Fortuna Detects Novel Splicing in Drosophila scRNASeq Data

B. Borozan*, L. Borozan*, D. Ševerdija*, D. Matijević*, S. Canzar†
* Department of mathematics, University of Osijek, Osijek, Croatia
† Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA
bborozan@mathos.hr lborozan@mathos.hr

***Abstract*—Recent developments in single-cell RNA sequencing techniques (scRNASeq) have made large quantities of sequenced data available across numerous species and tissues. Alternative splicing (AS) of pre-mRNA introns varies between tissues and even between cell-types and can be altered in disease. The study of novel AS, using standard RNASeq data, has been extensively studied for many years, while similar work on scRNASeq data has been scarce, despite its potential to offer a broader insight into cell-type specific processes. In this paper, we propose a novel pipeline that uses fortuna, a method that efficiently classifies and quantifies novel AS events, to process scRNASeq samples. Due to its short lifespan, high number of progeny, low maintenance cost, and intricate alternative splicing patterns similar in complexity to those of mammals, *Drosophila Melanogaster* (fruit fly) is a species of particular interest to researchers. Therefore, we experimentally evaluate our pipeline on real-world *Drosophila* single-cell data samples from the Fly Cell Atlas.**

Keywords*—*novel splicing events, scRNASeq, fortuna, drosophila, alignment

I. Introduction

Multiple different mRNA molecules (transcripts) can be transcribed from the same genomic region [1] [2]. These overlapping transcripts, which contain codes for protein synthesis, can be the result of the mechanisms that regulate alternative splicing (AS) [3]. Known AS events are recorded in the transcriptome annotation, while novel and aberrant AS events can occur in disease [4] and are often unannotated. Deviations from the regular splicing process may have a significant impact on the organism, as discussed in [5]. Therefore, analysis of AS events is an important topic in the field of computational molecular biology. With the advent of single-cell RNA sequencing (scRNASeq) [6], an opportunity to study AS events from a different perspective has presented itself to researchers. As it is discussed in [7], the study of AS in the context of scRNASeq is challenging and the related research has been scarce, despite its clear benefits. The authors of [8] have created a comprehensive reference atlas comprising of nearly 500,000 cells from 24 different tissues and organs and have dedicated a part of their research to AS events. A similar reference atlas has been created for the fruit fly (*Drosophila Melanogaster*) [9], called the Fly Cell Atlas. It contains approximately 580,000 cells from 15 different tissues, but no analysis of AS events has been conducted. In data obtained by the means of traditional RNA sequencing (RNASeq), novel AS events can be indentified and quantified [10] after conducting the computationally challenging alignment process. Pseudoalignment methods such as [11] and [12], significantly outperform traditional alignment methods such as [13], [14], and [15] in terms of running time. Though, they are usually limited to pseudoaligning short reads to annotated transcripts and are unable to detect novel AS events. Only a single tool exists, to our best knowledge, that combines the speed of a pseudoaligner with an ability to detect novel splicing [16]. In the scRNASeq case, additional computation is required to account for differentiation between cells, correcting for in-vitro [17] or PCR [18] amplification and related sequencing errors.

In this paper we propose a novel pipeline which can be used to catalog novel AS events in scRNASeq data. It uses fortuna [16], a tool originally developed for novel AS event quantification in bulk RNASeq data, in conjunction with 10x Genomics Cellranger [19] to offset for the specifics of the scRNASeq analysis. Additional scripts are used to extract the information from fortuna and Cellranger output files and adjust the counts of novel AS events in line with corrections for scRNASeq. We test our pipeline on 12 real-world scRNASeq 10x samples from Fly Cell Atlas, obtained from ArrayExpress with the accession number E-MTAB-10519. The aim of our experiments is to produce a catalog of novel AS events which could be expanded to include the entire dataset or to process entirely different data. Apart from the catalog, we provide a running time and memory usage analysis. The code for the additional scripts can be found at fortuna fly atlas github page[1], while fortuna can be obtained at its official github page[2].

