ESGq: Alternative Splicing Events Quantification across Conditions based on Event Splicing Graphs

Davide Cozzi[1], Paola Bonizzoni[1] and Luca Denti[1,*]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

Abstract

Alternative Splicing (AS) is a regulation mechanism that contributes to protein diversity and is also associated to many diseases and tumors. Alternative splicing events quantification from RNA-Seq reads is a crucial step in understanding this complex biological mechanism. However, tools for AS events detection and quantification show inconsistent results. This reduces their reliability in fully capturing and explaining alternative splicing. We introduce ESGq, a novel approach for the quantification of AS events across conditions based on read alignment against Event Splicing Graphs. By comparing ESGq to two state-of-the-art tools on real RNA-Seq data, we validate its performance and evaluate the statistical correlation of the results. ESGq is freely available at https://github.com/AlgoLab/ESGq.

Keywords

Alternative Splicing, RNA-Seq, Read Alignment, Splicing Graph

1. Introduction

Alternative Splicing (AS) is a post-transcription regulation mechanism that contributes to isoform and protein diversity in eukaryotes. Due to AS, depending on its environment, a single gene can produce multiple isoforms, hence complicating our understanding of the gene expression process. For instance, more than 95% of multi-exon human genes [1, 2] and more than 60% of multi-exon Drosophila Melanogaster genes [3] exhibit more than one isoform. Due to its association to aging [4], cancer [5], and neuro-degenerative diseases [6], the analysis of AS is of the utmost importance.

In the last decade, RNA-Sequencing has become the defacto standard for the analysis of alternative splicing and a plethora of tools have been proposed in the literature. From a very high level point of view, the approaches for the analysis of alternative splicing available in the literature can be divided in two groups, depending at which level they work: transcript-based [7, 8] and event-based approaches.

In this work, we will focus on the second category. This kind of approaches characterizes AS at the most fine-grained level by giving a detailed and strict description of what happens at the exon-exon (or splice) junction level. Although being more rigorous in the description of AS events, this kind of approaches resulted more accurate than the transcript-based competitors [9], thus potentially providing a more detailed characterization of alternative splicing. Classical AS events are grouped in 5 categories [10]: exon skipping, alternative 3' (acceptor) splice sites, alternative 5' (donor) splice sites, intron retention, and mutually exclusive exons. Many tools have been developed to perform AS events detection and quantification [11, 12, 13, 14, 15]. Recent works [16, 17] argue that the classic definition of AS events is not satisfactory and not adequate to fully capture the complexity of alternative splicing. To this aim, they introduce Local Splicing Variations, a novel concept that aims to represent complex AS patterns and then increase the expressive power of the classical and more strict classification. However, the detection and quantification of classical AS events is an already hard - and not fully solved - problem that does not need further complications. Indeed, tools and methodologies show several limitations [9]. For instance, although AS events exhibit a strict definition, tools available in literature inconsistent results, due to the different definitions and filtering criteria adopted. Moreover, every tool uses its own format to describe the AS events, making any downstream analysis quite complex.

In this context, we focus on the detection and quantification of non-novel AS events across conditions and provide an extensive comparison of two state-of-the-art tools, rMATS [11] and SUPPA2 [12]. Moreover, to validate our findings, we introduce ESGq, a novel graph-based methodology for the quantification of AS events across two conditions. Inspired by the recent progress and development in the field of pangenomic and graph algorithms [18, 19], ESGq models AS events as local splicing graphs, called *event splicing graphs*, and then quantify the events by aligning reads to them. The usage of graphs in the transcriptomic world is not new [13, 14, 20, 21, 22]

ITAT'23: Information technologies – Applications and Theory, September 22–26, 2023, Tatranské Matliare, Slovakia

*Corresponding author.
d.cozzi@campus.unimib.it (D. Cozzi); paola.bonizzoni@unimib.it (P. Bonizzoni); luca.denti@unimib.it (L. Denti)
0000-0003-2439-0608 (D. Cozzi); 0000-0001-7289-4988 (P. Bonizzoni); 0000-0001-8786-2276 (L. Denti)

but the use of simplified graph-based representation characterizing precise loci of the genome is something that was never investigated. This work provides a first exploratory investigation and may lay the foundations for a new generation of approaches for the efficient and accurate AS events quantification based on pantranscriptome graphs.

