Confronto tra le pipeline basate su HISAT e STAR per l'analisi dei dati RNA-Seq: un'esperienza reale

Andrea Bianchi, Antinisca Di Marco, Cristina Pellegrini
Università dell'Aquila, L'Aquila, Italia
{andrea.bianchi}@laurea.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-Una delle prime fasi dell'analisi dei dati di RNA-Sequencing (RNA-Seq) consiste nell'allineare le letture (Next Generation Sequencing) a un genoma di riferimento. In letteratura, esistono diversi strumenti implementati da professionisti e ricercatori per la fase di allineamento. Tuttavia, due strumenti sono il de-facto-standard utilizzati dai ricercatori bioinformatici nelle loro pipeline: HISAT (versione 2) e STAR (versione 2). Lo scopo di questo studio è determinare l'impatto dello strumento di allineamento sull'analisi RNA-Seq in termini di rilevanza biologica dei risultati e di tempo computazionale. Le due pipeline implementate restituiscono risultati diversi dal punto di vista biologico. Ciò è dovuto alle assunzioni fatte dagli strumenti utilizzati e alle caratteristiche specifiche dei modelli (statistici) sottostanti. Lo studio fornisce indicazioni preziose per i ricercatori interessati a ottimizzare le loro pipeline RNA-Seq e a prendere decisioni informate su quale pipeline utilizzare. Come lezione appresa, suggeriamo ai ricercatori bioinformatici di utilizzare più pipeline quando effettuano esperimenti per ridurre gli errori di predizione indotti dalle assunzioni di uno specifico strumento o metodo.

***Termini dell'indice-bioinformatica*, RNA-Sequencing, mielofibrosi, confronto, pipeline**

I. Introduzione

La bioinformatica ha registrato un'enorme crescita negli ultimi anni, in particolare con il progresso dell'analisi dei dati di sequenziamento ad alto rendimento. Gli esperimenti di RNA-Sequencing (RNA-Seq) sono un potente strumento per studiare i modelli di espressione genica e identificare i geni differenzialmente espressi. Affinché questi esperimenti abbiano successo, è essenziale che le pipeline bioinformatiche, utilizzate per elaborare e analizzare i dati, siano di buona qualità.

La fase di allineamento (o mappatura) dell'analisi dei dati RNA-Seq è il processo più intensivo dal punto di vista computazionale e che richiede molto tempo, poiché comporta l'allineamento delle letture generate da un esperimento di sequenziamento con un genoma di riferimento. La scelta del giusto strumento di mappatura per questo compito è fondamentale quando l'efficienza computazionale e l'accuratezza biologica dei risultati sono aspetti rilevanti.

Una mappatura accurata è essenziale per l'analisi a valle, ma la presenza di giunzioni di splice nelle letture RNA-seq rappresenta una sfida per l'accuratezza dell'allineamento. Per affrontare questa sfida, sono state sviluppate diverse piattaforme software per la mappatura su un genoma di riferimento, tra cui TopHat (versione 2, di seguito indicata come Tophat2) [15], HISAT (versione 2, di seguito indicata come HISAT2) [14] e STAR (versione 2, di seguito indicata come STAR2) [11]. TopHat2 è stata una scelta popolare, ma è stata sostituita da HISAT2 a causa della sua inefficienza computazionale. TopHat2 e HISAT2 sono costruiti sulla base di popolare strumento di mappatura a lettura corta Bowtie2 [16].

Sebbene tutti e tre gli allineatori siano considerati veloci, la scelta dell'allineatore ottimale può avere un impatto significativo sull'analisi a valle. Pertanto, è fondamentale valutare le prestazioni di allineamento di diversi allineatori per identificare lo strumento ottimale per il compito. Questo studio si propone di confrontare due diverse pipeline bioinformatiche utilizzando HISAT2 e STAR2 per valutarne le prestazioni e la qualità dell'output e per identificare lo strumento ottimale per una mappatura accurata e un'analisi a valle. La scelta di questi due allineatori è motivata dalla loro diffusione e dalla superiorità delle loro prestazioni rispetto a TopHat2 [2]. Oltre all'ampia diffusione di HISAT2 e STAR2, abbiamo scelto questi due allineatori per il confronto perchè sono le versioni più recenti e sono migliorate rispetto ai loro predecessori [14] [11] [8]. HISAT2 è una versione migliorata di HISAT e STAR2. È una versione aggiornata di STAR. Queste nuove versioni hanno risolto alcuni dei limiti dei loro predecessori, come l'aumento dell'accuratezza, della velocità e dell'efficienza della memoria.

HISAT2 e STAR2 sono strumenti di allineamento RNA-Sequenziamento (RNA-Seq) che differiscono per strategia di allineamento, dimensione dell'indice, sensibilità, velocità e ottimizzazione del tipo di lettura. HISAT2 utilizza il graph FM index (GFM) [14], ha una dimensione dell'indice più piccola, è più veloce ma ha capacità limitate di multi-threading ed è ottimizzato per letture single-end e paired-end.

STAR2 utilizza Spliced Transcripts Alignment [11] come algoritmo di riferimento, ha una dimensione dell'indice maggiore, una sensibilità più elevata e un migliore multi-threading, ma è più lento ed è ottimizzato per le letture spliced.

