site (A5SS), the Retained Intron (RI), the Exon Skipping (ES) and the Mutually Exclusive Exons (MXE). The alternative skipping exon group (EX) corresponds to the ES and MXE events, as Vast-tools does not allow to make the difference. **A** – Bar plot comparing the number of DAS events detected by rMATS with those detected by Vast-tools according to the alternative splicing pattern. **B** – Table showing the number of DAS events and DAS genes detected with rMATS and Vast-tools in the different alternative splicing categories. The values corresponding to ES for Vast-tools correspond to those of the EX group since the simple and multiple alternative exon skipping events from the Vast-tools output are indistinguishable.

rMATS redundancy and coverage between the tools

Since the alternative skipping exon pattern represents the main part of DAS pattern and the easiest to analyze, and given that it has been demonstrated that the DDX5 and DDX17 proteins control the splicing of a large number of exons (7), we focused on this type of event. In previous analyses conducted with rMATS, it was observed that a considerable

proportion of the DAS events identified with rMATS differ only by slight variations in nt coordinates.

As a first approach to estimate the part of unique differential exon skipping events, the events with strictly identical alternative exon coordinates were merged. Consequently, the numbers of differential ES and MXE events decreased from **4,120 to 3,700** (Fig. 6) and **1,302 to 1,217** events (corresponding to 10% and 6,5% of strict duplicates) respectively.

Furthermore, the **rMATS DAS events** were reduced by using the coordinates of corresponding alternative exons **from the Vast-tools inclusion table** (refer to the Materiel and Method section). This could be done only in cases where the events are annotated in the VastDB library, the inclusion table containing the coordinates of all the alternative splicing events that can be identified by Vast-tools according to the version of the VastDB library used. Among the **4,120 differential ES events** identified by rMATS, **3,319 differential ES events** are equivalent to those annotated in the VastDB library (outof a total of **176,246 alternative exon skipping (EX) events** in the inclusion table), corresponding to **2,930 unique ES events** after merging these events on the inclusion table events. Thus, **up to 770 ES events** could not be identified by Vast-tools (3,700 –2,930), representing 20% of the unique differential ES identified by rMATS (Fig.6).

Finally, the same operation was conducted on the **rMATS DAS events** using the coordinates of corresponding **Vast-tools DAS events** (instead of the inclusion table), to identify the proportion of common differential ES events detected. Before merging the rMATS output on the Vast-tools DAS events, **1,142 events** were considered as **common differential events**, corresponding to **906 unique and common differential exon skipping events** for **831 genes in common** (Fig.6). Interestingly, all the common differential exon skipping events have the same ΔPSI sign.

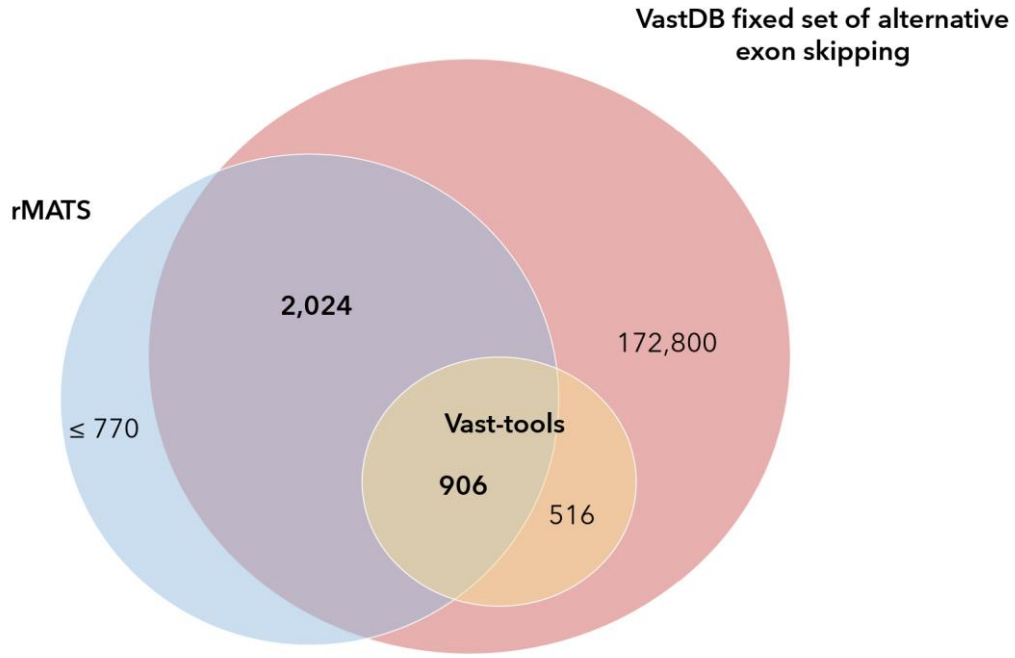


Figure 6 - Venn diagram illustrating the number of differential exon skipping events identified with rMATS and Vast-tools in common. The numbers were obtained by comparing the coordinates of the differential Exon Skipping (ES) events detected by rMATS with the coordinates of the differential alternative exon skipping (EX) events detected by Vast-tools, or with the EX events that can be detected by Vast-tools (VastDB fixed set of alternative exon skipping). The Vast-tools detectable events were obtained from the inclusion table generated by the combine module. The inclusion table contained 176,246 EX events. Vast-tools detected 1,422 differential EX events and rMATS identified a maximum of 3,700 unique ES events (estimation based on strictly identical rMATS coordinatesand correspondence with VastDB coordinates).