The structure of the paper is as follows. In the Methods section we will formulate the problem of building a novel AS catalog from scRNASeq data and present a novel pipeline which solves it. Then, in the Experiments section, we will present a catalog built from the real world data, analyze the novel splicing events found within and discuss the running time and memory usage of the tools that comprise our pipeline. Finally, in the Conclusion

[1] https://github.com/canzarlab/fortuna_flyatlas
[2] https://github.com/canzarlab/fortuna

section we will provide a discussion and a brief overview of our work.

II. Method

A. Problem formulation

There are a couple of challenges that need to be overcome in order to build a catalog of novel AS events from scRNASeq data. Reads coming from different cells are identified by their respective cell barcodes and each read is assigned a unique molecule identifier (UMI). During the sequencing process, both barcodes and UMIs may be incorrectly sequenced and, thus, have to be corrected. Furthermore, due to amplification, UMIs with the same corrected sequence have to be collapsed into a single UMI before further analysis can be conducted. Another issue that arises is the alignment of the reads to the transcriptome - a process in which we find the origin the read was likely sampled from. Each read that supports a novel AS event, i.e. splices over an unannotated intron, has to be recorded in the catalog. Finally, the catalog must only contain a list of novel AS events supported by the reads with corrected barcodes and UMIs.

B. Pipeline

Our proposed pipeline for classification and quantification of novel events in scRNASeq is comprised of two major steps:

1) sample processing, and
2) (pseudo-)alignment processing.

During the first step, we use fortuna to identify and quantify novel alternative splicing events, and Cellranger to perform barcode correction and UMI collapsing. During the second step, we adjust the counts of the events identified by fortuna according to Cellranger's corrections and output them into the final catalog. In the next two subsections we will explain each step in more detail.

C. Sample processing

The workflow of fortuna can be separated into three steps: indexing, pseudoalignment and postprocessing, out of which we will use only the first two. During the indexing step, fortuna supplements the annotation with a virtual transcriptome that enables the detection of novel AS events. An index constructed that way is used to pseudoalign an arbitrary number of samples and, at the same time, detect any potential novel splicing. We will use fortuna's pseudoalignments and novel AS catalog outputs (obtained by specifying *-bam* and *-alt* when running fortuna).

Cellranger is a software comprised of five pipelines intended for different steps of the single-cell analysis. We will use its *cellranger count* pipeline to correct cell barcodes and collapse UMIs. It stores known barcodes in a pre-defined whitelist and includes into the analysis all sampled barcodes that are a part of it. Each sampled barcode that has not been whitelisted, is included in the analysis if it is 1-Hamming-distance away from a whitlisted barcode, and additionally, if its probability of having originated from the whitelisted barcode is sufficiently high. Reads that contribute to UMI counts, i.e. are selected among other reads with the same UMI, are considered valid if their corresponding barcodes are valid and if they have sufficient mapping quality. Valid non-whitelisted barcodes have their UMIs merged with their respective whitelisted barcode. Cellranger outputs an alignment file with flags identifying the corrected values of UMIs and barcodes that we will use in the next step of our pipeline.

D. (Pseudo-)alignment processing

Cellranger's alignments are written in a binary file which can be converted to a plain text file using *samtools view* command [20]. On the other hand, similar binary file produced by fortuna has to be converted using *pseudo-to-genome-alignment* (p2g) tool provided as a part of our pipeline. This is a necessary step because pseudoalignments provided by fortuna are expressed in local fragment coordinates and have to be converted into standard genomic coordinates. The final part of our pipeline, *barcodeAS* tool, adjusts fortuna's catalog of novel AS events according to the information obtained from both pseudoalignment and alignment files by removing the contribution of the discarded reads (4th bit of the flag *xf:i:* in the alignment file).

III. Experiments

The experiments were conducted on dual 2.30 GHz Intel® Xeon® E5-2697 v4 processors, 320GB 2.40GHz DDR4 memory operating on Scientific Linux 7.5 (Nitrogen). C++ compiler used to compile fortuna and other scripts was GCC 4.8.5 20150623. GNU time command was used to record the running times and memory usage.

