Comparing HISAT and STAR-based pipelines for RNA-Seq Data Analysis: a real experience

Andrea Bianchi, Antinisca Di Marco, Cristina Pellegrini
University of L'Aquila, L'Aquila, Italy
{andrea.bianchi}@graduate.univaq.it
{antinisca.dimarco, cristina.pellegrini}@univaq.it

2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-1224-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CBMS58004.2023.00220

***Abstract*—One of the first step in RNA-Sequencing (RNA-Seq) data analysis consists of aligning (Next Generation Sequencing) reads to a reference genome. In literature, there are several tools implemented by practitioners and researchers for the alignment step. However, two tools are the de-facto-standard used by bioinformatics researchers in their pipelines: HISAT (version 2) and STAR (version 2). The aim of this study is to determine the impact of the alignment tool on the RNA-Seq analysis in terms of biological relevance of the results and computational time. The two implemented pipelines return different results on the biological side. This is due to assumptions the used tools made and to the specific characteristics of the underlying (statistical) models. The study provides valuable insights for researchers interested in optimizing their RNA-Seq pipelines and making informed decisions about which pipeline to use. As lesson learned, we suggest bioinformatics researchers to use more pipelines when make experiments to reduce the prediction errors induced by assumption of a specific tool or method.**

***Index Terms*—bioinformatics, RNA-Sequencing, myelofibrosis, comparison, pipelines**

I. Introduction

Bioinformatics has seen tremendous growth in recent years, particularly with the advancement of high-throughput sequencing data analysis. RNA-Sequencing (RNA-Seq) experiments are a powerful tool for studying gene expression patterns and identifying differentially expressed genes. For these experiments to be successful, it is essential that bioinformatics pipelines, used for processing and analyzing the data, are of good quality.

The alignment (or mapping) step of RNA-Seq data analysis is the most computationally intensive and time-consuming process, as it involves aligning reads generated from a sequencing experiment, with a reference genome. Choosing the right mapping tool for this task is fundamental when computational efficiency and biological accuracy of results are relevant aspects.

Accurate mapping is essential for downstream analysis, but the presence of splice junctions in RNA-seq reads poses a challenge for alignment accuracy. To address this challenge, several software platforms have been developed for mapping to a reference genome, including TopHat (version 2, hereafter referred to as Tophat2) [15], HISAT (version 2, hereafter referred to as HISAT2) [14], and STAR (version 2, hereafter referred to as STAR2) [11]. TopHat2 was a popular choice but has been superseded by HISAT2 due to its computational inefficiency. TopHat2 and HISAT2 are built on top of the popular short-read mapping tool Bowtie2 [16]. While all three aligners are considered fast, the choice of the optimal aligner can significantly impact downstream analysis. Therefore, it is crucial to evaluate the alignment performance of different aligners to identify the optimal tool for the task. This study aims to compare two different bioinformatics pipelines using HISAT2 and STAR2 in order to assess their performance and output quality and to identify the optimal tool for accurate mapping and downstream analysis. The choice of these two aligners is motivated by their widespread use and performance superiority over TopHat2 [2]. In addition to the widespread use of HISAT2 and STAR2, we selected these two aligners for comparison because they are the latest versions and have improved upon their predecessors [14] [11] [8]. HISAT2 is an improved version of HISAT, and STAR2 is an updated version of STAR. These newer versions have addressed some of the limitations of their predecessors, such as increased accuracy, speed, and memory efficiency.

HISAT2 and STAR2 are RNA-Sequencing (RNA-Seq) alignment tools that differ in their alignment strategy, index size, sensitivity, speed and read type optimization. HISAT2 uses graph FM index (GFM) [14], has a smaller index size, is faster but has limited multi-threading capabilities, and is optimized for both single-end and paired-end reads.

STAR2 uses Spliced Transcripts Alignment [11] as a reference algorithm, has a larger index size, higher sensitivity, and better multi-threading, but it is slower, and it is optimized for spliced reads.

