# Analisi di sequenziamento di SARS-CoV-2: Quality Control, Trimming, Mapping e Visualizzazione

Giovanni Arcangeli

1 febbraio 2025

# 1 Indice

# Indice

# 2 Introduzione

## 2.1 Contesto e Importanza della Bioinformatica

La bioinformatica è una disciplina in rapida espansione che combina l'informatica, la statistica e la biologia molecolare per analizzare e interpretare dati biologici complessi. Tra le sue applicazioni più rilevanti vi è l'analisi di sequenze genomiche, resa possibile dai progressi nelle tecnologie di sequenziamento di nuova generazione (NGS). Queste tecnologie producono enormi quantità di dati, richiedendo lo sviluppo di strumenti e metodi computazionali efficienti per la loro elaborazione e interpretazione.

## 2.2 Obiettivo del Progetto

Il presente progetto si inserisce in questo contesto e ha come obiettivo l'analisi di dati di sequenziamento del virus SARS-CoV-2, l'agente eziologico della COVID-19, una pandemia che ha avuto un impatto globale senza precedenti. In particolare, ci concentreremo sul processo di *mapping* (o allineamento), una fase cruciale dell'analisi di sequenziamento che consiste nell'allineare le brevi sequenze prodotte dal sequenziatore (reads) a un genoma di riferimento. Questo processo permette di identificare la posizione di origine di ciascuna read nel genoma, fornendo informazioni fondamentali per studi di genomica, trascrittomica e metagenomica.

## 2.3 Descrizione del Workflow e del Dataset

Il progetto prevede la replica di un workflow di analisi bioinformatica, basato su tutorial forniti dalla piattaforma Galaxy, un ambiente open-source per l'analisi di dati biologici. Il workflow comprende le fasi di controllo qualità delle reads, il mapping al genoma di riferimento e la visualizzazione dei risultati. Inizialmente, il workflow è stato eseguito utilizzando i dati forniti nei tutorial stessi, per poi essere applicato a un dataset indipendente di sequenze di SARS-CoV-2. Questo dataset, ottenuto dalla piattaforma Zenodo, è costituito da reads paired-end in formato FASTQ, prodotte dall'African Centre of Excellence for Genomics of Infectious Diseases (ACEGID). La scelta di questo dataset è motivata dalla sua disponibilità in un formato adatto all'analisi e dalla possibilità di applicare le conoscenze acquisite a un caso di studio reale e di rilevanza globale.

## 2.4 Competenze Acquisite e Struttura della Relazione

Attraverso questo progetto, si acquisiranno competenze pratiche nell'utilizzo di strumenti bioinformatici per l'analisi di dati di sequenziamento, con particolare attenzione alla fase di mapping. La relazione è strutturata come segue: il Capitolo 2 fornirà un background teorico approfondito sulle tecniche di sequenziamento, il formato FASTQ, il controllo qualità, gli algoritmi di mapping e la visualizzazione dei risultati; il Capitolo 3 descriverà in dettaglio i materiali e i

metodi utilizzati, inclusi i dataset, i software e il workflow di analisi; il Capitolo 4 presenterà i risultati ottenuti; il Capitolo 5 discuterà i risultati e, infine, il Capitolo 6 trarrà le conclusioni e delineerà possibili sviluppi futuri.

# 3 Background Teorico

## 3.1 Sequenziamento del DNA

### 3.1.1 Cos'è il Sequenziamento del DNA e Quali Sono i Suoi Scopi?

Il sequenziamento del DNA è il processo che permette di determinare l'ordine preciso dei nucleotidi (adenina (A), citosina (C), guanina (G) e timina (T)) all'interno di una molecola di DNA. Conoscere la sequenza di un frammento di DNA è fondamentale per una vasta gamma di applicazioni in biologia molecolare, medicina, biotecnologie e altre discipline. Gli scopi principali sono quelli di comprendere l'organizzazione e la funzione dei geni, studiare l'evoluzione degli organismi, identificare le basi molecolari delle malattie o le mutazioni genetiche.

### 3.1.2 Illumina: Piattaforma di Sequenziamento di Nuova Generazione

Illumina è una delle aziende leader nel settore del sequenziamento di nuova generazione (NGS). La tecnologia di sequenziamento Illumina si basa sul metodo "sequenziamento per sintesi" (sequencing by synthesis, SBS).

**Meccanismo di funzionamento (in breve):** Il DNA da sequenziare viene frammentato e legato ad adattatori. I frammenti vengono poi attaccati a una superficie solida (flow cell) e amplificati per formare cluster. Successivamente, nucleotidi marcati con fluorofori diversi vengono aggiunti uno alla volta. Dopo ogni aggiunta, un'immagine cattura l'emissione fluorescente, permettendo di identificare la base aggiunta in ciascun cluster.

**Tasso di errore**: Generalmente inferiore allo 0.1%.

**Applicazioni tipiche**: Sequenziamento di genomi interi, trascrittomica, studi di associazione, identificazione di varianti.

### 3.1.3 Sequenziamento Paired-End

Il sequenziamento paired-end è una tecnica di sequenziamento in cui entrambe le estremità di un frammento di DNA vengono sequenziate. Si ottengono quindi due reads per ogni frammento, una da ciascuna estremità, separate da una distanza nota.

**Vantaggi del sequenziamento paired-end rispetto al single-end:**

- **Migliore allineamento**: La conoscenza della distanza tra le due reads e del loro orientamento aiuta a risolvere ambiguità nell'allineamento, specialmente in regioni ripetitive del genoma.
- **Rilevamento di riarrangiamenti genomici:** Le reads paired-end possono identificare inserzioni, delezioni, inversioni e traslocazioni, confrontando la distanza e l'orientamento osservati con quelli attesi in base alla dimensione dell'inserto.

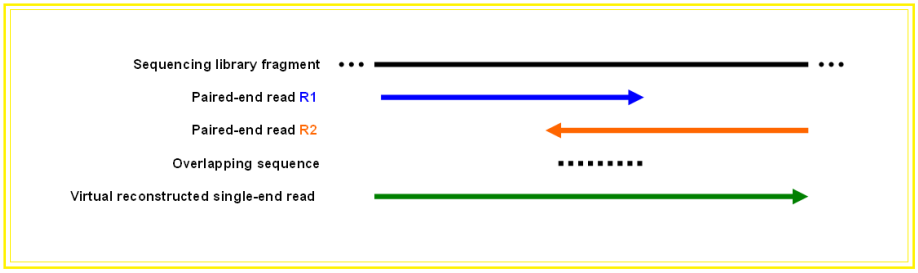


Figura 1: Schema del sequenziamento paired-end

- **Migliore risoluzione in studi di RNA-Seq:** Permette una migliore quantificazione di isoforme di splicing alternativo.

## 3.2 Formato FASTQ

### 3.2.1 Cos'è un file FASTQ e a cosa serve?

Un file FASTQ è un formato di file di testo utilizzato per memorizzare sia una sequenza biologica (di solito una sequenza nucleotidica) sia i suoi rispettivi punteggi di qualità. È diventato lo standard de facto per la memorizzazione dell'output dei sequenziatori di nuova generazione.

### 3.2.2 Struttura di un file FASTQ

Un file FASTQ è composto da record, ciascuno dei quali descrive una singola read ed è costituito da quattro righe:

1. **Identificatore della read:** Inizia con il carattere '@', seguito da un identificatore univoco della read e da eventuali informazioni aggiuntive (es. nome dello strumento, numero di run, coordinate sulla flow cell).
2. **Sequenza nucleotidica:** La sequenza della read, rappresentata dalle lettere A, C, G, T e N (per basi non determinate).
3. **Separatore:** Un singolo carattere '+', a volte seguito dallo stesso identificatore della riga.
4. **Punteggi di qualità**: Una stringa di caratteri che codificano i punteggi di qualità per ciascuna base della sequenza. Ogni carattere rappresenta la qualità della base corrispondente.

Di seguito un esempio di stringa FASTQ:


Figura 2: Esempio di sequenza FASTQ

#### 3.2.3 Phred Quality Score

Il Phred quality score (Q) è una misura della qualità di identificazione delle basi generate dal sequenziamento del DNA. È definito come:

$$Q = -10 \log_{10} P \tag{1}$$

dove P è la probabilità che la base chiamata sia errata.

**Codifica dei punteggi di qualità:** Per risparmiare spazio, i punteggi di qualità sono codificati utilizzando un singolo carattere ASCII. La codifica più comune è la Phred+33, in cui a ciascun punteggio Q viene sommato 33 e il risultato viene convertito nel carattere ASCII corrispondente.

### 3.3 Quality Control

#### 3.3.1 Importanza del Controllo Qualità

Il controllo qualità (QC) è una fase essenziale nell'analisi dei dati di sequenziamento. Permette di valutare la qualità delle reads prodotte dal sequenziatore e di identificare potenziali problemi che potrebbero compromettere le analisi a valle. Dati di bassa qualità possono portare a risultati errati o fuorvianti, come errori nell'assemblaggio del genoma, identificazione errata di varianti, o stime inaccurate dell'espressione genica.

**Problemi derivanti da dati di bassa qualità:**

- **Errori di mapping:** Reads con molti errori possono non allinearsi correttamente al genoma di riferimento o allinearsi in posizioni errate.
- **Identificazione errata di varianti:** Errori di sequenziamento possono essere scambiati per vere varianti genetiche (falsi positivi).
- **Assemblaggio frammentato:** Dati di bassa qualità possono portare a un assemblaggio del genoma più frammentato e meno accurato.
- **Stime inaccurate dell'espressione genica:** Errori nelle reads possono portare a una quantificazione errata dei trascritti in esperimenti di RNA-Seq.

### 3.3.2 FASTQC

FASTQC è un software ampiamente utilizzato per il controllo qualità di dati di sequenziamento. Fornisce una serie di analisi che permettono di valutare la qualità delle reads e di identificare potenziali problemi. FASTQC genera un report in formato HTML che include grafici e tabelle riassuntive.