In the next subsection we will describe the dataset we processed and highlight the importance of correcting barcodes and UMIs. After that, we will present a novel AS catalog and will focus on the classification of novel AS events and their coverage over chromosomes in the dataset. Finally, we will discuss running times and memory usage of the tools we have used.

A. Experimental data

We obtained 12 Fly Cell Atlas testis samples sequenced using 10x Genomics technologies from ArrayExpress. Genome and annotation files were downloaded from FlyBase FTP, release number FB2019_06. The annotation was preprocessed for usage by Cellranger in accordance to [9], and for usage by fortuna using *processGTF* script

available at the fortuna github repository.

Each raw sample has between 125.6 and 183.1 million reads spread across between 6 and 7.7 million barcodes. The number of unique barcode and UMI combinations in the raw data ranges between 93.3 and 123.9 million. Cellranger *count* was run using default settings, 16 cores (*–localcores*), and 64GB memory (*–localmemory*). After the corrections made by Cellranger, the number of barcodes dropped to between 1.41 and 1.82 million across all samples, while the number of unique barcode and UMI combinations ranged between 61.47 and 81.62 million. The latter number coincided with the number of reads which we considered in our analysis.

The distribution of UMIs across barcodes in raw and corrected sample *S44* is depicted in Fig. 1. The number of barcodes with less than 120 UMIs steadily drops after we run Cellranger due to it discarding a lot of invalid ones. On the other hand, on the interval between 120 and 300 UMIs per barcode, the total number of barcodes increases. While correcting barcodes, Cellranger reassigns their UMIs to others, thus increasing the number of UMIs per barcode. But, the number of discarded raw barcodes is significantly larger than the number of those that have been merged, as can be seen in Fig. 1. There is little difference between raw and corrected barcodes with more than 300 UMIs, so we have omitted them from the figure. Similar conclusions can be made for all other samples on different intervals.

B. Novel alternative splicing catalog

We constructed a T^{as} fortuna index (*–index*) limiting the number of skipped exons (*-exs*) to 8 and the number of transcript fragments per gene (*-Mc*) to 25000. The rest of the settings were set to default. Alignment and novel AS event analysis were done using fortuna *–quant* using 4 processing threads and default settings. (Pseudo-)alignment files were processed using *samtools* and *p2g*, whose outputs were passed to *barcodeAS* to obtain the catalog.

Across all 12 samples, 23078 novel introns supported by a total of 2318688 reads were identified. We observed 5 different kinds of novel AS events induced by these introns, namely intron-in-exon (IE), alternative donors (AD), alternative acceptors (AA), alternative donor-acceptor pairs (AP) and exon skipping (ES). We have included them in our analysis only if their respective number of supporting reads was at least 2. The definitions of these novel AS events are consistent with those in [16]. The number of novel AS events and their read support is summarized in Table I. Note that most events are detected in multiple samples so the sum over the number of events in Table I does not yield the total number of unique events. Furthermore, some novel introns can be classified in multiple ways, and thus increase the count in multiple columns. Approximately 93.47% of all novel AS events we have detected have been exon skipping events, supported by about 81.26% of the reads. Alternative donors amounted to 4.68% and 6.28%, alternative acceptors to 3.93% and 6.01%, intron-in-exon events to 2.63% and 5.75%, and alternative donor-acceptor pairs to 1.07% and 0.69% of the total events and reads, respectively. Fig. 2 depicts the total amount of novel splicing events and their supporting reads across combined samples.