Esistono solo pochi studi che confrontano diversi strumenti per l'analisi dei dati di sequenziamento dell'RNA su insiemi di dati tumorali. In [10], gli autori hanno riscontrato che STAR2 aveva prestazioni migliori in termini di percentuale di letture univocamente mappate (precisamente, 80%) rispetto a HISAT2 (70%), utilizzando assemblaggi genomici diversi (hg19 e hg38). Le percentuali di letture non mappate sono maggiori in HISAT2 e Tophat2 in entrambi gli assemblaggi genomici. Lo studio di [20] ha riscontrato che HISAT2 ha allineato un numero inferiore di letture e ha registrato tassi più elevati di allineamento a pseudogeni, compromettendo la fedeltà dell'allineamento e portando potenzialmente a risultati errati. A differenza degli studi precedenti, il nostro lavoro mira a estendere il confronto tra HISAT2 e STAR2 oltre l'analisi della percentuale di letture univocamente mappate, valutando le loro prestazioni in termini di rilevanza biologica dei risultati ottenuti. In particolare, utilizziamo l'assemblaggio del genoma hg38, che è l'assemblaggio

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

più recente e raccomandato per le analisi genomewide, per valutare i risultati di espressione differenziale ottenuti da ciascun genoma. Inoltre, abbiamo monitorato il tempo di calcolo delle due pipeline implementate per quantificare e confrontare la loro complessità temporale.

Questo lavoro è collegato a [6], dove abbiamo studiato i potenziali effetti del farmaco Ruxolitinib. Recentemente, questo farmaco è stato approvato dalla FDA per il trattamento dei pazienti affetti da mielofibrosi (MF), una malattia che colpisce il midollo osseo. Pur migliorando i sintomi, ruxolitinib non cura completamente la malattia né riduce significativamente il numero di cellule mutate. Questo perché alcune cellule MF sono resistenti al farmaco, forse a causa di geni o vie aggiuntive che promuovono la sopravvivenza delle cellule anche quando la via JAK2/STAT5, bersaglio del ruxolitinib, viene soppressa. Utilizzando una libreria di strumenti genetici, abbiamo scoperto che diversi membri della famiglia dei geni proteasomali sono importanti per la sopravvivenza cellulare e che la loro inibizione con il farmaco carfil-zomib ha reso le cellule MF più sensibili al ruxolitinib. Inoltre, la combinazione di ruxolitinib e carfilzomib ha ridotto l'espressione dei geni proteasomici nelle cellule MF, suggerendo che questo approccio potrebbe essere efficace nel trattamento dei pazienti MF. A differenza di [6], lo scopo di questo studio è presentare un confronto tra HISAT2 e STAR2 per l'elaborazione dei dati RNA-Seq in termini di tempo di esecuzione e di rilevanza biologica dei risultati ottenuti dalle fasi di allineamento, quantificazione e analisi dell'espressione differenziale.

Questo studio fornirà una risorsa inestimabile ai ricercatori che mirano a ottimizzare le loro pipeline di analisi dei dati RNA-Seq e a decidere con cognizione di causa quale pipeline si adatta meglio alle loro esigenze, fornendo indicazioni sull'equilibrio tra efficienza computazionale e accuratezza biologica.

Il documento procede come segue: La Sezione II descrive il flusso di lavoro generale dell'analisi RNA-Seq e gli strumenti utilizzati per implementare due pipeline. Nella Sezione III si riportano le impostazioni sperimentali e il set di dati utilizzato. La Sezione IV discute i risultati ottenuti in termini di tempo computazionale e rilevanza biologica delle due pipeline implementate. Infine, la Sezione V conclude l'articolo, evidenziando le principali intuizioni individuate e i possibili lavori futuri.

II. Flusso di lavoro generale di RNA-Seq e pipeline implementate

Negli esperimenti di RNA-Seq, non esiste una singola pipeline che funzioni meglio in ogni situazione. A seconda degli obiettivi della ricerca e degli organismi sequenziati, si possono prendere in considerazione diversi approcci con vari strumenti software disponibili [7], [12]. La Figura 1 riporta il flusso di lavoro generico per l'RNA-Sequencing e gli strumenti specifici che abbiamo utilizzato per implementare due pipeline distinte.

La velocità di allineamento è un aspetto critico delle prestazioni degli strumenti bioinformatici utilizzati nel processo di mappatura [19]. Per questo motivo, abbiamo deciso di variare gli strumenti utilizzati in questa fase, mantenendo costanti quelli utilizzati nelle altre fasi. Le nostre pipeline differiranno solo per quanto riguarda le rispettive parti di mappatura. Abbiamo scelto due strumenti diversi in base al loro approccio di lavoro simmetrico e all'affidabilità dei risultati delle prestazioni riportati in [2], che hanno dimostrato che attualmente sono tra le migliori opzioni disponibili.

Osservando la Figura 1, un flusso di lavoro RNA-Seq è composto da quattro fasi principali: *Controllo di qualità* (presentato nella Sezione II-A), Allineamento (descritto nella Sezione II-B), Quantificazione (presentato nella Sezione II-C) e Analisi dell'espressione differenziale (DE) (nella Sezione IV).

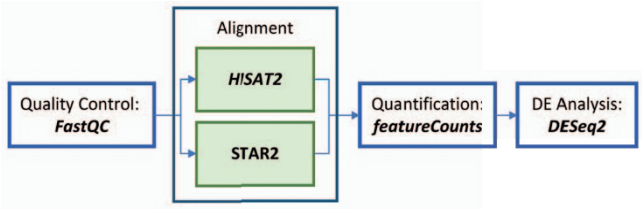

Figura 1. Flusso di lavoro generale per l'analisi RNA-Seq e pipeline implementate.

Di seguito descriviamo in dettaglio tutti i passaggi e gli strumenti utilizzati, mentre nella Sezione III ne riportiamo le impostazioni.