### 3.3.3 Interpretazione dei Grafici di FASTQC

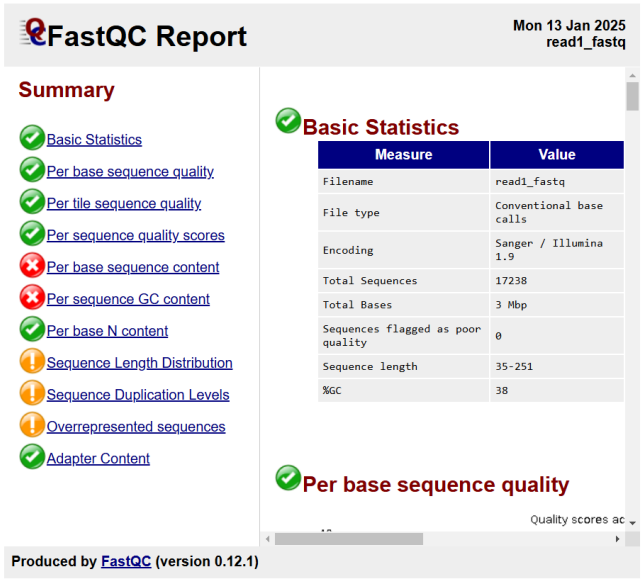


Figura 3: Report di FASTQC

- **Per Base Sequence Quality:** Mostra la distribuzione dei punteggi di qualità (Phred score) per ciascuna posizione lungo le reads. Idealmente, la maggior parte delle reads dovrebbe avere punteggi di qualità elevati (sopra Q30) lungo tutta la loro lunghezza. Un calo di qualità verso la fine delle reads è comune, ma un calo drastico o una qualità generalmente bassa possono indicare problemi durante il sequenziamento.

- **Quality Per Tile:** Mostra la qualità media per ciascuna "tile" (una regione della flow cell) per ciclo di sequenziamento. Questo grafico può aiutare a identificare problemi specifici di una regione della flow cell. Deviazioni significative dalla qualità media (rappresentate da colori più caldi come giallo o rosso) possono indicare problemi tecnici durante il sequenziamento in quella specifica area.

- **Per Base Sequence Content:** Mostra la proporzione di ciascuna base (A, C, G, T) per ciascuna posizione lungo le reads. In un sequenziamento casuale, ci si aspetta che le quattro basi siano presenti in proporzioni simili. Deviazioni significative da questa aspettativa, soprattutto all'inizio delle reads, possono indicare la presenza di sequenze sovrarappresentate (es. adattatori) o bias di sequenziamento.

- **Per Sequence GC Content:** Mostra la distribuzione del contenuto di GC (percentuale di basi G e C) tra tutte le reads. La forma di questa distribuzione dovrebbe generalmente corrispondere a una distribuzione normale e la sua posizione dovrebbe riflettere il contenuto GC atteso per l'organismo sequenziato. Deviazioni significative possono indicare contaminazioni o bias di sequenziamento.

- **Sequence Duplication Levels:** Mostra la percentuale di reads con diversi livelli di duplicazione. Un alto livello di duplicazione può indicare problemi di PCR durante la preparazione della libreria o una complessità insufficiente del campione. Tuttavia, in alcune applicazioni (es. RNA-Seq, ChIP-Seq) un certo livello di duplicazione è atteso.

- **Overrepresented Sequences:** Riporta le sequenze che appaiono più frequentemente del previsto nelle reads. La presenza di sequenze sovrarappresentate può indicare contaminazione da adattatori, primer o altre sequenze non target. FASTQC confronta le sequenze sovrarappresentate con un database di contaminanti comuni per aiutare nell'identificazione.

#### 3.3.4 Trimming

Il trimming è un processo di rimozione di porzioni di reads di bassa qualità o di sequenze indesiderate (es. adattatori). Viene effettuato per migliorare la qualità dei dati e ridurre il rischio di errori nelle analisi a valle.

### 3.3.5 Strategie di trimming

- **Trimming della qualità:** Rimozione delle estremità delle reads dove la qualità scende al di sotto di una certa soglia (es. Q20). Può essere effettuato utilizzando una finestra scorrevole (sliding window) che rimuove le basi se la qualità media all'interno della finestra è inferiore alla soglia.
- **Rimozione di adattatori:** Rimozione di sequenze di adattatori che possono essere presenti alle estremità delle reads. Queste sequenze non fanno parte del genoma di interesse e devono essere rimosse prima del mapping.
- **Trimming di lunghezza minima:** Rimozione di reads che, dopo il trimming della qualità e/o degli adattatori, risultano più corte di una certa lunghezza minima.

Strumenti comuni per il trimming includono **Trimmomatic**, Cutadapt e FASTX-Toolkit.

## 3.4 Mapping

### 3.4.1 Cos'è il Mapping?

Il mapping (o allineamento) è il processo di allineare le reads ottenute dal sequenziamento a un genoma di riferimento. Lo scopo è quello di identificare la posizione di origine di ciascuna read all'interno del genoma. Il mapping è una fase fondamentale in molte analisi di sequenziamento, tra cui l'identificazione di varianti, l'analisi dell'espressione genica (RNA-Seq), e lo studio di interazioni proteina-DNA (ChIP-Seq).

### 3.4.2 Difficoltà del Mapping

Il mapping delle reads al genoma di riferimento può essere complesso a causa di diversi fattori:

- **Errori di sequenziamento:** Le reads possono contenere errori introdotti durante il processo di sequenziamento.
- **Variazioni genomiche:** Il genoma sequenziato può differire dal genoma di riferimento a causa di varianti genetiche (SNPs, indels, varianti strutturali).
- **Regioni ripetitive:** Molti genomi contengono regioni ripetitive, che rendono difficile determinare l'esatta posizione di origine di una read.
- **Dimensioni del genoma:** I genomi di grandi dimensioni richiedono algoritmi di mapping efficienti in termini di tempo e di utilizzo della memoria.

### 3.4.3 Algoritmi di Mapping: Bowtie2

Bowtie2 è un algoritmo di mapping veloce ed efficiente, ampiamente utilizzato per allineare reads a genomi di riferimento. Si basa su due concetti chiave: il **Burrows-Wheeler Transform (BWT)** e l'**indice FM**.

**Burrows-Wheeler Transform (BWT):** Il BWT è una trasformazione reversibile che riorganizza una stringa di caratteri in modo da raggruppare caratteri simili. Questa riorganizzazione facilita la compressione e la ricerca efficiente all'interno della stringa.

**Indice FM:** L'indice FM è una struttura dati compatta che permette di effettuare ricerche rapide all'interno del testo trasformato con il BWT. Permette di contare in modo efficiente il numero di occorrenze di una determinata sequenza all'interno del testo e di localizzare le loro posizioni.

### 3.4.4 Principio di funzionamento di Bowtie2 (in breve)

1. **Indicizzazione del genoma di riferimento:** Bowtie2 costruisce un indice FM del genoma di riferimento, basato sulla sua trasformazione BWT.

2. **Ricerca delle reads nell'indice:** Per ciascuna read, Bowtie2 cerca nell'indice FM del genoma di riferimento per trovare potenziali allineamenti. Utilizza un algoritmo di ricerca "seed-and-extend", in cui prima cerca brevi sottosequenze della read (seeds) e poi estende gli allineamenti.

3. **Valutazione degli allineamenti:** Bowtie2 valuta la qualità degli allineamenti trovati, assegnando un punteggio in base al numero di mismatch, gap e altre caratteristiche.

4. **Output dei risultati:** Bowtie2 produce un file SAM (o BAM) che contiene gli allineamenti delle reads al genoma di riferimento, insieme ai relativi punteggi e ad altre informazioni.

### 3.4.5 Formato SAM/BAM

Il formato SAM (Sequence Alignment/Map) è un formato di file di testo tab-delimitato che contiene informazioni sull'allineamento delle reads al genoma di riferimento. Il formato BAM è la versione binaria compressa del formato SAM.

### 3.4.6 Struttura di un file SAM:

Un file SAM è costituito da due sezioni principali:

1. **Header:** Righe che iniziano con il carattere '@', contenenti informazioni sul genoma di riferimento, sul processo di allineamento e sui metadati del file.

2. **Alignment section:** Righe che descrivono l'allineamento di ciascuna read al genoma di riferimento. Ogni riga contiene 11 campi obbligatori, seguiti da un numero variabile di campi opzionali.

**Campi obbligatori di una riga di allineamento SAM:**

| Campo | Colonna | Descrizione |
|---|---|---|
| QNAME | 1 | Nome della read (identificatore) |
| FLAG | 2 | Combinazione di flag che descrivono l'allineamento (es. read paired, read mappata, strand, ecc.) |
| RNAME | 3 | Nome del cromosoma di riferimento a cui la read è allineata |
| POS | 4 | Posizione di inizio dell'allineamento sul cromosoma di riferimento (1-based) |
| MAPQ | 5 | Qualità di mapping (Phred-scaled) |
| CIGAR | 6 | Stringa CIGAR che descrive l'allineamento (M=match, I=inserzione, D=delezione, ecc.) |
| RNEXT | 7 | Nome del cromosoma di riferimento della read "mate" (se paired-end) |
| PNEXT | 8 | Posizione di inizio dell'allineamento della read "mate" |
| TLEN | 9 | Lunghezza del frammento (insert size) |
| SEQ | 10 | Sequenza della read |
| QUAL | 11 | Punteggi di qualità delle basi (Phred+33) |

Tabella 1: Formato di un file SAM (Sequence Alignment/Map).

**Interpretazione di alcuni campi importanti:**

- **FLAG:** Un valore numerico che rappresenta la somma di flag binarie. Ad esempio, un flag di 16 (0x10 in esadecimale) indica che la read è mappata sul reverse strand.
- **MAPQ:** Indica la probabilità che la posizione di mapping della read sia errata. Un valore di MAPQ elevato (es. 60) indica un'alta confidenza nell'allineamento.
- **CIGAR:** Descrive l'allineamento della read al genoma di riferimento. Ad esempio, '100M' indica un match di 100 basi, '2I' indica un'inserzione di 2 basi, '5D' indica una delezione di 5 basi.

### 3.4.7 Samtools Stats

Samtools è un insieme di utility per interagire con file in formato SAM e BAM. Il comando 'samtools stats' fornisce statistiche riassuntive sull'allineamento contenuto in un file BAM. Questo include informazioni come il numero totale di reads, il numero di reads mappate, la percentuale di reads mappate, il tasso di errore medio, la distribuzione della lunghezza degli inserti (per dati paired-end), e altre metriche utili per valutare la qualità e le caratteristiche dell'allineamento.

