Galaxy

Introduzione

Negli ultimi anni, l'analisi bioinformatica ha assunto un ruolo cruciale nella ricerca scientifica, consentendo di elaborare e interpretare enormi quantità di dati biologici. Tra gli strumenti più utilizzati per facilitare questi processi, la piattaforma Galaxy si distingue per la sua flessibilità e accessibilità. **Galaxy** è un ambiente open-source progettato per consentire agli utenti, anche senza competenze avanzate di programmazione, di eseguire analisi bioinformatiche complesse attraverso un'interfaccia grafica intuitiva.
Il presente lavoro si propone di illustrare il funzionamento di Galaxy applicandolo a un caso di studio specifico: il **trimming** di file FASTQ e l'utilizzo di strumenti di **allineamento** sui dati ripuliti. Inizialmente, verrà fornita una panoramica generale della piattaforma, descrivendone le caratteristiche principali e le funzionalità offerte. Successivamente, si passerà alla descrizione dettagliata del workflow impiegato per l'analisi dei dati biologici selezionati.
Attraverso questo studio, si intende evidenziare come Galaxy possa facilitare l'elaborazione dei dati genomici, riducendo la complessità computazionale e favorendo la riproducibilità delle analisi scientifiche.

La piattaforma Galaxy

Galaxy è una piattaforma open-source per l'analisi bioinformatica, sviluppata con l'obiettivo di rendere più accessibile l'elaborazione di dati biologici a ricercatori e professionisti del settore. Una delle sue principali caratteristiche è la possibilità di eseguire analisi attraverso un'interfaccia web-based, eliminando la necessità di installare software complessi o di possedere avanzate competenze di programmazione.
La piattaforma supporta un'ampia varietà di strumenti bioinformatici, consentendo di costruire pipeline analitiche personalizzate e riproducibili. Grazie alla sua architettura modulare, Galaxy permette l'integrazione di tool di terze parti e l'automazione dei flussi di lavoro, migliorando l'efficienza e la scalabilità delle analisi genomiche.

Funzionalità principali

Galaxy offre una serie di funzionalità che ne facilitano l'utilizzo in diversi ambiti della ricerca bioinformatica:

- **Interfaccia grafica intuitiva**: permette di eseguire analisi senza la necessità di scrivere codice, attraverso una gestione visuale dei workflow.
- **Gestione dei dati**: supporta il caricamento, l'organizzazione e la condivisione dei dati in diversi formati, inclusi FASTQ, BAM e VCF.
- **Riproducibilità delle analisi**: consente di salvare e condividere workflow, garantendo la trasparenza e la ripetibilità delle elaborazioni.
- **Scalabilità**: può essere utilizzato su infrastrutture locali, server dedicati o su cloud, adattandosi a diverse esigenze computazionali.

Ampia libreria di strumenti: include tool per il pre-processing dei dati, l'allineamento delle sequenze, l'analisi differenziale dell'espressione genica e molto altro.

Accessibilità e comunità

Galaxy è progettato per essere accessibile a diversi livelli di utenti, dai principianti ai bioinformatici esperti. La sua natura open-source favorisce il contributo della comunità scientifica, che può sviluppare e condividere nuovi strumenti e workflow. Inoltre, esistono diverse istanze pubbliche della piattaforma, come Galaxy Europe e Galaxy Main, che offrono accesso gratuito a risorse computazionali per la ricerca.
Nel contesto del presente lavoro, Galaxy verrà utilizzato per il trimming di file FASTQ e l'allineamento delle sequenze, dimostrando come la piattaforma possa semplificare e velocizzare questi processi bioinformatici.

Pipeline di Analisi dei Dati

In questa sezione descriviamo il **workflow** seguito per l'analisi dei dati di sequenziamento utilizzando la piattaforma Galaxy. L'analisi ha incluso il preprocessing delle sequenze, il trimming dei file FASTQ e l'allineamento al genoma di riferimento mediante Bowtie2.

Descrizione del dataset selezionato

Per questo progetto è stato utilizzato il dataset **SRX3733298**, contenente dati di **sequenziamento RNA-Seq** ottenuti da diversi ceppi di *Escherichia coli*, un microrganismo modello ampiamente impiegato in ambito microbiologico e biotecnologico. L'esperimento è stato concepito con l'obiettivo di confrontare l'**espressione genica** tra ceppi **resistenti**e **suscettibili** a **Magainin I**, un peptide antimicrobico noto per la sua attività antibiotica.

Tale confronto consente di identificare i geni coinvolti nella risposta batterica alla presenza del peptide, offrendo spunti utili per comprendere i meccanismi molecolari alla base della **resistenza agli antimicrobici**, un tema di grande rilievo nella ricerca microbiologica contemporanea.

Motivazioni della scelta

La selezione di questo dataset è stata guidata da diversi fattori:

1. **Compatibilità con la piattaforma Galaxy**
   Il dataset è strutturato in modo tale da poter essere analizzato interamente tramite Galaxy, una piattaforma bioinformatica open source che consente l'esecuzione di pipeline standard di RNA-Seq, come il trimming e l'allineamento.

2. **Rilevanza biologica e applicativa**
   Il tema della **resistenza antimicrobica** rappresenta una delle principali sfide della microbiologia moderna. L'analisi di questo dataset permette di affrontare un problema reale e attuale, fornendo una concreta applicazione dei metodi bioinformatici appresi.

3. **Dimensioni contenute del dataset**
   Con circa **7 milioni di letture**, il dataset presenta una dimensione bilanciata: sufficientemente ampia da garantire un'analisi affidabile, ma non eccessiva da richiedere infrastrutture computazionali avanzate.

4. **Possibilità di applicare strumenti bioinformatici fondamentali**
   Il dataset permette di impiegare strumenti essenziali per l'analisi di RNA-Seq, quali:
   – **Cutadapt** per il trimming delle letture,
   – **Bowtie2** per l'allineamento al genoma di riferimento,
   – **FeatureCounts** per il conteggio delle letture per gene,

5. **Accessibilità concettuale**
   Il dataset si presta a un'analisi bioinformatica completa anche senza un'approfondita conoscenza delle dinamiche biologiche sottostanti, rendendolo adatto a un progetto a forte componente computazionale.

Caratteristiche tecniche del dataset

- **Codice identificativo dell'esperimento:** *SRX3733298*
- **Organismo di riferimento:** *Escherichia coli*
- **Tecnologia di sequenziamento:** RNA-Seq
- **Tipo di dato:** dati di espressione genica (trascrittoma)
- **Piattaforma utilizzata:** *Illumina MiSeq*
- **Layout della libreria:** *Paired-end* (sequenziamento da entrambe le estremità dei frammenti)

- **Origine della libreria:** *Transcriptomic*, ovvero ottenuta dall'RNA trascritto attivamente nei campioni
- **Strategia di selezione:** *Random*, senza arricchimento di regioni specifiche del trascrittoma

Nozioni introduttive per la comprensione del dataset

Dataset SRX3733298

Il termine *dataset* si riferisce a un insieme strutturato di dati raccolti nel corso di un esperimento scientifico. In questo caso, **SRX3733298** è il codice identificativo assegnato a un esperimento archiviato all'interno del **European Nucleotide Archive (ENA)**. Tale identificativo consente di accedere in maniera univoca a questo specifico insieme di dati.

Sequenziamento RNA-Seq

RNA-Seq (RNA sequencing) è una tecnologia che consente di analizzare **quali geni sono attivamente trascritti** in un determinato momento all'interno di una cellula. In altre parole, permette di quantificare l'**espressione genica**.

Per comprendere meglio il concetto, si può ricorrere a una metafora: il **DNA** di un batterio può essere paragonato a un manuale d'istruzioni. Tuttavia, in ogni istante il batterio "legge" soltanto le parti rilevanti di questo manuale, ossia quei **geni** che devono essere attivati per far fronte a specifiche condizioni ambientali. L'**RNA** rappresenta la copia temporanea di queste parti attive, e l'RNA-Seq consente di rilevare quali porzioni del genoma vengono effettivamente trascritte e in quale quantità.

L'analisi RNA-Seq del dataset in questione permette dunque di determinare **quali geni vengono attivati nei batteri** in condizioni differenti, fornendo una visione dettagliata dei meccanismi di regolazione genica.