There are only a few studies that compare different tools for analyzing RNA sequencing data on cancer datasets. In [10], authors found that STAR2 had better performance in terms of the percentage of uniquely mapped reads (precisely, 80%) compared to HISAT2 (70%), using different genome assemblies (hg19 and hg38). The percentages of unmapped reads are larger in HISAT2 and Tophat2 in both genomic assemblies. The study in [20] found that HISAT2 aligned fewer reads and had higher rates of alignment to pseudogenes, compromising alignment fidelity and potentially leading to erroneous outputs. In contrast to the previous studies, our paper aims to extend the comparison of HISAT2 and STAR2 beyond the analysis of the percentage of uniquely mapped reads by assessing their performance in terms of the biological relevance of the obtained results. Specifically, we employ the hg38 genome assembly, which is the most recent and

2372-9198/23/$31.00 ©2023 IEEE
DOI 10.1109/CBMS58004.2023.00048

recommended assembly for genome-wide analyses, to evaluate the differential expression results obtained from each pipeline. Moreover, we monitor the computational time of the two implemented pipelines to quantify and compare their time complexity.

This work is related to [6], where we studied the potential effects of the Ruxolitinib drug. Recently, this drug has been approved by the FDA to treat patients with myelofibrosis (MF), a disease that affects the bone marrow. While ruxolitinib improves symptoms, it doesn't completely cure the disease or significantly reduce the number of mutated cells. This is because some MF cells are resistant to the drug, possibly due to additional genes or pathways that promote cell survival even when the JAK2/STAT5 pathway, the targets of the Ruxolitinib, are suppressed. Using a library of genetic tools we found that several members of the proteasomal genes' family are important for cell survival, and inhibiting them with carfilzomib drug made the MF cells more susceptible to ruxolitinib. Additionally, the combination of ruxolitinib and carfilzomib reduced proteasomal gene expression in MF cells, suggesting that this approach could be effective in treating MF patients. Unlike [6], the aim of this study is to present a comparison of HISAT2 and STAR2 for RNA-Seq data processing in terms of execution time as well as biological relevance of the results obtained from the alignment, quantification, and differential expression analysis steps.

This study will provide an invaluable resource to researchers aiming to optimize their RNA-Seq data analysis pipelines and to make informed decisions about which pipeline best fits their needs, by providing insights into the balance between computational efficiency and biological accuracy.

The paper proceeds as follows: Section II describes the general workflow of RNA-Seq analysis and the used tools to implement two pipelines. In Section III, we report the experimental settings and the used dataset. Section IV discusses the obtained results in terms of computational time and biological relevance of the two implemented pipelines. Finally, Section V concludes the paper, highlighting the main identified insights and possible future work.

II. RNA-Seq General Workflow and Implemented Pipelines

In RNA-Seq experiments, there is no a single pipeline that works better in every situation. Depending on research goals and sequenced organisms, different approaches could be taken into account with various available software tools [7], [12]. Figure 1 reports the generic workflow for RNA-Sequencing and the specific tools we used to implement two distinct pipelines.

The alignment speed is a critical performance aspect of bioinformatics tools used in the mapping process [19]. For this reason, we decided to vary the tool used in this phase while keeping constant the ones used in the others steps. Our pipelines will differ only concerning their respective mapping parts. We chose two different tools based both on their symmetrical working approaches and their reliability and performance results reported in [2], which showed that they are currently among the best options available.

Looking at Figure 1, a RNA-Seq workflow is composed by four main steps: *Quality control* (presented in Section II-A), Alignment (described in Section II-B), Quantification (presented in Section II-C) and the Differential Expression (DE) Analysis (in Section IV).

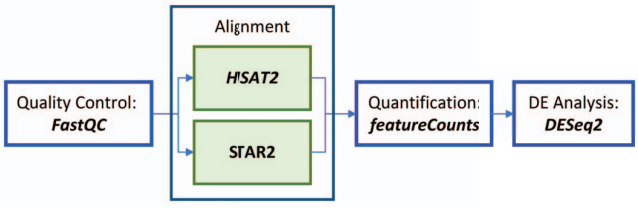

Fig. 1. General workflow for RNA-Seq analysis and the implemented pipelines.

In the following we describe in detail all the steps and the used tools while we report about their settings in Section III.