Experiments on a very recent real RNA-Seq dataset sequenced from Drosophila Melanogaster flies at two time points show that, by pairing simple graphs with accurate mapping, ESGq is able to achieve comparable results to state-of-the-art without sacrificing its efficiency. Moreover, our comparison proves once again the inconsistency of the results obtained by different methodologies for the detection and quantification of AS events.

2. Method

We introduce ESGq, a novel graph-based method for the differential quantification of alternative splicing events across conditions. ESGq takes as input a reference genome (in FASTA format), a gene annotation (in GTF format), and a two conditions RNA-Seq dataset with optional replicates (in FASTQ format), and computes the differential expression of annotated AS events (in custom text format). For each event ESGq provides the Percent-Spliced In (PSI, ψ) with respect to each input replicate and the $\Delta\psi$, summarizing the differential expression of each event across the two conditions. Current implementation focuses on four types of alternative splicing events (exon skipping, intron retention, alternative acceptor site, and alternative donor site) and supports both paired-end and single-end RNA-Seq datasets.

Differently from state-of-the-art approaches, which rely on spliced read alignment to a reference genome or quasi-mapping transcript quantification, ESGq relies on read alignment against a graph-based structure representing the events that need to be quantified. Instead of using a full splicing graph or a pantranscriptome, that are common structures in the literature [13, 14, 19], ESGq limits its computation to smaller and less complex graphs, the *Event Splicing Graphs*. An event splicing graph is a splicing graph which encodes only the exons and splice junctions involved in an alternative splicing event. Differently from splicing graphs commonly used in the literature, that represent all known transcripts of a gene, and differently from pantranscriptomes, where entire gene loci and intergenic regions are represented, an Event Splicing Graph encodes only the portions of the two transcripts involved in an event. By using this simpler representation, ESGq is able to achieve great efficiency without sacrificing its accuracy.

ESGq consists of three steps (also depicted in Figure 1):

1. event splicing graphs construction
2. read alignment against event splicing graphs
3. ψ and $\Delta\psi$ computation

ESGq starts its computation by extracting the annotated alternative splicing events from the input gene annotation. To this aim, it employs a module of the SUPPA2 tool [12]. The output of this module is a list of annotated alternative splicing events, that are events whose two isoforms are already annotated in the input gene annotation. For each event, SUPPA2 reports the type (one among SE, RI, A3, A5) and the genomic coordinates of the splice junctions involved in it. Starting from this list, ESGq builds the event splicing graphs, one per event. We note that multiple graphs can be built from the same gene. By exploiting the genomic coordinates of an event, ESGq retrieves the corresponding exons and adds them as nodes in the event splicing graph. Depending on the AS event type, ESGq adds edges between these nodes in order to represent the two isoforms involved in the event: the canonical isoform (that, in graph terms, is the path denoted as $\mathcal{P}_C$) and the alternative one (denoted as $\mathcal{P}_A$). More precisely, the four scenarios, one per event type, contemplated by ESGq are depicted in Figure 2 and can be formally defined as follow:

- to model an exon skipping event (SE), ESGq needs to take into account three exons and this implies that the corresponding event splicing graph is composed of three nodes n_1, n_2, n_3. In detail, n_2 represents the exon that is spliced out during the event. The canonical isoform is represented by the path involving all three nodes ($\mathcal{P}_C = n_1 \rightarrow n_2 \rightarrow n_3$) while the alternative isoform consists in the skip of n_2 ($\mathcal{P}_A = n_1 \rightarrow n_3$);
- to model an alternative acceptor site event (A3), ESGq needs to take into account three exons and accordingly three nodes n_1, n_2, n_3. In detail n_1 represents the upstream exon, n_2 the canonical downstream exon, and n_3 the downstream exon with the alternative acceptor splice site. The canonical isoform is represented by the path involving the shared upstream exon and the canonical downstream exon ($\mathcal{P}_C = n_1 \rightarrow n_2$) whereas the alternative isoform changes the downstream exon with the alternative one ($\mathcal{P}_A = n_1 \rightarrow n_3$);
- to model an alternative donor site event (A5), ESGq needs to take into account three exons and accordingly three nodes n_1, n_2, n_3. In detail, n_1 represents the canonical upstream exon, n_2 the canonical downstream exon, and n_3 the upstream exon with the alternative donor splicing site. The canonical isoform is represented by the path involving the canonical upstream exon and the shared downstream exon ($\mathcal{P}_C = n_1 \rightarrow n_2$) whereas the alternative isoform changes the up-