Diversi ceppi di *Escherichia coli*

Escherichia coli (*E. coli*) è un batterio Gram-negativo molto studiato nei laboratori di microbiologia, grazie alla sua rapidità di crescita e alla disponibilità di numerose informazioni genetiche. Il termine **ceppo** fa riferimento a una variante genetica all'interno della stessa specie batterica. Sebbene tutti i ceppi appartengano alla specie *E. coli*, possono presentare caratteristiche differenti, ad esempio una diversa sensibilità agli antibiotici.

Nel caso specifico del dataset SRX3733298, sono stati analizzati **ceppi di *E. coli* con fenotipi opposti** rispetto alla sensibilità a un peptide antimicrobico chiamato **Magainin I**: alcuni risultano **resistenti**, altri **suscettibili**.

L'**espressione genica** indica il processo attraverso il quale l'informazione contenuta in un gene viene utilizzata per produrre una molecola funzionale, solitamente una proteina. Analizzare l'espressione genica significa determinare **quali geni vengono attivati** e **in quale misura**, in risposta a determinati stimoli o condizioni ambientali.

L'obiettivo dell'esperimento è quello di **confrontare l'espressione genica tra ceppi resistenti e suscettibili a Magainin I**, al fine di identificare i **geni che contribuiscono alla resistenza** antimicrobica. Se, ad esempio, un certo gene risulta fortemente espresso nei ceppi resistenti ma non in quelli suscettibili, esso potrebbe avere un ruolo determinante nella capacità del batterio di sopravvivere al trattamento con Magainin I.

Ceppi resistenti e suscettibili a Magainin I

Magainin I è un peptide antimicrobico appartenente alla famiglia delle difensine, che agisce danneggiando le membrane cellulari dei batteri. Nel dataset analizzato, i ceppi di *E. coli* sono stati classificati in due gruppi:

- **Ceppi resistenti**, in grado di sopravvivere all'esposizione a Magainin I;
- **Ceppi suscettibili**, che risultano danneggiati o uccisi dallo stesso peptide.

L'analisi dell'espressione genica tra questi due gruppi consente di identificare i **meccanismi molecolari coinvolti nella resistenza**.

Obiettivo dell'esperimento

L'obiettivo principale dell'esperimento è quello di **identificare i geni coinvolti nella resistenza antimicrobica**, confrontando l'espressione genica tra ceppi resistenti e suscettibili. In particolare, si vuole:

- Determinare **quali geni sono maggiormente espressi nei ceppi resistenti**;
- Formulare ipotesi sui **meccanismi molecolari di difesa** messi in atto dai batteri;
- Fornire **possibili target terapeutici** per lo sviluppo di nuovi antimicrobici.

Ad esempio, se un gene codifica per una proteina in grado di impedire l'ingresso di Magainin I nella cellula, la sua espressione potrebbe rappresentare un elemento chiave nella sopravvivenza del batterio.

Rilevanza biologica e applicativa

Comprendere i meccanismi genetici alla base della resistenza antimicrobica è di fondamentale importanza nel contesto attuale, in cui il fenomeno della **resistenza agli antibiotici** rappresenta una crescente minaccia per la salute pubblica a livello globale. L'identificazione dei geni coinvolti può favorire:

- La progettazione di **nuove strategie terapeutiche**;
- Il miglioramento dell'**efficacia dei trattamenti antimicrobici**;
- Lo sviluppo di **diagnostiche molecolari** per l'identificazione precoce di ceppi resistenti.

Il sequenziamento Paired-End

Nel dataset SRX3733298 sono presenti **due file FASTQ**, in quanto i dati sono stati generati tramite **sequenziamento Paired-End**, una tecnologia ampiamente utilizzata negli strumenti di sequenziamento **Illumina**.

Il **Paired-End Sequencing** (sequenziamento a letture accoppiate) consiste nella **lettura di entrambe le estremità** di ogni frammento di DNA o RNA. A differenza del **Single-End Sequencing**, che legge solo un'estremità del frammento, il paired-end produce **due letture per ogni frammento**, consentendo una maggiore accuratezza e informazione.

Le letture prodotte vengono archiviate in **due file FASTQ distinti**:

- Un file per le **letture Forward** (in direzione 5' → 3')
- Un file per le **letture Reverse** (in direzione 3' → 5')

Nel caso del dataset analizzato:

- SRR6760835_1.fastq.gz → contiene le letture forward
- SRR6760835_2.fastq.gz → contiene le letture reverse

Fasi del sequenziamento Paired-End

1. **Frammentazione**: l'RNA (o il DNA, nei casi genomici) viene frammentato in segmenti di lunghezza definita, generalmente compresa tra 150 e 300 basi.
2. **Aggiunta degli adattatori**: agli estremi dei frammenti vengono legati adattatori necessari per il riconoscimento e la lettura da parte del sequenziatore.
3. **Sequenziamento bidirezionale**: il sequenziatore legge ogni frammento da entrambi i lati, generando:
    - **Read 1**: lettura della prima estremità (forward)
    - **Read 2**: lettura della seconda estremità (reverse)

Vantaggi del Paired-End Sequencing

Il sequenziamento paired-end presenta numerosi vantaggi rispetto al single-end, rendendolo particolarmente indicato per analisi complesse come l'**RNA-Seq**:

1. Maggiore accuratezza nel mapping

La disponibilità di due estremità facilita l'allineamento al genoma di riferimento, riducendo l'ambiguità nelle regioni ripetute o altamente conservate. Se una lettura singola può essere mappata in più posizioni, la lettura accoppiata consente di **restringere significativamente l'intervallo di posizionamento**.

2. Migliore copertura e ricostruzione

La lettura da entrambe le estremità aumenta la **copertura complessiva del trascrittoma**, contribuendo a una ricostruzione più completa delle sequenze, in particolare nelle aree difficili o frammentate.

3. Identificazione di varianti strutturali

Il paired-end è utile per rilevare:

- **Fusioni geniche**
- **Riarrangiamenti strutturali**
- **Inserzioni e delezioni**
- Errori di sequenziamento che potrebbero essere ignorati in modalità single-end

4. Vantaggi specifici per l'analisi RNA-Seq

In studi trascrittomici, il paired-end consente di:

- Discriminare tra **diverse isoforme** geniche
- Migliorare l'assemblaggio de novo dei trascritti in assenza di un genoma di riferimento
- Aumentare la precisione nella quantificazione dell'espressione genica

Il Trimming nei dati di sequenziamento

Il **trimming** rappresenta una fase preliminare fondamentale nell'analisi dei dati di sequenziamento ad alta processività (next-generation sequencing, NGS). Questa operazione consiste nella **rimozione di porzioni non informative, artefatti tecnici e basi di bassa qualità** dalle letture (reads) grezze contenute nei file **FASTQ**, al fine di migliorarne la qualità complessiva prima dell'allineamento al genoma di riferimento.

Le letture ottenute direttamente dal sequenziatore, infatti, possono contenere diverse componenti non desiderate, tra cui:

- **Adattatori di sequenziamento**: brevi sequenze artificiali aggiunte durante la preparazione della libreria.
- **Basi con bassa qualità di chiamata**: in particolare nelle estremità 3', dove gli errori sono statisticamente più frequenti.
- **Sequenze troppo corte** o contenenti un numero elevato di **"N"** (basi non identificate).
- **Contaminanti e artefatti tecnici**, come sequenze omopolimeriche (poli-A/T/C/G).

L'obiettivo del trimming è quindi quello di **"pulire" i dati grezzi**, ottenendo sequenze di maggiore affidabilità, che possono essere utilizzate in modo più efficiente nelle successive analisi bioinformatiche.

Importanza del Trimming: vantaggi e implicazioni

Il trimming svolge un ruolo cruciale per garantire **l'accuratezza e la robustezza delle analisi downstream**, ovvero le analisi successive al pre-processing, come l'allineamento, il conteggio dei trascritti o l'identificazione di varianti.

