
VALUTAZIONE DEGLI ASSEMBLAGGI GENOMICI CON QUAST

Progetto di Bioinformatica

Chiara Coscarelli
Matricola 0522501982
Marco Santoriello
Matricola 0522501992

Contents

1	Assemblaggio Genomico	2
1.1	Sfide degli Assemblaggi	2
1.2	Metodi Esistenti per il Confronto	2
1.3	Limiti dei metodi attuali	3
2	Il Contributo di QUASt	4
2.1	Metriche valutate da QUASt	4
3	Come QUASt valuta gli assemblatori	6
3.1	Limiti nell’N50 e alternative proposte	6
4	Misassemblaggi e variazioni strutturali	6
5	Tools genomici	7
6	Panoramica del codice	9
6.1	Utility generali	9
6.2	Generazione di statistiche e reports	9
6.3	Visualizzazione dei risultati	10
7	Analisi del codice	11
7.1	Flusso Operativo	11
7.1.1	Fase di Import e Configurazione	11
7.1.2	Processamento del genoma di riferimento	11
7.1.3	Processamento dei contigs	11
7.1.4	Allineamento e analisi	12
7.1.5	Analisi della metrica NG50	12
8	Galaxy Workflow	14
8.1	Risultati	15
8.1.1	Metriche legate al numero di contig	15
8.1.2	Errori di misassemblaggio	16
8.1.3	Rappresentazione genomica	18
8.1.4	Metriche basate su N50	18
8.1.5	Conclusioni	19

1 Assemblaggio Genomico

L'assemblaggio del genoma è il processo di unione di frammenti di sequenze biologiche (letture - *reads* - di DNA) per creare sequenze più lunghe chiamate **contig**. Questo processo è essenziale per analizzare la struttura e la funzionalità del genoma di un organismo, tuttavia la qualità dell'assemblaggio è difficile da misurare. Per farlo, vengono utilizzate diverse metriche, alcune delle quali si concentrano sulla **contiguità**, la **completezza** e la **correttezza** dell'assemblaggio. La **contiguità** si riferisce alla lunghezza e al numero di contig che risultano dall'assemblaggio. Idealmente, un assemblaggio dovrebbe riflettere la struttura del genoma reale, con pochi contig molto lunghi. Tuttavia, errori di contiguità possono verificarsi quando contig non correlati vengono uniti erroneamente. La **completezza**, invece, misura quanto l'assemblaggio copra tutto il genoma di riferimento, in particolare in termini di contenuti genici. Gli errori di completezza possono derivare da letture mancanti o sequenziamento incompleto. Infine, la **correttezza** riguarda l'ordine e la posizione dei contig. Se l'assemblaggio non riflette correttamente l'ordine del genoma di riferimento, si parla di errori di correttezza, come inversioni o traslocazioni di segmenti di DNA.

Mentre l'uso di metriche come l'N50 (una misura di contiguità) è comune, ci sono delle limitazioni. L'N50, ad esempio, può essere fuorviante, poiché un aumento artificiale del numero di contig grandi può gonfiare il punteggio, non riflettendo accuratamente la qualità dell'assemblaggio. Per questo motivo, alcuni ricercatori hanno proposto metodi alternativi, come l'**e-size**, che misura la dimensione di un contig contenente una base selezionata casualmente nel genoma, evitando l'influenza di unioni errate di contig.[1]

1.1 Sfide degli Assemblaggi

Le tecnologie di sequenziamento del DNA attuali non possono produrre la sequenza completa di un cromosoma in un'unica operazione. Al contrario, generano un grande numero di reads (sequenze di basi consecutive), che possono variare da decine a migliaia di basi, campionate da diverse parti del genoma. Il software di assemblaggio genomico combina tali reads in regioni più grandi chiamate **contig**. Tuttavia, le tecnologie e i software attuali affrontano numerose difficoltà che ostacolano la ricostruzione completa dei cromosomi, tra cui errori nelle reads e la presenza di grandi regioni ripetitive nel genoma.[2]

1.2 Metodi Esistenti per il Confronto

Negli ultimi anni sono stati sviluppati vari metodi per confrontare diversi assemblatori. Di seguito sono descritti alcuni dei principali strumenti e i loro limiti. **Plantagora** è una piattaforma web progettata per aiutare i ricercatori a visualizzare le caratteristiche delle strategie di sequenziamento più popolari (inclusi piattaforme di sequenziamento e software di assemblaggio) per i genomi delle piante. Sebbene Plantagora offra un'interfaccia ben progettata per consultare i risultati, il suo strumento di valutazione non fornisce un'interfaccia altrettanto user-friendly; gli utenti devono analizzare manualmente un file di log molto ampio.[3]

Assemblathon, invece è una competizione che ha confrontato 41 assemblaggi de novo su 4100 metriche di valutazione. Sebbene gli script di valutazione di Assemblathon siano liberamente disponibili, sono fortemente focalizzati sui genomi usati nella competizione, rendendo difficile la loro applicazione a genomi diversi.[4]

GAGE è uno strumento gratuito per la valutazione degli assemblaggi genomici. È stato utilizzato per confrontare diversi assemblatori di genomi su quattro dataset, valutando metriche come errori di assemblaggio (inversioni, rilocalizzazioni e traslocazioni). Tuttavia, GAGE è applicabile solo ai dataset con un genoma di riferimento noto e può analizzare un solo dataset alla volta, richiedendo agli utenti di combinare manualmente i risultati per confrontare più assemblatori.[5]

1.3 Limiti dei metodi attuali

I metodi come Plantagora e GAGE, pur essendo strumenti validi per la valutazione degli assemblaggi genomici, presentano alcune limitazioni significative che ne riducono l'efficacia in contesti specifici. Una delle principali restrizioni riguarda la loro inadeguatezza nell'analisi di specie mai sequenziate prima. Questo perché entrambi i metodi si basano su dati di riferimento già noti, rendendoli inutilizzabili in assenza di informazioni pregresse.