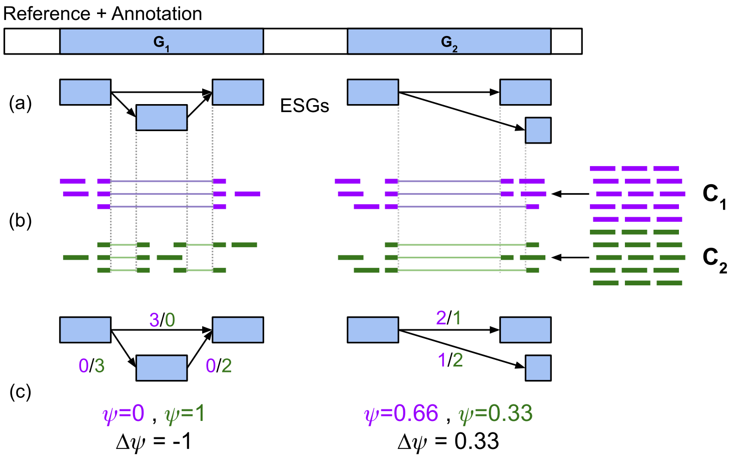


Figure 1: ESGq method. (a) From the reference genome and the gene annotation, ESGq builds the Event Splicing Graphs (ESGs in the figure). (b) RNA-Seq reads from the two conditions (C_1 and C_2) are aligned to the Event Splicing Graphs. (c) Graph alignments are used to weight junction edges and the weights are used to compute the ψ values (one per condition) and the $\Delta\psi$ value (one per dataset).

stream exon with the alternative one ($\mathcal{P}_A = n_3 \rightarrow n_2$);
- to model an intron retention event (RI), ESGq needs to take into account three exons. However, this case is harder than the previous ones: the three nodes n_1, n_2, n_3 of the graph do not closely correspond to three exons, but one of them (n_3) correspond to a portion of it. In detail, n_1 and n_2 represent the two upstream and downstream exons whereas n_3 represents the retained intron (i.e., the internal portion of the exon linking the upstream and downstream canonical exons). The canonical isoform is then represented by the path involving the upstream exon and the downstream exon ($\mathcal{P}_C = n_1 \rightarrow n_2$) whereas the alternative isoform includes the retained intron in the path ($\mathcal{P}_A = n_1 \rightarrow n_3 \rightarrow n_2$).

We note that, from a conceptual point of view, each event contributes to an event splicing graph, but, from a more practical point of view, ESGq builds a single graph with multiple connected components, one per event.

In the second step, ESGq indexes the graph constructed in the previous step. Since the goal is to align the input RNA-Seq reads to the graph using the Giraffe aligner [23], ESGq employs the VG toolkit [24] to build the gBWT (graph Burrows–Wheeler Transform) index [25]. Each input replicate is then independently aligned to the graph using Giraffe. We note that, since by default vg breaks each node longer than 32bp into smaller nodes (of length ≤ 32), in order to keep the association between the nodes in the input graph (that is the one built by ESGq) and the nodes in the indexed version, ESGq directly breaks the nodes while building the event splicing graphs and links them accordingly to maintain the same two paths, i.e., isoforms. Due to this, in an event splicing graph we can observe two kind of edges: edges linking the smaller ($\leq 32bp$) nodes internal to an exon and edges that represent the real splice junction of interest for the AS event. Although ESGq differentiates between these two kinds of edge, conceptually only the edges representing a splice junction are used by ESGq. For this reason, we decided to omit these edges from Figure 2.

In the third and last step, ESGq computes the ψ value of the events w.r.t. each replicate and then summarize these values by comparing the two conditions and computing a $\Delta\psi$ value per event. This value represent the differential expression of each event across the two input conditions.

To do so, ESGq analyses the graph alignments computed in the previous step and assign a weight to each edge that represents a splice junction. Since each read is aligned to a path of the graph, computing this weight is straightforward as increasing a counter per edge. Indeed, a read can be aligned to a single node of the graph, hence without using any edge, or to a sequence of nodes, hence using one or more edges. In such a case, ESGq checks every edge used by the alignment and, if an edge is a junction edge, it increases its weight by 1. In other words, since each junction edge represent a splice junction, its weight represents the number of reads that have been spliced aligned over it.

Finally, ESGq uses these weights to compute the ψ value of each AS event following its classical formulation, i.e., the proportion of reads supporting the standard isoform over the reads supporting both isoforms [26].