A. Quality Control

The first step of the RNA-Seq workflow is Quality Control. Specific actions must be taken during a quality control process to ensure that raw data are of the desired quality. This frequently entails adapter trimming, which involves removing any sequences that are not from the source organism and filtering out low-quality reads and uncalled bases. Because we have data from the Illumina platform, we use FastQC [1], one of the most popular software for this purpose, to evaluate the file quality. In case the quality control revels quality issues, the workflow adds a step aiming to remove adapter sequences and to trim low-quality bases (e.g., using Trimmomatic v0.39 [5]). In the conducted experiment, because the data was of sufficient quality, there was no need for adaptors removal or trimming, according to the findings.

B. Alignment

The alignment is a critical component of many bioinformatics analyses in which high-throughput sequencing data - in terms of individual reads - are compared with respect to a reference genome or transcriptome. The alignment step returns BAM files, passed as input to Quantification step. To implement the alignment step we considered two distinct tools: HISAT2 and STAR2 that we detail in the following.

1) Alignment Tool #1: HISAT2: HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts) [13] is a widely used RNA-Seq mapping tool that offers several advantages over other tools. It utilizes a hierarchical indexing approach to align reads to the reference genome, making it faster and more efficient than many other mapping tools. Additionally, its smaller index size requires less disk space and makes it quicker to build the index. HISAT2 is optimized for both single-end and paired-end reads and can be up to twice as fast in alignment due to its improved alignment strategy [13]. As a result, HISAT2 is an ideal candidate for analysing

transcriptomes and spliced exons. Because of its hierarchical indexing strategy, one of HISAT2's primary characteristics is its ability to reliably align readings even in areas with complicated splicing patterns. Thanks to its high sensitivity, HISAT2 is a helpful tool for large-scale RNA sequencing studies. However, the hierarchical indexing strategy utilised by HISAT2 consumes higher computational resources than other aligners, making it less appropriate on low-power computer platforms. Furthermore, HISAT2 is able to swiftly align data hence it invests a significant amount of time in constructing an extensive and complete index just once. As a consequence, it may align data in subsequent reads more quickly and with less memory utilisation.

2) Alignment Tool #2: STAR2: STAR2 (Spliced Transcripts Alignment to a Reference) [11] is one of the most popular RNA-Seq mapping tools. This algorithm utilizes a splicing transcripts alignment approach, which provides improved sensitivity compared to other methods and allows for more reads to be accurately aligned with the reference genome. STAR2 has superior multi-threading capabilities that allow it to make efficient use of multiple cores on systems with many processors. However, this requires larger index sizes than HISAT2. Since STAR2 is optimized for spliced reads, it's particularly advantageous when working with exon-exon junctions or gene fusions. The mechanism for splicing alignment differs between STAR2 and other aligners such as HISAT2. STAR2 employs a novel spliced alignment method, the *unique spliced alignment strategy*, whereas HISAT2 employs a hierarchical indexing method. Briefly, the *unique spliced alignment strategy* is a method for matching RNA sequencing reads to a reference genome in the context of spliced exons. It creates a genome index that detects spliced reads and then matches the reads to the genome using the index. The index is constructed by mapping splice junctions from the reference genome to anchors, and the alignment process uses the anchors to locate spliced exons in the reads. Even in locations with complicated splicing patterns, this leads in reliable matching of spliced reads to the reference genome. Another distinction is efficiency since STAR2 is optimised for large-scale RNA sequencing operations. STAR2's unique spliced alignment technique may also necessitate more computational resources than other aligners, making it unsuitable for usage on low-power computer platforms. Moreover, STAR2 does not provide pre-built indexes. This means that the user must create the index from scratch for each reference genome used in order to utilise the STAR2 tool. This required procedure can take a long time and require a lot of computational resources, especially for big genomes. In comparison to other alignment software, such as HISAT2, STAR2 lacks pre-built indexes compromising the efficiency and resource saving in case of pipeline reuse.

C. Quantification

Once the mapping step has been completed, the subsequent step is to count the number of reads associated with the characteristics of interest (genes in our experiment) since we aim to perform differential expression analysis on the genes, comparing them under different experimental conditions. We ran *featureCounts* [17] on all of the BAM files coming from the previous alignment step, in both pipelines at the same time.