Un altro limite cruciale è la dipendenza da un genoma di riferimento completo. Per ottenere risultati accurati, Plantagora e GAGE necessitano di un riferimento ben assemblato, il che può rappresentare una sfida quando si lavora su specie per cui tali dati non sono disponibili o sono incompleti. Questa dipendenza restringe il loro campo di applicazione, rendendoli poco adatti in contesti di esplorazione genomica o per specie meno studiate.

Infine, l'applicazione di questi metodi a più dataset contemporaneamente può risultare inefficace o estremamente complessa. La gestione simultanea di più assemblaggi richiede strumenti flessibili e scalabili, qualità che questi metodi non offrono pienamente. Di conseguenza, in scenari complessi o su larga scala, possono emergere difficoltà pratiche significative.

Queste limitazioni evidenziano la necessità di sviluppare approcci più versatili e adattabili, in grado di superare i vincoli imposti dall'assenza di riferimenti o dalla complessità del lavoro con grandi quantità di dati.

2 Il Contributo di QUASt

Con le limitazioni dei metodi tradizionali in mente, introduciamo **QUASt**, uno strumento che rappresenta un passo avanti significativo nella valutazione degli assemblaggi genomici. QUASt si distingue per una serie di caratteristiche innovative che lo rendono estremamente versatile ed efficace.

Una delle principali qualità di QUASt è la sua **ampia gamma di metriche**. Questo strumento integra tutte le metriche rilevanti per valutare un assemblaggio, ma evita di sovraccaricare l'utente con un numero eccessivo di parametri, che potrebbero complicarne l'interpretazione. In questo modo, garantisce un equilibrio tra completezza e chiarezza.

Inoltre, QUASt è stato progettato con un'attenzione particolare alla **facilità d'uso**. La sua interfaccia è intuitiva e le visualizzazioni fornite sono rappresentative, consentendo agli utenti di analizzare i risultati in modo immediato e comprensibile, senza richiedere competenze avanzate o una lunga curva di apprendimento. Un altro aspetto rivoluzionario è la capacità di QUASt di effettuare una **valutazione senza riferimento**. A differenza di molti metodi che richiedono un genoma di riferimento completo, QUASt può analizzare la qualità degli assemblaggi anche in sua assenza. Questa funzionalità lo rende ideale per studiare specie nuove o non ancora sequenziate, ampliando notevolmente il suo campo di applicazione.

Infine, QUASt si distingue per la sua **efficienza**. Grazie all'elaborazione parallela, sfrutta al massimo le capacità dei processori multi-core, rendendolo rapido anche nell'analisi di dataset di grandi dimensioni. Questa caratteristica lo rende uno strumento altamente performante per ricercatori che lavorano con volumi elevati di dati genomici.

In sintesi, QUASt si presenta come uno strumento completo, facile da usare e straordinariamente adattabile, rappresentando un vero e proprio punto di svolta nella valutazione degli assemblaggi genomici.

2.1 Metriche valutate da QUASt

Confrontare gli assemblatori è essenziale per scegliere lo strumento più adatto a un determinato progetto. **QUASt (Quality Assessment Tool for Genome Assemblies)** è uno strumento ampiamente utilizzato per valutare e confrontare assemblaggi genomici in modo sistematico. Rispetto ad altri strumenti come Plantago o GAGE, QUASt offre maggiore flessibilità e introduce metriche avanzate che rendono la valutazione più completa e precisa.

QUASt utilizza un insieme di metriche che coprono vari aspetti della qualità e accuratezza degli assemblaggi[2]:

1. Dimensioni dei contigs e scaffold:

- Numero totale di contigs e contig più lungo.
- Nx e NGx, che rappresentano la lunghezza minima dei contigs che coprono rispettivamente una percentuale dell'assemblaggio e del genoma di riferimento.

2. Errori di assemblaggio e variazioni strutturali:

- Numero di errori di assemblaggio, come inversioni, traslocazioni o spostamenti.
- Contigs non allineati o ambiguamente mappati.

3. Rappresentazione genomica ed elementi funzionali:

- Frazione del genoma coperta dall'assemblaggio.
- Rapporto di duplicazione, per identificare aree ripetitive o ridondanti.

- Numero di geni e operoni completamente o parzialmente coperti.

4. **Metriche basate su N50:**

- N_{Ax} e N_{GAx}, che combinano N50 con informazioni di allineamento, spezzando contigs in blocchi allineati al genoma di riferimento.

3 Come QUASt valuta gli assemblatori

QUAST valuta gli assemblaggi confrontandoli con un genoma di riferimento, quando disponibile, oppure utilizzando metriche intrinseche delle sequenze per assemblaggi de novo.[2]

I risultati vengono presentati in tabelle e grafici, che includono **distribuzioni cumulative** della lunghezza dei contigs, **allineamenti** con il genoma di riferimento, evidenziando errori e variazioni, **distribuzione del contenuto GC** per identificare contaminazioni o bias.

3.1 Limiti nell’N50 e alternative proposte

QUAST introduce metriche avanzate per una valutazione più accurata della qualità dell’assemblaggio genomico:

- **Numero di contigs:** Il numero totale di contigs nell’assemblaggio.
- **Contig più lungo:** La lunghezza del contig più lungo nell’assemblaggio.
- **Lunghezza totale:** Il numero totale di basi nell’assemblaggio.
- **Nx:** La lunghezza del contig più corto, che sommato ad altri più lunghi, copre almeno l’x% della lunghezza totale dell’assemblaggio.
- **NGx, Genome Nx:** La lunghezza del contig tale che usando contigs di uguale o maggiore lunghezza si ottiene il x% della lunghezza del genoma di riferimento, invece che x% della lunghezza dell’assemblaggio.