## 3.5 Visualizzazione

### 3.5.1 Importanza della Visualizzazione

La visualizzazione dei dati di mapping è fondamentale per:

- **Validare i risultati del mapping:** Permette di ispezionare visivamente l'allineamento delle reads al genoma di riferimento e di identificare potenziali problemi (es. regioni con bassa copertura, allineamenti errati, errori di sequenziamento).
- **Esplorare i dati:** Consente di navigare nel genoma e di esaminare in dettaglio regioni di interesse, come geni, varianti o siti di interazione proteina-DNA.

### 3.5.2 IGV e JBrowse

**IGV (Integrative Genomics Viewer):** è un visualizzatore genomico desktop scaricabile gratuitamente.

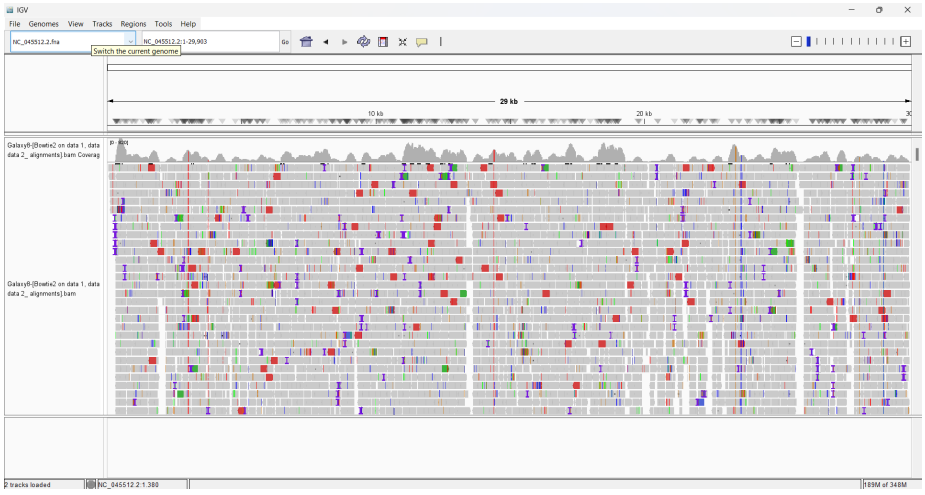


Figura 4: Screenshot di IGV

**Caratteristiche principali di IGV (Figura 4):**

- Visualizzazione di reads allineate a un genoma di riferimento.
- Supporto per diversi formati di file, tra cui BAM, BED, GTF, VCF.
- Visualizzazione di annotazioni genomiche (es. geni, trascritti, varianti).
- Strumenti per la navigazione e lo zoom del genoma.
- Possibilità di caricare dati locali o da server remoti.

**JBrowse:** è un visualizzatore genomico basato sul web.

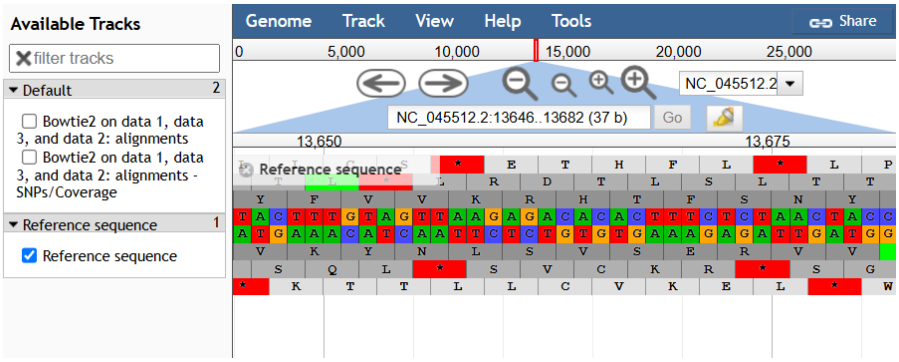


Figura 5: Screenshot di JBrowse

**Caratteristiche principali di JBrowse (Figura 5):**

- Interfaccia web interattiva e user-friendly.
- Visualizzazione fluida e veloce di dati genomici, anche di grandi dimensioni.
- Supporto per diversi formati di file, inclusi BAM, VCF, GFF3.
- Possibilità di personalizzare la visualizzazione e di aggiungere tracce dati custom.
- Facile condivisione dei dati tramite URL.

Sia IGV che JBrowse sono strumenti potenti e versatili per la visualizzazione di dati di sequenziamento e sono ampiamente utilizzati nella comunità scientifica. La scelta tra i due dipende spesso dalle preferenze personali, dalla necessità di un'applicazione desktop o web-based e da specifiche esigenze di visualizzazione.

# 4 Materiali e Metodi

## 4.1 Dati

I dati utilizzati in questo progetto provengono da due fonti: un set di dati fornito come parte di un tutorial sulla piattaforma Galaxy, utilizzati per una prima parte di prova e familiarizzazione con la piattaforma e un secondo dataset di sequenze di SARS-CoV-2 disponibile su Zenodo, che è quello che vedremo nei prossimi paragrafi.

### 4.1.1 Dati del Tutorial di Galaxy

Il primo set di dati è stato ottenuto seguendo i tutorial "Quality Control" e "Mapping" disponibili sulla piattaforma Galaxy Training Network:

**Quality Control:** `https://usegalaxy.eu/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html`

**Mapping:** `https://usegalaxy.eu/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html`

Questi tutorial utilizzano dati di sequenziamento paired-end di **Escherichia coli K-12**, ceppo MG1655. Le reads sono state generate con la piattaforma Illumina e sono disponibili in formato FASTQ. Lo scopo di questi tutorial è di illustrare le fasi di controllo qualità, trimming e mapping di dati di sequenziamento, utilizzando **Mouse (Mus musculus): mm10** come genoma di riferimento.

### 4.1.2 Dati di SARS-CoV-2

Il secondo set di dati, relativo a sequenze di SARS-CoV-2, è stato scaricato da Zenodo `https://zenodo.org/records/5189263`.

I file specifici utilizzati sono:

**CV1537_S1_L001_R1_001.fastq.gz (Read 1)**

**CV1537_S1_L001_R2_001.fastq.gz (Read 2)**

Questi file contengono reads paired-end in formato FASTQ, generate dall'African Centre of Excellence for Genomics of Infectious Diseases (ACEGID).

### 4.1.3 Genoma di Riferimento

Per l'allineamento delle reads di SARS-CoV-2 è stato utilizzato il genoma di riferimento del virus SARS-CoV-2, scaricato da NCBI:

**Accession Number:** NC_045512.2

**URL:** `https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2`

**Nome file:** NC_045512.2.fna

Il file è in formato FASTA e contiene la sequenza completa del genoma di riferimento.

## 4.2 Software e Strumenti

Tutte le analisi sono state eseguite utilizzando l'istanza pubblica di Galaxy (`usegalaxy.eu`). Di seguito sono elencati i principali strumenti utilizzati:

- **FASTQC:** Per il controllo qualità delle reads.
- **Trimmomatic:** Per il trimming delle reads.
- **Bowtie2:** Per l'allineamento delle reads al genoma di riferimento.
- **Samtools:** Per la manipolazione di file SAM/BAM.
- **IGV (Integrative Genomics Viewer):** Per la visualizzazione locale dei dati di mapping.
- **JBrowse:** Per la visualizzazione dei dati di mapping all'interno di Galaxy.

## 4.3 Workflow

Il workflow di analisi, sia per i dati del tutorial che per i dati di SARS-CoV-2, è stato eseguito come segue:

1. **Importazione dei dati:** I file FASTQ e il genoma di riferimento sono stati importati in Galaxy.
2. **Controllo Qualità:** FASTQC è stato utilizzato per valutare la qualità delle reads.
3. **Trimming:** Trimmomatic è stato utilizzato per rimuovere eventuali adattatori e basi di bassa qualità.
4. **Mapping:** Bowtie2 è stato utilizzato per allineare le reads al genoma di riferimento.
5. **Visualizzazione:** IGV e JBrowse sono stati utilizzati per visualizzare i risultati del mapping.

I dettagli specifici dei parametri utilizzati per ciascuno strumento e l'ordine esatto dei passaggi saranno descritti nelle sezioni successive.

## 4.4 Analisi dei Dati di SARS-CoV-2 su Galaxy

In questa sezione viene descritta in dettaglio l'analisi eseguita sui dati di sequenziamento di SARS-CoV-2, a partire dall'importazione dei file fino alla valutazione dei risultati del mapping, con e senza trimming.

### 4.4.1 Importazione dei Dati

I file FASTQ contenenti le reads paired-end di SARS-CoV-2 (CV1537_S1_L001_R1_001.fastq.gz e CV1537_S1_L001_R2_001.fastq.gz) e il file FASTA del genoma di riferimento (NC_045512.2.fna) sono stati importati nell'istanza pubblica di Galaxy (usegalaxy.eu). Per comodità, i file FASTQ sono stati rinominati rispettivamente in **read1** e **read2**, mentre il file del genoma di riferimento è stato rinominato in **genomaRiferimentoSARS-Cov-2**.

### 4.4.2 Controllo Qualità con FASTQC

Il controllo qualità delle reads è stato eseguito utilizzando lo strumento **FASTQC** (versione 0.12.1) disponibile su Galaxy. FASTQC è stato eseguito su entrambi i file 'read1' e 'read2' con i parametri di default. I report generati da FASTQC, in formato HTML, sono stati analizzati per valutare la qualità complessiva delle reads e per identificare potenziali problemi.

**Risultati di FASTQC per read1:** L'analisi di FASTQC su read1 ha evidenziato una buona qualità generale delle reads. I grafici "Per Base Sequence Quality" (Figura 6), "Per tile sequence quality" (Figura 7) e "Per Sequence Quality Scores" (Figura 8) hanno mostrato punteggi di qualità Phred prevalentemente superiori a 30, indicando un'elevata accuratezza delle chiamate di base.