D. DE Analysis

DE analysis is a critical stage in RNA-seq data processing that tries to discover genes that are expressed differently under different experimental circumstances. Following the normalization of the RNA-seq quantification data, differential expression analysis was performed to determine over- or down-expressed genes.

In this work, we used R programming language to perform differential expression analysis on RNA-seq data obtained by the HISAT2 and STAR2 pipelines. The *DESeq2* [18] tool was specifically utilised to evaluate the quantification data.

III. Experimental Settings

In this section we present the experimental settings we used in our experiments. In particular, in Section III-A we describe the used dataset, in Section III-B we describe the hardware we used. Finally, in Section III-C we describe the software configuration.

A. Dataset Description

In this study, we aimed to analyze the mRNA expression of CD34+ hematopoietic stem cells isolated from peripheral blood. The samples are from patients affected by myelofibrosis. In the dataset, we had samples from five patients, but during quality control, we discovered contamination by Escherichia coli in one patient's samples. To ensure accuracy and avoid any influence on our experiment, we decided to remove the contaminated samples. Hence, we worked with 32 fastq files coming from four patients, each providing two samples. The first one was extracted from CD34+ cells treated with Ruxolitinib drug, while the other was an untreated sample. Each patient has two replicates. Each file in the experiment is in the *fastq* format with an average size of 3 GB when unzipped. In the end, we worked on 96 GB of data in total. The RNA sequencing is then performed on the eight samples using a pair-ended design for duplication purposes. The samples were sequenced on an Illumina platform using the HiSeq 2500 system. The sequencing process produced reads ranging from 60 to 90 million bases for each sample. Almost all samples had at least 70 million reads.

The raw sequencing data generated in this study are available upon request, but, due to patient confidentiality concerns, they can be accessed only on-site.

B. Hardware Configuration

The research project is conducted on Caliban, a cluster environment comprising multiple nodes, with computations performed on a single node. The hardware configuration of this node is described in the following. The central processing unit (CPU) is comprised of 48 individual CPUs, each with a clock speed of 2.16 GHz. This configuration enables the node to perform a high volume of computations in parallel, resulting in

faster data processing and analysis times. The random access memory (RAM) of the node is 141.48 GB. In addition, the node is equipped with a local disk storage capacity of 1.5 TB, providing ample space to store intermediate and final results of computations and analyses. The node runs on the Linux 3.10.0 operating system. Although the node utilised in this research project has 48 CPUs, it should be noted that not all of the project's tools and software can efficiently leverage parallel processing. As a result, whenever available, multi-threading solutions were used to maximise computing performance.

C. Software Configuration

We chose Anaconda version 3 as the environment for bioinformatics tools and their setups. In particular, we used Anaconda's package manager (Conda) to create custom bash scripts to automate the execution of RNA-Seq workflow on the cluster. We report in Table I the commands used to invoke the selected tools and the related configurations. On the rows of the table, we report the used tools and for each of them we indicate the workflow *phase* they implement, the *command* and the relative *arguments* used in the experiments.

Both pipelines employed the GRCh38 genome, which was acquired from Ensembl (https://www.ensembl.org/Homo_sapiens/Info/Index), as the reference sequence for read alignment. In addition to the genome, the corresponding Gene transfer format (GTF) file, which encompasses information regarding splice sites and exons, was utilized to construct the index. The indexing process is known to be both time- and resource-intensive. To minimize the time required, the process was facilitated by utilizing a multi-core processor, with 40 threads being utilized both for the index construction and the actual alignment steps. In order to ensure a fair and accurate comparison between the two pipelines, the index was constructed from scratch instead of utilizing a pre-existing one. Because index formation is a critical element of each alignment tool, this choice was made to avoid possibly hiding the time necessary for index development and to avoid compromising the validity of the comparison. It is worth noting that the STAR2 does not have any pre-built indexes, therefore the user must create the index from scratch for each used reference genome.

