Algoritmo di mapping sul trascrittoma

Il trascrittoma

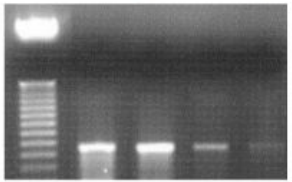
- *Il set completo di tutti gli mRNA di un organismo in un dato momento.*
- *Il trascrittoma è dinamico e cambia a seconda delle condizioni considerate. Differenti condizioni danno luogo a differenti profili di espressione genica.*

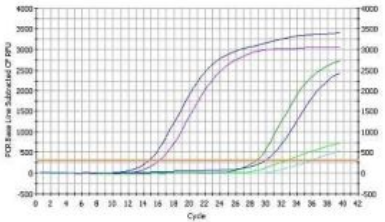
➔ *Trascrittomica: lo studio del trascrittoma; l'analisi del trascrittoma in diverse condizioni permette di inferire quali geni siano potenzialmente coinvolti in un dato processo di sviluppo, risposta a stress, ecc…*

Analisi di espressione genica

Prima delle tecnologie "omiche"

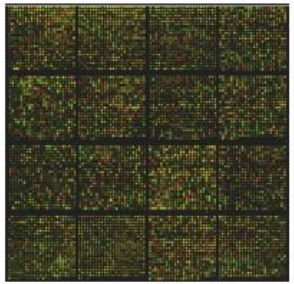
- Uno o pochi geni analizzati per volta tramite analisi Northern o PCR quantitativa/semiquantitativa





Oggi

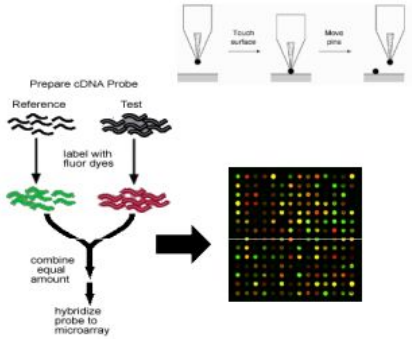
- Da poche migliaia di geni a trascrittomi completi analizzati in un singolo esperimento.

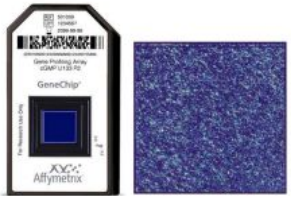


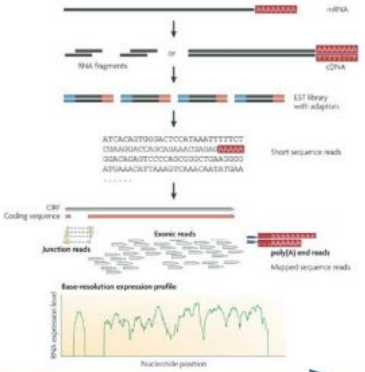
Microarray

Next Generation Sequecing (NGS)

Evoluzione delle tecnologie di analisi del trascrittoma







1995- Sviluppati i primi microarray basati su spotting di molecole di cDNA

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray- Schena et. al.

2002- High density oligo microarrays

2008- RNA-Seq: sequenziamento dei messaggeri basato su tecnologie NGS

Sequenziamento del trascrittoma

Campioni di interesse

Tessuto normale **Tessuto tumorale**

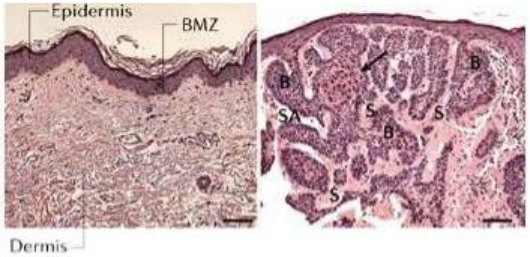
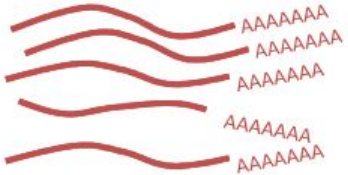


Immagine modificata da:
http://www.nature.com/nrc/journal/v6/n4/full/nrc1838.html

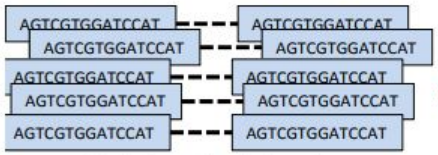
Isolamento dell'RNA/mRNA



Frammentazione chimica

Conversione a cDNA e ligazione degli adattatori

Sequenziamento



Milioni di read paired-end

Algoritmo di mapping sul trascrittoma

Panoramica sulla Trascrizione dell'RNA e il Trascrittoma:

La trascrizione è il processo mediante il quale le **informazioni genetiche** contenute nel DNA vengono copiate in molecole di **RNA**.

Queste molecole di RNA, chiamate **trascritti**, svolgono un ruolo cruciale nella sintesi delle proteine e in molti altri processi biologici.

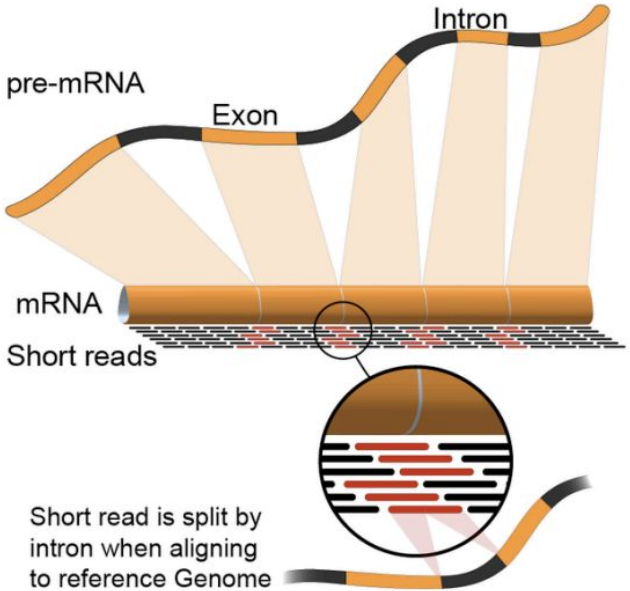
Il trascrittoma rappresenta l'insieme di tutti i trascritti presenti in una cellula o in un organismo in un dato momento.

Comprendere il trascrittoma è fondamentale per comprendere come vengano regolati i geni e come le cellule rispondano a cambiamenti ambientali o a malattie.

Algoritmo di mapping sul trascrittoma

Cos’è il mapping ?

Il mapping di dati RNA-seq sul genoma è il processo di allineamento delle sequenze di RNA ottenute mediante sequenziamento ad alto rendimento (RNA-Seq) sul genoma di riferimento. Questo passaggio è cruciale per identificare quali parti del genoma vengono trascritte e per quantificare l'espressione genica.



Algoritmo di mapping sul trascrittoma

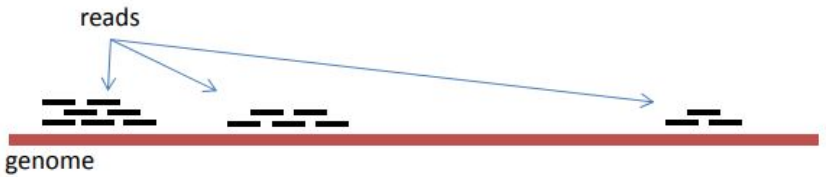
Cos’è il mapping ?