Principali vantaggi del trimming:

1. **Maggiore accuratezza nell'allineamento**
   Le sequenze pulite si mappano più correttamente sul genoma di riferimento, riducendo la percentuale di letture mal allineate o scartate.
2. **Miglioramento delle analisi a valle**
   Le fasi successive, come l'analisi dell'**espressione genica**, la **chiamata di varianti** o l'**assemblaggio de novo**, risultano più precise e meno soggette a distorsioni.
3. **Ottimizzazione delle risorse computazionali**
   La rimozione di sequenze non informative consente di **ridurre le dimensioni dei file** e il numero totale di basi da analizzare, **accorciando i tempi di elaborazione**.
4. **Eliminazione del rumore di fondo**
   La presenza di dati "sporchi" può generare **falsi positivi** o introdurre **bias** nell'interpretazione dei risultati. Il trimming contribuisce a limitare queste problematiche.

Elementi rimossi durante il trimming

Durante il processo di trimming vengono selettivamente eliminate le seguenti componenti:

1. **Adattatori**
   Sequenze artificiali aggiunte per facilitare il sequenziamento, che non appartengono al genoma o trascrittoma originale e devono essere rimosse.
2. **Basi a bassa qualità**
   Le basi con un punteggio **Phred score** inferiore a una soglia definita (es. Q20 o Q30) vengono tagliate, in quanto potenzialmente errate.
3. **Letture troppo corte**
   Dopo la rimozione delle porzioni di bassa qualità, le letture risultanti che non raggiungono una lunghezza minima (es. <30 bp) vengono scartate in quanto non informative.
4. **Sequenze anomale o contaminanti**
   Sequenze anomale (poli-A/T/G, stringhe di "N", ecc.), che non riflettono eventi biologici reali.

Cutadapt

Introduzione

Cutadapt è un tool bioinformatico altamente performante utilizzato per il **trimming e il filtraggio delle sequenze di DNA e RNA** provenienti da esperimenti di sequenziamento di nuova generazione (Next Generation Sequencing, NGS). È progettato per eseguire queste operazioni in modo **estremamente rapido ed efficiente**, sfruttando le **GPU (Graphics Processing Unit)** per accelerare l'elaborazione dei dati.

Il nome "Cutadapt" deriva dalla combinazione di **CUDA** (Compute Unified Device Architecture), una tecnologia sviluppata da NVIDIA per l'elaborazione parallela su GPU, e **Cutadapt**, un noto strumento per il trimming degli adattatori nelle sequenze.

Finalità e Utilizzo Principale

Cutadapt viene utilizzato nella **fase di preprocessing dei dati NGS**, in particolare per:

1. **Rimuovere adattatori** e altri artefatti di sequenziamento.
2. **Tagliare le basi di bassa qualità** dalle estremità delle letture.
3. **Filtrare** le letture troppo corte o contenenti caratteri ambigui.
4. Ottimizzare i dati per le analisi downstream, come **l'allineamento al genoma** o l'analisi dell'espressione genica.
5. **Caratteristiche Principali**

Caratteristica	Descrizione
Alta velocità	Grazie all'uso della GPU, Cutadapt può processare milioni di letture al secondo, rendendolo uno degli strumenti più veloci nel campo del preprocessing NGS.
Supporto per letture paired-end	È compatibile con sequenziamenti **paired-end**, come quelli prodotti da piattaforme Illumina.
Basato su Cutadapt	Riutilizza molte funzionalità di Cutadapt, ma le implementa in modo ottimizzato per l'hardware GPU.
Output compatibile	I file generati sono pienamente compatibili con i formati standard bioinformatici (FASTQ, FASTA).
Scalabilità	Adatto sia per dataset piccoli che per progetti di larga scala, tipici degli studi di trascrittomica o metagenomica.

Requisiti Tecnici

Per funzionare al meglio, Cutadapt richiede:

- **GPU compatibile NVIDIA** con supporto CUDA.
- **Driver aggiornati** e installazione del toolkit CUDA.
- Sistemi operativi Linux o compatibili (in genere, non disponibile nativamente su Windows).
- **Ambiente bioinformatico** preconfigurato oppure piattaforme come **Galaxy** che lo integrano senza la necessità di installazioni complesse.

Funzionalità Tecniche

1. **Adattatori e linkers**: Rimozione precisa delle sequenze adattatrici fornite dall'utente o identificate automaticamente.
2. **Trimming di qualità**: Basato su soglie definite (es. Q20 o Q30), rimuove le basi dalle estremità delle letture se la qualità è troppo bassa.
3. **Filtraggio per lunghezza**: Permette di escludere letture troppo corte dopo il trimming (es. < 30 bp).
4. **Supporto per formati compressi**: Gestisce anche file .gz, evitando la necessità di decompressione manuale.
5. **Parallelizzazione massiva**: Esegue operazioni su migliaia di letture contemporaneamente grazie alla struttura parallela delle GPU.

Integrazione in Galaxy

Su Galaxy, Cutadapt è disponibile come tool preinstallato. I vantaggi principali dell'utilizzo su questa piattaforma includono:

- Nessun bisogno di installare manualmente CUDA o driver NVIDIA.
- Interfaccia user-friendly per definire parametri di trimming.
- Log dei processi e confronto diretto con altri strumenti.
- Output immediatamente utilizzabile per tool downstream.

Cutadapt rappresenta una delle soluzioni più moderne ed efficienti per il trimming dei dati NGS. Grazie alla sua **velocità, precisione e compatibilità con GPU**, è particolarmente indicato per analisi bioinformatiche intensive, come quelle legate all'**espressione genica** o alla **metagenomica**.

Nel tuo progetto, l'uso di Cutadapt ha permesso di **migliorare la qualità delle letture** e **ridurre i tempi di elaborazione**, facilitando l'allineamento al genoma di *Escherichia coli* e ponendo solide basi per un'analisi differenziale affidabile.

Parametri di configurazione del tool Cutadapt

1. Tipo di letture: Single-end o Paired-end

- **Valore selezionato**: *Paired-end*
- **Motivazione**: Il dataset comprende due file FASTQ (SRR6760835_1.fastq.gz e SRR6760835_2.fastq.gz), corrispondenti alle letture forward e reverse. L'opzione *paired-end* è quindi corretta.

2. Minimum Overlap Length

- **Valore**: *3*
- **Descrizione**: Definisce la lunghezza minima di sovrapposizione tra la sequenza adattatrice e la lettura. Un valore di 3 basi consente di evitare trimming erronei dovuti a sovrapposizioni casuali, mantenendo al contempo una buona sensibilità.

3–4. Wildcards Matching

- **Match Wildcards in Reads**: *No*
- **Match Wildcards in Adapters**: *Yes*
- **Motivazione**: Le wildcard sono ammesse solo negli adattatori, dove possono comparire varianti. Il matching nelle letture è stato disattivato per evitare rimozioni non specifiche.

5. Ricerca di adattatori nel complemento inverso

- **Valore**: *No*
- **Motivazione**: L'analisi è stata condotta considerando adattatori posizionati in orientamento diretto. Non è stato necessario analizzare il complemento inverso.

6. Azione in caso di match con un adattatore

- **Valore**: *Trim (adattatore e sequenza a monte/valle)*
- **Motivazione**: La rimozione sia dell'adattatore che delle porzioni adiacenti assicura un trimming completo, utile per evitare distorsioni nelle fasi successive di analisi.

7. Tasso massimo di errore (Maximum Error Rate)

- **Valore**: *0.1*
- **Motivazione**: Consente fino al 10% di mismatch nella regione di allineamento con l'adattatore, un valore generalmente considerato bilanciato tra specificità e sensibilità.

8. Consentire indels

- **Valore**: *No*
- **Motivazione**: Non sono stati utilizzati adattatori ancorati al 5'; l'opzione di consentire indels è stata disabilitata per evitare allineamenti ambigui.

9. Numero massimo di match dell'adattatore per lettura (Match Times)

- **Valore**: *1*
- **Motivazione**: È stata consentita una singola rimozione per lettura, sufficiente per la maggior parte dei dataset standard.

10–14. Filtri opzionali

- **Maximum Length (R2)**, **Max N**, **Max Expected Errors**, **Max Average Expected Errors**: *Non impostati*
- **Discard CASAVA-filtered Reads**: *No*
- **Motivazione**: Non sono stati applicati filtri restrittivi specifici su lunghezza massima, contenuto di "N" o tasso di errore previsto, al fine di mantenere la massima copertura possibile. Nessun filtro CASAVA è stato richiesto.