We analyzed events supported by at least 2 reads, and cells in which we have detected at least one novel AS event, for all combined samples. In 90.5% of the cells we

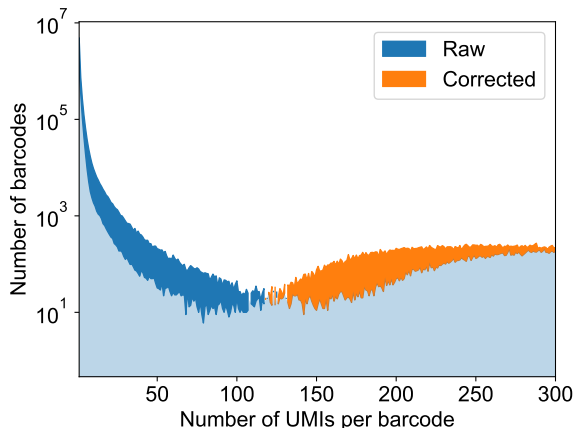

Fig. 1: Overlapped log-scaled histograms representing the number of barcodes having between 1 and 300 unique UMIs in raw (blue) and corrected (orange) sample *S44*. Overlap between bars has been colored light blue.

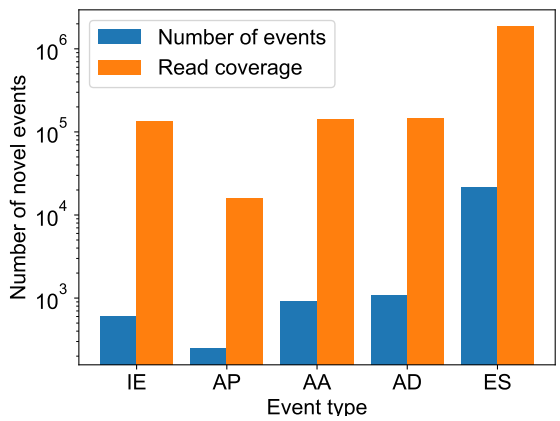

Fig. 2: Log-scaled amount of detected novel AS events separated by their type (intron-in-exon, alternative donor-acceptor pairs, alternative acceptors, alternative donors, exon skipping) and their supporting reads from twelve combined *Drosophila* samples.

Sample	Number of events					Read coverage				
#	IE	AP	AA	AD	ES	IE	AP	AA	AD	ES
S44	350	116	485	583	9577	10319	1333	11641	11935	151275
S48	391	121	509	628	10694	12861	1615	14192	14642	183180
S49	370	109	481	584	9934	11052	1382	10955	12541	161098
S56	337	98	454	573	9377	9665	1177	9655	10277	143079
S59	366	109	463	581	9739	10650	1269	11718	11617	150414
S60	389	111	478	590	9911	11825	1330	11353	12434	159158
S61	370	117	472	607	9820	10692	1272	11836	11970	152819
S65	364	99	473	564	9419	10861	1175	10465	11506	147321
S67	353	104	464	574	9514	10002	1214	10067	11002	143471
S68	357	112	489	583	9773	11039	1255	12209	12313	154495
S69	395	116	520	657	10707	13659	1586	14637	14817	183609
S70	365	114	479	587	9792	10807	1345	10703	10602	154297

TABLE I: Number of detected novel AS events separated by their type and their supporting reads in each of the twelve *Drosophila* samples.

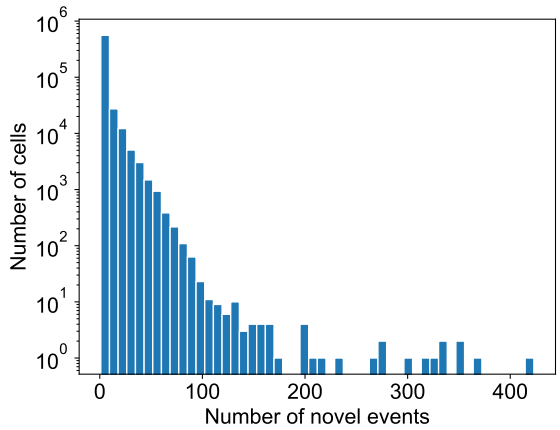

Fig. 3: A histogram of the log-scaled number of cells in relation to the number of novel AS events detected in them for the twelve combined *Drosophila* samples.

found at most 8 different novel AS events, and in 54.8% no more than 1 novel AS event. The highest number of events detected in a single cell was 423. We present the distribution of the number of novel AS events across the cells in Fig. 3.