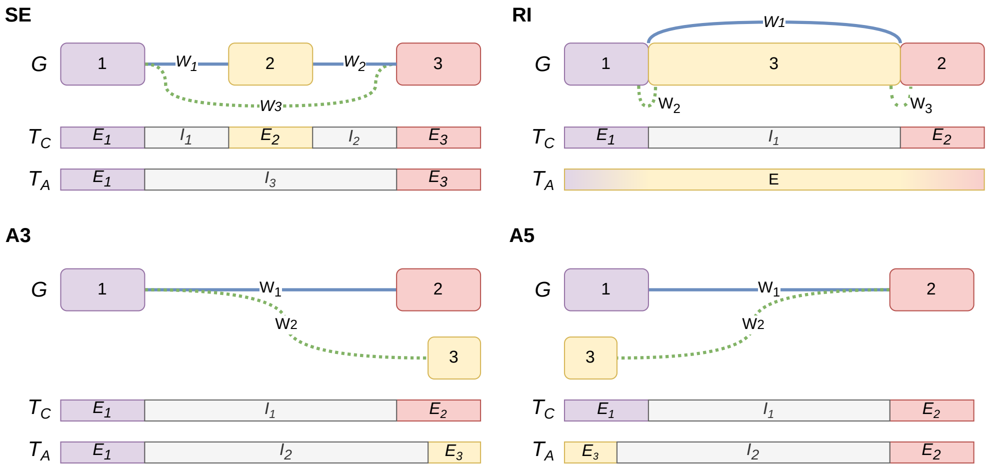


Figure 2: Event splicing graphs computed by ESGq. For each AS event type (exon skipping, SE, alternative acceptor site, A3, alternative donor site, A5, and intron retention, RI), we report the event splicing graph G and the two annotated isoforms involved in the event: the canonical transcript T_C (represented in G by blue edges) and the alternative transcript T_A (represented in G by dotted green edges). In T_C and T_A, E blocks represent exons and I blocks represent introns. The edge labels W represent the weights computed by ESGq after the read alignment step.

Differently from other approaches, which rely on both spliced and not spliced reads, ESGq ψ computation is based only on spliced reads counts, hence the support of an isoform is approximated using only these values and does not take into account its full coverage. We believe that this is a good approximation of the correct ψ value and this is also confirmed by our experimental evaluation. ψ calculation can be summarized as follows (we also refer the reader to Figure 2):

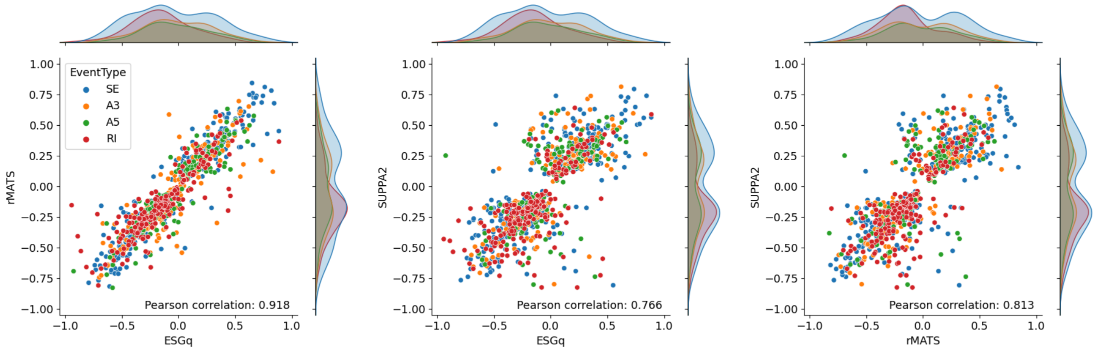
- $\psi_{SE} = \frac{\frac{w_1+w_2}{2}}{\frac{w_1+w_2}{2}+w_3}$ considers the mean of the weights w_1, w_2 of the canonical isoform and the weight w_3 of the alternative isoform (in a similar fashion to [27]);
- $\psi_{A3} = \frac{w_1}{w_1+w_2}$ and $\psi_{A5} = \frac{w_1}{w_1+w_2}$ consider the weight w_1 of the canonical isoform and the weight w_2 of the alternative isoform
- $\psi_{RI} = \frac{w_1}{w_1+\frac{w_2+w_3}{2}}$ considers the weight w_1 of the canonical isoform and the mean of the weights w_2, w_3 of the alternative isoform

Starting from these ψ values (one per event per replicate), which summarize the event quantification for each input replicate, ESGq computes the differential quantification across the two input conditions ($\Delta\psi$) as the difference between the absolute value of the ψ means in the two conditions. Differently from other approaches, ESGq does not assign a *p value* to the $\Delta\psi$. This is mainly a consequence of the simplified AS events quantification based only on spliced reads counts. Future works will be devoted to improve the statistical validation of ESGq results.