15. Pair Filter

- **Valore**: *Any*
- **Motivazione**: La coppia di letture viene scartata se **una delle due** non soddisfa i criteri di qualità. Questa impostazione previene l'inserimento di coppie incomplete o inaffidabili nell'analisi downstream.

16. Read Modification Options

- **Valore**: *Select all*
- **Motivazione**: Sono state selezionate tutte le opzioni di modifica disponibili per massimizzare l'output informativo e agevolare strumenti come **MultiQC**.

17. Rimozione di basi iniziali da R1

- **Valore**: *0*
- **Motivazione**: Nessuna base è stata rimossa a priori. Le rimozioni sono state effettuate solo in base a qualità e adattatori.

18. Quality Cutoff (R1 e R2)

- **Valore**: *Non impostato*
- **Motivazione**: Non è stata definita una soglia di qualità manuale; Cutadapt ha operato utilizzando i parametri di trimming automatico, già efficaci per il tipo di dati analizzati.

19. Lunghezza minima (Minimum Length per R1 e R2)

- **Valore**: *1*

- **Motivazione**: Sono state mantenute anche le letture molto corte, per non perdere dati potenzialmente informativi. Tuttavia, si può considerare l'impostazione di un filtro più restrittivo (es. > 30 bp) nelle fasi successive.

Bowtie2

Introduzione

Bowtie2 è uno dei software di riferimento per l'**allineamento delle letture di sequenziamento** su un genoma o trascrittoma di riferimento. Viene largamente impiegato nelle pipeline di analisi di dati **RNA-Seq**, **DNA-Seq** e **metagenomici**, per progetti su vasta scala grazie alla sua **alta velocità e accuratezza**.

Bowtie2 è il successore migliorato di **Bowtie**, pensato per allineare **letture di lunghezza variabile**, anche lunghe (fino a centinaia di basi), ottenute da tecnologie come **Illumina**, **Ion Torrent** e **BGI**.

Obiettivo di Bowtie2

L'obiettivo principale di Bowtie2 è quello di **mappare** milioni di brevi sequenze (read) ottenute da esperimenti di sequenziamento sul **genoma di riferimento**, restituendo come output:

- La **posizione di allineamento** di ogni lettura.
- Il **numero di mismatch**, inserzioni o delezioni.
- File in formato **SAM/BAM** compatibili con gli strumenti downstream (es. featureCounts, HTSeq, IGV, etc.).

Come Funziona: Principio Algoritmico

BWT + FM-index

Bowtie2 è basato sull'algoritmo **Burrows-Wheeler Transform (BWT)** e sull'**FM-index**, una struttura dati altamente compressa ed efficiente per la ricerca di stringhe.

Questa architettura consente a Bowtie2 di:

- Effettuare ricerche estremamente veloci.
- Allineare letture con **piccolo consumo di memoria**.
- Gestire **grandi genomi** (es. umano, murino, batterici).

Caratteristiche Tecniche Principali

Caratteristica	Descrizione
Compatibilità	Funziona su Linux, Mac, Windows; integrato in piattaforme come Galaxy.
Input	File FASTQ (anche compressi), single-end o paired-end.

Caratteristica	Descrizione
Output	File SAM/BAM contenenti le letture allineate.
Efficienza	Estremamente veloce e con basso consumo RAM (~4 GB per genoma umano).
Personalizzazione	Parametri configurabili: mismatch ammessi, tipo di allineamento, numero di thread, ecc.
Supporto per paired-end	Calcola l'allineamento simultaneo delle coppie di letture.
Multi-threading	Supporta il parallelismo per sfruttare al massimo le CPU disponibili.

Configurazione dello strumento Bowtie2 per l'allineamento delle letture

1. **Selezione del genoma di riferimento**
   - **Scelta effettuata**: caricamento manuale di un file FASTA dal proprio workspace.
   - **Descrizione**: anziché utilizzare un indice preconfigurato disponibile nella piattaforma Galaxy, è stato preferito scaricare e utilizzare un **genoma di riferimento specifico**, coerente con l'organismo e la versione utilizzata nel dataset di sequenziamento.
   - **File utilizzato**: Escherichia_coli_str_k_12_substr_mg1655_gca_000005845.ASM584v2.dna.chromosome.Chromosome
   - **Fonte**: scaricato dal portale **Bacteria Ensembl**, versione **ASM584v2**, corrispondente al ceppo *Escherichia coli*K-12 substrain MG1655, in linea con il dataset analizzato.
2. **Tipo di libreria**
   - **Opzione selezionata**: *Paired-end*
   - **Motivazione**: Il dataset è stato generato con tecnologia **paired-end**, che produce due letture per ciascun frammento. L'impostazione corretta garantisce che Bowtie2 gestisca le coppie in modo coordinato, migliorando la qualità dell'allineamento.
3. **Modalità di analisi (Analysis Mode)**
   - **Impostazione**: *No, just use defaults*
   - **Motivazione**: sono state utilizzate le **impostazioni di default**, adatte alla maggior parte dei dataset RNA-Seq e in grado di fornire un buon compromesso tra **velocità** e **accuratezza** dell'allineamento, senza richiedere ottimizzazioni specifiche.
4. **Scrittura delle letture non allineate in file FASTQ separati**

- **Impostazione**: *No*
- **Motivazione**: non è stato ritenuto necessario salvare le letture non allineate in un file separato, poiché l'analisi si concentra sulle letture che mappano correttamente al genoma.

5. **Scrittura delle letture allineate in file FASTQ separati**
   - **Impostazione**: *No*
   - **Motivazione**: analogamente alla voce precedente, l'output standard di Bowtie2 (in formato SAM/BAM) è sufficiente per le successive fasi di analisi. La generazione di un FASTQ separato non è richiesta.
6. **Utilizzo di preset**
   - **Preset utilizzato**: *Impostazioni di default*
   - **Nota**: Sebbene Bowtie2 offra preset specifici (es. *very-fast*, *very-sensitive*), per questa analisi è stato scelto di mantenere le impostazioni predefinite per assicurare un equilibrio ottimale tra **prestazioni computazionali** e **sensibilità di allineamento**.

Confronto dei tempi di esecuzione: effetto del trimming sull'allineamento con Bowtie2

Per valutare l'impatto del **trimming delle letture** sulle performance computazionali dell'allineamento, è stato eseguito un confronto diretto tra i tempi di esecuzione di **Bowtie2** applicato sia al dataset **grezzo** (non trimmato) che al dataset **preprocessato** tramite trimming. I dati riportati derivano direttamente dalle **Job Metrics** fornite dalla piattaforma Galaxy.

Dataset preprocessato con trimming

Parametro	**Valore**
CPU usage time	48 minuti
CPU user time	44 minuti
CPU system time	3 minuti
Max memory usage	4.7 GB
Numero core allocati	8
Memoria allocata	20.480 MB (20 GB)
Job Runtime (Wall Clock)	**7 minuti**
Data/ora inizio lavoro	01/04/2025 – 16:05:24
Data/ora fine lavoro	01/04/2025 – 16:12:29

Dataset grezzo (senza trimming)

Parametro	Valore
CPU usage time	1 ora e 44 minuti
CPU user time	1 ora e 36 minuti
CPU system time	8 minuti
Max memory usage	4.4 GB
Numero core allocati	8
Memoria allocata	20.480 MB (20 GB)
Job Runtime (Wall Clock)	**15 minuti**
Data/ora inizio lavoro	29/03/2025 – 20:45:41
Data/ora fine lavoro	29/03/2025 – 21:00:43

Analisi del confronto

L'**applicazione del trimming** ha comportato una **riduzione significativa del tempo di esecuzione** complessivo dell'allineamento. In particolare, il **Wall Clock Time** si è dimezzato, passando da **15 minuti** per il dataset grezzo a **7 minuti** per il dataset trimmato.

Questa differenza è attribuibile alla **riduzione della lunghezza media delle letture** e alla **rimozione di sequenze di bassa qualità o contenenti adattatori**, che possono ostacolare il processo di allineamento e richiedere risorse computazionali maggiori. Inoltre, il trimming ha probabilmente ridotto anche il numero complessivo di letture da processare, eliminando quelle troppo corte o contenenti ambiguità (es. "N").

Dal punto di vista della gestione delle risorse, il trimming non ha comportato un aumento significativo della **memoria utilizzata**, ma ha permesso di ottimizzare il carico computazionale, con una maggiore efficienza di esecuzione.