Le seguenti metriche (eccetto **NGx**) possono essere valutate sia con che senza un genoma di riferimento. Sono anche disponibili versioni filtrate di queste metriche, limitate ai contigs di lunghezza superiore a una dimensione minima specificata, per escludere contigs troppo corti che potrebbero non essere particolarmente utili.[2] Molti ricercatori hanno sollevato preoccupazioni sul fatto che l’**N50** non rifletta accuratamente la qualità di un assemblaggio. Un aumento del punteggio N50 può, infatti, verificarsi anche quando i contig vengono uniti erroneamente, portando a una falsa impressione di un assemblaggio di alta qualità, quando in realtà ci sono errori significativi. Per affrontare questa limitazione, sono state proposte alternative più robuste, come l’e-size, che misura la dimensione di un contig contenente una base selezionata casualmente, evitando distorsioni dovute ad assemblaggi errati.

QUAST introduce anche una nuova metrica, **NA50**, che si calcola come l’N50, ma includendo solo i contig che soddisfano specifici criteri di qualità o lunghezza, ad esempio quelli sopra una determinata soglia dimensionale. Questa modifica consente di ottenere una valutazione più affidabile, riducendo il rischio di sovrastimare la qualità dell’assemblaggio.

4 Misassemblaggi e variazioni strutturali

Quando si assemblano le sequenze di DNA (contigs), può capitare che ci siano errori rispetto al genoma reale dell’organismo. QUAST è un software che permette di valutare questi errori confrontando i contigs con un genoma di riferimento conosciuto.

Se il genoma di riferimento è esattamente lo stesso di quello assemblato, ogni differenza osservata è probabilmente dovuta a errori, come problemi del software o dati di sequenziamento sbagliati (ad esempio, letture chimere, che sono sequenze combinate male). Tuttavia, a volte si utilizza un genoma di riferimento simile ma non identico. In questi casi, le differenze potrebbero essere errori oppure variazioni reali tra i genomi,

come riarrangiamenti o modifiche nelle sequenze ripetute.

Le principali metriche che QUASt calcola sono:

- **Numero di misassemblaggi:** avviene quando una parte del contig non corrisponde correttamente al genoma di riferimento. Ad esempio una parte del contig si allinea in una posizione molto distante dall'altra parte oppure le estremità del contig si sovrappongono troppo, o ancora le estremità si trovano su filamenti opposti o addirittura su cromosomi diversi. QUASt segnala quanti misassemblaggi di ogni tipo sono stati trovati.
- **Numero di contigs misassemblati:** indica quanti contigs contengono almeno un errore di misassemblaggio.
- **Lunghezza dei contigs misassemblati:** rappresenta il numero totale di basi (nucleotidi) presenti nei contigs che contengono errori.
- **Numero di contigs non allineati:** numero di contigs che non riescono a trovare alcuna corrispondenza nel genoma di riferimento. Questo può succedere se il contig è errato o rappresenta una sequenza non presente nel genoma di riferimento.
- **Numero di contigs mappati ambiguamente:** i contigs che si allineano bene a più punti nel genoma di riferimento, il che può accadere in presenza di sequenze ripetute o molto simili.

Oltre a queste metriche generali, QUASt fornisce anche rapporti dettagliati per ogni contig, classificandoli in categorie come non allineato, mappato ambiguamente, misassemblato o corretto.[2]

5 Tools genomici

L'assemblaggio genomico rappresenta un passaggio fondamentale per ricostruire l'intera sequenza di un genoma a partire da frammenti di DNA generati dalle moderne tecnologie di sequenziamento, come Illumina e PacBio. Questo processo è essenziale per studiare la struttura e la funzione dei genomi, e le tecnologie di sequenziamento di nuova generazione (NGS) hanno reso possibile produrre quantità enormi di dati a costi contenuti. Tuttavia, la brevità delle letture prodotte (25-75 basi nel caso di Illumina) e l'eterogeneità di errori o copertura rappresentano una sfida significativa. Per affrontare queste difficoltà, sono stati sviluppati numerosi strumenti di assemblaggio progettati per diversi tipi di dati e finalità. Ogni assemblatore è progettato per rispondere a specifiche esigenze, sfruttando approcci computazionali avanzati.[2]

Tra i più noti troviamo **Velvet** basato su grafi di de Bruijn, pensato per assemblare letture brevi gestendo con efficienza le ripetizioni genomiche e correggendo errori comuni, come nodi isolati o percorsi paralleli ridondanti. Questo lo rende particolarmente utile per genomi batterici o di dimensioni contenute.

SPAdes, estremamente versatile, è ottimizzato per progetti complessi come il sequenziamento a singola cellula. SPAdes utilizza grafi di de Bruijn multisized, adattando la lunghezza dei parametri in base alla copertura locale, migliorando così l'accuratezza nelle regioni difficili.

SOAPdenovo, invece, adatto a genomi di grandi dimensioni. SOAPdenovo impiega una pipeline modulare per correggere errori, costruire grafi di de Bruijn e collegare contigs in scaffold utilizzando paired-end reads. È noto per la sua capacità di assemblare regioni ripetitive e creare scaffold di notevole lunghezza.

ABYSS è progettato per dataset di grandi dimensioni, utilizza un approccio distribuito che consente di parallelizzare i calcoli su cluster, rendendolo ideale per assemblaggi su larga scala.

Canu e **Flye** entrambi progettati per letture lunghe, affrontano tassi di errore elevati tipici delle tecnologie come PacBio e Oxford Nanopore, offrendo assemblaggi accurati di genomi complessi.

Unicycler e MaSuRCA questi, invece, sono assemblatori ibridi che combinano i vantaggi delle letture brevi e lunghe, producendo assemblaggi completi e risolvendo con efficacia regioni ripetitive.

6 Panoramica del codice

Di seguito viene riportata un'analisi di più alto livello delle varie componenti che implementano il progetto QUAST. Tale analisi è stata effettuata sulla repository ufficiale del progetto QUAST, disponibile al seguente link: <https://github.com/ablab/quast>. Per un'analisi approfondita del flusso operativo del programma si rimanda al paragrafo successivo.