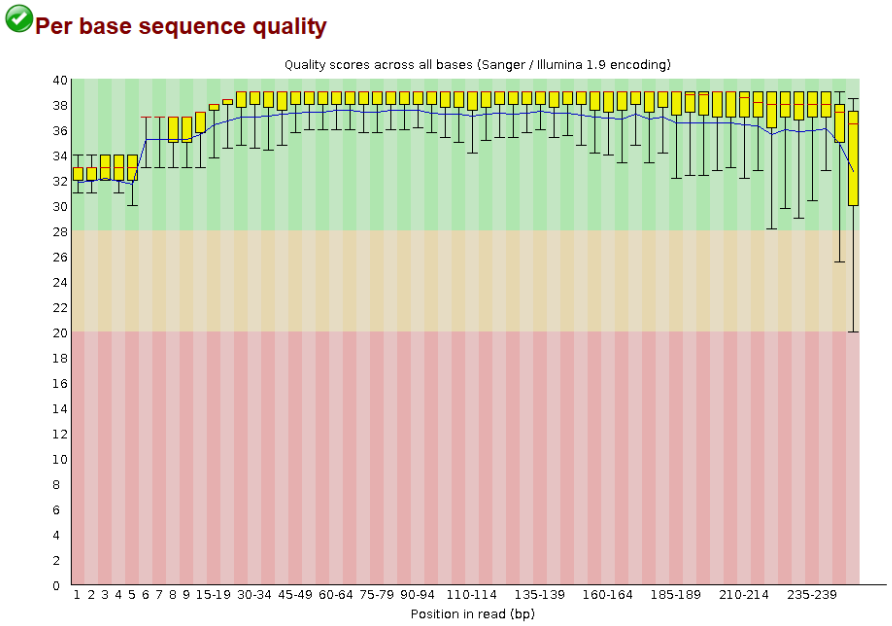


Figura 6: Grafico "Per Base Sequence Quality" di FASTQC per read1. Il grafico mostra la distribuzione dei punteggi di qualità Phred per ciascuna posizione nelle reads.

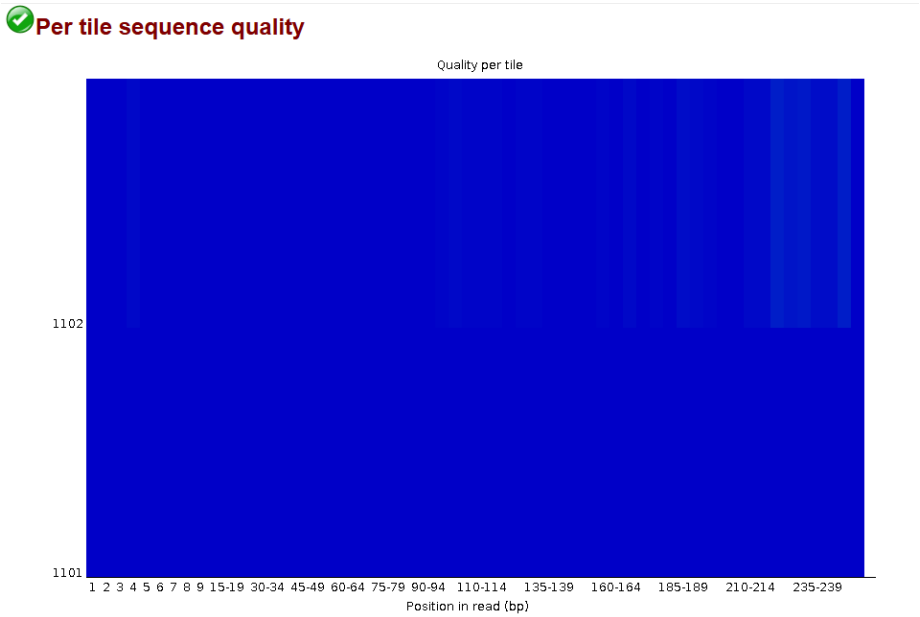


Figura 7: Grafico "Per Tile Sequence Quality" di FASTQC per read1. Il grafico non evidenzia alcun valore anomalo.

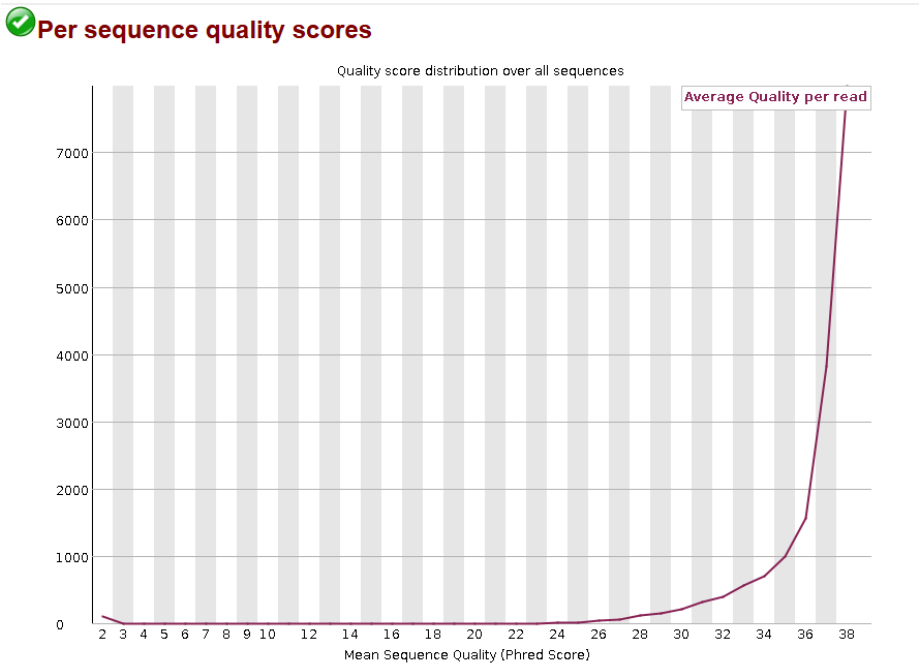


Figura 8: Grafico "Per Sequence Quality Scores" di FASTQC per read1. Il grafico mostra la distribuzione dei punteggi di qualità medi per ciascuna read.

Tuttavia, il grafico "Per Base Sequence Content" (Figura 9) ha rivelato una distorsione nella composizione delle basi nelle prime 15 posizioni, suggerendo la presenza di sequenze sovrarappresentate. Questa ipotesi è stata confermata dal modulo "Overrepresented Sequences" (Figura 10), che ha riportato diverse sequenze con una frequenza superiore all'atteso, tra cui una sequenza di sole N nelle prime posizioni. Inoltre è stata rilevata una possibile presenza di sequenze Nextera. Per quanto riguarda il grafico "Per Sequence GC content" (Figura 11), seppur viene segnalato un fail, la funzione si discosta di poco da quella prevista.

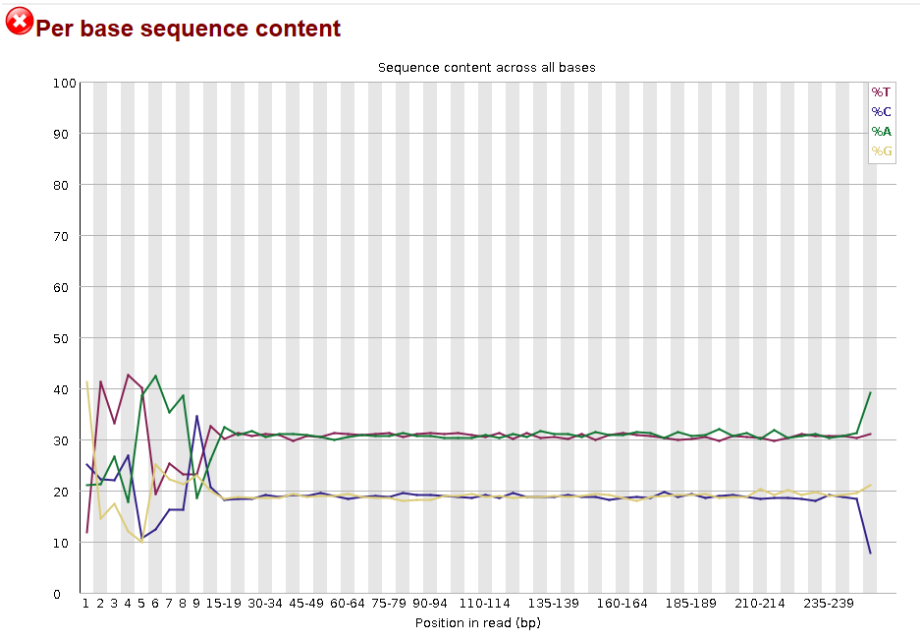


Figura 9: Grafico "Per Base Sequence Content" di FASTQC per read1. Il grafico mostra la proporzione di ciascuna base (A, T, G, C) in ciascuna posizione lungo le reads.