L'analisi del trascritti mappati sul genoma di riferimento consente di rispondere a domande chiave nella ricerca biologica, come:

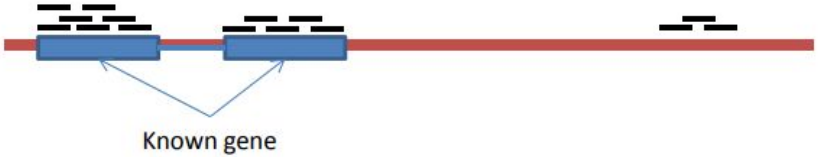
- Quanti trascritti vengono prodotti da un gene specifico?
- Come cambia l'espressione genica in diverse condizioni o in risposta a trattamenti?
- Quali varianti di splicing sono presenti in un gene?
- Quali geni sono coinvolti in una specifica via metabolica o processo biologico?

Protocollo di analisi dati RNA-Seq

per organismi modello

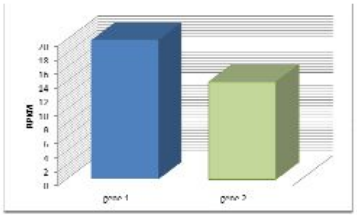


Allineamento su un genoma di riferimento



Assegnamento delle read ai geni annotati

Rilevazione di eventuali geni "nuovi" non annotati



Quantificazione dell'espressione e analisi statistica

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-8. doi: 10.1038/nmeth.1226.

Principi di base della RNA-seq:

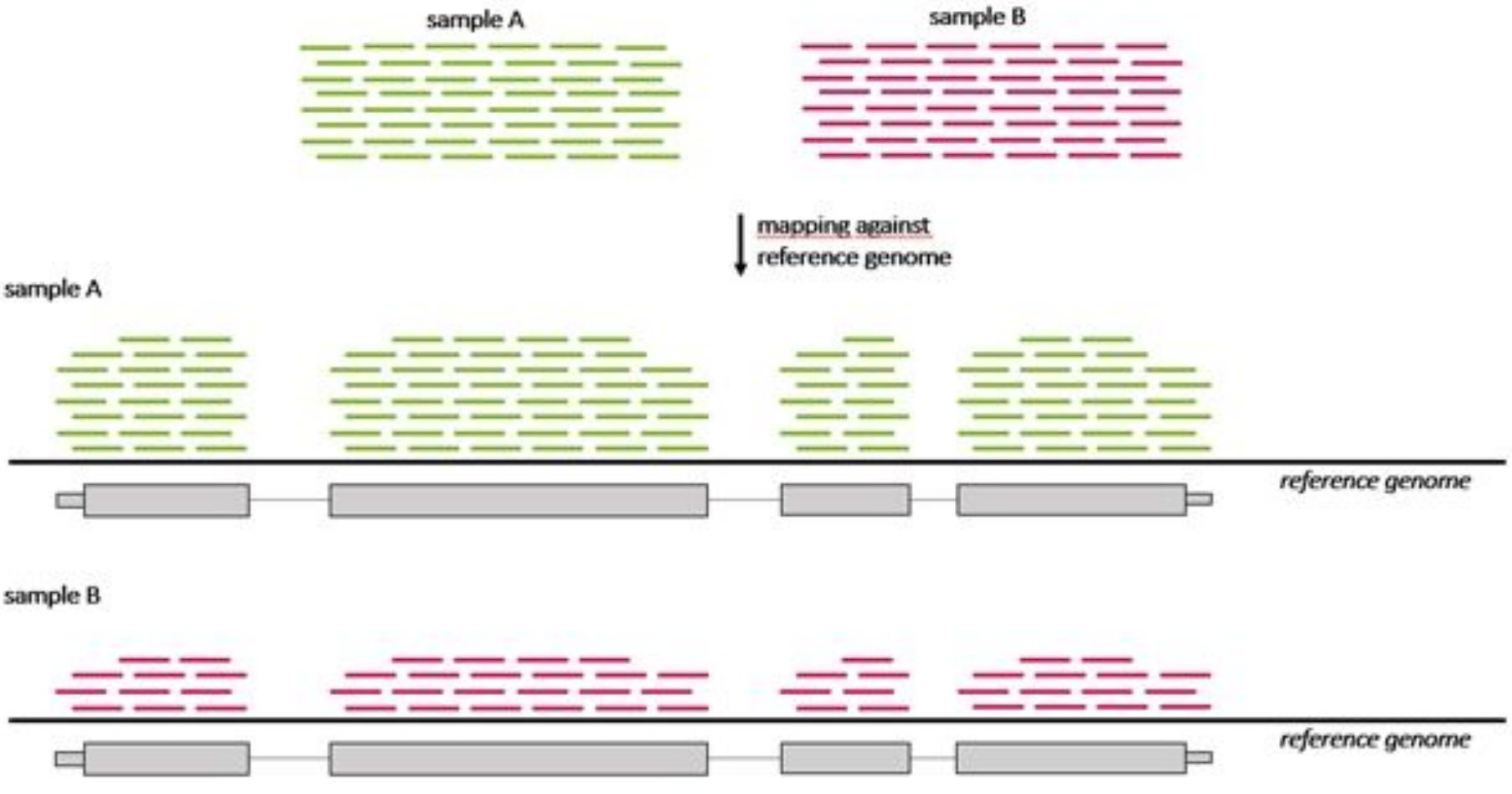
Durante un esperimento di RNA-seq, l'RNA totale viene convertito in una libreria di cDNA, che viene successivamente sequenziata.

Poiché il numero di molecole di cDNA deriva dal numero di molecole di RNA presenti inizialmente, la quantità di sequenze (o reads) ottenute da un particolare gene dovrebbe riflettere il suo livello di espressione iniziale.

Validazione con tecniche alternative:

I risultati ottenuti dalla RNA-seq sono stati confrontati con tecniche di quantificazione dell'RNA ben stabilite, come qRT-PCR. Questi studi hanno generalmente mostrato una forte correlazione tra i livelli di espressione misurati con RNA-seq e qRT-PCR.

Replicabilità: Gli esperimenti di RNA-seq replicati su campioni identici o molto simili mostrano una forte correlazione tra il numero di reads mappati e il livello di espressione. Questa replicabilità è un indicatore che la RNA-seq è una tecnica affidabile per misurare l'espressione genica.


sample A
sample B
mapping against
reference genome
sample A
reference genome
sample B
reference genome

Studi di saturazione: Man mano che si sequenziano più reads da un campione, il numero di reads che mappano su un gene tende a stabilizzarsi, suggerendo che si sta raggiungendo una rappresentazione completa dell'RNA presente. Questa saturazione è coerente con l'idea che il numero di reads è proporzionale al livello di espressione.

Risoluzione: La RNA-seq può identificare differenze sottili nell'espressione genica che altre tecniche potrebbero non rilevare. La capacità di distinguere livelli di espressione molto simili tra loro suggerisce che la RNA-seq è sensibile alle variazioni nella quantità di RNA.