3. Experimental evaluation

We implemented the ESGq pipeline in Python and the code is freely available at https://github.com/AlgoLab/ESGq. We note that the most computationally intensive steps of the pipeline (i.e., graph indexing and read-to-graph alignment) rely on the VG toolkit [24], that is implemented in C++. We assessed ESGq efficacy and efficiency on a real dataset of RNA-Seq reads (SRA BioProject ID: PRJNA718442) that comes from a recent study [28] on the correlation between ageing and differential gene expression in Drosophila Melonogaster. More precisely, the study conducted a genome-wide differential expression analysis at two different time points: day 1 and day 60 flyes. The dataset consists of three replicates for two conditions (the two time points), for a total of 6 Illumina Hiseq samples. All samples are paired-end and consist of 151bp-long reads (see Table 1 for more details).

Differently from the aforementioned study, where the focus was the analysis of differential gene expression, in this work, we analyze the same dataset from the perspective of differential quantification of alternative splicing events. To do so, we applied ESGq and two other state-of-the-art approaches for the differential quantification of alternative splicing events across multiple conditions: rMATS [11] (version 4.1.2) and SUPPA2 [12] (version 2.3). The former performs differential quantification starting from read alignment to the reference genome whereas the latter starts from the quasi-mapping transcript quantification of Salmon [29]. In this way, we have been able



(a) ESGq vs rMATS

(b) ESGq vs SUPPA2

(c) rMATS vs SUPPA2

Figure 3: Correlation plots between the $\Delta\psi$ values computed by ESGq, rMATS, and SUPPA2. These results refer to our analysis of $151bp$ paired-end sample and $k = 31$ (for Salmon and SUPPA2).

Condition	Replicate	n.Pairs	Size (GB)
Day 1	SRR14101759	26 658 610	19.2
	SRR14101760	25 474 257	18.4
	SRR14101761	28 339 185	22
Day 60	SRR14101762	24 985 317	18
	SRR14101763	25 569 084	18.6
	SRR14101764	24 605 265	17.8

Table 1
Real dataset used in our experimental evaluation (SRA BioProject ID: PRJNA718442). Table reports the number of reads and the size in GigaByte of each paired-end replicate.

to compare three methodologies based on completely different frameworks: graph-based alignment, reference-based alignment, and transcript-based quasi-mapping. rMATS was run starting from the alignments produced by STAR aligner [30] (version 2.7.10b) whereas SUPPA2 was run starting from Salmon transcript quantification (version 1.10.1). All tools were run using their default parameters and 16 threads (when possible). In our analyses, we considered the reference, gene annotation, and transcripts provided by FlyBase [31], release 6.52.

We compared the $\Delta\psi$ values reported by the three considered tools. ESGq reported 3 276, rMATS reported 3 699, and SUPPA2 reported 1 619 AS events correctly quantified. We note that both ESGq and SUPPA2 start from a list of events and quantify them: each time an event can not be correctly quantified (due to, for instance, no coverage support), they assign $\Delta\psi = NaN$ to that event. In that case, the event is not considered as correctly quantified, thus excluded from the analysis. The huge difference in the number of reported events proves the complexity of detecting AS events from RNA-Seq reads and highlights the inconsistency between different methodologies based on different filtering criteria [9]. In our analysis we included all those events correctly quantified by all three tools and considered statistically significant by both rMATS and SUPPA2, i.e., events with p value ≤ 0.05. A total of 933 events resulted from this filter: 374 exon skipping (40%), 190 alternative 3' (20%), 154 alternative 5' (17%), and 215 intron retention (23%). We note that we also tried to include Whippet [15] in our evaluation but, due to a different AS event representation that cannot be easily compared to the representation given by the other tools, we ended up omitting it from the analysis.