In Fig. 4 we present the combined coverage of novel AS events (blue bars) across different chromosomes. X-axis represents genomic coordinates, while the left y-axis depicts the number of detected novel AS events. Additionally, the red line signifies the number of different cells in which we have detected the aforementioned novel AS events. Its corresponding values are located on the right y-axis. As can be seen in Fig. 4, there exist a few regions on each chromosome where certain novel AS events have been observed in a large amount of different cells, e.g. red line reaching a value over 60000 on chromosome 2R. This could indicate the existence of AS events that are common across multiple cells or tissues. Such cases could be investigated in the data comprised of multiple samples of different tissues and any suitable AS events could be considered for patching into the annotation. On the other hand, high coverage of novel AS events present in a low number of cells, visualized by tall blue bars and low red values, might point to processes affecting only specific cells that are worth investigating. Such examples can be visually detected in Fig. 4 on, e.g. chromosome 4.

We have investigated the outlier on chromosome 2R and found that a single novel AS event at genomic coordinates *2R:17478508-17485002* is supported by 135285 reads and is detected in 64511 different cells across all samples. It might suggest that the particular intron is a part of a common unannotated transcript. Investigation of this as well as other similar outliers through the genome might prove an interesting topic for further research.

C. Running time and memory usage

The experiments section will be concluded with a running time and memory usage analysis. There are five essential components of our pipeline - Cellranger, fortuna, samtools, *p2g*, and *barcodeAS*. Their memory usages and running times are presented in Table II. Cellranger and fortuna ran, on average, for 48m50s and 26m43s, respectively, while consuming 3.32 and 6.90 GB of memory. Increasing the amount of available threads would decrease the running time for both tools with some diminishing returns, likely due to disk-related limitations. Average samtools and *p2g* running times were 11m28s and 15m4s, while their memory consumtion was 1.29 and 310.39 MB. Finally *barcodeAS* ran for 24m56s while using 14.02 GB. That would mean that an average run of the entire pipeline would last slightly less than 130m and consume 14.02 GB of memory. Notice that *barcodeAS* consumed the most memory out of all listed tools. Reducing its memory usage could be an important part of our future work.

IV. Conclusion

We have presented a novel pipeline for building a catalog of novel alternative splicing events from scRNASeq data which uses fortuna and Cellranger in its core. Then, we have tested the pipeline using 12 real-world *Drosophila Melanogaster* testis samples from