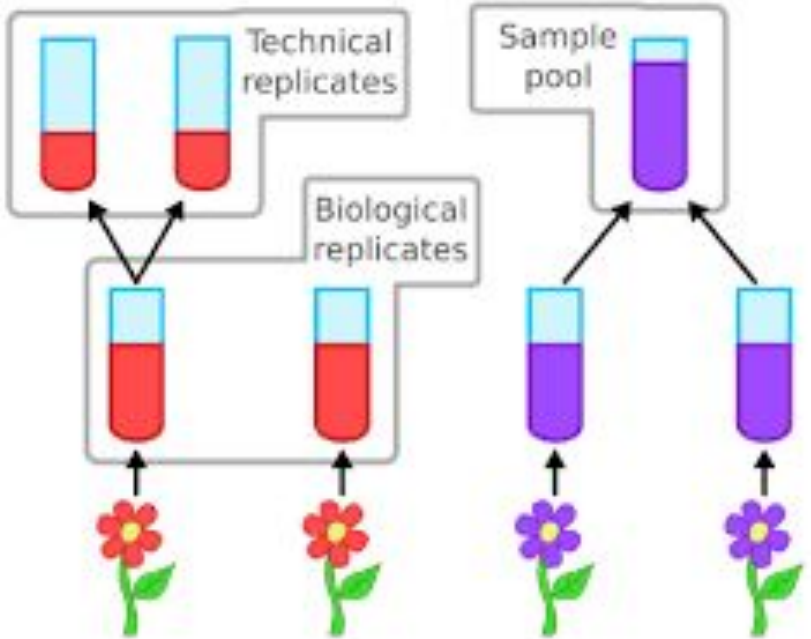
Assegnamento delle read ai geni e quantificazione dell'espressione genica

- **Il numero di read che mappano su un gene è proporzionale al livello di espressione**
- I valori di espressione ottenuti dall'RNA-Seq deriva dalla conta diretta delle read che mappano su un gene: **misura digitale**
- Non richiede la conoscenza a priori delle posizioni dei geni

RNA-seq: disegno sperimentale

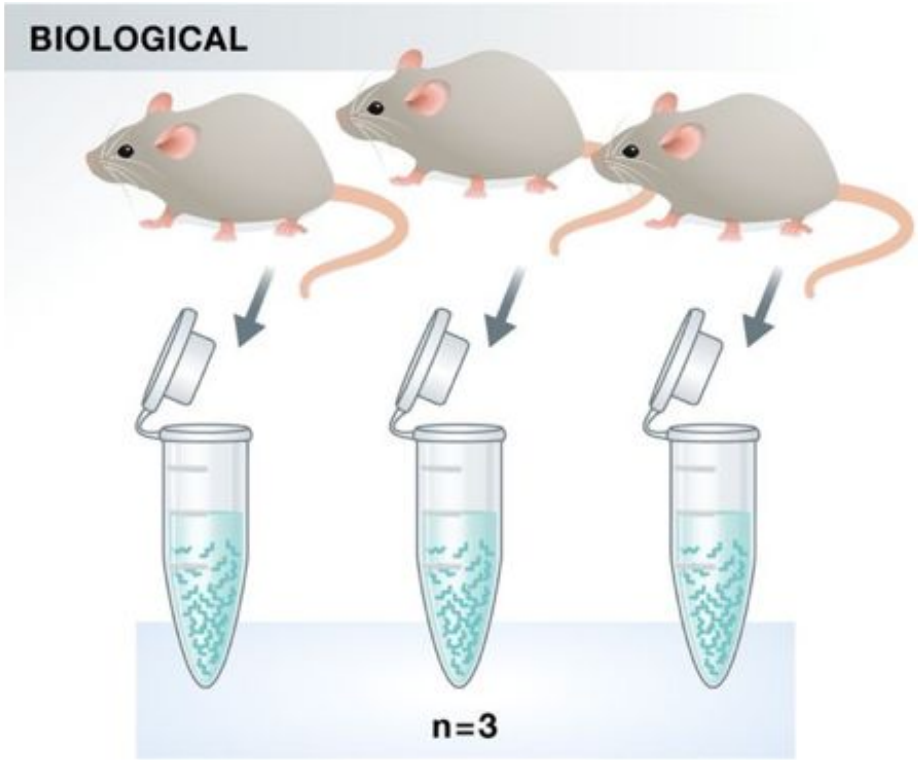
Nel contesto degli esperimenti di RNA-seq, le repliche sono fondamentali per determinare la variazione biologica e tecnica, e per garantire che i risultati ottenuti siano robusti e riproducibili.

Ci sono due tipi principali di repliche in un esperimento di RNA-seq: le repliche BIOLOGICHE e le repliche TECNICHE



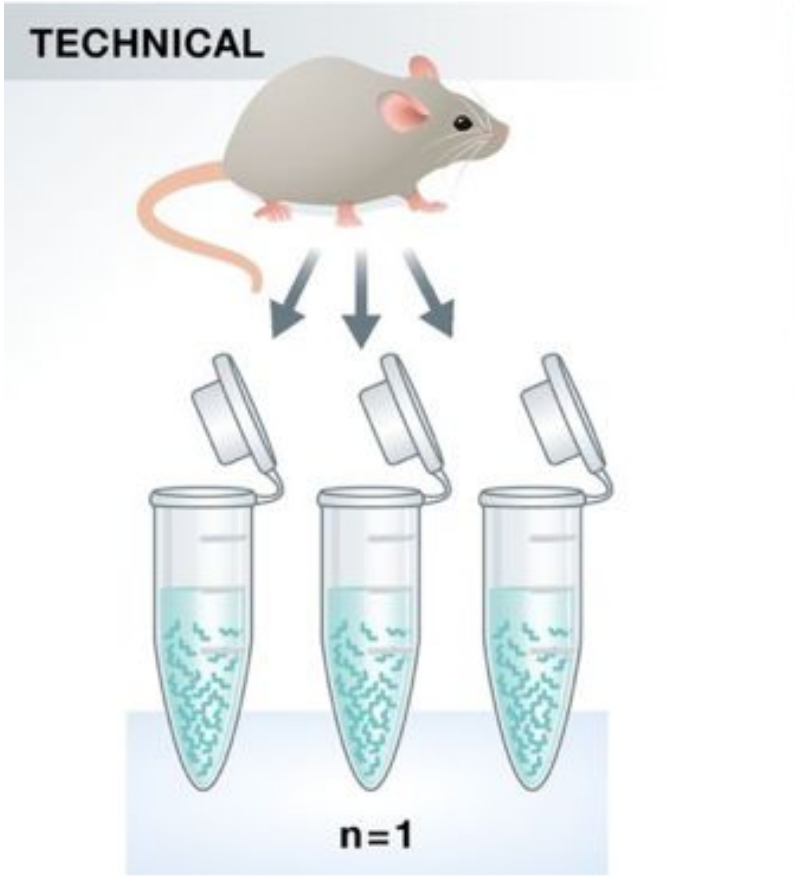
RNA-seq: disegno sperimentale

Repliche Biologiche: Si tratta di campioni separati provenienti da sorgenti biologiche distinte ma sottoposte alle stesse condizioni sperimentali. Ad esempio, potrebbero essere cellule prelevate da diversi individui o organismi, o culture cellulari cresciute separatamente. Le repliche biologiche aiutano a determinare la variabilità tra individui o esperimenti separati e forniscono una stima della variazione biologica. Sono fondamentali per distinguere le variazioni reali nell'espressione genica da quelle che potrebbero derivare da fluttuazioni casuali o da errori sperimentali.



RNA-seq: disegno sperimentale

Repliche Tecniche: Queste repliche provengono dalla stessa sorgente biologica, ma vengono processate separatamente. Possono derivare, ad esempio, da diverse librerie di RNA-seq preparate dalla stessa preparazione di RNA o da sequenziamenti ripetuti della stessa libreria. Le repliche tecniche sono utili per valutare la variabilità introdotta dai processi di laboratorio, come la preparazione della libreria o il sequenziamento stesso.