Samtools flagstat

Per valutare l'efficienza dell'allineamento delle letture al genoma di riferimento, è stato utilizzato il tool `Samtools flagstat`, disponibile su Galaxy. Questo strumento fornisce un riepilogo statistico dettagliato del file BAM risultante dall'allineamento, permettendo di analizzare rapidamente la qualità e l'efficacia del processo.

Nel nostro caso, sono state analizzate **14.481.266 letture** totali. Di queste, circa **il 96,16%** risultano mappate correttamente sul genoma, un dato che indica un **alto livello di qualità dell'allineamento**. Inoltre, il **91,65% dei reads** è mappato in modo **corretto come coppia** (*properly paired*), confermando la buona riuscita dell'allineamento *paired-end*.

Non sono presenti letture duplicate, secondarie o supplementari, e anche questo rappresenta un **ottimo segnale di pulizia e accuratezza del dataset**. Solo una piccola frazione dei reads (1,40%) risulta come *singleton*, ovvero mappata singolarmente senza il proprio *mate*.

Complessivamente, questi dati confermano che l'allineamento con `Bowtie2` ha prodotto risultati di alta qualità, garantendo l'affidabilità delle analisi successive sull'espressione genica.

```
14481266 + 0 in total (QC-passed reads + QC-failed reads)
14481266 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
13925080 + 0 mapped (96.16% : N/A)
13925080 + 0 primary mapped (96.16% : N/A)
14481266 + 0 paired in sequencing
7240633 + 0 read1
7240633 + 0 read2
13272302 + 0 properly paired (91.65% : N/A)
13722256 + 0 with itself and mate mapped
202824 + 0 singletons (1.40% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Andando ad utilizzare lo stesso tool sul dataset allineato senza trimming abbiamo notato che non ci sono differenze. Ciò ci fa trarre come conclusione che:
il trimming ci ha aiutato ad avere un alleggerimento del file e soprattutto ad avere una pulizia del file, quindi ci permette di avere tempi di esecuzione molto brevi. Ma per quanto riguarda il risultato atteso è lo stesso poichè probabilmente Cutadapt ha tagliato solo le code di bassa qualità ma senza eliminare le letture intere. Questo migliora la qualità delle basi mappate ma non cambia il numero di letture allineate.

FeatureCounts

Introduzione

FeatureCounts è uno dei programmi più utilizzati e affidabili per **quantificare l'espressione genica** nei dati di sequenziamento RNA-Seq. Fa parte del pacchetto **Subread**, sviluppato dal Walter and Eliza Hall Institute of Medical Research, ed è ampiamente integrato nelle **pipeline bioinformatiche standard** (come Galaxy, R/Bioconductor, Snakemake).

Il suo scopo principale è **contare il numero di letture mappate su ciascuna feature genomica** (come geni, esoni o trascritti), a partire dai file BAM generati dall'allineamento.

Il tool serve per trasformare i dati di sequenziamento allineati (in formato BAM) in **dati numerici strutturati**, che rappresentano il livello di espressione di ogni gene. Questi dati sono essenziali per:

- **Analisi dell'espressione differenziale**
- **Clustering** e **visualizzazione**
- Studi di **regolazione genica**, **effetto di trattamenti**, **profilazione tissutale**, ecc.

Funzionamento

Input necessari:

1. **File BAM/SAM**: contenenti le letture mappate (es. output di Bowtie2).
2. **File di annotazione**: in formato **GTF** o **GFF**, che descrive la posizione dei geni ed esoni sul genoma.
3. **Opzioni configurabili**: ad es. tipo di feature da contare (esone, gene, CDS), modalità paired-end, strandedness, soglie di qualità, ecc.

Output:

Un file tabellare CSVcontenente:

- **Riga = gene o feature**
- **Colonna = conteggio delle letture mappate su quella feature**

Problema di compatibilità tra file GTF e BAM

Descrizione del problema

Durante l'utilizzo del tool **featureCounts** per la quantificazione dell'espressione genica, è emersa una **discordanza tra i nomi dei contig** presenti nel file di annotazione **GTF** e quelli riportati nel file di allineamento **BAM** generato da **Bowtie2**.

- Nel file **BAM**, derivato dall'allineamento contro il genoma FASTA scaricato da Ensembl Bacteria, il nome del contig è indicato come:
  → **Chromosome**
- Nel file **GTF** originale fornito da Ensembl per lo stesso genoma, invece, il nome del contig è:
  → **U00096.3**

Questa **incongruenza nei nomi** impedisce a featureCounts di associare correttamente le letture allineate (dal BAM) con le feature genomiche (dal GTF), bloccando di fatto il processo di quantificazione.

Soluzione adottata

Per risolvere il problema, è stata effettuata una **modifica manuale** del file GTF, sostituendo tutti i riferimenti a U00096.3 con Chromosome, in modo da uniformare i nomi dei contig tra i due file.

Passaggi eseguiti:

1. **Download del file GTF** originale da Ensembl.
2. **Apertura del file** con un editor di testo (es. Visual Studio Code).
3. **Sostituzione globale (Find & Replace)**:
   - **Trova**: U00096.3
   - **Sostituisci con**: Chromosome
4. **Salvataggio** del file modificato.
5. **Ricaricamento** del GTF aggiornato su Galaxy.
6. **Riesecuzione di featureCounts** con il nuovo file di annotazione compatibile.

Dopo la correzione, il tool **featureCounts** ha potuto elaborare correttamente i dati, associando le letture alle rispettive feature geniche. Questa fase ha evidenziato l'importanza di **verificare la corrispondenza tra file BAM e GTF**, soprattutto quando si utilizzano genomi personalizzati o scaricati manualmente.

Configurazione del tool featureCounts

1. Minimum Mapping Quality per read

- **Valore impostato**: *0* (valore predefinito)
- **Motivazione**: Nessuna soglia di qualità di mapping è stata applicata, in modo da includere **tutte le letture mappate**, anche quelle con qualità bassa. Questo consente di ottenere una panoramica completa dell'espressione genica.

2. Filter Split Alignments

- **Valore impostato**: *No filtering*
- **Motivazione**: Sono stati considerati **sia gli allineamenti split che non split**, senza filtri specifici. Questo approccio è coerente con la natura dell'RNA-Seq, dove possono verificarsi allineamenti su più esoni.

3. **Only Count Primary Alignments**
   - **Valore impostato**: *No*
   - **Motivazione**: Sono stati inclusi anche **gli allineamenti secondari**, per non escludere letture che si allineano in più regioni. Questo è utile per analizzare geni ripetitivi o regioni omologhe.

4. **Minimum Fraction (of read) Overlapping a Feature**
   - **Valore impostato**: *0*
   - **Motivazione**: È stata richiesta solo una **minima sovrapposizione** tra la lettura e la feature affinché la lettura venisse conteggiata. Questo garantisce massima sensibilità.

5. **Minimum Fraction (of Feature) Overlapping a Read**
   - **Valore impostato**: *0*
   - **Motivazione**: Anche in questo caso, è stato sufficiente che **una piccola porzione della feature** fosse coperta dalla lettura affinché il conteggio fosse effettuato.

6. **Read 5' / 3' Extension**
   - **Valore impostato**: *0* per entrambe
   - **Motivazione**: Non è stata applicata alcuna estensione delle letture. Le letture sono state analizzate esattamente come riportate nei file BAM, senza espansione artificiale dei confini.

7. **Reduce Read to Single Position**