La repository è strutturata attorno a diversi componenti chiave. In particolare, nella cartella *quast_libs* sono presenti le librerie e i moduli Python utilizzati dal programma principale. Questi moduli contengono funzioni per il calcolo delle metriche, la gestione degli errori, gestione degli input, gestione degli output e dei report finali, e tutto ciò che necessita il programma per funzionare correttamente. Il nucleo principale del programma è costituito dal file *quast.py*: tale script gestisce l'intero flusso operativo e accetta come input file in formato FASTQ, oppure SAM/BAM, per le reads e file in formato FASTA per le sequenze; inoltre, opzionalmente, prende in input un genoma di riferimento con cui confrontarli. Tali file possono essere compressi con *zip*, *gzip* o *bzip2*. Lo script esegue il calcolo delle metriche di qualità (già discusse in precedenza nell'articolo) e di altre statistiche che aiutano a valutare quanto il montaggio sia completo ed accurato, restituendo, infine, un report dettagliato per l'utente. Lo script *quast-lg.py* implementa un'estensione di QUAST, in particolare l'estensione QUAST-LG, specifica per genomi di grandi dimensioni. Altra estensione di QUAST è rappresentata da MetaQUAST, implementata dallo script *metaquast.py*, e specifica per il mondo delle analisi metagenomiche.

Le funzionalità specifiche dell'applicazione sono raccolte in *quast_libs*: in questa libreria vi sono sia script nativi che librerie esterne. In particolare, alcune delle librerie esterne sono *minimap2*, ovvero un allineatore di sequenze *general-purpose*[6], utilizzato di default nella pipeline di QUAST, *GeneMarkS*, che è un algoritmo di predizione genetica[7], *Barrnap*, uno strumento progettato per prevedere la posizione dei geni dell'rRNA nei genomi, *BUSCO*, utilizzato per valutare la qualità e la completezza delle annotazioni genomiche[8] e *Glimmer*, un sistema avanzato per l'identificazione di geni in DNA microbico, che utilizza Interpolated Markov Models per distinguere le regioni codificanti da quelle non codificanti nel DNA[9]. Vi è, ancora, *bedtools*, ovvero un insieme di strumenti utilitari applicabili ad una ampia gamma di task genomici, utilizzato da QUAST per calcolare la copertura delle letture, sia *raw* che fisica, sugli assemblaggi genomici. Infine, per l'identificazione delle variazioni strutturali nel genoma assemblato, QUAST può utilizzare *BWA*, *Sambamba*, oppure *GRIDSS*. Verranno, di seguito, illustrati i principali script nativi presenti nella suddetta libreria.

6.1 Utility generali

I principali script che implementano funzioni utilitarie o di configurazione sono:

- **qconfig.py**: controlla la versione di Python e gestisce le informazioni della configurazione di QUAST.
- **quils.py**: offre funzioni per la verifica della validità dei percorsi delle directory, per la gestione dei file di lettura e di riferimento, funzioni per la correzione dei nomi dei contigs (possono esservi dei caratteri speciali che fanno fallire determinati tools) e numerose altre funzioni utilitarie.
- **options_parser.py**: si occupa dell'analisi e della gestione delle opzioni della riga di comando.
- **log.py**: configura e gestisce il logging dell'applicazione.

6.2 Generazione di statistiche e reports

Relativamente al calcolo delle metriche, alla generazione delle statistiche e dei reports, i principali script sono:

- **N50.py**: offre le funzioni per il calcolo della metrica N50.
- **plotter.py**, *plotter_data.py*: responsabili della creazione di grafici e visualizzazione dei dati.
- **reads_analyzer.py**: analizza i file contenenti le *reads* per generare informazioni sulla copertura (*coverage*).
- **contigs_analyzer.py**: fornisce funzionalità per analizzare i contig genomici, identificare misassemblaggi, calcolare statistiche di copertura e generare report dettagliati.
- **genome_analyzer.py**: analizza le caratteristiche genomiche come operoni (insieme di geni che vengono regolati in modo strettamente coordinato[10]), funzionalità codificanti e altre proprietà del genoma.
- **reporting.py**: genera report dettagliati sui risultati delle analisi.
- **unique_kmers.py**: analizza k-mer unici per valutare la qualità e l'unicità dell'assemblaggio.

6.3 Visualizzazione dei risultati

Per la visualizzazione dei risultati, vi sono *icarus.py* e *circos.py*. Icarus è uno strumento interattivo responsabile della visualizzazione dei risultati dell'assemblaggio genomico. È, fondamentalmente, un browser di contig che permette di esaminare visivamente l'allineamento dei contig al genoma di riferimento. Circos, invece, è uno strumento per la visualizzazione dei dati genomici in formato circolare. Questo tipo di rappresentazione è molto utile per grandi quantità di dati, poiché possono essere visualizzate in maniera compatta e intuitiva.

7 Analisi del codice

In questa sezione verrà analizzato il flusso operativo del programma, gestito interamente da *quast.py*, che si avvale delle librerie appena presentate per lo svolgimento del lavoro.

7.1 Flusso Operativo

Come già accennato, *quast.py* è il nucleo principale di QUASt. Tale script gestisce le opzioni di configurazione ed inizializzazione, il processing dei genomi di riferimento, dei contigs, è responsabile dell'allineamento contigs al genoma di riferimento utilizzando *Minimap2*, della predizione dei geni e del calcolo, nonché della visualizzazione delle statistiche e dei reports.

7.1.1 Fase di Import e Configurazione

La prima sezione dello script è responsabile dell'importazione delle librerie esterne e delle librerie native di QUASt, implementate in *quast_libs*. In questa sezione, inoltre, viene controllata la versione di Python mediante la funzione `check_python_version()` definita in *qconfig.py*, e vengono processati gli argomenti, che vengono convertiti in azioni del programma. Infine, vengono inizializzati i report.