RNA-seq: disegno sperimentale

Alcune considerazioni sulle repliche nei campioni RNA-seq:

- **Numero di Repliche**: Per gli esperimenti di RNA-seq, l'inclusione di un numero adeguato di repliche biologiche è cruciale. Questo non solo aumenta la potenza statistica per identificare differenze significative nell'espressione genica, ma fornisce anche una rappresentazione più accurata della variazione biologica. Mentre il numero esatto di repliche necessarie può dipendere dall'obiettivo dello studio, molte pubblicazioni suggeriscono almeno tre repliche biologiche per condizione come punto di partenza.

- **Importanza delle Repliche Biologiche**: Sebbene le repliche tecniche possano aiutare a identificare e correggere il rumore tecnico, le repliche biologiche sono generalmente considerate più critiche in un esperimento di RNA-seq. Ciò perché la variazione biologica è spesso maggiore della variazione tecnica, e senza repliche biologiche, non è possibile determinare se un cambiamento nell'espressione genica è significativo o se è semplicemente dovuto alla variazione casuale tra campioni.

RNA-seq: disegno sperimentale

- **Analisi Statistica**: La presenza di repliche consente analisi statistiche robuste, come i test di significatività per identificare geni differentemente espressi tra le condizioni.

In sintesi, le repliche, specialmente le repliche biologiche, sono essenziali nei campioni di RNA-seq per fornire risultati accurati, affidabili e riproducibili.

Problematiche connesse all’analisi dei dati RNA-seq

Gli esperimenti di RNA-seq, nonostante le loro innumerevoli applicazioni e potenzialità, presentano diverse sfide e problematiche. Ecco alcune delle principali:

1. **Qualità del RNA**: La degradazione dell'RNA o la contaminazione da DNA possono influenzare negativamente la qualità dei dati. È quindi fondamentale valutare la qualità dell'RNA, ad esempio attraverso il valore RIN (RNA Integrity Number*), prima di procedere con la preparazione della libreria.

2. **Bias nella preparazione della libreria**: La conversione di RNA in cDNA e l'amplificazione possono introdurre bias, favorendo alcune sequenze rispetto ad altre. Ad esempio, l'efficienza della trascrittasi inversa può variare a seconda della sequenza e della struttura dell'RNA.

*L’RNA Integrity Number è un parametro che indica la qualità dell'RNA totale estratto da un campione. È stato introdotto come uno standard per valutare l'integrità dell'RNA, aiutando a determinare se un campione di RNA è idoneo per ulteriori analisi, come la RNA-seq o la RT-PCR.

Problematiche connesse all’analisi dei dati RNA-seq

3. **Normalizzazione**: Poiché il numero totale di reads può variare da un campione all'altro, è necessario normalizzare i dati tra le diverse librerie per rendere comparabili i livelli di espressione.

4. **Mappatura dei reads**: La mappatura dei reads al genoma di riferimento può presentare sfide, in particolare in regioni del genoma con duplicazioni, ripetizioni o alta omologia tra i geni.

5. **Espressione di geni a basso livello**: I geni espressi a livelli molto bassi possono essere difficili da rilevare e analizzare, in quanto la copertura potrebbe essere insufficiente per fornire una stima affidabile dell'espressione.

6. **Isoforme e splicing alternativo**: L'RNA-seq può rilevare diverse isoforme di un gene, ma determinare l'espressione di singole isoforme può essere complesso a causa della sovrapposizione di reads tra diverse varianti di isoforme.

Problematiche connesse all’analisi dei dati RNA-seq

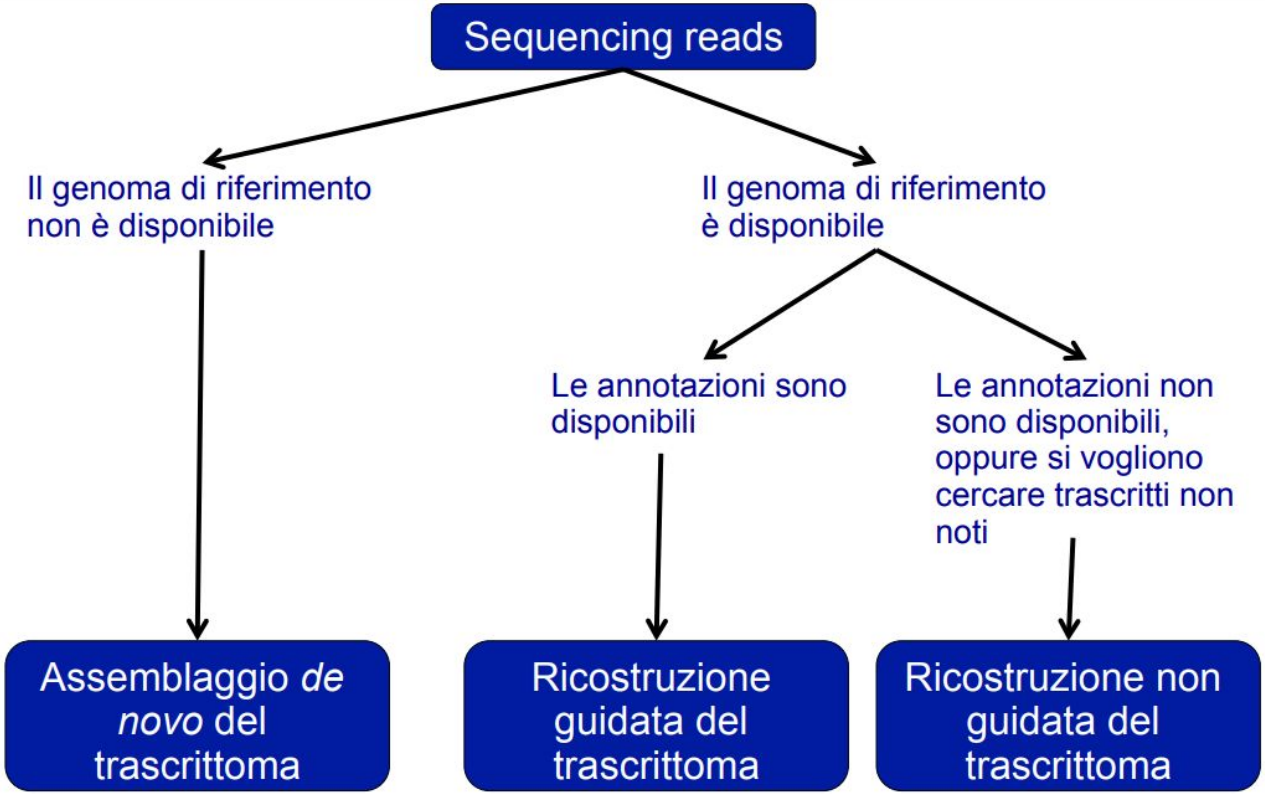
7. **Interferenza da trascritti non codificanti**: L'RNA-seq può rilevare non solo mRNA, ma anche una varietà di RNA non codificanti, che potrebbero complicare l'analisi se non si è interessati a questi trascritti.

8. **Variazione tecnica e biologica**: Come menzionato in precedenza, è cruciale distinguere la variazione dovuta a differenze reali nell'espressione genica da quella causata da rumore tecnico o variazione biologica casuale.

9. **Risoluzione temporale**: Gli esperimenti di RNA-seq rappresentano uno "snapshot" dell'espressione genica in un dato momento. La dinamica dell'espressione genica nel tempo potrebbe richiedere esperimenti di sequenziamento multipli e complessi.