RT-PCR validated events comparison with Vast-tools and rMATS events

Based on previous genome-wide analyses, the inclusion of ten selected alternative exons from nine genes was validated by RT-PCR (9). These exons in the genes *CTTN, SEMA6C, GPR175, ABLIM1, NRCAM, NCS1* (for two exons), *PAM, PRMT2* and *URB1* have been identified by the FaRLine tools (11).

Vast-tools identified five of these exons as differentially alternatively spliced (*CTTN,GPR175, ABLIM1, NCS1* x2). The ΔPSI of these five events are consistent with the RT-PCR results. Nevertheless, a discordance in exon coordinates was noticed between the Vast- tools output and VastDB for the event relative to the *CTTN* gene. Indeed, the coordinates corresponding to the end of the alternative exon differ for the same VastID. **rMATS identified nine of the ten exons as differentially alternatively spliced** (*CTTN, SEMA6C, GPR175, ABLIM1, NRCAM, NCS1* x2, *PRMT2, URB1*). As for Vast-tools, the ΔPSI values of these events are consistent with the inclusion levels measured by RT-PCR. Thus, five ofthe events of interest were identified by both tools, whereas the exon of the *PAM* gene was not identified as alternatively spliced by either tool.

Discussion

Differential alternative splicing events were detected using Vast-tools and rMATS from three replicates per condition (siDDX5/17 vs Control) of neuroblastoma cells.

Contrary to what has been described in the literature, rMATS detected many more events than Vast-tools. *Jiang and al* reported that Vast-tools detects more DAS events than rMATS from simulated RNA-seq data (10). Although our results do not follow the trend described by the others authors, the difference could be explained, at least in part, by the human reference genome used. Indeed, the authors used the hg38 reference genome instead of the hg19, and the inclusion table corresponding to the hg38 reference genome contains twice as many annotated alternative splicing events (721,551) as the table based on hg19 (364,004). Thus, more DAS events are expected using the most recent version of the human genome. Moreover, the RNA-seq dataset used differs from those employed in the comparative study. In addition, the dataset does not respect the recommended number of reads per sample.

Indeed, a minimum of 70 million reads per sample is advised for conducting alternative splicing analyses (github.com/vastgroup/vast-tools). However, the previous alternative splicing analyses with rMATS and FaRLine proved sufficient to detect differential events, and a subset of these events was experimentally validated (9).

Concerning the number of DAS events detected by Vast-tools and rMATS, the difference can be attributed to the different statistical approaches employed by the tools, namely the Bayesian inference and the likelihood-ratio test. Furthermore, one particularity of Vast-tools is its use of a fixed set of alternative splicing events resulting in an underestimation of the real number of DAS events. In contrast, rMATS overestimates this quantity through the detection of DAS events in duplicates.

Eliminating duplicates is a crucial step to accurately apprehend the rMATS output. In this study, a portion of the duplicates for the ES events was removed using the Vast-tools coordinates to facilitate the comparison of the outputs of both tools. However, this approach makes possible the removal of duplicates only for the ES annotated in VastDB, resulting in underestimation of the differential ES events identified by rMATS. Moreover, changes in coordinates between the output of the Vast-tools diff module and VastDB for a same event ID (VastID) were noticed, but it seems to be an inherent Vast-tools issue. Nevertheless, a range of the real number of differential ES events can be determined by combining the approach using the coordinates from the Vast-tools inclusion table and the rMATS ouput without the strict duplicates (i.e. events with the exact same coordinates for the exons involved). Another interesting approach would have been the merging of rMATS output based on rMATS coordinates to create event groups. Nevertheless, this raises questions about how to determine the limiting coordinates of these groups.

Despite these differences, the same proportion of DAS events depending on the splicing type, as well as the same inclusion direction for common DAS events identified with both tools were noticed. Finally, some of DAS events validated through RT-PCR in the previous

study (9) were also identified with Vast-tools and rMATS, with consistent ΔPSI values, confirming a certain level of reliability for these tools. Nevertheless, Vast-tools identified less of these events than rMATS, whereas *Tapial and al* described a comparable number of differential ES events identified by both tools in the simulated data tested, corresponding to about 80% of simulated events (12). However, only ten ES events were tested in our study, contrary to their study that tested more than 4,000 simulated differential ES events.