7.1.2 Processamento del genoma di riferimento

Se al programma è stato fornito il genoma di riferimento (il path del file contenente tale genoma), ne viene innanzitutto effettuata la correzione (se non è stata specificata l'opzione `--no-check-meta` all'avvio del programma) mediante la funzione `correct_reference()` di *qutils.py*. Tale correzione consiste nella rimozione di sequenze che sono più corte di una certa soglia (che può essere specificata mediante l'opzione `--min-contig <int>`) e che contengono caratteri non validi (diversi da A, C, G, T, N). Se il file non contiene sequenze valide, registra un errore e interrompe l'esecuzione. Successivamente, viene controllato se l'opzione relativa all'*optimal assembly* è attiva. In tal caso viene verificata la presenza di tutti i dati necessari per la generazione dell'assemblaggio ottimale, in particolare, deve essere presente almeno uno tra *mate pairs* e *long reads* (quest'ultime provenienti da PacBio, oppure da Oxford Nanopore), altrimenti viene lanciato un messaggio di errore. Se non è stato generato alcun errore, il codice prosegue con la generazione dell'assemblaggio ottimale mediante la funzione `optimal_assembly.do()` della libreria *optimal_assembly* di *quast_libs*, che prende come argomenti il file del genoma di riferimento (la versione corretta dalla funzione citata poc'anzi), il file originale del genoma di riferimento e il percorso di output, costruito combinando la directory di output con il nome base specificato in `qconfig.optimal_assembly_basename`. Infine, se l'assemblaggio ottimale è stato creato con successo, il percorso del file di tale assemblaggio viene aggiunto all'inizio della lista dei contig e viene inserita la label *UpperBound* all'inizio della lista delle label, che vengono rielaborate mediante la funzione `qutils.process_labels(contigs_fpaths, labels)`.

7.1.3 Processamento dei contigs

A questo punto, il programma si occupa della correzione dei contigs e della loro preparazione per la successiva fase di analisi. In particolare, viene impiegata la funzione `qutils.correct_contigs()`. Tale funzione, principalmente rimuove eventuali caratteri speciali dai nomi dei contigs per evitare errori durante l'elaborazione, infatti tali caratteri potrebbero dare problemi con strumenti embedded o con strumenti come Nucmer. La peculiarità di questa funzione è che permette di eseguire le elaborazioni dei contigs in **parallelo**, permettendo di suddividere il carico di lavoro e di sfruttare al meglio le risorse del sistema. È infatti possibile specificare il numero di thread che è possibile utilizzare per il processing tramite l'opzione `--threads <int>`.

Se sono presenti delle reads, inoltre, ne effettua un'analisi mediante il modulo `reads_analyzer`, calcola le statistiche di copertura (ad esempio, il numero di basi coperte nel genoma di riferimento, il numero di SNPs, ovvero variazioni a singolo nucleotide, numero di inserizioni e delezioni) e salva il file BED generato per la visualizzazione della copertura. Infine, conta quanti file di contigs sono stati processati e salva questo valore in `qconfig.assemblies_num`, che tiene traccia, appunto, del numero di assemblaggi genomici che devono essere analizzati.

7.1.4 Allineamento e analisi

Se è stato fornito un genoma di riferimento, si procede con l'allineamento dei contigs a quest'ultimo. Il modulo responsabile è `quast_libs.contigs_analyzer`. Innanzitutto viene determinato se il genoma è ciclico e, successivamente, viene invocata la funzione `do()` del modulo appena importato per effettuare l'allineamento effettivo. Vengono restituiti due dizionari: `aligner_statuses`, che indica per ogni contig se l'allineamento è andato a buon fine, e `aligned_lengths_per_fpath`, che contiene le lunghezze degli allineamenti per ciascun contig. Il programma, dunque, itera su tutti i contigs per verificare se l'allineamento è andato a buon fine, e aggiunge quelli il cui riscontro è stato positivo alla lista `aligned_contigs_fpaths`. Se non ci sono contigs allineati, non ha senso proseguire con l'analisi, che viene terminata. A questo punto si passa al calcolo delle metriche **Nx** e **NGx** basate sugli allineamenti. Viene importato il modulo `aligned_stats`, che calcola le metriche di allineamento mediante la funzione `do()`. In particolare, **NAx** (Aligned Nx) misura la lunghezza minima dei contigs allineati che coprono x% dell'assemblaggio allineato (dove x è la percentuale dell'assemblaggio, o del genoma di riferimento, costituita da contigs di una certa lunghezza o più lunghi), mentre **NGAx** (Aligned NGx) è simile alla precedente metrica, ma riferita al genoma di riferimento. Se è stata fornita l'opzione `--glimmer`, viene utilizzato il software Glimmer per la predizione genica. Se è stata, invece, fornita l'opzione `--gene-finding`, viene utilizzato GeneMark. I geni risultanti vengono salvati in `predicted_genes`. A questo punto, se l'opzione `--rna-gene-finding` è attivata, viene utilizzato Barrnap, un tool specifico per l'identificazione dei geni codificanti per RNA ribosomiale. Successivamente, se il sistema operativo in cui si sta eseguendo il programma è Linux, viene eseguito BUSCO sui contigs assemblati per valutarne la completezza, altrimenti continua senza usare BUSCO.

Infine, nell'ultima sezione di codice, vengono rappresentati tutti i risultati ottenuti dalle precedenti fasi.