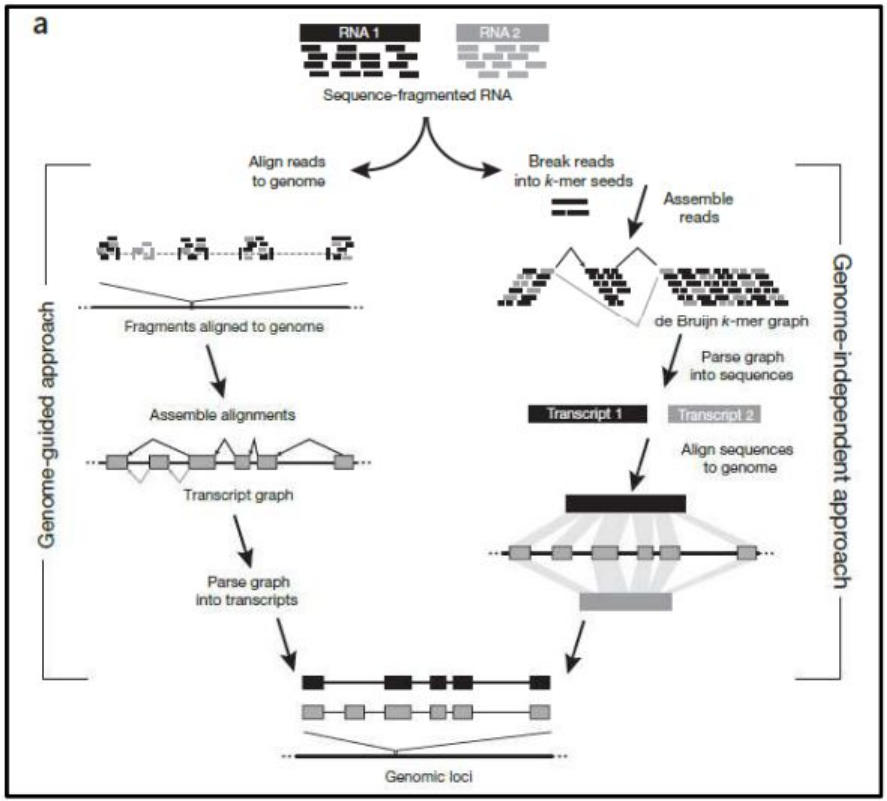
Nonostante queste sfide, l'RNA-seq è uno strumento potentissimo che ha rivoluzionato la nostra capacità di studiare l'espressione genica e la biologia trascrizionale. Con una pianificazione e analisi attente, molte di queste problematiche possono essere affrontate e gestite efficacemente.

Algoritmo di mapping sul trascrittoma



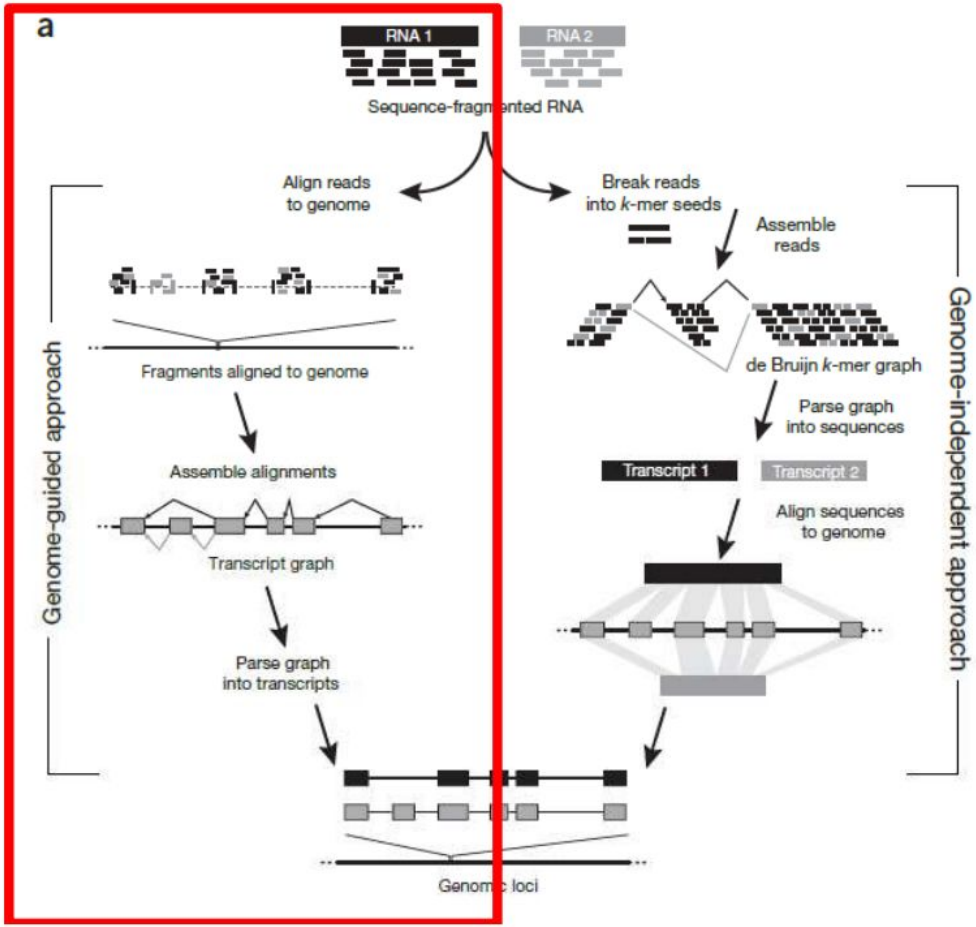
Metodi di ricostruzione del trascrittoma