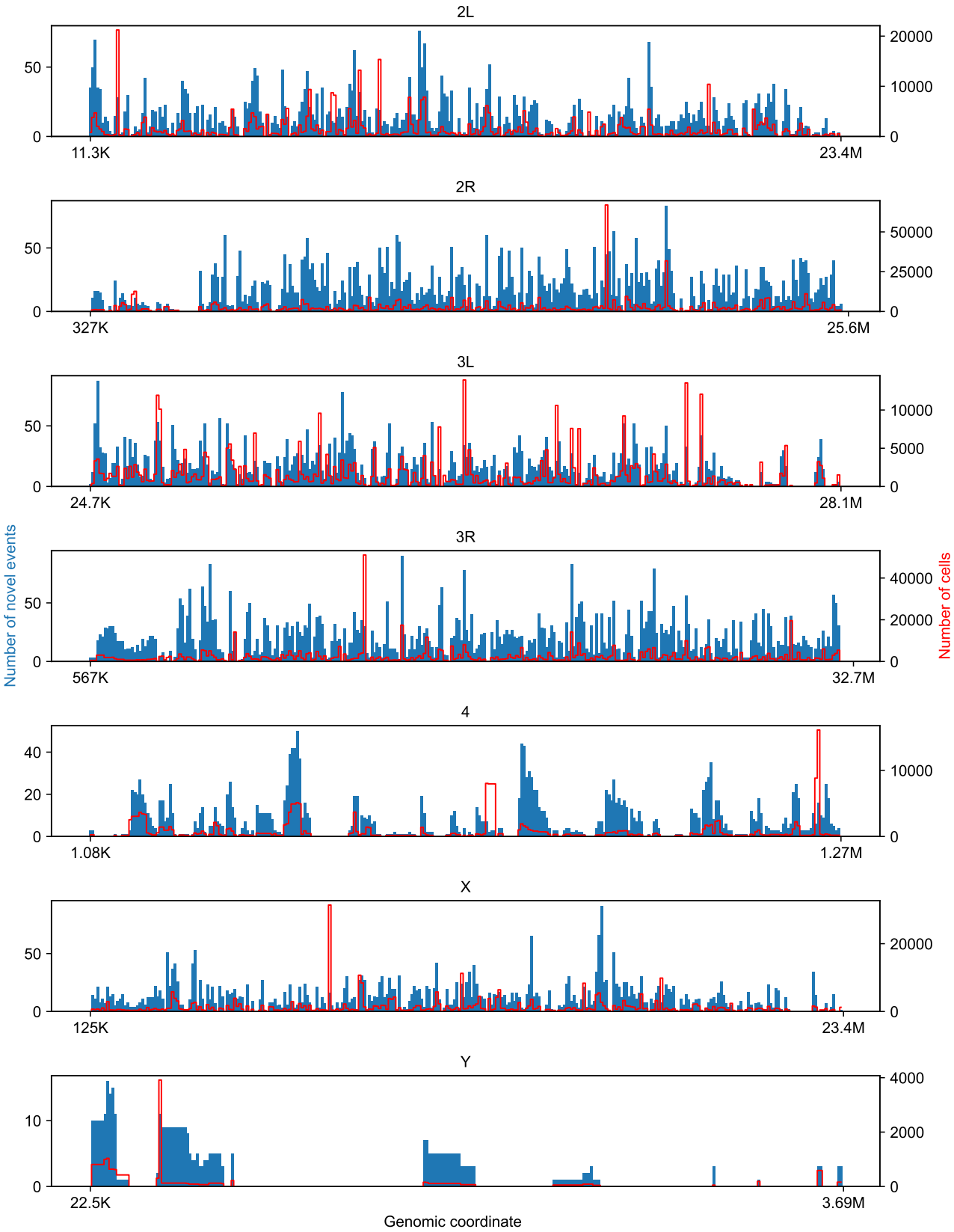

Fig. 4: The coverage of novel AS events across the chromosomes (blue) and the number of cells in which these events have been detected (red) for the twelve combined *Drosophila* samples.

Sample	Running time					Memory consumption (GB)			
#	fortuna	Cellranger	p2g	samtools	barcodeAS	fortuna	Cellranger	p2g	barcodeAS
S44	26m39s	50m54s	15m05s	8m47s	24m32s	6.88	3.26	0.30	13.85
S48	34m42s	60m08s	19m41s	12m04s	35m41s	6.90	3.50	0.30	16.79
S49	29m19s	55m39s	17m20s	10m03s	28m30s	6.90	3.32	0.30	14.76
S56	23m52s	44m18s	14m12s	11m46s	25m27s	6.88	3.17	0.30	13.03
S59	25m00s	44m33s	13m36s	14m24s	26m18s	6.90	3.29	0.30	13.37
S60	25m35s	47m25s	14m11s	9m20s	25m08s	6.90	3.36	0.30	13.98
S61	24m33s	49m28s	14m20s	15m10s	25m01s	6.90	3.34	0.30	13.38
S65	23m48s	42m08s	13m16s	8m26s	29m52s	6.90	3.16	0.30	12.94
S67	23m14s	41m25s	13m02s	13m47s	22m10s	6.90	3.27	0.30	12.65
S68	26m26s	45m20s	14m01s	9m26s	26m02s	6.90	3.20	0.30	13.57
S69	32m23s	56m01s	17m56s	11m47s	32m29s	6.91	3.50	0.30	16.16
S70	25m00s	48m40s	14m09s	12m37s	30m39s	6.90	3.40	0.30	13.71

TABLE II: Running times and memory usage of all tools in our pipeline recorded for each *Drosophila* sample. Samtools have been omitted from the memory consumption due to using less than 2 MB of memory.