Figure 3 and Table 3 report the results of our analysis. All tools achieved comparable results. Remarkably, ESGq and rMATS achieved a very good correlation, with a Pearson correlation coefficient equal to 0.918 (Figure 3a). Although both approaches are based on read alignment, they show two main methodological differences that slightly affect their results. Firstly, rMATS uses read counts coming from read alignment to a reference genome whereas ESGq uses (spliced) read counts coming from alignment to event splicing graphs, that are a reduced representation. Secondly, the quantification step implemented in ESGq is quite simplistic and not elaborate as the probabilistic framework implemented in rMATS. However, none of these two differences (that are a wanted restriction and a current - undesired - limitation) seems to affect the results of ESGq. On the other hand, SUPPA2 resulted less correlated to the two other approaches, showing a Pearson correlation coefficient ranging from 0.766 to 0.813 (Figure 3b and 3c). This was somewhat expected since it is based on transcript quantification, and not on spliced read alignment. As proven in the literature, such a difference is expected since current transcript annotation models may result inaccurate [15].

Event type	Tool1	Tool2	Pearson
SE	ESGq	rMATS	0.952
	ESGq	SUPPA2	0.808
	rMATS	SUPPA2	0.836
A3	ESGq	rMATS	0.868
	ESGq	SUPPA2	0.786
	rMATS	SUPPA2	0.862
A5	ESGq	rMATS	0.920
	ESGq	SUPPA2	0.666
	rMATS	SUPPA2	0.708
RI	ESGq	rMATS	0.859
	ESGq	SUPPA2	0.677
	rMATS	SUPPA2	0.760

Table 2
Pearson correlation coefficients between ESGq, rMATS, and SUPPA2 (over the transcript quantification of Salmon ran with $k = 31$) broken down by event type on the 151bp paired-end dataset.

Table 2 reports the correlation between the three considered tools broken down by event type. Surprisingly there is no clear trend that can be observed. ESGq and rMATS exhibit the highest correlation on exon skipping events and the lowest on intron retentions. ESGq and SUPPA2, instead, show higher correlation on exon skippings and lower correlation on alternative donor events. Finally, rMATS and SUPPA2 exhibit higher correlation on alternative acceptor site and lower correlation on alternative donor events. These results are somewhat unexpected and require a further investigation.

ESGq also resulted very computationally efficient, completing the analysis in half an hour requiring 1GB of RAM. Similarly, SUPPA2 ran in less than 10 minutes and used 1.5GB of RAM. On the contrary, rMATS resulted the most expensive approach, requiring more than 5 hours and 8GB of RAM. The most expensive step is read alignment with STAR, that required from half an hour to two hours per sample. By pairing simple and precise graph representation of well localized loci of the genome (i.e., the event splicing graphs) with fast and accurate read alignment, ESGq is able to achieve results comparable to the other alignment-based approach, while being 10x faster.

Since SUPPA2 is based on the k-mer based quasi-mapping of Salmon, we also analyzed how k-mer size affects its results. For this reason, we ran Salmon (and, consequently, SUPPA2) two additional times with $k \in \{13, 21\}$ (we note that 31 is the default value used in the previous results). As shown in Table 3, the results of SUPPA2 seems to be unaffected by the choice of the k parameter. Indeed, the correlation between SUPPA2 (ran with different k value) and other tools changes marginally.

Moreover, to evaluate if the results of the considered methodologies are affected by read length, starting from the 151-bp paired-end dataset, we manually trimmed the input reads to 51bp and 101bp using seqtk and created two additional datasets. Table 3 reports the results of this analysis. Even with very short reads (i.e., 51bp reads), the three tools achieved the same correlation (with a very marginal difference of < 0.015).

Finally, to evaluate how much paired-end information may improve the accuracy of the tools, we merged the two pairs of each replicate into a single sample, in order to simulate a single-end dataset. Surprisingly, there is small to none difference between the results on paired-end and single-end dataset, highlighting the robustness of the considered approaches. For instance, the Pearson correlation coefficient between ESGq and rMATS decreased by a very marginal 0.009 whereas correlation between SUPPA2 and the other approaches decreased by 0.041 (w.r.t. ESGq) and 0.027 (w.r.t. rMATS).

The experimental evaluation has been implemented as a Snakemake workflow [32], thus it is fully reproducible and easily replicable. Scripts and instructions are available at https://github.com/AlgoLab/ESGq. All the experiments were performed on a 64bit Linux (Kernel 5.15.0) system equipped with two 16-core AMD EPYC 7301 2.2GHz processors and 128GB of RAM.