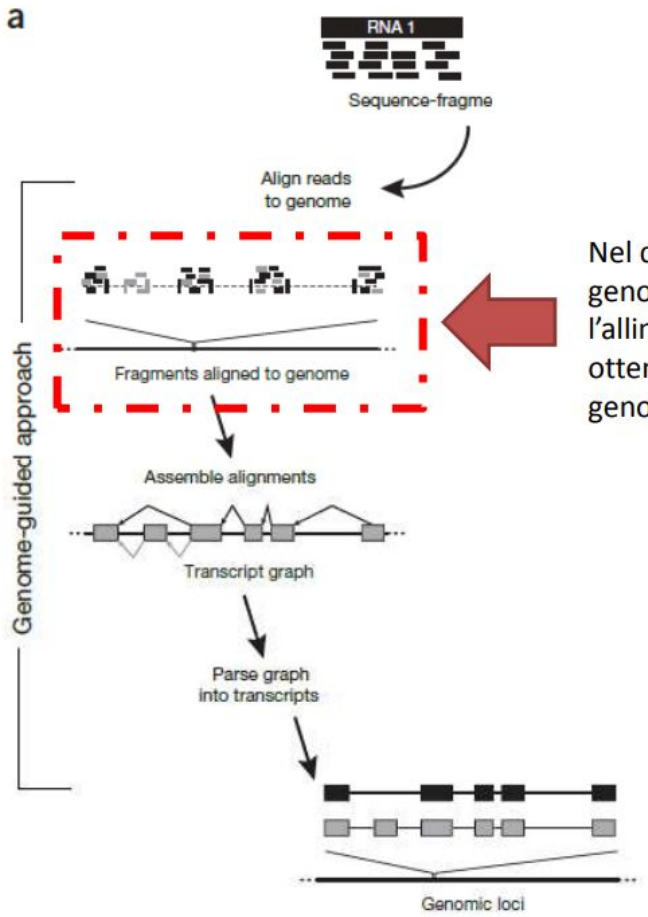
Metodi guidati dal genoma



Metodi indipendenti dal genoma

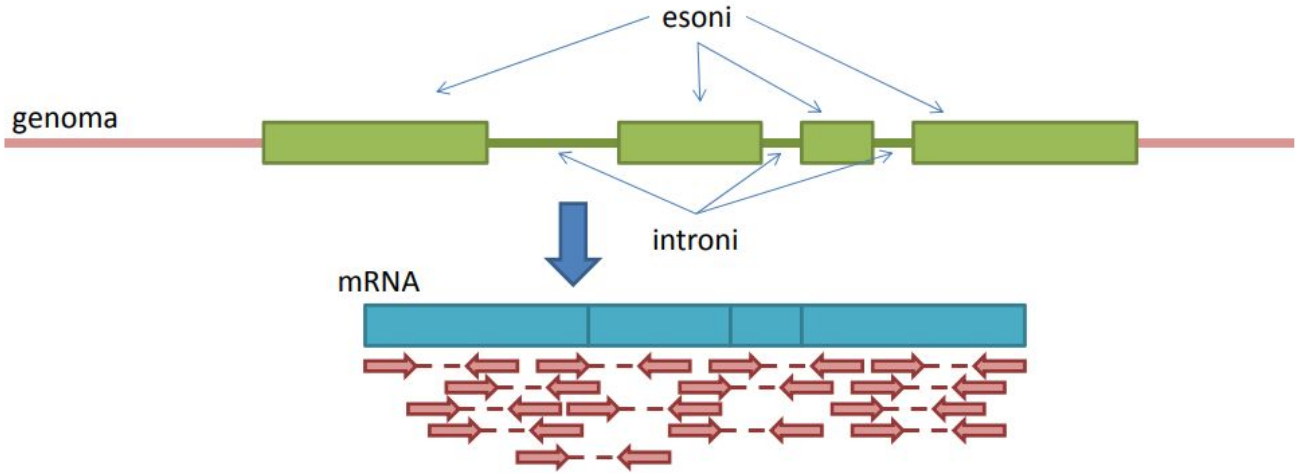
Metodi guidati dal genoma





Nel caso di metodi basati sul genoma il primo passaggio è l'allineamento delle read ottenute dai frammenti al genoma di riferimento

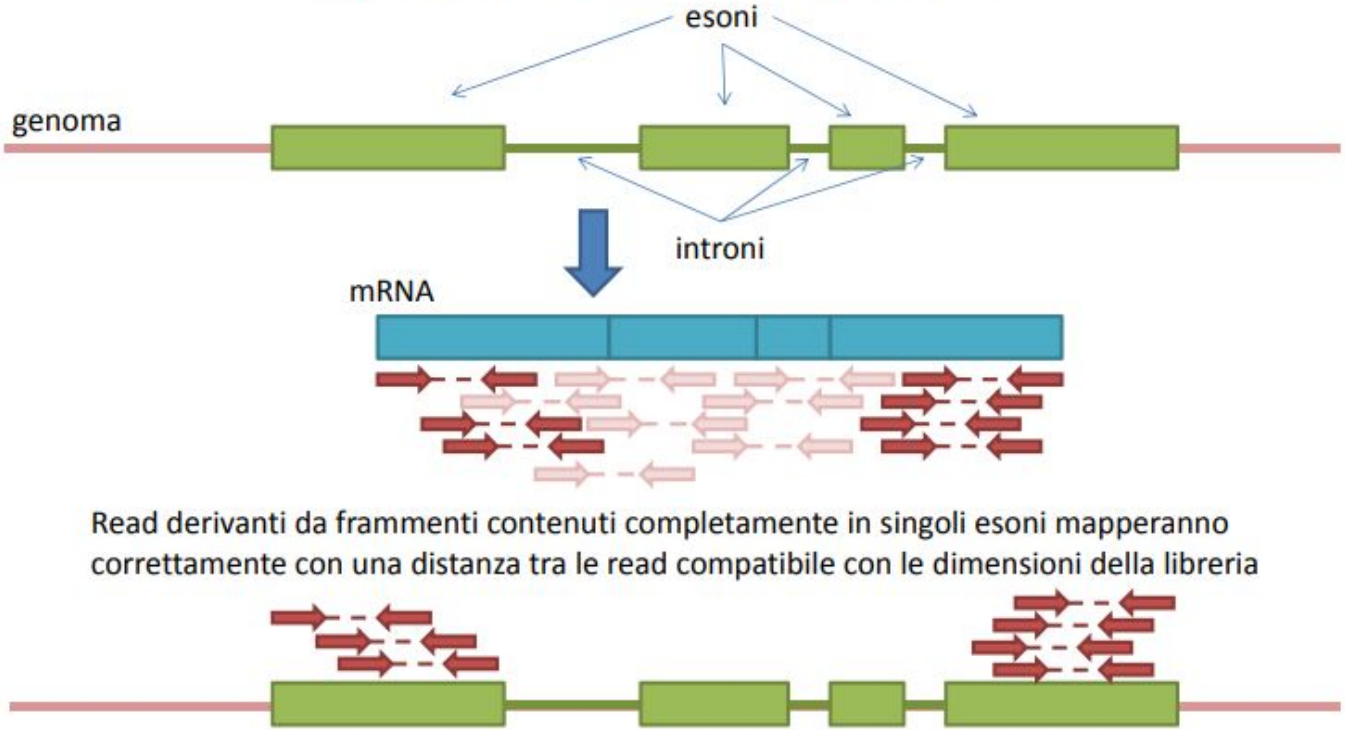
Allineamento di read RNA-Seq ad un genoma di riferimento



In un esperimento RNA-Seq le read vengono generate dal sequenziamento delle estremità di frammenti da 200-300 bp dell'RNA messaggero da cui le sequenze introniche sono state rimosse dal macchinario di splicing durante la maturazione dell'mRNA.

➔ Alcuni frammenti saranno a cavallo delle giunzioni esone-esone

Allineamento di read RNA-Seq ad un genoma di riferimento



Mapping dei dati RNA-seq

Il mapping dei dati RNA-seq, cioè l'allineamento dei reads trascrittomici ad un genoma o trascrittoma di riferimento, è un passaggio cruciale nell'analisi RNA-seq. Esistono numerosi algoritmi e software sviluppati specificamente per affrontare le sfide uniche presentate dal RNA-seq, come la presenza di esoni, introni e splice junctions.

Di seguito alcuni degli algoritmi e software più popolari e comunemente utilizzati per il mapping dei dati RNA-seq:

1. STAR (Spliced Transcripts Alignment to a Reference):
   - E' uno degli allineatori più veloci e viene utilizzato ampiamente nella comunità di ricerca per il suo equilibrio tra velocità e accuratezza.
   - E' in grado di rilevare eventi di splicing alternativo.

2. **HISAT e HISAT2** (Hierarchical Indexing for Spliced Alignment of Transcripts):

- Successore di TopHat2, utilizza un approccio di indexing gerarchico.
- E' molto efficiente in termini di memoria e velocità. E’ usato per mappare dati RNA-seq contro un genoma di riferimento.

3. **SpliceMap**:

- Pensato per la rilevazione di giunzioni senza dipendere da annotazioni di trascrizione esistenti.

4. **MapSplice**:

- Progettato per mappare accuratamente i reads di RNA-seq alle giunzioni di splicing.

5. **GMAP/GSNAP** (Genomic Mapping and Alignment Program):

- Può mappare reads di sequenza sia su genomi che su trascrittomi.

- E' noto per la sua capacità di gestire reads di varie lunghezze, che lo rende utile per dati provenienti da diverse piattaforme di sequenziamento.