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN | 120 | 0.6961364427427775 | No Hit |
| GTGCCAAGCTCGTCGCCTAAGTCAAATGACTTTAGATCGGCGCCGTAACT | 58 | 0.3364659473256758 | No Hit |
| CTCCTAGCACCATCATCATACACAGTTCTTGCTGTCATAAGGATTAGTAA | 53 | 0.30746026221139344 | No Hit |
| GTCTAAAGTAGCGGTTGAGTAAACAAAAGAGGCCAAAGTAACAAGTACAA | 38 | 0.22044320686854624 | No Hit |
| CCCTTGGAGAGTGCTAGTTGCCATCTCTTTTTGAGGGTTATGATTTTGGA | 37 | 0.21464206984568973 | No Hit |
| GCCTTACATTAAGTGGGATTTGTTAAAATATGACTTCACGGAAGAGAGGT | 27 | 0.15663069961712495 | No Hit |
| CTCCAGGGACCACCTGGTACTGGTAAGAGTCATTTTGCTATTGGCCTAGC | 25 | 0.14502842557141202 | No Hit |
| GTGCATAGCAGGGTCAGCAGCATACACAAGTAATTCCTTAAAACTAAGTC | 25 | 0.14502842557141202 | No Hit |
| GAATTAGACTCATTCAAGGAGGAGTTAGATAAATATTTTAAGAATCATAC | 25 | 0.14502842557141202 | No Hit |
| GTACTATTACCGTTGAAGAGCTTAAAAAGCTCCTTGAACAATGGAACCTA | 24 | 0.13922728854855554 | No Hit |
| CATCACAACCAGGCAAGTTAAGGTTAGATAGCACTCTAGTGTCAAATCTA | 23 | 0.13342615152569903 | No Hit |
| GTATATGAGATCTCTCAAAGTGCCAGCTACAGTTTCTGTTTCTTCACCTG | 22 | 0.12762501450284255 | No Hit |
| GCTTTGAGTTGACATCTATGAAGTATTTTGTGAAAATAGGACCTGAGCGC | 22 | 0.12762501450284255 | No Hit |
| ATTCTGTGCTGGTAGTACATTTATTAGTGATGAAGTTGCGAGAGACTTGT | 22 | 0.12762501450284255 | No Hit |
| AACCTATACTGTTACTAGATCAGGCATTAGTGTCTGATGTTGGTGATAGT | 22 | 0.12762501450284255 | No Hit |
| TGTACACATAGTGCTTAGCACGTAATCTGGCATTGACAACACTCAAATCA | 21 | 0.12182387747998608 | No Hit |
| CCTATACTGTTACTAGATCAGGCATTAGTGTCTGATGTTGGTGATAGTGC | 21 | 0.12182387747998608 | No Hit |
| ACATTATACTTAAACCAGCAAATAATAGTTTAAAAATTACAGAAGAGGTT | 21 | 0.12182387747998608 | No Hit |
| GGCATACACCATCTGTGAATTTGTCAGAATGTGTGGCATAAGAATAGAAT | 20 | 0.1160227404571296 | No Hit |
| TGCCACAACTGCTTATGCTAATAGTGTTTTTAACATTTGTCAAGCTGTCA | 20 | 0.1160227404571296 | No Hit |
| GTATTAATGCTAACCAAGTCATCGTCAACAACCTAGACAAATCAGCTGGT | 20 | 0.1160227404571296 | No Hit |
| GTCTTACTCTCAGTTTTGCAACAACTCAGAGTAGAATCATCATCTAAATT | 20 | 0.1160227404571296 | No Hit |
| CTATTAAACAGCCTGCACGTGTTTGAAAAACATTAGAACCTGTAGAATAA | 20 | 0.1160227404571296 | No Hit |
| CAACAGGTGCGCTCAGGTCCTATTTTCACAAAATACTTCATAGATGTCAA | 20 | 0.1160227404571296 | No Hit |

Figura 10: Tabella delle sequenze overrepresented di read1

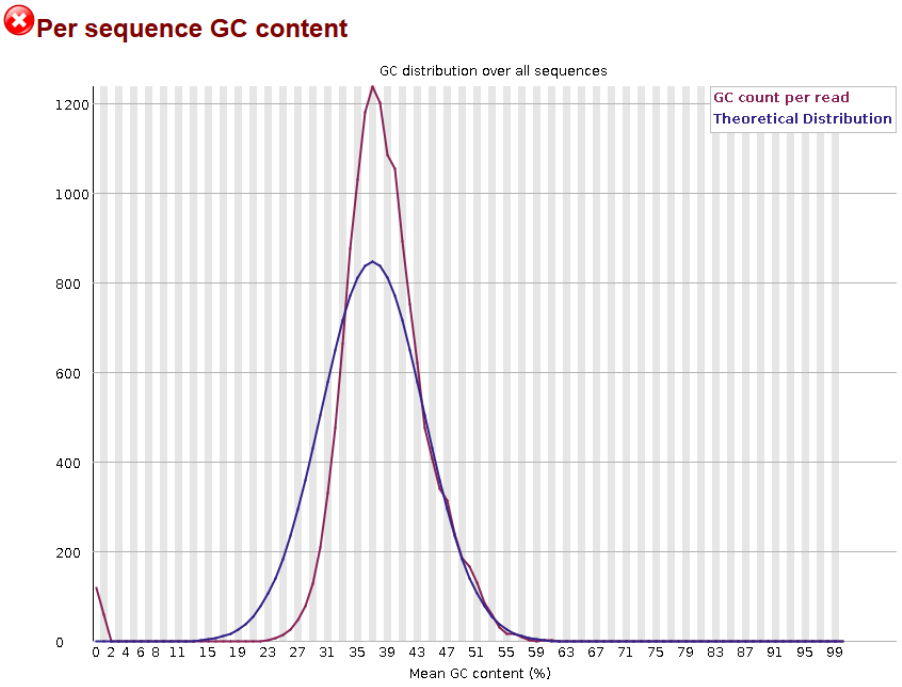


Figura 11: Grafico "Per Sequence GC Content" di FASTQC per read1. Il grafico mostra una leggera differenza con la funzione prevista.

**Risultati di FASTQC per read2:** L'analisi di FASTQC su read2 ha prodotto risultati simili a quelli ottenuti per read1. La qualità generale delle reads è risultata buona, come mostrato dai grafici "Per Base Sequence Quality" (Figura 12), "Per Tile Sequence Quality" (Figura 13) e "Per Sequence Quality Scores" (Figura 14). Analogamente a read1, il grafico "Per Base Sequence Content" (Figura 15) ha evidenziato una distorsione nelle prime 15 posizioni (spesso non significativa in quanto non influisce negativamente sulle analisi) e il modulo "Overrepresented Sequences" (Fig 16) ha confermato la presenza di sequenze sovrarappresentate, sebbene diverse da quelle trovate in read1. Anche in questo caso è stata rilevata una possibile presenza di sequenze Nextera, inferiore rispetto al caso precedente. Per quanto riguarda il grafico "Per Sequence GC Content" (Figura: 17), come per read1 si discosta di poco da quello previsto.

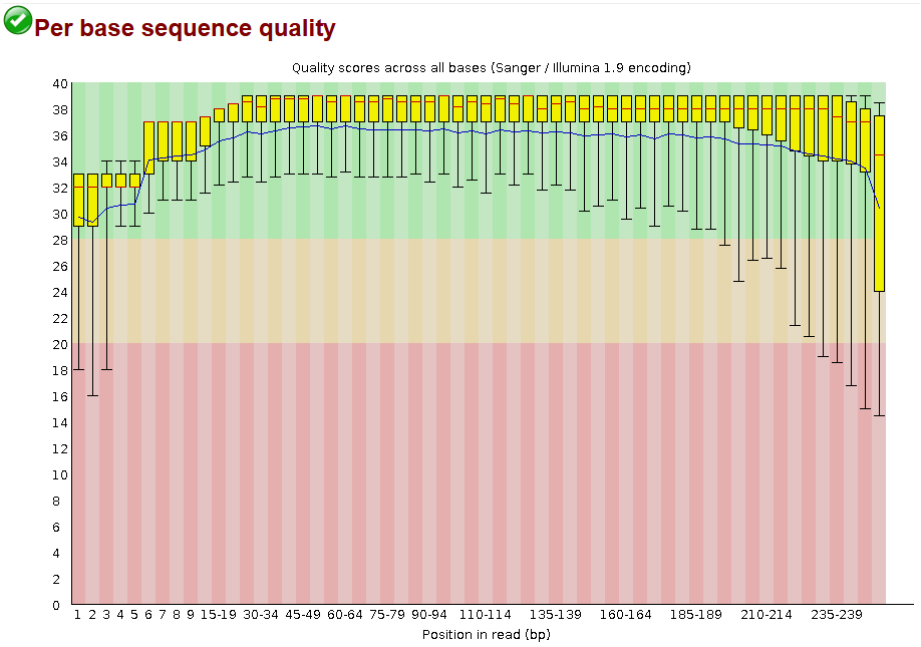


Figura 12: Grafico "Per Base Sequence Quality" di FASTQC per read2. Il grafico mostra la distribuzione dei punteggi di qualità Phred per ciascuna posizione nelle reads.

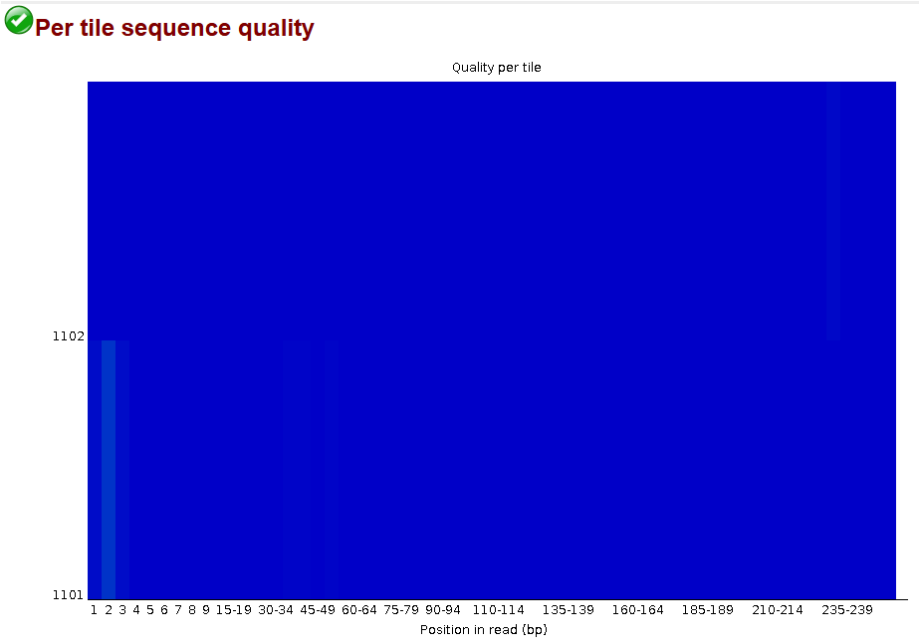


Figura 13: Grafico "Per Tile Sequence Quality" di FASTQC per read2. Il grafico non evidenzia alcun valore anomalo.