Fly Cell Atlas. In our experiments, we have analyzed and discussed potential benefits of computing the coverage of novel AS events across the chromosomes, split novel AS events and their supporting reads by the type of AS, and briefly commented the running times and memory usage of the each part of our pipeline.

We acknowledge several new directions we could take in our future work. Working directly with binary alignment files without converting them into plain text, thus omitting samtools, could significantly speed up the process of building a catalog. Expanding the experiments to encompass the entire Fly Cell Atlas or even scRNASeq datasets of other species could result in potentially interesting and important biological findings. Finally, the inclusion of novel intron retention events could prove useful for downstream analysis.

REFERENCES

[1] E.T. Wang et al., "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.

[2] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.

[3] D. Brett, H. Pospisil, J. Valcárcel, J. Reich, and P. Bork, "Mechanisms of alternative pre-messenger rna splicing," *Nature Genetics*, vol. 30, no. 1, pp. 29–30, 2002.

[4] K. Jaganathan et al., "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535–548, 2019.

[5] O. Kelemen et al., "Function of alternative splicing," *Gene*, vol. 514, no. 1, pp. 1–30, 2013.

[6] T. K.Olsen and N. Baryawno, "Introduction to single-cell rna sequencing," *Curr Protoc Mol Biol*, vol. 122, no. 1, 2018.

[7] A. Arzalluz-Luque and A. Conesa, "Single-cell rnaseq for the study of isoforms—how is that possible?" *Genome Biol*, vol. 19, no. 110, 2018.

[8] The Tabula Sapiens Consortium and S. R. Quake, "The tabula sapiens: a multiple organ single cell transcriptomic atlas of humans," *bioRxiv*, 2022. [Online]. Available: https://www.biorxiv.org/content/early/2022/03/04/2021.07.19.452956

[9] H. Li and J. Janssens et al., "Fly cell atlas: A single-nucleus transcriptomic atlas of the adult fruit fly." *Science*, vol. 375, no. 6584, 2022.

[10] A. Kahles, C. S. Ong, Y. Zhong, and G. Ratsch, "Spladder: identification, quantification and testing of alternative splicing events from rna-seq data," *Bioinformatics*, vol. 32, no. 12, pp. 1840–1847, 2016.

[11] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic rna-seq quantification," *Nature Biotechnology*, vol. 34, pp. 525–527, 2016.

[12] R. Patro, G. Duggal, and M. Love et al., "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods*, vol. 14, pp. 417–419, 2017.

[13] A. Dobin et al., "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[14] D. Kim, J. M. Paggi, and C. Park et al., "Graph-based genome alignment and genotyping with hisat2 and hisat-genotype," *Nature Biotechnology*, vol. 37, pp. 907–915, 2019.

[15] D. Kim, G. Pertea, C. Trapnell, R. K. H. Pimentel, and S. L. Salzberg, "Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, p. R36, 2013.

[16] L. Borozan, F. Rojas Ringeling, S. Kao, E. Nikonova, P. Monteagudo-Mesas, D. Matijević, M. L. Spletter, and S. Canzar, "Counting pseudoalignments to novel splicing events," *article in revision*.

[17] J. Eberwine et al., "Analysis of gene expression in single live neurons," *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 3010–3014, 1992.

[18] G. Brady, M. Barbara, and N. N. Iscove, "Representative in vitro cdna amplification from individual hemopoietic cells and colonies," *Methods Mol. Cell Biol.*, vol. 2, pp. 17–25, 1990.

[19] G. X. Y. Zheng et al., "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, pp. 1–12, 2017.

[20] H. Li et al., "The sequence alignment/map (sam) format and samtools," *Bioinformatics*, vol. 25, pp. 2078–2079, 2009.