6. **Salmon e Kallisto**:

- Questi sono allineatori quasi-liberi, che quantificano rapidamente l'abbondanza trascrizionale senza richiedere un passaggio di allineamento tradizionale. Essi utilizzano un approccio basato sull'hashing per allineare pseudo-aleatoriamente le reads ai trascritti di riferimento.

Viene usato per mappare i dati RNA-seq contro un trascrittoma m(de novo).

Questa lista di programmi non è esaustiva, dato che l'analisi dell'RNA-seq è un campo in rapida e continua evoluzione e nuovi strumenti e metodi continuano a emergere. Tuttavia, questi sono alcuni degli strumenti più consolidati e ampiamente utilizzati nella comunità di ricerca. La scelta dell'allineatore appropriato dipenderà spesso dalla natura specifica del dataset e dagli obiettivi dell'analisi.

STAR

STAR (Spliced Transcripts Alignment to a Reference) è uno degli algoritmi di allineamento più popolari e veloci utilizzati per mappare reads di RNA-seq su un genoma di riferimento. L'algoritmo STAR è stato specificamente progettato per gestire splicing di trascritti, rendendolo ideale per l'allineamento di reads derivati da RNA.

Di seguito riportiamo una sintesi del funzionamento dell'algoritmo STAR:

1. **Generazione dell'indice del genoma**:
   - Prima di poter allineare reads su un genoma di riferimento, STAR genera un indice del genoma utilizzando l'annotazione dei geni (se disponibile).
   - Questo indice è essenzialmente una struttura dati chiamata "hash table" o "suffix array", che consente ricerche efficienti per allineare reads rapidamente.

2. **Scansione del Read**:

- STAR inizia a mappare ciascun read cercando le corrispondenze esatte di brevi segmenti del read (detti "seeds") nel genoma di riferimento.
- Questi "seeds" sono utilizzati come punto di partenza per tentare di estendere l'allineamento al resto del read.

3. **Gestione dello splicing**:

- Una delle principali forze di STAR è la sua capacità di gestire reads che coprono siti di splicing.
- Se un read non può essere allineata in modo contiguo, STAR cerca potenziali siti di splicing che corrispondono a quelli noti (se sono fornite annotazioni) o tenta di identificarne di nuovi.
- Utilizza una serie di regole e pattern per identificare donatori e accettori di splicing, permettendo l'allineamento di reads che attraversano tali siti.

4. Scelta degli allineamenti multipli:
   - Alcuni reads possono essere mappati in più posizioni nel genoma con una simile affidabilità. In tali casi, STAR può scegliere un allineamento basato su vari criteri o può segnalare il read come "multi-mapping".

5. **Generazione di output**:
   - STAR produce un file SAM/BAM che contiene tutti gli allineamenti. Questo file può poi essere utilizzato per ulteriori analisi, come la quantificazione dell'espressione genica.

6. **Ottimizzazioni**:
   - Una delle ragioni per cui STAR è così veloce è a causa delle sue molte ottimizzazioni, come l'uso efficiente della memoria e l'allineamento parallelo.

In termini generali, la forza dell'algoritmo STAR risiede nella sua capacità di allineare reads con grande efficienza, tenendo conto dello splicing, e nel suo utilizzo di annotazioni geniche per migliorare l'accuratezza dell'allineamento. Sebbene questa sia una panoramica semplificata, spero che fornisca una comprensione generale del funzionamento di STAR.

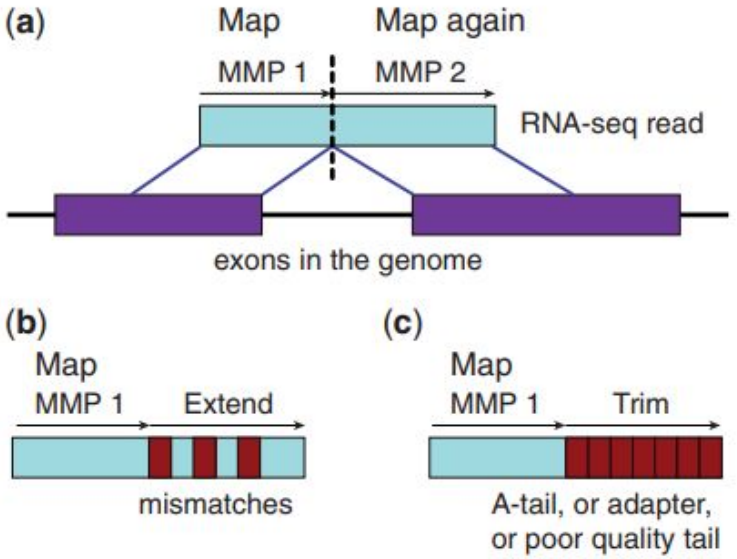


Fig. 1. Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (**a**) splice junctions, (**b**) mismatches and (**c**) tails

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/pdf/bts635.pdf

HISAT2

HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2) è un altro popolare algoritmo di allineamento utilizzato per mappare reads di RNA-seq su un genoma di riferimento, simile a STAR. Tuttavia, HISAT2 utilizza un approccio unico basato su un sistema di indicizzazione gerarchica per gestire l'ampia dimensione del genoma dei mammiferi e le complessità associate allo splicing.

DI seguito riporto una panoramica semplificata del funzionamento dell'algoritmo HISAT2:

1. **Indicizzazione gerarchica**:
   - HISAT2 suddivide il genoma di riferimento in molte piccole parti o "bin".
   - Per ogni bin, viene costruita la struttura di indicizzazione di "Burrows-Wheeler Transform" (BWT) e "FM-index".
   - L'indicizzazione gerarchica consente a HISAT2 di localizzare rapidamente la posizione di un read nel genoma riducendo gradualmente la regione di ricerca attraverso questa struttura multi-livello.

HISAT2

2. **Globali e locali FM-index**:
   - L'indicizzazione principale del genoma viene effettuata tramite il "global FM-index".
   - Per gestire lo splicing, HISAT2 utilizza anche ciò che chiama "local FM-index" per le regioni che circondano i noti siti di splicing, consentendo allineamenti efficienti attraverso questi siti.

3. **Gestione dello splicing**:
   - HISAT2, similmente a STAR, può allineare efficacemente reads che coprono siti di splicing.
   - La combinazione di indici globali e locali di tipo FM-index permette a HISAT2 di gestire efficacemente reads che coprono regioni esoniche e introniche, identificando correttamente le giunzioni tra esoni.

4. **Allineamento di reads**:
   - HISAT2 inizia cercando corrispondenze esatte o quasi esatte per segmenti dei reads usando l'FM-index.
   - Una volta identificato un potenziale allineamento, HISAT2 prova a estendere l'allineamento per tutto il read, tenendo conto delle potenziali discrepanze come mismatch, inserzioni o cancellazioni.

HISAT2

5. Scelta degli allineamenti multipli:
   - Se una read può essere mappata in più posizioni nel genoma, HISAT2 può segnalare il read come "multi-mapping" o scegliere una posizione basata su vari criteri.

6. Generazione di output:
   - HISAT2 produce un file SAM/BAM come output, che può essere ulteriormente elaborato o utilizzato per analisi downstream.

Un vantaggio di HISAT2 è la sua efficienza in termini di memoria e velocità. L'uso di una struttura di indicizzazione gerarchica consente a HISAT2 di gestire genomi di grandi dimensioni, come quello umano, con una quantità relativamente piccola di memoria.