A. Controllo qualità

La prima fase del flusso di lavoro RNA-Seq è il controllo di qualità. Durante il processo di controllo della qualità è necessario intraprendere azioni specifiche per garantire che i dati grezzi abbiano la qualità desiderata. Ciò comporta spesso il trimming dell'adattatore, che implica la rimozione di tutte le sequenze che non provengono dall'organismo di origine e il filtraggio delle letture di bassa qualità e delle basi non chiamate. Poiché disponiamo di dati provenienti dalla piattaforma Illumina, per valutare la qualità dei file utilizziamo FastQC [1], uno dei software più diffusi a questo scopo. Nel caso in cui il controllo di qualità riveli problemi di qualità, il flusso di lavoro aggiunge una fase volta a rimuovere le sequenze di adattamento e a tagliare le basi di bassa qualità (ad esempio, utilizzando Trimmomatic v0.39 [5]). Nell'esperimento condotto, poiché i dati erano di qualità sufficiente, non è stato necessario rimuovere o tagliare gli adattatori.

B. Allineamento

L'allineamento è una componente critica di molte analisi bioinformatiche in cui i dati di sequenziamento high-throughput - in termini di singole letture - sono confrontati rispetto a un genoma o trascrittoma di riferimento. La fase di allineamento restituisce file BAM, passati come input alla fase di quantificazione. Per implementare la fase di allineamento abbiamo considerato due strumenti distinti: HISAT2 e STAR2, che illustriamo di seguito.

1) Strumento di allineamento #1: HISAT2: HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts) [13] è uno strumento di mappatura RNA-Seq ampiamente utilizzato che offre diversi vantaggi rispetto ad altri strumenti. Utilizza un approccio di indicizzazione gerarchica per allineare le letture al genoma di riferimento, rendendolo più veloce ed efficiente di molti altri strumenti di mappatura. Inoltre, le dimensioni ridotte dell'indice richiedono meno spazio su disco e rendono più rapida la costruzione dell'indice. HISAT2 è ottimizzato per le letture single-end e paired-end e può essere fino a due volte più veloce nell'allineamento grazie alla sua strategia di allineamento migliorata [13]. Di conseguenza, HISAT2 è un candidato ideale per analizzare trascrittomi ed esoni splicati.

Grazie alla sua strategia di indicizzazione gerarchica, una delle caratteristiche principali di HISAT2 è la capacità di allineare in modo affidabile le letture anche in aree con modelli di splicing complicati.
Grazie alla sua elevata sensibilità, HISAT2 è uno strumento utile per studi di sequenziamento dell'RNA su larga scala. Tuttavia, la strategia di indicizzazione gerarchica utilizzata da HISAT2 consuma maggiori risorse computazionali rispetto ad altri allineatori, rendendolo meno adatto alle piattaforme informatiche a bassa potenza. Inoltre, HISAT2 è in grado di allineare rapidamente i dati, quindi investe una quantità significativa di tempo nella costruzione di un indice esteso e completo solo una volta. Di conseguenza, può allineare i dati delle letture successive in modo più rapido e con un minore utilizzo di memoria.

2) Strumento di allineamento #2: STAR2: STAR2 (Spliced Transcripts Alignment to a Reference) [11] è uno degli strumenti di mappatura RNA-Seq più popolari. Questo algoritmo utilizza un approccio di allineamento dei trascritti di splicing, che offre una maggiore sensibilità rispetto ad altri metodi e consente di allineare accuratamente un maggior numero di letture al genoma di riferimento. STAR2 ha capacità multi-threading superiori che gli consentono di utilizzare in modo efficiente più core su sistemi con molti processori. Tuttavia, ciò richiede indici di dimensioni maggiori rispetto a HISAT2. Poiché STAR2 è ottimizzato per le letture di splicing, è particolarmente vantaggioso quando si lavora con giunzioni esone-esone o fusioni geniche. Il meccanismo di allineamento dello splicing è diverso tra STAR2 e altri allineatori come HISAT2. STAR2 impiega un nuovo metodo di allineamento degli spliced, la *strategia di allineamento unique spliced*, mentre HISAT2 impiega un metodo di indicizzazione gerarchica. In breve, la *strategia di allineamento a splicing unico* è un metodo per far corrispondere le letture di sequenziamento dell'RNA a un genoma di riferimento nel contesto degli esoni splicati. Crea un indice del genoma che individua le letture con splicing e quindi abbina le letture al genoma utilizzando l'indice. L'indice è costruito mappando le giunzioni di splicing dal genoma di riferimento ad ancore e il processo di allineamento utilizza le ancore per localizzare gli esoni splicati nelle letture. Anche in luoghi con schemi di splicing conformi, questo porta a una corrispondenza affidabile delle letture di splicing con il genoma di riferimento. Un'altra distinzione è l'efficienza, poiché STAR2 è ottimizzato per operazioni di sequenziamento di RNA su larga scala. L'esclusiva tecnica di allineamento degli spliced di STAR2 può anche richiedere più risorse computazionali rispetto ad altri allineatori, rendendolo inadatto all'uso su piattaforme informatiche a bassa potenza. Inoltre, STAR2 non fornisce indici precostituiti. Ciò significa che l'utente deve creare l'indice da zero per ogni genoma di riferimento utilizzato al fine di utilizzare lo strumento STAR2. Questa procedura può richiedere molto tempo e molte risorse computazionali, soprattutto per i genomi di grandi dimensioni. Rispetto ad altri software di allineamento, come HISAT2, STAR2 non dispone di indici precostituiti, compromettendo l'efficienza e il risparmio di risorse in caso di riutilizzo della pipeline.