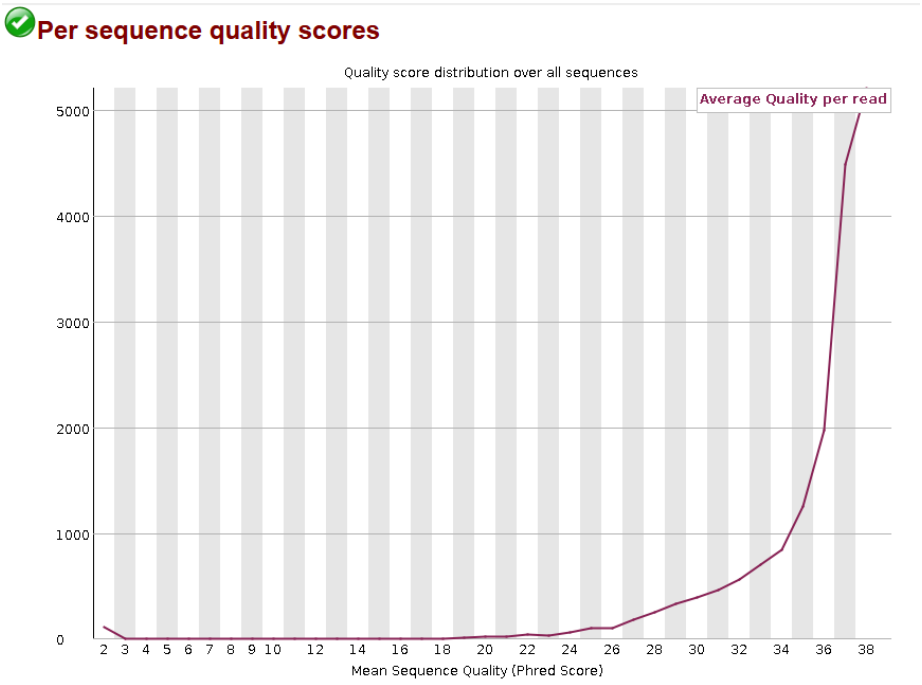


Figura 14: Grafico "Per Sequence Quality Scores" di FASTQC per read2. Il grafico mostra la distribuzione dei punteggi di qualità medi per ciascuna read.

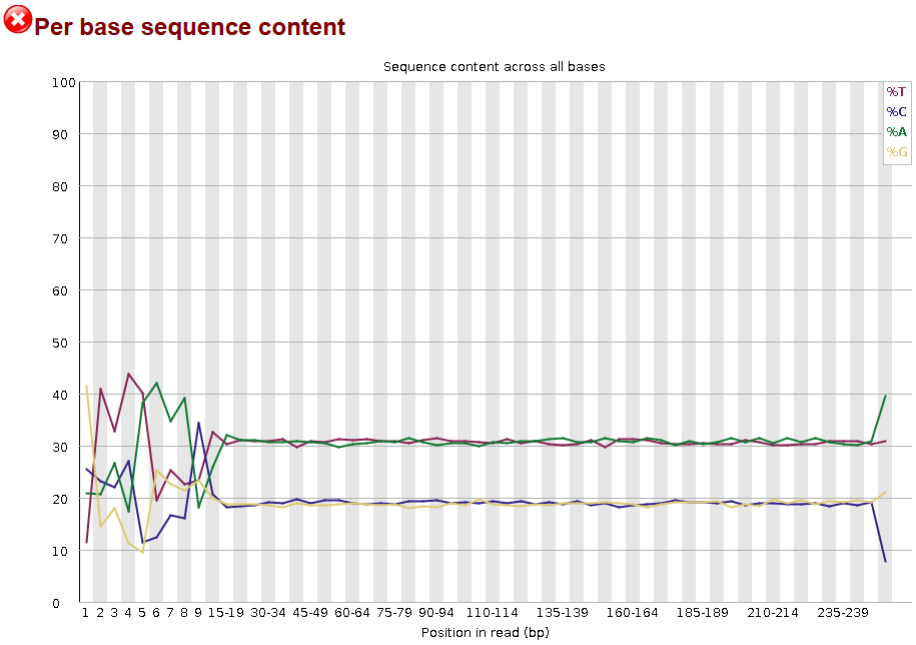


Figura 15: Grafico "Per Base Sequence Content" di FASTQC per read2. Il grafico mostra la proporzione di ciascuna base (A, T, G, C) in ciascuna posizione lungo le reads.

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN | 120 | 0.6961364427427775 | No Hit |
| GTGCCAAGCTCGTCGCCTAAGTCAAATGACTTTAGATCGGCGCCGTAACT | 64 | 0.3712727694628147 | No Hit |
| CTCCTAGCACCATCATCATACACAGTTCTTGCTGTCATAAGGATTAGTAA | 39 | 0.22624434389140274 | No Hit |
| GTCTAAAGTAGCGGTTGAGTAAACAAAAGAGGCCAAAGTAACAAGTACAA | 38 | 0.22044320686854624 | No Hit |
| GACAACAGGTGCGCTCAGGTCCTATTTTCACAAAATACTTCATAGATGTC | 35 | 0.20303979579997677 | No Hit |
| CCCTTGGAGAGTGCTAGTTGCCATCTCTTTTTGAGGGTTATGATTTTGGA | 28 | 0.16243183663998143 | No Hit |
| CTCCAGGGACCACCTGGTACTGGTAAGAGTCATTTTGCTATTGGCCTAGC | 26 | 0.1508295625942685 | No Hit |
| GCTTTGAGTTGACATCTATGAAGTATTTTGTGAAAATAGGACCTGAGCGC | 25 | 0.14502842557141202 | No Hit |
| GTCATGTAGTTGCCTTTAATACTTTACTATTCCTTATGTCATTCACTGTA | 24 | 0.13922728854855554 | No Hit |
| GTGTATACAGCTTGCTCTCATGCCGCTGTTGATGCACTATGTGAGAAGGC | 24 | 0.13922728854855554 | No Hit |
| GGCATACACCATCTGTGAATTTGTCAGAATGTGTGGCATAAGAATAGAAT | 22 | 0.12762501450284255 | No Hit |
| GCTATAAGACATGTACGTGCATGGATTGGCTTCGATGTCGAGGGGTGTCA | 22 | 0.12762501450284255 | No Hit |
| ACCTGGTACTGGTAAGAGTCATTTTGCTATTGGCCTAGCTCTCTACTACC | 20 | 0.1160227404571296 | No Hit |
| GTATATGAGATCTCTCAAAGTGCCAGCTACAGTTTCTGTTTCTTCACCTG | 20 | 0.1160227404571296 | No Hit |
| GCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTT | 20 | 0.1160227404571296 | No Hit |
| AATGTACACATAGTGCTTAGCACGTAATCTGGCATTGACAACACTCAAAT | 20 | 0.1160227404571296 | No Hit |
| GTGTTTAAACCGTGTTTGTACTAATTATATGCCTTATTTCTTTACTTTAT | 19 | 0.11022160343427312 | No Hit |
| GTACAGACTGTGTTTTTAAGTGTAAAACCCACAGGGTCATTAGCACAAGT | 19 | 0.11022160343427312 | No Hit |
| CCTATACTGTTACTAGATCAGGCATTAGTGTCTGATGTTGGTGATAGTGC | 19 | 0.11022160343427312 | No Hit |
| GAATTAGACTCATTCAAGGAGGAGTTAGATAAATATTTTAAGAATCATAC | 19 | 0.11022160343427312 | No Hit |
| GGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAG | 18 | 0.10442046641141664 | No Hit |
| GTGCATAGCAGGGTCAGCAGCATACACAAGTAATTCCTTAAAACTAAGTC | 18 | 0.10442046641141664 | No Hit |

Figura 16: Tabella delle sequenze overrepresented di read2

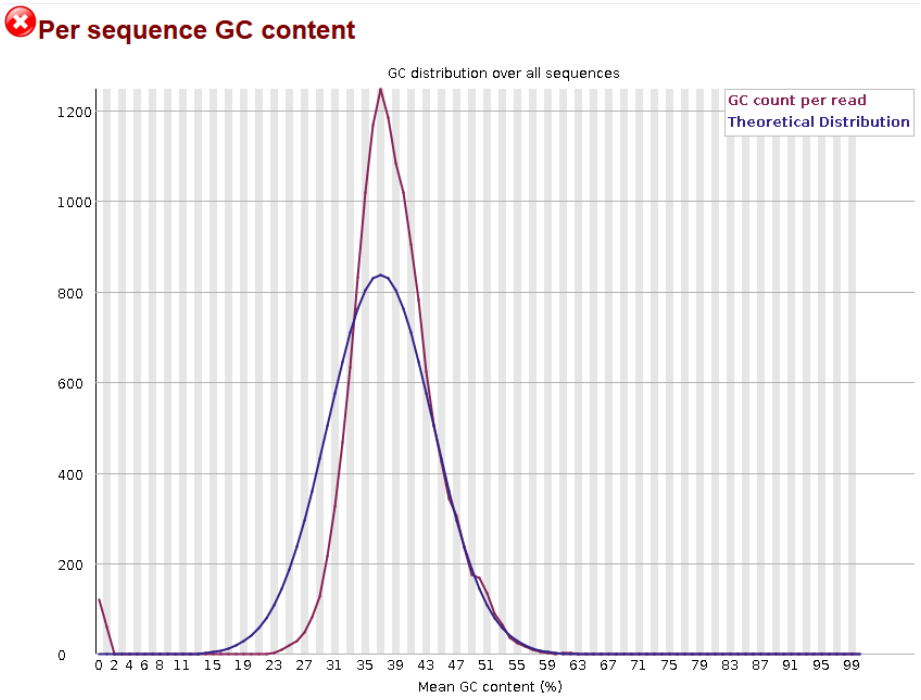


Figura 17: Grafico "Per Sequence GC Content" di FASTQC per read2. Il grafico mostra una leggera differenza con la funzione prevista.

**Tentativo di Trimming con Trimmomatic** Sulla base dei risultati di FASTQC, si è deciso di effettuare un tentativo di trimming delle reads utilizzando lo strumento Trimmomatic (versione 0.38), con l'obiettivo di rimuovere le sequenze sovrarappresentate (probabilmente adattatori) e le porzioni di bassa qualità alle estremità delle reads.