Finally, in DE analysis, performed to determine over- or down-expressed genes, two different thresholds were considered. Both such thresholds enable setting biological and statistical relevance. They are:

- the *padj* threshold: it represents the level of statistical significance of the differential gene expression; it corresponds to an adjusted p-value and it has been set to 0.05 value;
- the *log fold change* threshold: $\log_2$ fold change threshold reflects the magnitude of the biological differences between the experimental conditions; it has been set to 1 (in absolute value).

These thresholds were applied consistently across both pipelines to ensure that the results were comparable and biologically relevant.

Code Availability

In this study, the pipelines were implemented as bash scripts and executed on a Linux Cluster. The associated code is accessible at the provided reference [4] and is governed by the Creative Commons Attribution 4.0 International license.

IV. Results

In this section, we perform a comprehensive evaluation of the distinct bioinformatics pipelines to compare their execution time and biological results. The computing time analysis focuses on the three main steps of the pipelines, namely the index creation, alignment and quantification, to understand the overhead introduced by the alignment tools. On the other side, the biological results evaluation aims to compare the pipelines in terms of the biological relevance of the Alignment and Quantification steps (in Section IV-2) on one side, and of the Differential Expression step, (in SectionIV-3) in the other side. Our aim is to provide a thorough comparison of the two pipelines highlighting their strengths and limitations.

1) Computing time: In the experiment, we configured the tools of the two pipelines with consistent configuration parameters (e.g., the same number of threads, the considered mapping regions), as reported in Section III.

The results in Table II suggest that while HISAT2 requires more time for index creation, it performs similarly in terms of alignment time and slightly slower in terms of quantification time compared to STAR2. The discrepancies seen in the quantification stage between the two pipelines might be traced to the methods utilised by HISAT2 and STAR2 during alignment. Because the software and hardware setup was identical, the observed discrepancies in processing time may be attributed *i)* to intrinsic differences in the algorithms and how they handle data and/or *ii)* to different implementation of multithreading options between HISAT2 and STAR2 and/or to their models and optimization techniques.

2) Biological Relevance of Alignment and Quantification: Another dimension to consider in the comparison is the alignment overall and uniquely mapped reads (Table III). The overall alignment rate between the two pipelines was similar: HISAT2 aligned 98.03% of the reads and STAR2 aligned 98.78% of the reads. Although the difference in overall alignment rate was relatively small, the difference in the number of uniquely mapped reads was more substantial. HISAT2 aligned 80.47% of the reads as uniquely mapped, while STAR2 aligned 81.66% of the reads as uniquely mapped. This resulted in a difference of over 300,000 uniquely mapped reads: HISAT2 aligned 24,309,436 reads while STAR2 aligned 24,667,395 reads. It is important to note that this difference in terms of the number of uniquely mapped reads could have implications for the next steps in the pipeline and/or in the final biological results and should be taken into consideration when interpreting them. We want to highlight that HISAT2 has been reported to have a slightly higher rate of multi-mapped

TABLE I
Commands used for each tool in the bioinformatics pipelines and the related configurations

Tool	Phase	Command	Arguments
FASTQc	Quality Control	fastqc	-t 40 -o output_dir input_file
HISAT2	Alignment (Indexing)	hisat2-build	– p 40 – ss splice_sites.txt – exon exons.txt
STAR2	Alignment (Indexing)	star	– runThreadN 40 – runMode genomeGenerate – sjdbGTFfile Homo_sapiens.GRCh38.97.gtf – genomeDir GRCh38 – genomeFastaFiles GRCh38.dna.primary.fa
STAR2	Alignment	star	– runThreadN 40 – genomeDir GRCh38 – sjdbGTFfile Homo_sapiens.GRCh38.97.gtf - readFilesIn Sample_R1.fastq Sample_R2.fastq - outSAMtype BAM SortedByCoordinate - outSAMunmapped Within - outSAMattributes Standard - quantMode GeneCounts - outFileNamePrefix AlignmentSample1 - twopassMode Basic
featureCounts	Quantification	featureCounts	-T 40 -p -t exon -g gene_name -a Homo_sapiens.GRCh38.97.gtf -o countmatrix.txt S1.bam ... Sn.bam
DESeq2	DE Analysis	R Script (custom)	Official Code: [3]