C. Quantificazione

Una volta completata la fase di mappatura, il passo successivo consiste nel contare il numero di letture associate alle caratteristiche di interesse (i geni nel nostro esperimento), poiché si intende eseguire l'analisi dell'espressione differenziale sui geni, confrontandoli in diverse condizioni sperimentali. Abbiamo eseguito *featureCounts* [17] su tutti i file BAM provenienti dalla precedente fase di allineamento, in entrambe le pipeline contemporaneamente.

D. Analisi DE

L'analisi DE è una fase critica dell'elaborazione dei dati RNA-seq che cerca di scoprire i geni che sono espressi in modo diverso in diverse circostanze sperimentali. Dopo la normalizzazione dei dati di quantificazione dell'RNA-seq, è stata eseguita l'analisi dell'espressione differenziale per determinare i geni sovra o sottoespressi.
In questo lavoro, abbiamo utilizzato il linguaggio di programmazione R per eseguire l'analisi dell'espressione differenziale sui dati RNA-seq ottenuti dalle pipeline HISAT2 e STAR2. Lo strumento *DESeq2* [18] è stato utilizzato specificamente per valutare i dati di quantificazione.

III. Impostazioni sperimentali

In questa sezione presentiamo le impostazioni sperimentali utilizzate nei nostri esperimenti. In particolare, nella Sezione III-A descriviamo il set di dati utilizzato, nella Sezione III-B descriviamo l'hardware impiegato. Infine, nella Sezione III-C descriviamo la configurazione del software.

A. Descrizione del set di dati

In questo studio abbiamo voluto analizzare l'espressione dell'mRNA delle cellule staminali ematopoietiche CD34+ isolate dal sangue periferico. I campioni provengono da pazienti affetti da mielofibrosi. Nel set di dati avevamo i campioni di cinque pazienti, ma durante il controllo di qualità abbiamo scoperto una contaminazione da Escherichia coli nei campioni di un paziente. Per garantire l'accuratezza ed evitare qualsiasi influenza sul nostro esperimento, abbiamo deciso di rimuovere i campioni contaminati. Abbiamo quindi lavorato con32 file fastq provenienti da quattro pazienti, ciascuno dei quali ha fornito due campioni. Il primo è stato estratto da cellule CD34+ trattate con il farmaco Ruxolitinib, mentre l'altro era un campione non trattato. Ogni paziente ha due repliche. Ogni file dell'esperimento è in formato fastq con una dimensione media di 3 GB una volta decompresso. Alla fine, abbiamo lavorato su 96 GB di dati in totale. Il sequenziamento dell'RNA è stato quindi eseguito sugli otto campioni utilizzando un disegno a coppie per scopi di duplicazione. I campioni sono stati sequenziati su una piattaforma Illumina utilizzando il sistema HiSeq 2500. Il processo di sequenziamento ha prodotto letture comprese tra 60 e 90 milioni di basi per ciascun campione. Quasi tutti i campioni avevano almeno 70 milioni di letture.
I dati di sequenziamento grezzi generati in questo studio sono disponibili su richiesta, ma, per motivi di riservatezza dei pazienti, possono essere consultati solo in loco.

B. Configurazione hardware

Il progetto di ricerca è condotto su Caliban, un ambiente cluster composto da più nodi, con calcoli eseguiti su un singolo nodo. La configurazione hardware di questo nodo è descritta di seguito. L'unità di elaborazione centrale (CPU) è composta da 48 CPU individuali, ciascuna con una velocità di clock di 2,16 GHz. Questa configurazione permette al nodo di eseguire un elevato volume di calcoli in parallelo, con conseguenti tempi di elaborazione

e analisi dei dati più rapidi. La memoria ad accesso casuale (RAM) del nodo è di 141,48 GB. Inoltre, il nodo è dotato di una capacità di archiviazione su disco locale di 1,5 TB, che offre ampio spazio per memorizzare i risultati intermedi e finali dei calcoli e delle analisi. Il nodo gira su sistema operativo Linux 3.10.0. Sebbene il nodo utilizzato in questo progetto di ricerca disponga di 48 CPU, va notato che non tutti gli strumenti e i software del progetto possono sfruttare in modo efficiente l'elaborazione parallela. Di conseguenza, quando disponibili, sono state utilizzate soluzioni multi-threading per massimizzare le prestazioni di calcolo.

C. Configurazione del software

Abbiamo scelto la versione 3 di Anaconda come ambiente per gli strumenti bioinformatici e le loro configurazioni. In particolare, abbiamo utilizzato il gestore di pacchetti di Anaconda (Conda) per creare script bash personalizzati per automatizzare l'esecuzione del flusso di lavoro RNA-Seq sul cluster. Nella Tabella I sono riportati i comandi utilizzati per richiamare gli strumenti selezionati e le relative configurazioni. Nelle righe della tabella, riportiamo gli strumenti utilizzati e per ciascuno di essi indichiamo la *fase* del flusso di lavoro che implementano, il *comando* e i relativi *argomenti* utilizzati negli esperimenti.

Entrambe le pipeline hanno utilizzato il genoma di GRCh38, acquisito da Ensembl (https://www.ensembl.org/Homo_sapiens/Info/Index), come sequenza di riferimento per l'allineamento delle letture.

Oltre al genoma, per costruire l'indice è stato utilizzato il corrispondente file Gene transfer format (GTF), che contiene informazioni sui siti di splice e sugli esoni. È noto che il processo di indicizzazione richiede molto tempo e risorse. Per ridurre al minimo il tempo richiesto, il processo è stato facilitato dall'utilizzo di un processore multi-core, con 40 thread utilizzati sia per la costruzione dell'indice sia per le fasi di allineamento. Per garantire un confronto equo e accurato tra le due pipeline, l'indice è stato costruito da zero invece di utilizzare un indice preesistente. Poiché la formazione dell'indice è un elemento critico di ogni strumento di allineamento, questa scelta è stata fatta per evitare di nascondere il tempo necessario allo sviluppo dell'indice e per non compromettere la validità del confronto. Vale la pena notare che STAR2 non dispone di indici precostituiti, pertanto l'utente deve creare l'indice da zero per ogni genoma di riferimento utilizzato.