Salmon

Salmon è un tool progettato per la quantificazione rapida e accurata dei livelli di espressione trascrizionale a partire dai dati di RNA-seq. A differenza di STAR o HISAT2, che sono allineatori contro un genoma di riferimento, Salmon si focalizza sulla quantificazione dei trascritti, cioè stima quante copie di ogni trascritto sono presenti nel campione. Un punto chiave di Salmon è che può funzionare sia in modalità di allineamento che in modalità quasi-allineamento, che è molto più veloce.

Ecco una panoramica dell'algoritmo alla base di Salmon:

1. **Modalità Quasi-Allineamento**:
   - Nella modalità quasi-allineamento, Salmon non allinea i reads direttamente sul genoma o sui trascritti. Invece, esso mappa rapidamente i reads sulle sequenze dei trascritti usando un approccio di "quasi-mapping".
   - Questa operazione identifica la posizione approssimativa dei reads sui trascritti senza l'overhead computazionale di un allineamento completo.

2. **Costruzione dell'Indice**:
   - Prima di eseguire la quantificazione, è necessario costruire un indice delle sequenze dei trascritti utilizzando il comando `salmon index`. Questo indice usa una struttura chiamata "hash table" per una rapida ricerca.

3. **Quantificazione dei Trascritti**:

Salmon

Salmon è un tool progettato per la quantificazione rapida e accurata dei livelli di espressione trascrizionale a partire dai dati di RNA-seq. **A differenza di STAR o HISAT2, che sono allineatori, Salmon si focalizza sulla quantificazione dei trascritti, cioè stima quante copie di ogni trascritto sono presenti nel campione**. Un punto chiave di Salmon è che può funzionare sia in modalità di allineamento che in modalità quasi-allineamento, che è molto più veloce.

Di seguito riporto una panoramica dell'algoritmo alla base di Salmon:

1. **Modalità Quasi-Allineamento**:
   - Nella modalità di quasi-allineamento, Salmon non allinea i reads direttamente sul genoma o sui trascritti. Invece, esso mappa rapidamente le reads sulle sequenze dei trascritti usando un approccio di "**quasi-mapping**".
   - Questa operazione identifica la posizione approssimativa dei reads sui trascritti senza l'overhead computazionale di un allineamento completo.

Salmon: approccio di "quasi-mapping".

Il concetto di "quasi-mapping" è centrale per comprendere l'efficienza di alcuni moderni strumenti di quantificazione, come Salmon.

Il mapping o allineamento completo di reads su un riferimento (ad esempio, un genoma o un set di trascritti) richiede che ogni read venga allineato sul riferimento, identificando esattamente dove e come si adatta. Questo processo implica di considerare:

- Mismatches (differenze tra il read e il riferimento).
- Indels (inserzioni o cancellazioni nel read rispetto al riferimento).
- Un punteggio per ogni possibile allineamento per determinare l'allineamento "migliore" o più probabile.

L'overhead computazionale di questo processo può essere significativo, soprattutto quando si hanno milioni o miliardi di reads.

Salmon: approccio di "quasi-mapping".

Il quasi-mapping, al contrario, non cerca di determinare l'allineamento esatto di un read sul riferimento. Invece, rapidamente identifica la posizione approssimativa o le posizioni in cui un read potrebbe essere allineato. Ciò elimina la necessità di gestire mismatches e indels in dettaglio o di calcolare punteggi di allineamento complessi.

Ecco cosa significa nella pratica:

- Velocità: Poiché il quasi-mapping non si preoccupa degli allineamenti esatti, è molto più veloce dell'allineamento tradizionale.
- Minore uso della memoria: L'overhead di memoria per conservare i dettagli sugli allineamenti esatti viene ridotto.
- Focus sulla quantificazione: Per la quantificazione dell'espressione dei trascritti, spesso ciò che è veramente importante è sapere su quale trascritto si mappa un read, piuttosto che la sua esatta posizione o allineamento. Il quasi-mapping fornisce queste informazioni senza il dettaglio extra che potrebbe non essere necessario per questa specifica applicazione.

Salmon: approccio di "quasi-mapping".

In sintesi, il "quasi-mapping" permette di identificare rapidamente e con un overhead computazionale ridotto dove una read potrebbe mappare su un set di trascritti, fornendo le informazioni essenziali necessarie per la quantificazione dell'espressione senza l'overhead e la complessità degli allineamenti completi.

2. **Costruzione dell'Indice**:
   - Prima di eseguire la quantificazione, è necessario costruire un indice delle sequenze dei trascritti utilizzando il comando `salmon index`. Questo indice usa una struttura chiamata "hash table" per una rapida ricerca.

3. **Quantificazione dei Trascritti**:
   - Usando l'indice dei trascritti e i reads (o i fragmenti, nel caso di dati paired-end), Salmon stima l'abbondanza dei trascritti utilizzando un modello probabilistico.
   - Questo modello tiene conto di vari fattori, come le sequenze dei trascritti, la lunghezza dei reads, i potenziali errori di sequenziamento e il bias di sequenziamento.

Salmon

Nel contesto della sequenziamento, specialmente nell'RNA-seq, il termine "bias" si riferisce a qualsiasi tendenza sistematica che può influenzare la rappresentatività e l'interpretazione dei dati sequenziati. Questi bias possono derivare da vari passaggi nel protocollo di sequenziamento e possono influenzare l'apparente abbondanza di particolari sequenze o trascritti. Due tipi comuni di bias nel contesto della sequenziamento sono il "**bias di sequenza**" (preferenza nell'amplificazione e sequenziamento di frammenti di DNA/RNA basati sulla loro sequenza nucleotidica: può essere introdotto durante vari passaggi, come la frammentazione, la ligation degli adattatori o la PCR. Per esempio, certi primer possono avere una maggiore affinità per sequenze specifiche, portando a una sovrarappresentazione o sottorappresentazione di determinate sequenze nei dati finali) e il "**bias di CG**" (tendenza sistematica nella quale frammenti con diversi contenuti di guanina e citosina (GC) sono amplificati o sequenziati con efficienze diverse. I frammenti ricchi di GC potrebbero, ad esempio, formare strutture secondarie stabili che influenzano l'efficienza della PCR, mentre i frammenti poveri di GC potrebbero non essere amplificati altrettanto efficacemente in certe condizioni. Questo può portare a una distorsione nell'abbondanza apparente di trascritti basata sul loro contenuto di GC).

Salmon

4. **Correzione del Bias**:
   - Salmon implementa metodi per correggere vari tipi di bias, come il bias di sequenza e il bias GC, che possono influenzare la quantificazione dell'espressione.

5. **Bootstrapping e Estimazione della Variabilità:**
   - Una caratteristica distintiva di Salmon è la sua capacità di eseguire il "bootstrapping" sui dati, che produce multiple stime replicate dell'abbondanza dei trascritti. Questo consente di stimare la variabilità nell'abbondanza dei trascritti.

6. **Output**:
   - Salmon produce un file di output con le stime dell'abbondanza per ogni trascritto. Queste abbondanze sono tipicamente espresse in TPM (Transcripts Per Million) o conteggi grezzi.

Salmon

7. **Integrazione con altri tools**:

- Gli output di Salmon possono essere facilmente integrati con altri tool di analisi downstream, come tximport o DESeq2, per ulteriori analisi, come la determinazione dei geni differentemente espressi.

Un vantaggio significativo di Salmon è la sua velocità. L'uso di quasi-mapping anziché di allineamenti tradizionali consente a Salmon di quantificare l'espressione trascrizionale in pochi minuti, pur mantenendo un'altissima accuratezza del risultato.