4. Conclusions

In this paper we introduced ESGq, a novel graph-based approach for the AS event quantification across two conditions. Differently from state-of-the-art tools, ESGq is based on read alignment against local graph structures, introduced here as *event splicing graphs*, that represent AS events precisely represented in a given gene annotation. An extensive exploratory analysis on real RNA-Seq dataset showed that ESGq is able to achieve comparable results with respect to other approaches based on the alignment of reads to the reference genome, while being 10x faster.

Future works will be devoted to improving the statistical framework behind the ψ and $\Delta\psi$ computation and to extending the experimental evaluation by assessing the actual accuracy of ESGq (e.g., by computing performance metrics on simulated and RT-PCR validated events). An interesting future direction consists in extending the notion of event splicing graphs to include information on known genetic variations (SNPs and indels) in order to improve the quality of read alignment, as proven in a recent work on pantrascriptomes [19]. Moreover, the detection and quantification of novel AS events from pantrancriptomes remain another interesting open problem.

Dataset	Reads	Tool1	Tool2	Pearson
PE	51	ESGq	rMATS	0.907
			SUPPA2 (k13)	0.764
			SUPPA2 (k21)	0.766
			SUPPA2 (k31)	0.764
		rMATS	SUPPA2 (k13)	0.807
			SUPPA2 (k21)	0.809
			SUPPA2 (k31)	0.808
	101	ESGq	rMATS	0.921
			SUPPA2 (k13)	0.763
			SUPPA2 (k21)	0.763
			SUPPA2 (k31)	0.761
		rMATS	SUPPA2 (k13)	0.797
			SUPPA2 (k21)	0.796
			SUPPA2 (k31)	0.797
	151	ESGq	rMATS	0.918
			SUPPA2 (k13)	0.766
			SUPPA2 (k21)	0.766
			SUPPA2 (k31)	0.766
		rMATS	SUPPA2 (k13)	0.811
			SUPPA2 (k21)	0.812
			SUPPA2 (k31)	0.813
SE	151	ESGq	rMATS	0.909
			SUPPA2 (k13)	0.728
			SUPPA2 (k21)	0.724
			SUPPA2 (k31)	0.725
		rMATS	SUPPA2 (k13)	0.789
			SUPPA2 (k21)	0.785
			SUPPA2 (k31)	0.786

Table 3
Pearson correlation coefficients between any combination of tools, i.e., ESGq, rMATS, and SUPPA2 (over the transcript quantification of Salmon ran with different k values) and experimental settings, i.e., different read lengths (51, 101, and 151bp) and paired/single-end dataset (PE/SE).

Acknowledgements

The authors thank Simone Ciccolella and Yuri Pirola for insightful discussion and technical advice.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Staff Exchange programme under the Marie Skłodowska-Curie grant agreement No. 872539.

References

[1] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, C. B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature 456 (2008) 470–476.

[2] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nature genetics 40 (2008) 1413–1415.

[3] B. R. Graveley, A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. Van Baren, N. Boley, B. W. Booth, et al., The developmental transcriptome of drosophila melanogaster, Nature 471 (2011) 473–479.

[4] M. Bhadra, P. Howell, S. Dutta, C. Heintz, W. B. Mair, Alternative splicing in aging and longevity, Human genetics 139 (2020) 357–369.

[5] S. C. Bonnal, I. López-Oreja, J. Valcárcel, Roles and mechanisms of alternative splicing in cancer—implications for care, Nature reviews Clinical oncology 17 (2020) 457–474.

[6] G. Biamonti, A. Amato, E. Belloni, A. Di Matteo, L. Infantino, D. Pradella, C. Ghigna, Alternative splicing in alzheimer's disease, Aging clinical and experimental research 33 (2021) 747–758.

[7] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with rna-seq, Nature biotechnology 31 (2013) 46–53.

[8] Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson, A. Monroy, P.-F. Kuan, S. M. Hammond, L. Makowski, et al., Diffsplice: the genome-wide detection of differential splicing events with rna-seq, Nucleic acids research 41 (2013) e39–e39.

[9] A. Fenn, O. Tsoy, T. Faro, F. L. Rößler, A. Dietrich, J. Kersting, Z. Louadi, C. T. Lio, U. Völker, J. Baumbach, et al., Alternative splicing analysis benchmark with dicast, NAR Genomics and Bioinformatics 5 (2023) lqad044.