**Parametri di Trimmomatic:** Per il trimming sono stati utilizzati i seguenti parametri:

1. **ILLUMINACLIP**: con le opzioni 'palindrome', 'seed mismatches' impostato a 2 e 'simple clip threshold' impostato a 10.
2. Il predefinito di **Nextera** come adattatore
3. **SLIDINGWINDOW**: windowSize:4 requiredQuality:25
4. **LEADING**: 5
5. **TRAILING**: 5

6. **MINLEN**: 75

Dopo il trimming, è stata eseguita una nuova analisi con FASTQC sui file trimmati per verificare l'efficacia dell'operazione. I grafici "Per Base Sequence Quality" (Figura: 18) e "Per Sequence GC Content" (Figura: 19) hanno mostrato un miglioramento della qualità, mentre i grafici "Per Base Sequence Content" (Figura: 20) hanno evidenziato una riduzione, ma non l'eliminazione completa, della distorsione nelle prime posizioni. Anche le sequenze sovrarappresentate sono state ridotte ma non eliminate.

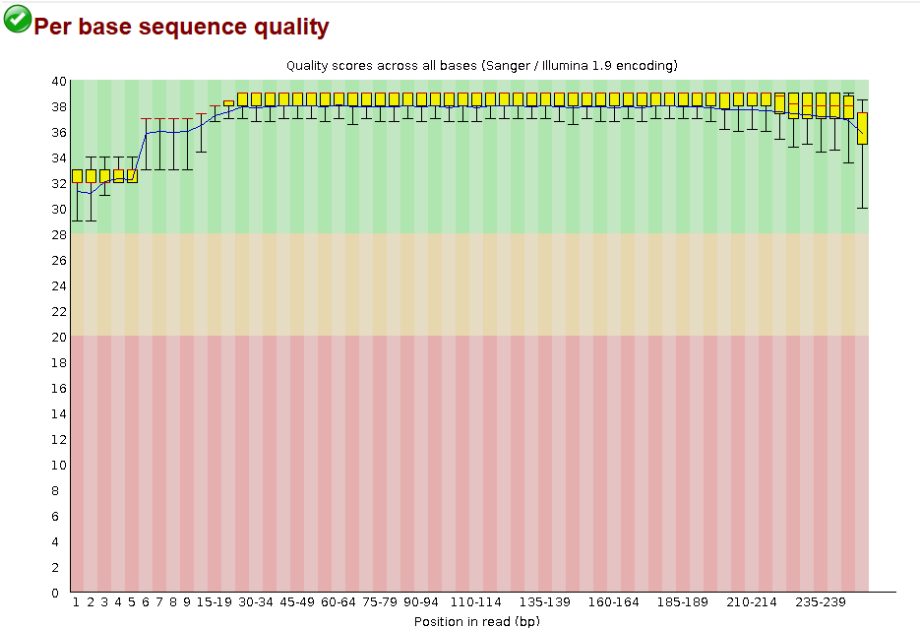


Figura 18: Grafico "Per Base Sequence Quality" di FASTQC per read2 post trimming. Il grafico mostra un netto miglioramento.

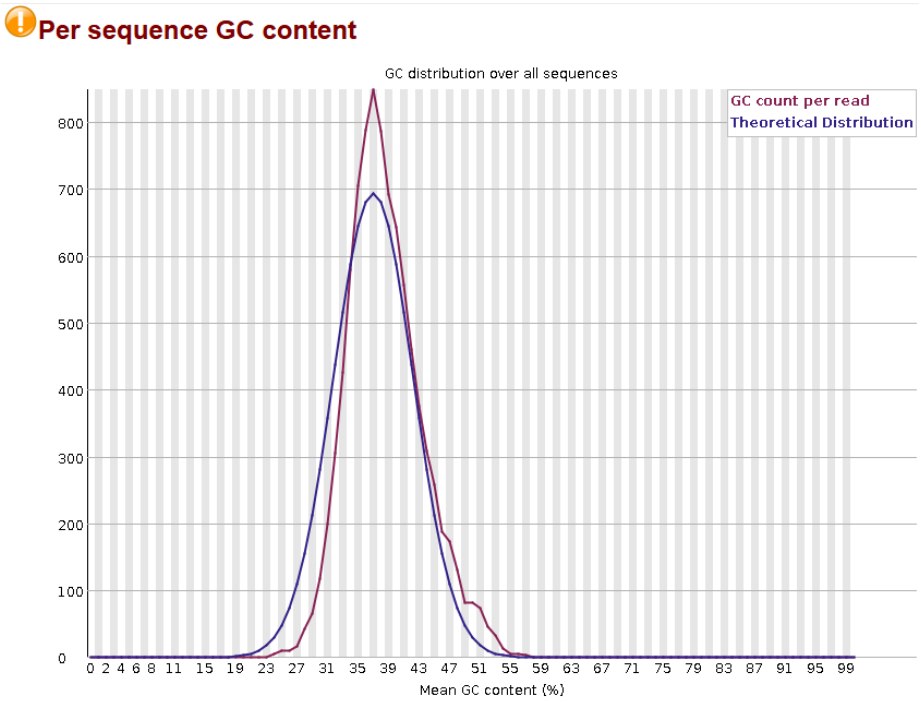


Figura 19: Grafico "Per Sequence GC Content" di FASTQC per read2 post trimming. Il grafico mostra un netto miglioramento.

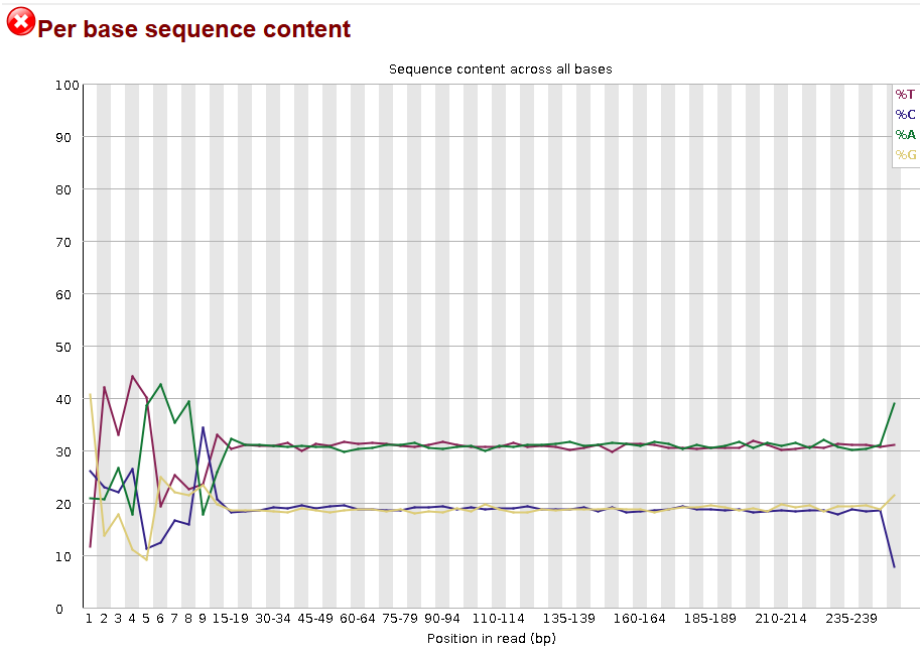


Figura 20: Grafico "Per Base Sequence Content" di FASTQC per read2 post trimming. Il grafico mostra un leggero miglioramento, ma non ancora sufficiente.

### 4.4.3 Mapping con e senza Trimming

Successivamente, è stato eseguito il mapping delle reads sul genoma di riferimento (NC_045512.2) utilizzando **Bowtie2** con i parametri di default, sia per i dati originali non trimmati ('read1' e 'read2') sia per i dati trimmati.

**Risultati del mapping:**

- **Senza trimming:** La percentuale di allineamento ottenuta è stata del **98.39%**.
- **Con trimming:** La percentuale di allineamento è scesa drasticamente al **49.89%**.

A causa del peggioramento significativo delle prestazioni di mapping dopo il trimming, si è deciso di proseguire le analisi successive utilizzando i **dati originali non trimmati**.

### 4.4.4 Utilizzo di Bowtie2

Il mapping delle reads di SARS-CoV-2 sul genoma di riferimento è stato effettuato utilizzando lo strumento Bowtie2 (versione 2.4.2), disponibile su Galaxy. Sono state utilizzate le **reads originali non trimmate** ('read1' e 'read2').

**Parametri di Bowtie2:** Bowtie2 è stato eseguito in modalità **paired-end**, utilizzando i file 'read1' e 'read2' come input per le due reads e il file 'genomaRiferimentoSARS-Cov-2' (contenente il genoma di riferimento NC_045512.2) come genoma di riferimento. Sono stati utilizzati i parametri di default di Bowtie2.

### 4.4.5 Risultati del Mapping e Samtools

I risultati del mapping, riportati dal file di output di Bowtie2, sono i seguenti:
**Reads totali:** 17238
**Reads paired:** 17238 (100.00%)
**Allineate concordantemente 0 volte:** 458 (2.66%)
**Allineate concordantemente esattamente 1 volta:** 16780 (97.34%)
**Allineate concordantemente almeno 1 volta:** 0 (0.00%)
**Coppie allineate concordantemente 0 volte:** 458
**Allineate discordantemente 1 volta:** 101 (22.05%)
**Coppie allineate 0 volte conc. o discordantemente:** 357
**Mates che compongono le coppie:** 714
**Allineati 0 volte:** 554 (77.59%)
**Allineati esattamente 1 volta:** 160 (22.41%)
**Allineati almeno 1 volta:** 0 (0.00%)
**Percentuale di allineamento complessiva: 98.39%**
Questi risultati indicano che la stragrande maggioranza delle reads (97.34%) si è allineata in modo univoco e concorde al genoma di riferimento. Una piccola percentuale di reads (2.66%) non si è allineata, mentre una frazione ancora minore (22.05% di 458) si è allineata in modo discorde. Ulteriori analisi sono state effettuate usando **samtools**, da cui è risultato un tasso di errore del **0.004%**.

### 4.4.6 Visualizzazione con IGV e JBrowse

Per visualizzare i risultati del mapping, sono stati utilizzati i visualizzatori genomici IGV (Integrative Genomics Viewer) e JBrowse.