Perspectives

Although the approaches used in this study remove a considerable number of rMATS duplicate events, further improvements are necessary to obtain more relevant results. Specific approaches to create event groups from rMATS coordinates according to the alternative splicing pattern could be implemented in the form of processes in a Nextflow pipeline. Concerning Vast-tools, it is an interesting toolset for detecting reliable differential alternative splicing events while providing a simpler and more user friendly output than rMATS. Nevertheless, since it uses a limited set of annotated alternative splicing, it should be used along with a complementary tool to expand the number of detectable events. Moreover, the newest human genome version should be used to increase this number of detectable events.

Furthermore, as the results obtained in this study do not follow the trends described in the literature, additional tests must be conducted using inputs that conform to the recommendations. Simulated data sets could be interesting to improve further the benchmarking of Vast-tools. However, alternative splicing events will have to be generated on the basis of annotated differential alternative splicing events.

References

1. Ren P, Lu L, Cai S, Chen J, Lin W, Han F. Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease. Front Immunol. 2021;12:713540.

2. Muller IB, Meijers S, Kampstra P, van Dijk S, van Elswijk M, Lin M, et al. Computational comparison of common event-based differential splicing tools: practical considerations for laboratory researchers. BMC Bioinformatics. 2021;22:347.

3. Bourgeois CF, Mortreux F, Auboeuf D. The multiple functions of RNA helicases as drivers and regulators of gene expression. Nat Rev Mol Cell Biol. 2016;17:426–38.

4. Ameur LB, Marie P, Thenoz M, Giraud G, Combe E, Claude J-B, et al. Intragenic recruitment of NF-κB drives splicing modifications upon activation by the oncogene Tax of HTLV-1. Nat Commun. 2020;11:3045.

5. Kokolo M, Bach-Elias M. P68 RNA Helicase (DDX5) Required for the Formation of Various Specific and Mature miRNA Active RISC Complexes. Microrna Shariqah United Arab Emir. 2022;11:36–44.

6. Dardenne E, Pierredon S, Driouch K, Gratadou L, Lacroix-Triki M, Espinoza MP, et al. Splicing switch of an epigenetic regulator by RNA helicases promotes tumor-cell invasiveness. Nat Struct Mol Biol. 2012;19:1139–46.

7. Dardenne E, Polay Espinoza M, Fattet L, Germann S, Lambert M-P, Neil H, et al. RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. Cell Rep. 2014;7:1900–13.

8. Lambert M-P, Terrone S, Giraud G, Benoit-Pilven C, Cluet D, Combaret V, et al. The RNA helicase DDX17 controls the transcriptional activity of REST and the expression of proneural microRNAs in neuronal differentiation. Nucleic Acids Res. 2018;46:7686–700.

9. Terrone S, Valat J, Fontrodona N, Giraud G, Claude J-B, Combe E, et al. RNA helicase-dependent gene looping impacts messenger RNA processing. Nucleic Acids Res. 2022;50:9226–46.

10. Jiang M, Zhang S, Yin H, Zhuo Z, Meng G. A comprehensive benchmarking of differential splicing tools for RNA-seq analysis at the event level. Brief Bioinform. 2023;24:bbad121.

11. Benoit-Pilven C, Marchet C, Chautard E, Lima L, Lambert M-P, Sacomoto G, et al. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. Sci Rep. 2018;8:4307.

12. Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017;27:1759–68.

13. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell. 2014;159:1511–23.

14. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39:D19-21.

15. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.

16. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. Nature Publishing Group; 2010;28:511–5.

17. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24:1774–86.

18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

19. Gohr A, Mantica F, Hermoso-Pulido A, Tapial J, Márquez Y, Irimia M. Computational Analysis of Alternative Splicing Using VAST-TOOLS and the VastDB Framework. Methods Mol Biol Clifton NJ. 2022;2537:97–128.

20. Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, et al. MBNL proteins repress embryonic stem cell-specific alternative splicing and reprogramming. Nature. 2013;498:241–5.

21. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022;31:8–22.

22. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

23. Slurm Workload Manager - Documentation [Internet]. [cited 2024 May 31]. Available from: https://slurm.schedmd.com/documentation.html

24. Priedhorsky R, Randles T. Charliecloud: unprivileged containers for user-defined software stacks in HPC. Proc Int Conf High Perform Comput Netw Storage Anal [Internet]. New York, NY, USA: Association for Computing Machinery; 2017 [cited 2024 May 31]. page 1–10. Available from: https://dl.acm.org/doi/10.1145/3126908.3126925

25. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29:15–21.

27. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002;12:656–64.

28. Raney BJ, Barber GP, Benet-Pagès A, Casper J, Clawson H, Cline MS, et al. The UCSC Genome Browser database: 2024 update. Nucleic Acids Res. 2024;52:D1082–8.

29. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. New York, NY: Springer; 2009 [cited 2023 May 22]. Available from: https://link.springer.com/10.1007/978-0-387-98141-3