[10] Y. Wang, J. Liu, B. Huang, Y.-M. Xu, J. Li, L.-F. Huang, J. Lin, J. Zhang, Q.-H. Min, W.-M. Yang, et al., Mechanism of alternative splicing and its regulation, Biomedical reports 3 (2015) 152–158.

[11] S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, Y. Xing, rmats: robust and flexible detection of differential alternative splicing from replicate rna-seq data, Proceedings of the National Academy of Sciences 111 (2014) E5593–E5601.

[12] J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, E. Eyras, Suppa2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions, Genome biology 19 (2018) 1–11.

[13] A. Kahles, C. S. Ong, Y. Zhong, G. Rätsch, Spladder: identification, quantification and testing of alternative splicing events from rna-seq data, Bioinformatics 32 (2016) 1840–1847.

[14] L. Denti, R. Rizzi, S. Beretta, G. D. Vedova, M. Previtali, P. Bonizzoni, Asgal: aligning rna-seq data to a splicing graph to detect novel alternative splicing events, BMC bioinformatics 19 (2018) 1–21.

[15] T. Sterne-Weiler, R. J. Weatheritt, A. J. Best, K. C. Ha, B. J. Blencowe, Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop, Molecular Cell 72 (2018) 187–200.e6.

[16] J. Vaquero-Garcia, A. Barrera, M. R. Gazzara, J. Gonzalez-Vallinas, N. F. Lahens, J. B. Hogenesch, K. W. Lynch, Y. Barash, A new view of transcriptome complexity and regulation through the lens of local splicing variations, elife 5 (2016) e11752.

[17] Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, J. K. Pritchard, Annotation-free quantification of rna splicing using leafcutter, Nature genetics 50 (2018) 151–158.

[18] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, et al., A draft human pangenome reference, Nature 617 (2023) 312–324.

[19] J. A. Sibbesen, J. M. Eizenga, A. M. Novak, J. Sirén, X. Chang, E. Garrison, B. Paten, Haplotype-aware pantranscriptome analyses using spliced pangenome graphs, Nature Methods (2023) 1–9.

[20] S. Beretta, P. Bonizzoni, L. Denti, M. Previtali, R. Rizzi, Mapping rna-seq data to a transcript graph via approximate pattern matching to a hypertext, in: Algorithms for Computational Biology: 4th International Conference, AlCoB 2017, Aveiro, Portugal, June 5-6, 2017, Proceedings 4, Springer, 2017, pp. 49–61.

[21] M. F. Rogers, J. Thomas, A. S. Reddy, A. Ben-Hur, Splicegrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data, Genome biology 13 (2012) 1–17.

[22] S. Beretta, P. Bonizzoni, G. D. Vedova, Y. Pirola, R. Rizzi, Modeling alternative splicing variants from rna-seq data with isoform graphs, Journal of Computational Biology 21 (2014) 16–40.

[23] J. Sirén, J. Monlong, X. Chang, A. M. Novak, J. M. Eizenga, C. Markello, J. A. Sibbesen, G. Hickey, P.-C. Chang, A. Carroll, et al., Pangenomics enables genotyping of known structural variants in 5202 diverse genomes, Science 374 (2021) abg8871.

[24] E. Garrison, J. Sirén, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, et al., Variation graph toolkit improves read mapping by representing genetic variation in the reference, Nature biotechnology 36 (2018) 875–879.

[25] J. Sirén, E. Garrison, A. M. Novak, B. Paten, R. Durbin, Haplotype-aware graph indexes, Bioinformatics 36 (2020) 400–407.

[26] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, et al., The evolutionary landscape of alternative splicing in vertebrate species, Science 338 (2012) 1587–1593.

[27] K.-T. Lin, A. R. Krainer, Psi-sigma: a comprehensive splicing-detection method for short-read and long-read rna-seq analysis, Bioinformatics 35 (2019) 5048–5054.

[28] M. Bajgiran, A. Azlan, S. Shamsuddin, G. Azzam, M. A. Halim, Data on rna-seq analysis of drosophila melanogaster during ageing, Data in brief 38 (2021) 107413.

[29] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nature methods 14 (2017) 417–419.

[30] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, Star: ultrafast universal rna-seq aligner, Bioinformatics 29 (2013) 15–21.

[31] J. Thurmond, J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, S. J. Marygold, B. B. Matthews, G. Millburn, G. Antonazzo, V. Trovisco, et al., Flybase 2.0: the next generation, Nucleic acids research 47 (2019) D759–D765.

[32] J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine, Bioinformatics 28 (2012) 2520–2522.