**IGV:** è un visualizzatore desktop che permette di esplorare dati genomici in modo interattivo. Il file **BAM** prodotto da Bowtie2, contenente gli allineamenti delle reads al genoma di riferimento, è stato caricato su IGV insieme al file FASTA del genoma di riferimento (NC_045512.2).

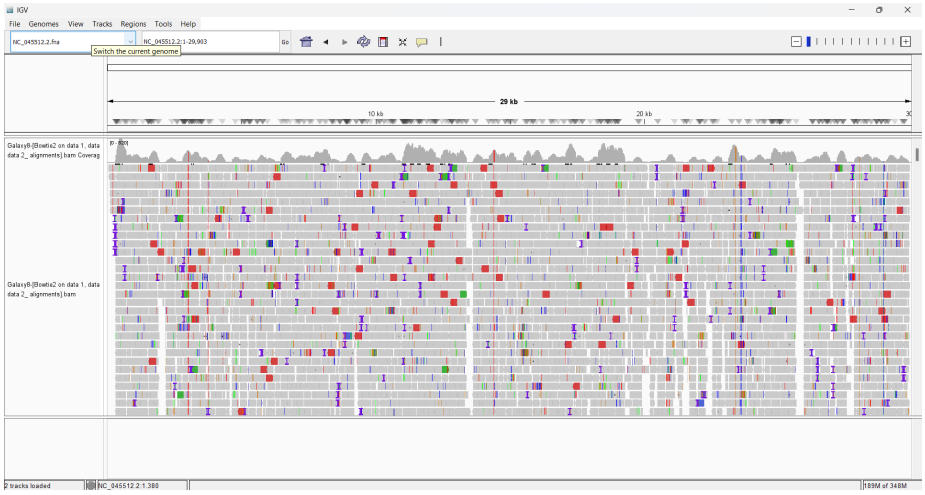


Figura 21: Visualizzazione dell'allineamento delle reads di SARS-CoV-2 al genoma di riferimento (NC_045512.2) in IGV.

In linea generale, dalla rappresentazione in IGV (Figura: 21) si nota che alcune regioni presentano una copertura elevata, mentre altre mostrano una ridotta profondità di sequenziamento.

Analizzando le letture allineate, che sono rappresentate in grigio, si osservano diverse basi colorate, le quali segnalano variazioni rispetto alla sequenza di riferimento. In particolare, il rosso indica la presenza di polimorfismi a singolo nucleotide (SNP) o possibili errori di sequenziamento, il verde è associato a inserzioni, mentre il blu e il viola segnalano delezioni o regioni caratterizzate da una qualità di lettura più bassa. La distribuzione di queste variazioni suggerisce la presenza di mutazioni puntiformi localizzate in specifiche posizioni lungo il genoma.

Dal punto di vista biologico, la presenza di mutazioni ricorrenti potrebbe indicare variazioni caratteristiche del dataset analizzato, ad esempio campioni clinici di SARS-CoV-2 con specifiche mutazioni. Inoltre, alcune aree del genoma mostrano un'elevata variabilità, suggerendo la possibilità di hotspot mutazionali.

In conclusione, l'analisi dell'allineamento in IGV evidenzia un buon livello di copertura generale, con alcune regioni caratterizzate da una minore profondità e da variazioni rispetto alla sequenza di riferimento.

**JBrowse:** è un visualizzatore genomico web-based integrato in Galaxy. Il file BAM prodotto da Bowtie2 è stato visualizzato direttamente in JBrowse all'interno dell'ambiente Galaxy.

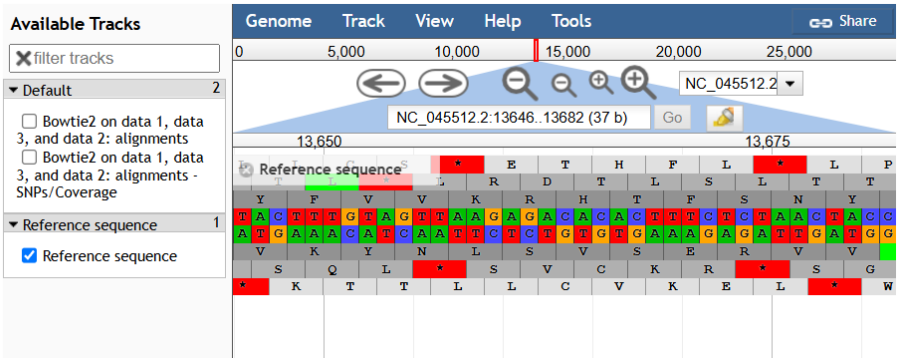


Figura 22: Visualizzazione dell'allineamento delle reads di SARS-CoV-2 al genoma di riferimento (NC_045512.2) in JBrowse.

La sezione evidenziata (Figura: 22) comprende le basi da 13,646 a 13,682, con la sequenza nucleotidica e la relativa traduzione in amminoacidi. Le basi sono colorate secondo la convenzione standard (A: verde, T: rosso, C: blu, G: giallo), mentre le stelle rosse indicano possibili mutazioni o anomalie nella sequenza.

L'analisi di questa regione suggerisce la presenza di variazioni nucleotidiche che potrebbero influenzare la sequenza proteica risultante.

# 5 Risultati

In questo capitolo sono presentati i risultati delle analisi di quality control, trimming e mapping delle reads di SARS-CoV-2, ottenute utilizzando il workflow descritto nel Capitolo 3.

## 5.1 Controllo Qualità e Trimming

I risultati dettagliati del controllo qualità, effettuato con FASTQC prima e dopo il tentativo di trimming con Trimmomatic, sono stati descritti nel Capitolo 3 (Sezione 3.4.2 e 3.4.3). In sintesi, l'analisi di FASTQC ha evidenziato una buona qualità generale delle reads, ma ha anche rilevato la presenza di sequenze sovrarappresentate, probabilmente adattatori, nelle prime posizioni di entrambe le reads (Figura 9, 15, 10, 16). Il tentativo di trimming con Trimmomatic ha migliorato la qualità delle reads e ridotto la presenza di sequenze sovrarappresentate, ma ha portato a una drastica diminuzione della percentuale di allineamento in fase di mapping.

## 5.2 Mapping delle Reads Non Trimmate

Il mapping delle reads originali non trimmate di SARS-CoV-2 ('read1' e 'read2') sul genoma di riferimento (NC_045512.2) è stato eseguito con Bowtie2, ottenendo una percentuale di allineamento complessiva del **98.39%**. La maggior parte delle reads (97.34%) si è allineata in modo concorde ed univoco al genoma di riferimento. I dettagli numerici del mapping sono riportati nella Sezione 3.4.4 del Capitolo 3. Ulteriori analisi con samtools hanno portato a un tasso di errore dello **0.004%**.

## 5.3 Visualizzazione del Mapping

L'allineamento delle reads al genoma di riferimento è stato visualizzato con i visualizzatori genomici IGV (Figura 21) e JBrowse (Figura 22). Le immagini mostrano una regione rappresentativa del genoma di SARS-CoV-2, con una buona copertura di reads, a conferma dell'elevata percentuale di allineamento ottenuta con Bowtie2.

# 6 Discussione

I risultati ottenuti dimostrano che è possibile ottenere un'elevata percentuale di allineamento (**98.39%**) delle reads di SARS-CoV-2 al genoma di riferimento (NC_045512.2) utilizzando Bowtie2, anche senza ricorrere al trimming delle reads. Questo risultato è particolarmente significativo in considerazione della presenza di sequenze sovrarappresentate, rilevate da FASTQC, nelle prime posizioni delle reads.

Il tentativo di trimming, effettuato con Trimmomatic per rimuovere tali sequenze, ha portato a un drastico peggioramento delle prestazioni di mapping, con una riduzione della percentuale di allineamento al 49.89%. Questo suggerisce che le sequenze rimosse, pur essendo sovrarappresentate, potrebbero non essere semplici adattatori, ma potrebbero avere un ruolo biologico o derivare da artefatti specifici del processo di preparazione del campione o di sequenziamento. Un'altra possibilità è che i parametri di trimming utilizzati fossero troppo aggressivi per questo specifico dataset, causando la rimozione di porzioni informative delle reads.

La visualizzazione del mapping con IGV e JBrowse ha permesso di confermare visivamente l'elevata copertura del genoma di riferimento ottenuta con l'allineamento delle reads non trimmate. Le immagini (Figura 21 e Figura 22) mostrano chiaramente l'allineamento delle reads lungo il genoma, con una buona concordanza tra le due reads di ciascuna coppia.

Questi risultati suggeriscono che, per questo specifico dataset di SARS-CoV-2, il trimming con Trimmomatic, con i parametri utilizzati, non è necessario e può addirittura essere controproducente. Tuttavia, è importante sottolineare che la necessità e l'efficacia del trimming possono variare a seconda del dataset e del protocollo di sequenziamento utilizzato. In altri casi, il trimming potrebbe essere essenziale per ottenere un buon mapping e per rimuovere artefatti che potrebbero compromettere le analisi a valle.

Ulteriori analisi, potrebbero fornire maggiori informazioni sulla qualità del mapping e sull'eventuale impatto delle sequenze sovrarappresentate. Inoltre, un'analisi più approfondita delle sequenze sovrarappresentate, potrebbe aiutare a chiarirne l'origine e la natura.

# 7 Conclusioni

In questo progetto è stato analizzato un dataset di sequenze di SARS-CoV-2, eseguendo il controllo qualità delle reads con FASTQC, il trimming con Trimmomatic e il mapping al genoma di riferimento con Bowtie2. I risultati ottenuti hanno dimostrato che, in questo caso specifico, il trimming delle reads non è necessario per ottenere un'elevata percentuale di allineamento (**98.39%**). La visualizzazione del mapping con IGV e JBrowse ha confermato la buona qualità dell'allineamento.

Questo lavoro dimostra l'importanza di valutare attentamente la necessità del trimming in base alle caratteristiche specifiche del dataset e ai risultati del controllo qualità. Inoltre, evidenzia come l'utilizzo di strumenti di visualizzazione genomica come IGV e JBrowse possa essere utile per la validazione dei risultati di mapping.

Le analisi future su questo dataset potrebbero includere l'identificazione di varianti, l'analisi dell'espressione genica e un'indagine più approfondita sull'origine delle sequenze sovrarappresentate rilevate da FASTQC.