7.1.5 Analisi della metrica NG50

Di seguito verrà approfondita in dettaglio la metrica NGx, in particolare la **NG50**. Innanzitutto, tale metrica rappresenta la **lunghezza** del contig **più corto** che, sommato ai contigs di lunghezza **maggiore o uguale**, copre almeno l'*x* percento della lunghezza del genoma di riferimento.[2]

La funzione che calcola tale metrica è la funzione `NG50_and_LG50`, definita nel file `N50.py`, ed invocata dalla funzione `do()` del file `aligned_stats.py`. Come argomento prende la lista delle lunghezze dei contigs, `numlist` (misurata in coppie di basi, bp), la lunghezza totale del genoma di riferimento, `reference_length`, la percentuale della lunghezza del genoma che si intende coprire, `percentage` (impostata di default a 50, trattandosi di NG50), e un valore booleano, `need_sort`, che specifica se la lista delle lunghezze dei contigs debba essere ordinata, oppure no. Tale valore è impostato di default a `False`, siccome viene già passata una lista ordinata in ordine decrescente di lunghezza.

Innanzitutto, la funzione controlla che il valore percentuale sia compreso tra 0 e 100 e, se la lista necessita di ordinamento, ordina la lista in ordine decrescente mediante la funzione `sort` di Python.

A questo punto, assegna alla variabile `s` la lunghezza totale del genoma di riferimento. Questa variabile verrà utilizzata per tenere traccia del valore residuo da confrontare con la variabile `limit`, ovvero il valore

di **s** che è necessario raggiungere. Quest'ultimo valore, in particolare, rappresenta la lunghezza necessaria per coprire *almeno* l' $x\%$ della lunghezza del genoma di riferimento, calcolata, banalmente, con la seguente formula:

$$\text{limit} = \text{genLen} \times (100 - 50)$$

dove **genLen** rappresenta la lunghezza del genoma di riferimento, espresso in bp.

Viene, inoltre, inizializzata la variabile **lg50** a 0, che conterrà il numero di contigs necessari per raggiungere NG50.

Infine, viene eseguito il ciclo **for** che calcola la metrica nel seguente modo: scorre tutti i contigs in ordine decrescente, dove **l** è la lunghezza del contig corrente, sottrae la lunghezza del contig corrente da **s** ed incrementa il valore di **lg50**. Se il valore di **s** scende al di sotto il **limit**, allora il 50% della lunghezza del genoma è stata coperta, dunque NG50 sarà il valore della lunghezza del contig corrente, e verranno restituiti NG50 e LG50. Se, invece, il ciclo termina senza che il valore di **s** sia sceso al di sotto della soglia, la funzione restituisce **None**, **None**. Di seguito viene riportato il ciclo **for**, nonché il cuore del calcolo della metrica:

```
for l in numlist:
    s -= l
    lg50 += 1
    if s <= limit:
        ng50 = l
        return ng50, lg50

return None, None
```

8 Galaxy Workflow

L'obiettivo di questa analisi è confrontare le performance di due strumenti di assemblaggio genomico, **SPAdes** e **Velvet**, valutandone l'accuratezza e la qualità nella ricostruzione del genoma di *Escherichia coli*. Per condurre questa valutazione, è stato seguito un workflow ben definito, comprendente la selezione e il download dei dati di sequenziamento, l'assemblaggio delle letture mediante i due algoritmi e, infine, la valutazione degli assemblaggi ottenuti attraverso **QUAST**.

La prima fase del lavoro ha riguardato l'acquisizione dei dati di sequenziamento. A tal fine, è stato selezionato un dataset di *Escherichia coli* proveniente dal database pubblico **ENA (European Nucleotide Archive)**, nello specifico il dataset **SRR941218**. Questo dataset è stato scelto per diversi motivi. In primo luogo, contiene letture *paired-end*, che permettono di ottenere informazioni complementari su entrambe le estremità dei frammenti di DNA, migliorando così la qualità dell'assemblaggio rispetto alle letture single-end. Inoltre, il sequenziamento *paired-end* consente di ricostruire con maggiore accuratezza le regioni ripetute e di colmare eventuali gap tra le sequenze assemblate, riducendo gli errori e migliorando la qualità complessiva dell'assemblaggio.

Una volta individuato il dataset, sono stati scaricati i file contenenti le letture forward e reverse, ovvero **SRR941218_1.fastq.gz** e **SRR941218_2.fastq.gz**. Questi file, essendo compressi nel formato **.gz**, sono stati mantenuti in tale formato per facilitare l'elaborazione. La piattaforma Galaxy, utilizzata per l'analisi, permette infatti di gestire direttamente i file compressi senza necessità di decompressione manuale.

Per valutare l'accuratezza degli assemblaggi, è stato impiegato un **genoma di riferimento** con cui confrontare le sequenze assemblate. A tal fine, è stato selezionato il genoma completo CP135227, disponibile nel database ENA, che rappresenta la sequenza di riferimento di *Escherichia coli*. L'utilizzo di un genoma di riferimento consente di calcolare parametri chiave per la valutazione della qualità dell'assemblaggio, tra cui la percentuale del genoma ricostruita, il numero di errori e la qualità complessiva delle sequenze ottenute.

Una volta acquisiti tutti i dati necessari, questi sono stati caricati sulla piattaforma **Galaxy**, scelta per la sua capacità di gestire in modo intuitivo e riproducibile flussi di lavoro bioinformatici complessi, evitando la necessità di installare software dedicati a livello locale. I file caricati su Galaxy includono:

- I file **FASTQ.gz** contenenti le letture di sequenziamento;
- Il file **FASTA** del genoma di riferimento;
- I file di output degli assemblaggi generati.

L'assemblaggio delle sequenze è stato eseguito utilizzando **SPAdes** e **Velvet**. Questi due strumenti sono tra i più utilizzati per l'assemblaggio di genomi batterici, ma adottano strategie differenti. **SPAdes** utilizza un approccio **multi-kmer**, che migliora la qualità dell'assemblaggio variando la lunghezza dei k-mer utilizzati nella costruzione del grafo, mentre **Velvet** si basa su un approccio a grafi **de Bruijn**, più rigido nella selezione dei parametri ma comunque efficace. Il confronto tra i due algoritmi consente quindi di valutare quale produca un assemblaggio più accurato, caratterizzato da contigs più lunghi e meno frammentati.