TABLE II
Computing time of index creation, alignment and quantification steps.

pipeline name	index creation (minutes)	alignment (minutes)	quantification (minutes)
HISAT2	54	10.19	3.34
STAR2	28	10.43	2.18

reads (17.12% for HISAT2 compared to 15.9% for STAR2). As multi-mapped reads can lead to inaccurate mapping and affect downstream analyses [9], such as differential expression analysis, correcting for this issue can potentially lead to the identification of more genes by HISAT2. This could be a possible explanation for why HISAT2 identified more differentially expressed genes than STAR2.

TABLE III
Percentage and number of mapped reads obtained after the alignment.

pipeline name	overall alignment rate	uniquely mapped read rate	no. of (uniquely) mapped read (in millions)
HISAT2	98,03%	80,47%	24.309.436
STAR2	98,78%	81,66%	24.667.395

3) Biological Relevance of Differential Expression: In Figure 2, we report, by means of Venn Diagrams, the number of differentially expressed genes found by HISAT2 and STAR2 pipelines as total numbers of found genes (first Venn diagram), and details of over-expressed and down-expressed genes (respectively in the second and third diagram).

Based on our findings in Figure 2, the HISAT2 pipeline identified more expressed genes (n = 197) than the STAR2 pipeline (n = 147) within our statistical and biological significance thresholds (reported in Section III). A deeper look at the Venn diagrams, however, revealed that 138 genes were identified as differentially expressed in both algorithms. In other words, practically all of the genes found in the STAR2 pipeline were also found in the HISAT2 pipeline. Of the 138 found genes, 48 were over-expressed and 90 were down-expressed. On the other hand, 59 genes (of which 35 were overexpressed and 24 down-expressed) identified in the HISAT2 pipeline were not recognised as differentially expressed in the STAR2 pipeline, whereas the STAR2 pipeline had just 9 unique genes (5 over-expressed and 4 down-expressed) that were not present in the HISAT2 pipeline. This finding shows that, while HISAT2 may have discovered a greater number of expressed genes, STAR2 may have missed others. We hypothesize that this could be attributed to the fact that HISAT2 has more alignment regions with respect to pseudogenes, compared to STAR2. Pseudogenes are non-functional copies of genes that are often similar to functional genes, and their presence in the genome can complicate the alignment process [20].

In order to better comprehend the biological meaning of the data, it is usual practice in the field of bioinformatics to focus on the most differentially expressed genes (Figure 3). We selected the top 30 differentially expressed genes, regardless of whether they were over- or down-expressed, in our study. We discovered that 27 of the top 30 genes were expressed in both processes. This finding implies that the two pipelines are essentially consistent, and that, while

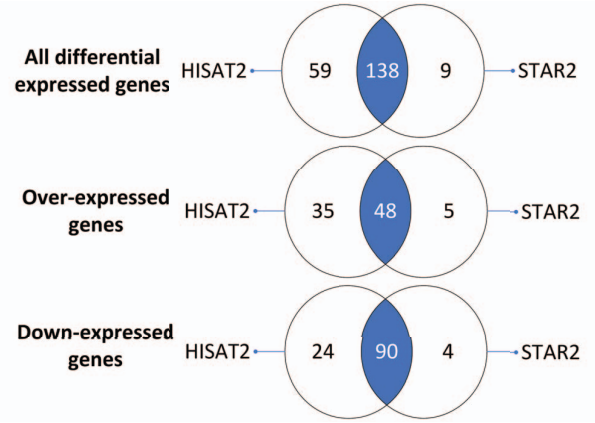

Fig. 2. Comparison of gene expression results across pipelines and gene sets. The first Venn diagram shows the overlap and differences in overall differentially expressed genes between two pipelines. The second diagram focuses on over-expressed genes, while the third diagram compares down-expressed genes.

there are some false positives among the total collection of expressed genes, the most physiologically important genes are consistently expressed in both pipelines. As a consequence, the top 30 expressed genes give a trustworthy and can be used for further investigation and interpretation.