Infine, nell'analisi DE, eseguita per determinare i geni sovra o sottoespressi, sono state considerate due soglie diverse.

Entrambe le soglie consentono di stabilire la rilevanza biologica e rilevanza statistica.

Esse sono:

- la soglia *padj*: rappresenta il livello di significatività statistica dell'espressione genica differenziale; corrisponde a un valore di p aggiustato ed è stato impostato su un valore di 0,05;
- la soglia di *log fold change*: la soglia di $\log_2$ fold change riflette l'entità delle differenze biologiche tra le condizioni sperimentali; è stata impostata a 1 (in valore assoluto).

Queste soglie sono state applicate in modo coerente in entrambe le pipeline per garantire che i risultati fossero comparabili e biologicamente rilevanti.

DISPONIBILITÀ DEL CODICE

In questo studio, le pipeline sono state implementate come script bash ed eseguite su un cluster Linux. Il codice associato è accessibile al riferimento fornito [4] ed è regolato dalla licenza Creative Commons Attribution 4.0 International.

IV. RISULTATI

In questa sezione, eseguiamo una valutazione completa delle diverse pipeline bioinformatiche per confrontare i loro tempi di esecuzione e i risultati biologici. L'analisi dei tempi di calcolo si concentra sulle tre fasi principali delle pipeline, ossia la creazione dell'indice, l'allineamento e la quantificazione, per comprendere l'overhead introdotto dagli strumenti di allineamento. Dall'altro lato, la valutazione dei risultati biologici mira a confrontare le pipeline in termini di rilevanza biologica delle fasi di allineamento e quantificazione (nella Sezione IV-2) da un lato, e della fase di espressione differenziale (nella Sezione IV-3) dall'altro. Il nostro obiettivo è quello di fornire un confronto approfondito delle due pipeline, evidenziandone i punti di forza e i limiti.

1) Tempo di calcolo: Nell'esperimento, abbiamo configurato gli strumenti delle due pipeline con parametri di configurazione coerenti (ad esempio, lo stesso numero di thread, le regioni di mappatura considerate), come riportato nella Sezione III.

I risultati della Tabella II suggeriscono che HISAT2, pur richiedendo più tempo per la creazione degli indici, ha prestazioni simili in termini di tempo di allineamento e leggermente più lente in termini di tempo di quantificazione rispetto a STAR2. Le discrepanze riscontrate nella fase di quantificazione tra le due pipeline potrebbero essere ricondotte ai metodi utilizzati da HISAT2 e STAR2 durante l'allineamento. Poiché il software e l'hardware erano identici, le discrepanze osservate nei tempi di elaborazione possono essere attribuite *i)* a differenze intrinseche negli algoritmi e nel modo in cui gestiscono i dati e/o *ii)* alla diversa implementazione delle opzioni di multithreading tra HISAT2 e STAR2 e/o ai loro modelli e tecniche di ottimizzazione.

2) Rilevanza biologica dell'allineamento e della quantificazione: Un'altra dimensione da considerare nel confronto è l'allineamento complessivo e le letture univocamente mappate (Tabella III). Il tasso di allineamento complessivo tra le due pipeline è simile: HISAT2 ha allineato il 98,03% delle letture e STAR2 il 98,78% delle letture. Sebbene la differenza nel tasso di allineamento complessivo fosse relativamente piccola, la differenza nel numero di letture univocamente mappate era più sostanziale. HISAT2 ha allineato l'80,47% delle letture come univocamente mappate, mentre STAR2 ha allineato l'81,66% delle letture come univocamente mappate. Il risultato è una differenza di oltre 300.000 letture mappate in modo univoco: HISAT2 ha allineato 24.309.436 letture, mentre STAR2 ha allineato 24.667.395 letture. È importante notare che questa differenza in termini di numero di letture mappate in modo univoco potrebbe avere implicazioni per le fasi successive della pipeline e/o per i risultati biologici finali e dovrebbe essere presa in considerazione quando li si interpreta. Vogliamo sottolineare che HISAT2 è stato segnalato per avere un tasso leggermente più alto di letture multimappate.

TABELLA I
Comandi utilizzati per ogni strumento delle pipeline bioinformatiche e relative configurazioni

Strumento	Fase	Comando	Argomenti
FASTQc	Controllo qualità	fastqc	-t 40 -o output dir file di input
HISAT2	Allineamento (indicizzazione)	hisat2-build	- p 40 - siti di giunzione ss.txt - esoni exon.txt
STELLA2	Allineamento (indicizzazione)	stella	- runThreadN 40 - runMode genomeGenerate - sjdbGTFfile Homo sapiens.GRCh38.97.gtf - genomeDir GRCh38 - file genomeFasta GRCh38.dna.primary.fa
STELLA2	Allineamento	stella	- runThreadN 40 - genomeDir GRCh38 - sjdbGTFfile Homo sapiens.GRCh38.97.gtf - readFilesIn Campione R1.fastq Campione R2.fastq - outSAMtipo BAM Ordinato per coordinata - outSAMunmapped All'interno - outSAMattributi Standard - quantMode GeneCounts - outFileNamePrefix AllineamentoCampione1 - twopassMode Basic
caratteristicheConti	Quantificazione	caratteristicheConti	-T 40 -p -t esone -g nome del gene -a Homo sapiens.GRCh38.97.gtf -o countmatrix.txt S1.bam ... Sn.bam
DESeq2	Analisi DE	Script R (personalizzato)	Codice ufficiale: [3]