L'assemblaggio effettuato su Galaxy ha prodotto due file in formato **FASTA**, contenenti rispettivamente i contigs ottenuti da **SPAdes** e quelli ottenuti da **Velvet**. Per un confronto oggettivo dei risultati, è stato impiegato **QUAST** (Quality Assessment Tool for Genome Assemblies), uno strumento che calcola diverse metriche per valutare la qualità dell'assemblaggio.

Nell'analisi con **QUAST**, sono stati selezionati i file **FASTA** generati dagli assemblatori e il genoma di riferimento precedentemente scaricato. Tra le metriche calcolate, le più significative includono **N50**, che indica la lunghezza media dei contigs e rappresenta un parametro della frammentazione dell'assemblaggio;

Genome fraction, che misura la percentuale del genoma di riferimento coperta dai contigs ottenuti e infine il **Numero di misassemblaggi**, che rappresenta gli errori di assemblaggio rispetto alla sequenza di riferimento.

8.1 Risultati

Analizziamo adesso i risultati ottenuti confrontando le prestazioni di **SPAdes** e **Velvet** secondo diverse metriche di qualità dell'assemblaggio.

8.1.1 Metriche legate al numero di contig

Un parametro chiave per valutare la frammentazione dell'assemblaggio è il **numero totale di contig generati**. Un numero più elevato di contig indica una maggiore frammentazione, mentre un valore più basso suggerisce un assemblaggio più continuo e affidabile.

L'analisi mostra che **Velvet ha prodotto 318 contig**, mentre **SPAdes ne ha generati solo 190**. Questa differenza evidenzia come l'assemblaggio ottenuto con SPAdes sia meno frammentato e abbia una maggiore capacità di ricostruire il genoma in sequenze più lunghe e ben collegate. Al contrario, il numero più elevato di contig in Velvet suggerisce una maggiore dispersione delle sequenze, rendendo l'interpretazione genomica più complessa.

Un altro parametro importante è la **lunghezza del contig più lungo ottenuto dall'assemblaggio**. **Velvet ha prodotto un contig lungo 289.097 bp**, mentre **SPAdes ha raggiunto 313.898 bp**. Questo risultato conferma che SPAdes è stato in grado di ricostruire regioni più ampie del genoma senza frammentarle, migliorando la qualità dell'assemblaggio complessivo. La capacità di generare contig più lunghi è un indicatore di una maggiore efficienza dell'algoritmo di assemblaggio nel collegare le sequenze in modo accurato.

Analizzando la **lunghezza totale delle sequenze ottenute**, emerge che **Velvet ha prodotto un assemblaggio leggermente più lungo (5.445.933 bp) rispetto a SPAdes (5.293.223 bp)**. Tuttavia, questa differenza non deve essere interpretata come un vantaggio per Velvet. Una maggiore lunghezza totale può infatti essere dovuta alla frammentazione dell'assemblaggio, piuttosto che a una migliore copertura del genoma. Il fatto che Velvet abbia una lunghezza totale superiore, ma al tempo stesso un numero di contig molto più elevato, indica che l'assemblaggio è meno efficiente e più disgregato rispetto a quello di SPAdes.

Per valutare la **continuità dell'assemblaggio**, si analizzano metriche come **N50** e **N90**, che indicano la lunghezza del contig più corto tra quelli che, sommati, coprono rispettivamente il 50% e il 90% dell'assemblaggio. **SPAdes ha ottenuto un N50 di 124.687 bp, più del doppio rispetto ai 59.880 bp di Velvet**. Questo significa che le sequenze ottenute da SPAdes sono mediamente più lunghe e meno frammentate, contribuendo a una ricostruzione più continua del genoma. Anche il valore di **N90 conferma questa tendenza: SPAdes ha un N90 di 28.229 bp, mentre Velvet si ferma a 9.283 bp**. Un valore più alto indica che una maggiore porzione del genoma è stata assemblata in contig più lunghi, riducendo la frammentazione.

Un ulteriore parametro utile per valutare la qualità dell'assemblaggio è **NG50**, che, a differenza di N50, considera la lunghezza attesa del genoma di riferimento invece della lunghezza totale dell'assemblaggio. Questo permette di ottenere un'indicazione più realistica della completezza e della continuità dell'assemblaggio rispetto al genoma originale.

L'analisi mostra che **SPAdes ha ottenuto un NG50 di 127.550 bp, mentre Velvet ha raggiunto solo 64.016 bp**, confermando ancora una volta la maggiore continuità e qualità dell'assemblaggio di SPAdes. Anche il valore di **NG90**, che indica la lunghezza del contig più corto tra quelli che coprono almeno il 90%

della lunghezza attesa del genoma, riflette una netta superiorità di SPAdes, con un valore di 37.419 bp, più del doppio rispetto ai 18.041 bp di Velvet.

Questi dati indicano che SPAdes produce un assemblaggio non solo più lungo e meno frammentato, ma anche più vicino alla lunghezza attesa del genoma di riferimento.

Statistics without reference		Statistics without reference	
# contigs	190	# contigs	318
# contigs (>= 0 bp)	1113	# contigs (>= 0 bp)	747
# contigs (>= 1000 bp)	134	# contigs (>= 1000 bp)	258
Largest contig	313 898	Largest contig	289 097
Total length	5 293 223	Total length	5 445 933
Total length (>= 0 bp)	5 414 009	Total length (>= 0 bp)	5 553 552
Total length (>= 1000 bp)	5 253 566	Total length (>= 1000 bp)	5 402 552
N50	124 687	N50	59 880
N90	28 229	N90	9283