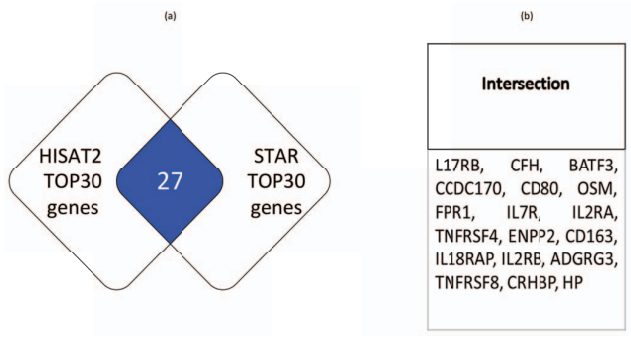

Fig. 3. On the left (figure a), the comparison of gene expression results across pipelines and gene sets in terms of most important differentially expressed genes (top 30). On the right (figure b), the focus is on the intersection and the expressed genes that were found down-regulated

Let us make an example of this concern: the top 30 differentially expressed genes are physiologically meaningful and congruent with clinical data. We found a collection of genes involved in the disease and targeted by the medicine that are switched off when treated with the drug, resulting in a decrease in inflammation, a particular sign of myelofibrosis (Figure 3). All of the genes in this list were down-regulated, indicating that the medication was successful in reducing their expression. Interestingly, all of the genes that were down-regulated by the drug are related to the inflammation that characterizes the disease. These findings suggest that the drug is having a targeted effect on the biological processes that are driving the disease. Such a finding indicates that the therapy utilised in this study is working as predicted, altering the critical genes in the illness pathway.

A. Limitations

Despite the promising results obtained in this study, there are certain limitations to our approach that should be taken into consideration. In this section, we discuss these limitations and their potential impact on the interpretation of our findings.

- The dataset we used in our study is a proprietary dataset that was made available to us by our collaborators and cannot be freely distributed. While we recognize that our current dataset is not ideal for reproducing our findings, we want to emphasize that obtaining publicly available myelofibrosis datasets is challenging due to the rarity of the disease. However, using more accessible datasets in future studies will help validate our results and provide a more general overview of the tools' performance in RNA-seq analysis. Therefore, as future work, we plan to explore a wider range of freely available datasets to further validate and generalize our findings
- Although we performed differential gene expression analysis using DESeq2 and compared the results, we did not conduct downstream analysis such as pathway and functional enrichment analysis. These additional analyses are necessary to validate the effectiveness of the pipelines and identify novel biological mechanisms involved in myelofibrosis development and progression. We plan to perform these analyses in future work to assess which pipeline is more effective
- As a means of improving the robustness of our results, we aim to integrate other methods for differential expression analysis in future work. Although we used the DESeq2 method in this study, the integration of other related methods will enable us to evaluate potential false positives or false negatives that may arise from using only one method. Such a comprehensive evaluation will provide a more accurate and reliable comparison of the different cleaning methods.

V. Conclusion

In conclusion, our study highlights the importance of carefully selecting tools to implement pipelines for RNA-sequencing analysis. Our comparison of two popular alignment tools, HISAT2 and STAR2, has shown that different alignment tools lead to different results in terms of computational time, number of aligned reads and number of expressed and differentially expressed genes.

By utilizing appropriate statistical thresholds, we were able to detect a significant group of genes that display differential expression. Moreover, we found that the detected genes are linked to the response of the commonly used medication for myelofibrosis. Interestingly, all of the genes that were at the intersection of the pipelines and down-regulated by the drug are related to the inflammation that characterizes the disease. These findings suggest that the drug is having a targeted effect on the biological processes that are driving the disease.