   - **Valore impostato**: *Leave the read as it is*
   - **Motivazione**: Le letture non sono state ridotte a una singola posizione (es. 5' o 3'), al fine di mantenere l'integrità delle informazioni di mappatura.

8. **Long Reads**
   - **Valore impostato**: *No*
   - **Motivazione**: Non essendo utilizzate tecnologie long-read (es. Nanopore, PacBio), questa opzione è stata lasciata disattivata.

9. **Count Reads by Read Group**
   - **Valore impostato**: *No*

- **Motivazione**: Le letture non sono state separate per gruppi (*read group*), poiché il dataset analizzato non prevedeva suddivisioni sperimentali multiple.

10. Ignore Reads Marked as Duplicate

- **Valore impostato**: *No*
- **Motivazione**: Le letture duplicate non sono state escluse, dato che nel contesto RNA-Seq la duplicazione può riflettere reali eventi biologici (es. alta espressione genica) e non necessariamente artefatti.

11. Allow Reads to Map to Multiple Features

- **Valore impostato**: *Disabled*
- **Motivazione**: Le letture **sono state assegnate a una sola feature** per evitare ambiguità nei conteggi. Questo migliora la specificità nella quantificazione.

12. Exon-exon Junctions

- **Valore impostato**: *Do not count*
- **Motivazione**: Le **giunzioni esone-esone** non sono state conteggiate separatamente. Lo scopo principale dell'analisi era la **quantificazione per gene**, piuttosto che la ricostruzione dell'intero trascrittoma o delle isoforme.

Analisi dei geni maggiormente espressi

Obiettivo dell'analisi

Dopo la quantificazione dell'espressione genica mediante **featureCounts**, è stata condotta un'analisi esplorativa per identificare i **20 geni con il maggior numero di letture mappate**, ovvero quelli con **livelli di espressione più elevati** nel campione analizzato.

Questa analisi fornisce una prima panoramica dell'attività trascrizionale della cellula e consente di evidenziare **quali geni sono funzionalmente più rilevanti** nella condizione sperimentale esaminata.

Rilevanza biologica

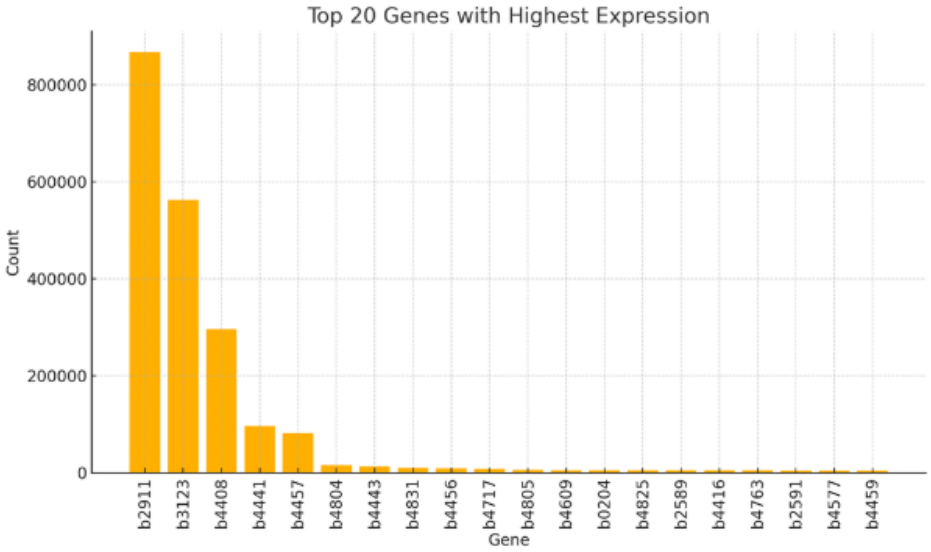
L'identificazione dei geni maggiormente espressi è utile per:

- Comprendere **quali processi cellulari sono attivi** nel momento in cui è stato effettuato il sequenziamento.
- Individuare **geni candidati** potenzialmente coinvolti nella **risposta alla presenza dell'antibiotico Magainin I**, elemento centrale del progetto.

- Porre le basi per un'**interpretazione funzionale** dei dati, attraverso analisi di annotazione genica, pathway o arricchimento GO (Gene Ontology).

Risultati

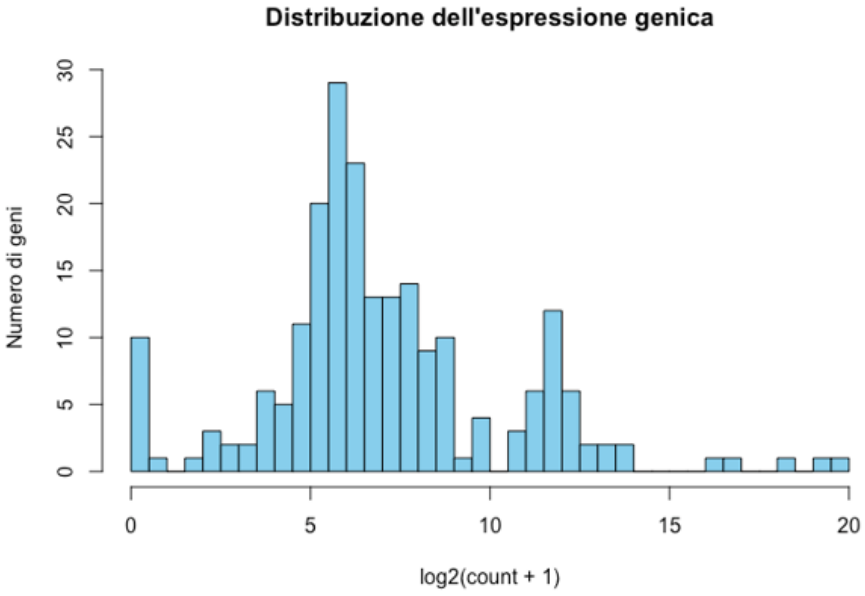
Tramite R siamo andati a generare grafici per ricavare i geni più espressi nel dataset:



Distribuzione dell'espressione genica

Obiettivo dell'analisi

Una fase preliminare fondamentale nell'analisi di dati RNA-Seq consiste nell'osservare la **distribuzione dell'espressione genica** complessiva all'interno del campione. Questo passaggio consente di valutare la **variabilità trascrizionale** e di confermare la presenza di un pattern tipico dei dati RNA-Seq, ovvero una forte asimmetria nei livelli di espressione tra i diversi geni.



Descrizione del grafico

L'istogramma riportato di seguito mostra la distribuzione dell'espressione genica trasformata tramite $\mathbf{log_2(count + 1)}$:

- **Asse X** – $\log_2(\mathbf{count} + 1)$: rappresenta il livello di espressione genica trasformato per mezzo della funzione logaritmica in base 2. Questa trasformazione è comunemente utilizzata per **normalizzare i dati**, riducendo la distorsione provocata da valori estremamente elevati. L'aggiunta di 1 consente di evitare errori nel calcolo logaritmico in presenza di zeri.
- **Asse Y – Numero di geni**: indica quanti geni rientrano in ciascun intervallo di espressione, ovvero **la frequenza**dei valori di espressione trasformati.

Interpretazione biologica

L'andamento della distribuzione mostra una tendenza tipica dei dati RNA-Seq:

- **La maggior parte dei geni presenta un'espressione bassa o moderata**, con un picco attorno a valori $\log_2$ compresi tra 5 e 6.
- **Solo una frazione limitata di geni è altamente espressa**, evidenziata dalle code più esterne a destra della distribuzione.

Questo pattern riflette il comportamento fisiologico della cellula, che tende ad **attivare selettivamente solo una parte del genoma trascrivibile**, mentre la maggior parte dei geni rimane scarsamente espressa o inattiva.

Utilità del grafico nella pipeline RNA-Seq

- **Verifica qualitativa dei dati**: la distribuzione conferma che l'espressione genica non è uniforme, ma piuttosto concentrata su un numero limitato di geni ad alta trascrizione.
- **Punto di partenza per l'analisi differenziale**: comprendere la struttura di base dei dati di espressione consente di impostare correttamente i parametri dei metodi statistici successivi (es. DESeq2 o EdgeR).
- **Supporto alla normalizzazione e al filtraggio**: i dati che mostrano una distribuzione coerente come questa sono candidati ideali per procedere con trasformazioni ulteriori e test di significatività.

Analisi funzionale del gene B2911 (RNA 6S)

Identità e funzione del gene

Il gene **B2911** è risultato il **più espresso nel dataset**, con un conteggio di **867.693** letture. Esso codifica per l'**RNA 6S**, un **piccolo RNA regolatore** noto per il suo ruolo nella regolazione trascrizionale in *Escherichia coli*.

L'RNA 6S interagisce con la **RNA polimerasi associata al fattore sigma70**, bloccandone temporaneamente l'attività. Questo meccanismo è particolarmente attivo nella **fase stazionaria** del ciclo di crescita, in cui la cellula rallenta il metabolismo per adattarsi a condizioni ambientali sfavorevoli, come carenza di nutrienti o stress.

Interpretazione dell'elevata espressione di B2911

L'elevata espressione del gene B2911 nel nostro campione suggerisce che i batteri si trovino in una **fase di stress o crescita stazionaria**. Questa osservazione è coerente con il contesto sperimentale, in cui i batteri sono esposti a **Magainin I**, un peptide antimicrobico in grado di compromettere l'integrità della membrana cellulare.

B2911 e la resistenza a Magainin I

Sebbene **l'RNA 6S non sia direttamente coinvolto nei meccanismi di resistenza agli antibiotici**, la sua sovraespressione può essere **funzionalmente collegata** a una risposta generale allo stress indotto dall'antimicrobico. Alcuni aspetti rilevanti:

- **Adattamento trascrizionale**: L'RNA 6S riduce la trascrizione globale modulando l'attività della RNA polimerasi, promuovendo l'espressione di geni associati alla **sopravvivenza in condizioni avverse**.
- **Effetto indiretto sulla resistenza**: La sua espressione può facilitare l'attivazione di circuiti regolatori più ampi che coinvolgono **geni di stress**

response, alcuni dei quali potrebbero avere un ruolo **protettivo** nei confronti degli effetti citotossici di Magainin I.

- **Rallentamento metabolico e persistenza**: In fase stazionaria, il batterio riduce la propria attività metabolica, **diventando meno vulnerabile** all'azione di alcuni antibiotici. Questo stato può contribuire a una **resistenza funzionale o transitoria**, anche in assenza di meccanismi specifici di resistenza.

Conclusione biologica