TABELLA II
Tempo di calcolo delle fasi di creazione dell'indice, allineamento e quantificazione.

nome della conduttura	creazione dell'indice dell'indice (minuti)	allineamento (minuti)	quantificazione (minuti)
HISAT2	54	10.19	3.34
STELLA2	28	10.43	2.18

(17,12% per HISAT2 rispetto al 15,9% per STAR2). Poiché le letture multimappate possono portare a una mappatura imprecisa e influenzare le analisi a valle [9], come l'analisi dell'espressione differenziale, la correzione di questo problema può potenzialmente portare all'identificazione di un maggior numero di geni da parte di HISAT2. Questa potrebbe essere una possibile spiegazione del motivo per cui HISAT2 ha identificato più geni differenzialmente espressi rispetto a STAR2.

TABELLA III
Percentuale e numero di letture mappate ottenute dopo l'allineamento.

conduttura nome	complessivo tasso di allineamento	univocamente velocità di lettura mappata	n. di (unicamente) letti mappati (in milioni)
HISAT2	98,03%	80,47%	24.309.436
STELLA2	98,78%	81,66%	24.667.395

3) Rilevanza biologica dell'espressione differenziale: Nella figura 2, riportiamo, mediante diagrammi di Venn, il numero di geni differenzialmente espressi trovati dalle pipeline HISAT2 e STAR2 come numero totale di geni trovati (primo diagramma di Venn) e i dettagli dei geni sovraespressi e sottoespressi (specularmente nel secondo e terzo diagramma).

Sulla base dei risultati riportati nella Figura 2, la pipeline HISAT2 ha identificato un maggior numero di geni espressi (n =197) rispetto alla pipeline STAR2 (n = 147) entro le nostre soglie di significatività statistica e biologica (riportate nella Sezione III).
Un'analisi più approfondita dei diagrammi di Venn, tuttavia, ha rivelato che 138 geni sono stati identificati come differenzialmente espressi in entrambi gli algoritmi. In altre parole, praticamente tutti i geni trovati nella pipeline STAR2 sono stati trovati anche nella pipeline HISAT2.
Dei 138 geni trovati, 48 erano sovraespressi e 90 sottoespressi. D'altra parte, 59 geni (di cui 35 sovraespressi e 24 sottoespressi) identificati nella pipeline HISAT2 non sono stati riconosciuti come differenzialmente espressi nella pipeline STAR2, mentre quest'ultima presentava solo 9 geni unici (5 sovraespressi e 4 sottoespressi) che non erano presenti nella pipeline HISAT2. Questo dato dimostra che, mentre HISAT2 potrebbe aver scoperto un maggior numero di geni espressi, STAR2 potrebbe averne tralasciati altri. Si ipotizza che ciò possa essere attribuito al fatto che HISAT2 presenta un maggior numero di regioni di allineamento per quanto riguarda gli pseudogeni, rispetto a STAR2. Gli pseudogeni sono copie non funzionali di geni spesso simili a geni funzionali e la loro presenza nel genoma può complicare il processo di allineamento [20].

Per comprendere meglio il significato biologico dei dati, è prassi abituale nel campo della bioinformatica concentrarsi sui geni più differenzialmente espressi (Figura 3). Abbiamo selezionato i 30 geni più espressi in modo differenziato, indipendentemente dal fatto che fossero sovra o sottoespressi, nel nostro studio. Abbiamo scoperto che 27 dei 30 geni principali erano espressi in entrambi i processi. Questo risultato implica che le due pipeline sono sostanzialmente coerenti e che, sebbene vi siano alcuni falsi positivi nella collezione

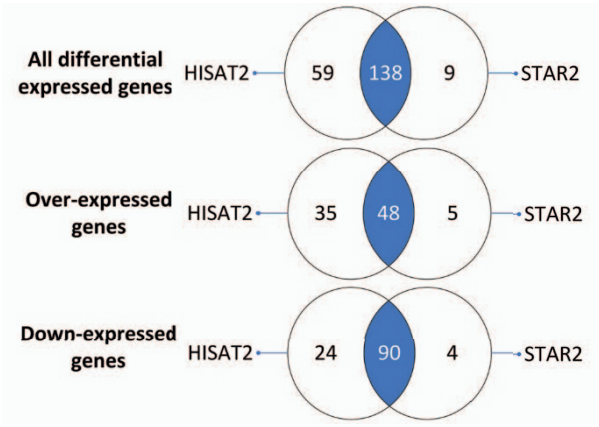

Fig. 2. Confronto dei risultati dell'espressione genica tra pipeline e set di geni. Il primo diagramma di Venn mostra la sovrapposizione e le differenze dei geni complessivamente espressi in modo differenziato tra due pipeline. Il secondo diagramma si concentra sui geni sovra-espressi, mentre il terzo diagramma confronta i geni down-espressi.

totale di geni espressi, ma i geni più importanti dal punto di vista fisiologico sono costantemente espressi in entrambe le pipeline. Di conseguenza, i 30 geni più espressi sono affidabili e possono essere utilizzati per ulteriori indagini e interpretazioni.