Having an overall view of the results obtained, we can summarize some further insights:

- Based on the results of our study, we found that STAR2 had better alignment accuracy than HISAT2. Therefore, we recommend using HISAT2 for the purpose of identifying novel putative genes to investigate, while using STAR2 when a more precise differential expression (DE) analysis is required. Further investigation (false positives and mapping on pseudogenes regions) is needed to understand the impact of the alignment tool on downstream analyses and to fully evaluate the implications of our findings;
- in spite of expectations, STAR2 showed better execution time even if it lacks pre-built indices;
- Considering the number of unique mapped reads, STAR2 showed better performance aligning more than 300,000 reads. Such difference however resulted in lower identified genes. Further investigation is needed to understand the effect of the genes' differential expression of the higher number of unique mapped reads;
- the identified genes not belonging to the intersection (i.e. 9 for STAR2 and 59 for HISAT2) can reveal new aspects about myelofibrosis disease, Ruxolitinib drug and their relationship that are not yet identified or studied.

CRediT authorship contribution statement

Andrea Bianchi: Methodology, Software, Validation, Investigation, Visualization, Writing - Original Draft, Writing - Review & Editing. Antinisca Di Marco: Methodology, Writing - Review & Editing, Supervision, Project administration. Cristina Pellegrini: Review & Editing, Resources.

Acknowledgment

This work is funded by the project LIFEMAP-Dalla patologia pediatrica alle malattie cardiovascolari e neoplastiche nell'adulto: mappatura genomica per la medicina e prevenzione personalizzata Traiettoria 3 "Medicina rigenerativa, predittiva e personalizzata" - Linea di azione 3.1 "Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale" of the Ministry of Health

European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021

All the numerical simulations have been realized on the Linux HPC cluster Caliban of the High-Performance Computing Laboratory of the Department of Information Engineering, Computer Science and Mathematics (DISIM) at the University of L'Aquila.

References

[1] Andrews. Fastqc: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

[2] Giacomo Baruzzo, Katharina Hayer, Eun Kim, Barbara Camillo, Garret FitzGerald, and Gregory Grant. Simulation-based comprehensive benchmarking of rna-seq aligners. *Nature Methods*, 14, 2016.

[3] Andrea Bianchi. Deseq2 script on myelofibrosis dataset (hisat2 and star2 pipelines), March 2023.

[4] Andrea Bianchi. Hisat2 - star2 pipelines on myelofibrosis, May 2023.

[5] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[6] Simone Claudiani, Clinton C Mason, Dragana Milojkovic, Andrea Bianchi, Cristina Pellegrini, Antinisca Di Marco, Carme R Fiol, Mark Robinson, Kanagaraju Ponnusamy, Katya Mokretar, et al. Carfilzomib enhances the suppressive effect of ruxolitinib in myelofibrosis. *Cancers*, 13(19):4863, 2021.

[7] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Szcześniak, Daniel Gaffney, Laura Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17, 2016.

[8] Juliana Costa-Silva, Douglas S Domingues, David Menotti, Mariangela Hungria, and Fabricio M Lopes. Temporal progress of gene expression analysis with rna-seq data: A review on the relationship between computational methods. *Computational and Structural Biotechnology Journal*, 2022.

[9] Gabrielle Deschamps-Francoeur, Joël Simoneau, and Michelle S Scott. Handling multi-mapped reads in rna-seq. *Computational and structural biotechnology journal*, 18:1569–1576, 2020.

[10] S Akila Parvathy Dharshini, Y-H Taguchi, and M Michael Gromiha. Identifying suitable tools for variant detection and differential gene expression using rna-seq data. *Genomics*, 112(3):2166—2172, 2020.

[11] Alexander Dobin, Carrie Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics (Oxford, England)*, 29, 2012.

[12] Pallavi Gaur and Anoop Chaturvedi. *A Survey of Bioinformatics-Based Tools in RNA-Sequencing (RNA-Seq) Data Analysis*, pages 223–248. 2017.

[13] Daehwan Kim, Ben Langmead, and Steven Salzberg. Hisat: A fast spliced aligner with low memory requirements. *Nature methods*, 12, 2015.

[14] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915, 2019.

[15] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven Salzberg. Tophat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14, 2013.

[16] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[17] Yang Liao, Gordon Smyth, and Wei Shi. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30, 2013.

[18] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. 2014.

[19] Paul McGettigan. Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, 17, 2013.

[20] Isaac D. Raplee, Alexei V. Evsikov, and Caralina Marín de Evsikova. Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *Journal of Personalized Medicine*, 9(2), 2019.