Il dato osservato suggerisce che l'**alta espressione del gene B2911** rifletta un adattamento della cellula a uno stato di **stress ambientale**, probabilmente indotto da Magainin I. Ciò supporta l'ipotesi che, oltre ai geni direttamente implicati nella resistenza, anche **piccoli RNA regolatori come l'RNA 6S possano contribuire indirettamente** a una maggiore **tolleranza o persistenza** in ambienti ostili.

Approfondimento bibliografico

Uno studio pubblicato su *Microbiology* conferma che l'esposizione a Magainin I **modifica significativamente il profilo trascrizionale dei batteri**, attivando **vie di risposta allo stress**, **adattamenti metabolici** e **soppressione della motilità cellulare**. Queste modifiche possono concorrere a una maggiore sopravvivenza in presenza del peptide antimicrobico.

Fonte: https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.000725

Analisi funzionale del gene B3123 (RnpB, RNA della RNasi P)

Identità e funzione del gene

Il gene **B3123** è risultato il **secondo più espresso** nel dataset, con **562.936 letture**. Esso codifica per **RnpB**, l'**RNA catalitico della RNasi P**, un enzima fondamentale per la maturazione dei **precursori del tRNA**. In particolare, la RNasi P taglia le sequenze leader dei pre-tRNA, rendendoli funzionalmente attivi per la sintesi proteica.

Oltre ai tRNA, RNasi P può anche agire su **altri RNA regolatori**, e alcuni studi suggeriscono un ruolo nella **regolazione della lisogenia e della lisi nei batteriofagi**, contribuendo così a dinamiche cellulari complesse in condizioni di stress.

Significato dell'elevata espressione di B3123

L'alta espressione di RnpB suggerisce che la cellula si trovi in una **fase di elevata attività biosintetica**, con una forte **domanda di tRNA maturi** per

supportare la sintesi proteica. Questo può indicare che i batteri, nonostante la presenza dell'antimicrobico **Magainin I**, stanno mantenendo un **metabolismo attivo**, forse per **contrastare o riparare** i danni indotti dal trattamento.

B3123 e la resistenza a Magainin I: ipotesi funzionali

Sebbene **RnpB non sia un gene di resistenza canonico**, la sua sovraespressione può essere interpretata come parte di una **risposta cellulare adattativa**. Alcune possibili correlazioni:

- **Adattamento allo stress antimicrobico**: Magainin I danneggia le membrane cellulari, generando stress e possibili danni metabolici. L'aumento della sintesi di tRNA e proteine potrebbe riflettere **un tentativo della cellula di riparare le strutture danneggiate** o di produrre **proteine difensive**.
- **Ruolo regolatorio**: RNasi P, attraverso RnpB, può influenzare anche RNA non codificanti implicati nella **regolazione di percorsi di stress** o nella modulazione di **risposte fagiche**, che potrebbero impattare sulla **sopravvivenza del batterio in ambienti ostili**.
- **Contrasto con RNA 6S (B2911)**: mentre l'elevata espressione di B2911 suggerisce una **fase di quiescenza o adattamento stazionario**, la sovraespressione di B3123 indica una condizione di **attività cellulare sostenuta**, suggerendo **strategie differenti** di risposta tra ceppi batterici.

Conclusione biologica

Il gene B3123, codificante per RnpB, mostra un'elevata espressione nei nostri dati, coerente con una forte richiesta di maturazione dei tRNA. Questo supporta l'ipotesi di una **risposta metabolica attiva** da parte dei batteri, potenzialmente diretta a fronteggiare lo stress indotto da **Magainin I**. Sebbene RnpB non sia direttamente implicato nella resistenza antimicrobica, la sua sovraespressione può indicare **una strategia di sopravvivenza basata sul mantenimento della sintesi proteica**, e quindi una **resistenza funzionale** o **adattamento compensativo**.

Approfondimento bibliografico

Uno studio recente pubblicato sul *Journal of Biological Chemistry* conferma che la RNasi P svolge **funzioni chiave oltre la semplice maturazione dei tRNA**, contribuendo a regolare **la stabilità e l'elaborazione di altri RNA regolatori**, spesso coinvolti nella **risposta allo stress** e nell'equilibrio lisogenia/lisi nei batteri.

Fonte: https://doi.org/10.1093/jambio/lxae116

Analisi funzionale del gene B4408 (csrB)

Identità e funzione del gene

Il gene **B4408**, noto anche come **csrB**, codifica per un **RNA regolatore non codificante** che svolge un ruolo centrale nel controllo post-trascrizionale del metabolismo in *Escherichia coli*. La sua funzione principale è quella di **sequestrare la proteina CsrA**, un regolatore che modula la **traduzione, la stabilità e la degradazione di diversi mRNA** coinvolti in numerosi processi biologici, tra cui:

- **Metabolismo del carbonio** (es. regolazione della sintesi del glicogeno),
- **Formazione di biofilm**,
- **Motilità batterica**,
- **Adattamento allo stress ambientale**,
- **Risposta SOS**.

La molecola csrB agisce come **decoy RNA**: interagisce con molte copie di CsrA, inibendone l'attività e alterando la regolazione trascrizionale e traduttiva di numerosi geni bersaglio.

Significato dell'elevata espressione di csrB nel nostro esperimento

Nel dataset ottenuto dall'esperimento RNA-Seq, **csrB risulta fortemente espresso**, indicando un'intensa attività regolatoria a livello post-trascrizionale. Questo suggerisce che *E. coli* stia attivando **meccanismi di adattamento metabolico e risposta allo stress**, coerenti con l'esposizione a un agente antimicrobico come **Magainin I**.

csrB e la resistenza a Magainin I

Sebbene **csrB non sia un gene di resistenza diretta**, la sua sovraespressione può avere **effetti significativi sull'assetto cellulare**, contribuendo alla sopravvivenza in ambienti ostili. Alcuni possibili collegamenti:

- **Conservazione delle risorse energetiche**: inibendo CsrA, csrB favorisce l'accumulo di glicogeno e una riduzione della degradazione di mRNA, contribuendo alla **conservazione dell'energia**.
- **Adattamento allo stress**: l'attività di CsrA influenza anche la risposta SOS e altri pathway di stress. La sua inibizione da parte di csrB può favorire una **modulazione più favorevole della risposta cellulare** a danni indotti da Magainin I.
- **Ristrutturazione trascrizionale**: l'asse csrB–CsrA agisce come un **hub regolatorio**, coordinando metabolismo, motilità e biofilm, tutti processi che possono essere **alterati per aumentare la tolleranza agli antimicrobici**.

Conclusione biologica

Il gene csrB (B4408), altamente espresso nel nostro dataset, suggerisce che *E. coli* stia attivando meccanismi regolatori per fronteggiare lo stress indotto da Magainin I. La sua funzione di RNA antagonista della proteina CsrA permette al batterio di **riprogrammare il metabolismo**, ridurre l'attività trascrizionale non essenziale e **ottimizzare l'adattamento cellulare**, favorendo così una **sopravvivenza più efficace** in condizioni sfavorevoli.

Analisi funzionale del gene B4441 (GlmY)

Identità e funzione del gene

Il gene **B4441** codifica per **GlmY**, un **piccolo RNA regolatore** che gioca un ruolo chiave nella regolazione della sintesi dell'enzima **GlmS** (glucosamina-6-fosfato sintetasi), fondamentale per la biosintesi della parete cellulare batterica.