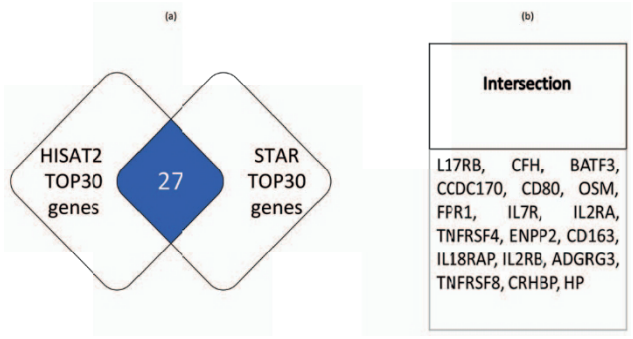

Fig. 3. A sinistra (figura a), il confronto dei risultati dell'espressione genica tra pipeline e set di geni in termini di geni differenzialmente espressi più importanti (top 30). A destra (figura b), l'attenzione si concentra sull'intersezione e sui geni espressi che sono stati trovati down-regulated.

Facciamo un esempio di questa preoccupazione: i primi 30 geni differenzialmente espressi sono fisiologicamente significativi e congruenti con i dati clinici. Abbiamo individuato un insieme di geni coinvolti nella malattia e presi di mira dal farmaco che vengono disattivati quando vengono trattati con il farmaco, con conseguente diminuzione dell'infiammazione, un segno particolare della mielofibrosi (Figura 3). Tutti i geni di questo elenco sono stati ridotti, a indicare che il farmaco è riuscito a ridurne l'espressione. È interessante notare che tutti i geni regolati dal farmaco sono legati all'infiammazione che caratterizza la malattia. Questi risultati suggeriscono che il farmaco sta avendo un effetto mirato sui processi biologici che guidano la malattia. Questa scoperta indica che la terapia utilizzata in questo studio funziona come previsto, alterando i geni critici del percorso della malattia.

A. Limitazioni

Nonostante i risultati promettenti ottenuti in questo studio, il nostro approccio presenta alcune limitazioni che devono essere prese in considerazione. In questa sezione discutiamo tali limitazioni e il loro potenziale impatto sull'interpretazione dei nostri risultati.

- Il set di dati utilizzato nel nostro studio è un set di dati proprietario che ci è stato messo a disposizione dai nostri collaboratori e non può essere distribuito liberamente. Pur riconoscendo che il nostro set di dati attuale non è ideale per riprodurre i nostri risultati, vogliamo sottolineare che ottenere set di dati sulla mielofibrosi disponibili al pubblico è difficile a causa della rarità della malattia. Tuttavia, l'utilizzo di set di dati più accessibili in studi futuri contribuirà a convalidare i nostri risultati e a fornire una panoramica più generale delle prestazioni degli strumenti nell'analisi RNA-seq. Pertanto, come lavoro futuro, abbiamo in programma di esplorare una gamma più ampia di set di dati liberamente disponibili per convalidare e generalizzare ulteriormente i nostri risultati.
- Anche se abbiamo eseguito un'analisi dell'espressione genica differenziale DESeq2 e abbiamo confrontato i risultati, non abbiamo condotto analisi a valle, come l'analisi dei percorsi e l'arricchimento funzionale. Queste analisi aggiuntive sono necessarie per convalidare l'efficacia delle pipeline e identificare nuovi meccanismi biologici coinvolti nello sviluppo e nella progressione della mielofibrosi. Abbiamo in programma di eseguire queste analisi nel lavoro futuro per valutare quale pipeline è più efficace.
- Per migliorare la robustezza dei nostri risultati, ci proponiamo di integrare altri metodi per l'analisi dell'espressione differenziale nel lavoro futuro. Sebbene in questo studio abbiamo utilizzato il metodo DESeq2, l'integrazione di altri metodi correlati ci consentirà di valutare i potenziali falsi positivi o falsi negativi che possono derivare dall'uso di un solo metodo. Una valutazione così completa fornirà un confronto più accurato e affidabile dei diversi metodi di pulizia.

V. Conclusione

In conclusione, il nostro studio evidenzia l'importanza di selezionare attentamente gli strumenti per implementare le pipeline per l'analisi del sequenziamento dell'RNA. Il nostro confronto tra due strumenti di allineamento popolari, HISAT2 e STAR2, ha dimostrato che strumenti di allineamento diversi portano a risultati diversi in termini di tempo computazionale, numero di letture allineate e numero di geni espressi e differenzialmente espressi.

Utilizzando soglie statistiche appropriate, siamo stati in grado di individuare un gruppo significativo di geni che mostrano un'espressione differenziale. Inoltre, abbiamo scoperto che i geni rilevati sono legati alla risposta dei farmaci comunemente usati per la mielofibrosi. È interessante notare che tutti i geni che si trovavano all'incrocio delle pipeline e che sono stati sotto-regolati dal farmaco sono legati all'infiammazione che caratterizza la malattia. Questi risultati suggeriscono che il farmaco ha un effetto mirato sui processi biologici che guidano la malattia.

Avendo una visione d'insieme dei risultati ottenuti, possiamo riassumere alcuni approfondimenti:

- Sulla base dei risultati del nostro studio, abbiamo riscontrato che STAR2 ha una migliore accuratezza di allineamento rispetto a HISAT2. Pertanto, raccomandiamo di utilizzare HISAT2 per identificare nuovi geni putativi da studiare, mentre utilizziamo STAR2 quando è necessaria un'analisi più precisa dell'espressione differenziale (DE). Sono necessarie ulteriori indagini (falsi positivi e mappatura su regioni di pseudogeni) per comprendere l'impatto dello strumento di allineamento sulle analisi a valle e per valutare appieno le implicazioni dei nostri risultati;
- nonostante le aspettative, STAR2 ha mostrato un'esecuzione migliore anche se non ha indici precostituiti;
- Considerando il numero di letture uniche mappate, STAR2 ha mostrato prestazioni migliori allineando più di 300.000 letture. Tale differenza, tuttavia, si è tradotta in un numero inferiore di geni identificati. Sono necessarie ulteriori indagini per comprendere l'effetto dell'espressione differenziale dei geni sul maggior numero di letture mappate uniche;
- i geni identificati non appartenenti all'intersezione (ad esempio 9 per STAR2 e 59 per HISAT2) possono rivelare nuovi aspetti sulla malattia della mielofibrosi, sul farmaco Ruxolitinib e sulla loro relazione che non sono ancora stati identificati o studiati.

DICHIARAZIONE DI CONTRIBUTO ALLA PATERNITÀ CREDIT

Andrea Bianchi: Metodologia, software, validazione, indagine, visualizzazione, scrittura - bozza originale, scrittura - revisione ed editing. Antinisca Di Marco: Metodologia, Scrittura - Revisione ed Editing, Supervisione, Amministrazione del progetto. Cristina Pellegrini: Revisione ed editing, Risorse.

RICONOSCIMENTO

Questo lavoro è finanziato dal progetto LIFEMAP-Dalla patologia pediatrica alle malattie cardiovascolari e neoplastiche nell'adulto: mappatura genomica per la medicina e pre- venzione 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" del Ministero della Salute Unione Europea - NextGenerationEU (Piano di Ripresa Nazionale e Resilienza Plan).1 "Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale" del Ministero della Salute Unione Europea - NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Progetto: "SoBigData.it - Rafforzamento della RI italiana per il Social Mining e Big Data Analytics" - Prot. IR0000013

- Avviso n. 3264 del 28/12/2021

Tutte le simulazioni numeriche sono state realizzate sul cluster HPC Linux Caliban del Laboratorio di Calcolo ad Alte Prestazioni del Dipartimento di Ingegneria dell'Informazione, Informatica e Matematica (DISIM) dell'Università dell'Aquila.

RIFERIMENTI

[1] Andrews. Fastqc: uno strumento di controllo della qualità per i dati di sequenza ad alta velocità. http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

[2] Giacomo Baruzzo, Katharina Hayer, Eun Kim, Barbara Camillo, Garret FitzGerald e Gregory Grant. Valutazione comparativa completa di allineatori rna-seq basata sulla simulazione. *Nature Methods*, 14, 2016.

[3] Andrea Bianchi. Script Deseq2 su dataset di mielofibrosi (pipeline hisat2 e star2), marzo 2023.

[4] Andrea Bianchi. Hisat2 - star2 pipeline sulla mielofibrosi, maggio 2023.

[5] Anthony M. Bolger, Marc Lohse e Bjoern Usadel. Trimmo- matic: un trimmer flessibile per i dati di sequenza Illumina. *Bioinformatica*, 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 potenzia l'effetto soppressivo di ruxolitinib nella mielofibrosi. *Cancers*, 13(19):4863, 2021.

[7] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Szczes'niak, Daniel Gaffney, Laura Elo, Xuegong Zhang e Ali Mortazavi. Un'indagine sulle migliori pratiche per l'analisi dei dati rna-seq. *Genome Biology*, 17, 2016.

[8] Juliana Costa-Silva, Douglas S Domingues, David Menotti, Mariangela Hungria e Fabricio M Lopes. Progressi temporali dell'analisi dell'espressione genica con dati rna-seq: Una revisione del rapporto tra i metodi computazionali. *Computational and Structural Biotechnology Journal*, 2022.

[9] Gabrielle Deschamps-Francoeur, Joe¨l Simoneau e Michelle S Scott. Gestione di letture multimappate in rna-seq. *Computational and structural biotechnology journal*, 18:1569-1576, 2020.

[10] S Akila Parvathy Dharshini, Y-H Taguchi e M Michael Gromiha. Identificazione di strumenti adeguati per l'individuazione di varianti e l'espressione genica differenziale utilizzando dati rna-seq. *Genomics*, 112(3):2166-2172, 2020.

[11] Alexander Dobin, Carrie Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson e Thomas Gin- geras. Star: allineatore universale ultrarapido di rna-seq. *Bioinformatics (Oxford, Inghilterra)*, 29, 2012.

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

[13] Daehwan Kim, Ben Langmead e Steven Salzberg. Hisat: Un allineatore veloce di spliced con bassi requisiti di memoria. *Nature methods*, 12, 2015.

[14] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett e Steven L Salzberg. Allineamento del genoma e genotipizzazione basati su grafi con hisat2 e hisat-genotype. *Nature biotechnology*, 37(8):907-915, 2019.

[15] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kel- ley e Steven Salzberg. Tophat2: Allineamento accurato dei trascrittomi in presenza di inserzioni, delezioni e fusioni geniche. *Genome biology*, 14, 2013.

[16] Ben Langmead e Steven L Salzberg. Allineamento veloce di gapped-read con bowtie 2. *Nature methods*, 9(4):357-359, 2012.

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

[18] Michael I. Love, Wolfgang Huber e Simon Anders. Stima moderata di fold change e dispersione per dati rna-seq con deseq2. 2014.

[19] Paul McGettigan. La trascrittomica nell'era dell'rna-seq. *Current opinion in chemical biology*, 17, 2013.

[20] Isaac D. Raplee, Alexei V. Evsikov e Caralina Mar´n de Evsikova. Allineamento degli allineatori: Confronto tra strumenti di allineamento dei dati di sequenziamento del rna e di quantificazione dell'espressione genica per la ricerca clinica sul cancro al seno. *Journal of Personalized Medicine*, 9(2), 2019.