GlmY non agisce direttamente sull'mRNA di glmS, ma esercita il suo effetto regolando **indirettamente l'RNA GlmZ**, che stabilizza l'mRNA di **glmS**. In particolare, **GlmY sequestra la proteina RapZ**, impedendole di degradare GlmZ, e garantendo così la stabilizzazione dell'mRNA di **glmS**.

Questo sistema di regolazione **a triplo livello** (GlmY–GlmZ–glmS) consente al batterio di adattare con precisione la produzione di GlmS **in risposta alle variazioni ambientali**, come ad esempio un abbassamento nei livelli intracellulari di glucosamina-6-fosfato.

Espressione di GlmY nel nostro esperimento

Nel nostro dataset RNA-Seq, **GlmY risulta fortemente espresso**. Questo suggerisce che *Escherichia coli* **stia attivamente cercando di aumentare la produzione dell'enzima GlmS**, probabilmente per rafforzare la **sintesi della parete cellulare** in risposta a **condizioni di stress ambientale**, come l'esposizione a **Magainin I**.

GlmY e la resistenza a Magainin I

L'alta espressione di GlmY può essere interpretata come parte di una **risposta adattativa al danno della membrana** indotto da Magainin I. Alcune possibili implicazioni funzionali:

- **Rafforzamento della parete cellulare**: GlmS è coinvolto nella produzione di precursori fondamentali per la sintesi del **peptidoglicano**, elemento strutturale della parete. Una sua maggiore produzione può **aumentare la resistenza meccanica** della cellula contro l'azione litica del peptide antimicrobico.
- **Regolazione fine tramite RNA non codificanti**: la via GlmY–GlmZ dimostra come *E. coli* possa impiegare **meccanismi RNA-dipendenti** per

adattarsi rapidamente a stress esterni, regolando in modo dinamico i livelli di enzimi essenziali.

- **Interazione con percorsi di resistenza noti**: Studi recenti hanno evidenziato che i ceppi resistenti a Magainin I presentano **sovraregolazione di geni coinvolti nella biosintesi della parete cellulare**, nella **formazione di biofilm** e nell'**adattamento metabolico**. Il ruolo di GlmY potrebbe quindi essere **funzionalmente integrato** in questi processi di sopravvivenza.

Conclusione biologica

Il gene B4441 (GlmY) è fortemente espresso nel nostro esperimento, suggerendo una risposta attiva alla presenza di Magainin I. La sua funzione di regolatore della sintesi di **GlmS**, essenziale per la parete cellulare, indica che *E. coli*potrebbe star cercando di **rinforzare la propria struttura per contrastare i danni indotti dal peptide antimicrobico**.

Questo supporta l'ipotesi che **gli RNA regolatori non codificanti**, come GlmY, **contribuiscano indirettamente alla tolleranza o resistenza** agli antimicrobici, agendo attraverso percorsi metabolici chiave e meccanismi strutturali di difesa.

Approfondimento bibliografico

Uno studio pubblicato su *Scientific Reports* ha mostrato che ceppi resistenti a Magainin I presentano **modificazioni dell'espressione genica associate all'omeostasi cellulare e alla formazione di biofilm**, percorsi in cui RNA regolatori come **GlmY e CsrB** sono fortemente coinvolti.

Fonte: https://www.nature.com/articles/s41598-017-04181-y

Analisi funzionale del gene B4457 (CsrC)

Identità e funzione del gene

Il gene **B4457** codifica per **CsrC**, un **piccolo RNA regolatore non codificante** che, analogamente a **CsrB**, svolge un ruolo centrale nella regolazione dell'attività della proteina **CsrA**. CsrA è un **regolatore post-trascrizionale globale**coinvolto nel controllo del **metabolismo del carbonio**, della **motilità**, della **formazione di biofilm** e di altri processi legati alla fisiologia batterica.

CsrC agisce **sequestrando CsrA**, impedendole di legarsi a specifici mRNA bersaglio. In tal modo, regola **la stabilità e la traduzione degli mRNA**, favorendo l'espressione di geni implicati nella risposta a condizioni ambientali sfavorevoli.

La **degradazione di CsrC** è mediata dalla proteina **CsrD** ed è influenzata dalla disponibilità della fonte di carbonio: in presenza di glucosio, **CsrC viene**

degradato più rapidamente, mentre in condizioni di crescita limitata o nella **fase stazionaria**, la sua espressione aumenta.

Espressione di CsrC nel nostro esperimento

Nel nostro dataset RNA-Seq, **CsrC risulta altamente espresso**, suggerendo che i batteri si trovino in uno stato di **crescita povera o di stress ambientale**, condizioni che promuovono l'attività di questo RNA regolatore. La cellula sembra dunque impegnata in un processo di **adattamento metabolico**, probabilmente in risposta alla **presenza dell'antimicrobico Magainin I**.

CsrC e la resistenza a Magainin I: implicazioni funzionali

L'elevata espressione di CsrC può contribuire a rafforzare la **resilienza batterica**, agendo su più fronti:

- **Adattamento metabolico**: regolando l'attività di CsrA, CsrC modula il metabolismo del carbonio e ottimizza **l'uso delle risorse** in condizioni ostili.
- **Formazione di biofilm**: CsrA inibisce la formazione di biofilm; la sua inibizione tramite CsrC può favorire la **formazione di strutture biofilmiche**, che offrono **protezione fisica** e aumentano la **tolleranza agli antimicrobici**.
- **Controllo della motilità**: l'interferenza con CsrA modifica l'espressione di geni legati alla motilità (es. flagelli), influenzando la **strategia comportamentale** della cellula di fronte allo stress.

Conclusione biologica

Il gene B4457 (CsrC) è fortemente espresso nel nostro esperimento, evidenziando una risposta adattativa da parte di *E. coli* a condizioni di stress ambientale. Attraverso la modulazione di **CsrA**, CsrC contribuisce a **riprogrammare il metabolismo cellulare**, a regolare la **formazione del biofilm** e a gestire risposte chiave per la **sopravvivenza in presenza di Magainin I**.

Il ruolo combinato di **CsrC e CsrB** evidenzia l'importanza degli **RNA regolatori non codificanti** nella **tolleranza o resistenza indiretta** agli antimicrobici, nonché nella capacità del batterio di adattarsi dinamicamente all'ambiente.

Conclusioni

L'analisi condotta ha permesso di esplorare, attraverso la piattaforma Galaxy, l'intero workflow di pre-processing, allineamento e quantificazione dell'espressione genica su dati RNA-Seq di *Escherichia coli* esposti al peptide antimicrobico Magainin I. Dopo un'accurata fase di trimming, l'allineamento con Bowtie2 ha

mostrato un'elevata percentuale di letture correttamente mappate, confermando l'efficacia del pre-processing nella qualità dell'output.

Tramite `featureCounts` sono stati identificati i geni maggiormente espressi, tra cui RNA regolatori come **B2911 (RNA 6S)**, **B3123 (RnpB)**, **csrB**, **GlmY** e **CsrC**. L'analisi funzionale ha evidenziato che questi geni, pur non essendo direttamente associati a meccanismi classici di resistenza, partecipano a vie di regolazione dello stress, adattamento metabolico e rafforzamento strutturale della cellula.

Tali risultati suggeriscono che la resistenza a Magainin I non dipenda solo da specifici geni di difesa, ma da un network complesso di risposte trascrizionali che coinvolgono anche RNA non codificanti. Questo studio dimostra l'importanza dell'integrazione tra strumenti bioinformatici e interpretazione funzionale per comprendere meglio i meccanismi molecolari della resistenza antimicrobica.