Figure 1: SPAdes

Figure 2: Velvet

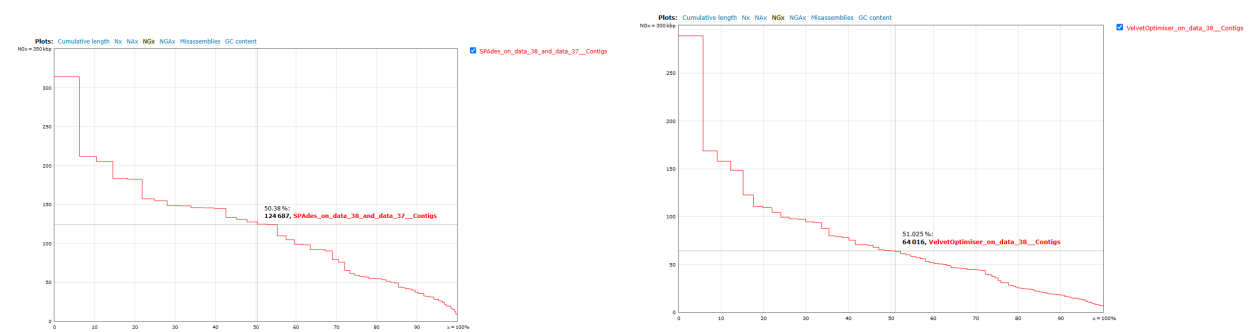


Figure 3: Confronto tra la distribuzione di NGx per gli assemblaggi ottenuti con SPAdes e Velvet.

8.1.2 Errori di misassemblaggio

In questa analisi confrontiamo quattro parametri chiave che ci permettono di valutare la correttezza dell’assemblaggio. Il **numero di misassemblaggi** è un indicatore chiave della qualità strutturale dell’assemblaggio. In generale, meno errori significano un’assemblaggio più affidabile e accurato.

L’analisi mostra che **SPAdes ha generato 133 misassemblaggi, mentre Velvet ne ha prodotti 150**. Anche se la differenza non è enorme, SPAdes risulta più preciso, con 17 errori in meno. Questo suggerisce che l’algoritmo di SPAdes è stato più efficace nel ricostruire le sequenze genomiche senza introdurre errori significativi.

Oltre al numero totale di errori, è importante valutare **quanti contig sono stati coinvolti nei misassemblaggi**. Un numero elevato di contig errati può compromettere la qualità dell’intero assemblaggio e rendere più difficile l’interpretazione del genoma.

I dati mostrano che **SPAdes ha prodotto 41 contig misassemblati, mentre Velvet ne ha generati 74**. Questo significa che gli errori in SPAdes si concentrano su un numero più limitato di contig, mentre

in Velvet sono distribuiti su un numero maggiore di frammenti. Un assemblaggio con meno contig errati è generalmente più stabile e affidabile, confermando la migliore qualità di SPAdes rispetto a Velvet.

Alcuni contig possono **non trovare una corretta corrispondenza con il genoma di riferimento**, segnalando possibili errori nell’assemblaggio o la presenza di sequenze spurie.

In questo caso, **SPAdes ha prodotto solo 12 contig non allineati, mentre Velvet ne ha generati 22**. Questo è un altro punto a favore di SPAdes, che ha fornito un’assemblaggio con meno sequenze problematiche. Una minore presenza di contig non allineati indica una migliore precisione nel ricostruire la struttura del genoma.

Infine, è utile valutare la **lunghezza complessiva delle sequenze che contengono errori di misassemblaggio**. Inaspettatamente, **SPAdes ha una lunghezza totale di contig misassemblati leggermente superiore a quella di Velvet (4.055.419 bp contro 3.759.719 bp)**.

Questa apparente discrepanza può essere spiegata dal fatto che, mentre Velvet ha un numero maggiore di contig errati, in SPAdes gli errori sono concentrati in un minor numero di contig più lunghi. Questo è un dettaglio importante: avere pochi contig con errori è preferibile rispetto ad avere tanti frammenti con piccoli errori, poiché i primi possono essere più facilmente corretti o migliorati con strumenti di post-processing. Dall’analisi delle metriche relative ai misassemblaggi emerge chiaramente che SPAdes ha fornito un assemblaggio più preciso e strutturalmente più corretto rispetto a Velvet

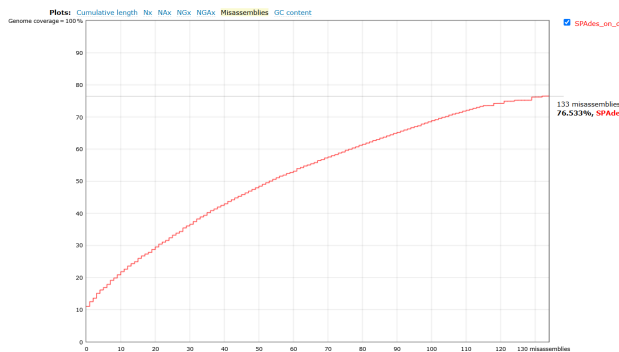


Figure 4: Misassemblaggio SPAdes

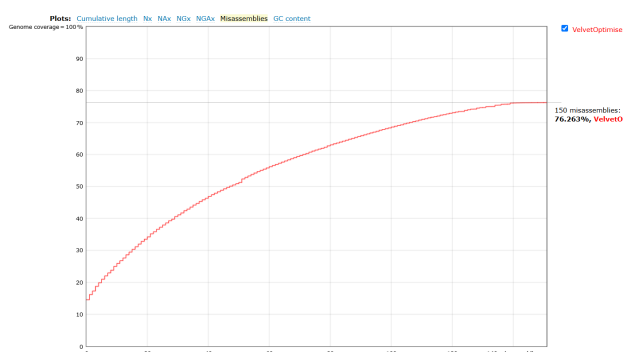


Figure 5: Misassemblaggio Velvet

Misassemblies	
# misassemblies	133
# relocations	133
# translocations	0
# inversions	0
# misassembled contigs	41
Misassembled contigs length	4 055 419
# local misassemblies	79
# scaffold gap ext. mis.	0
# scaffold gap loc. mis.	0
# unaligned mis. contigs	12

Figure 6: SPAdes

Misassemblies	
# misassemblies	150
# relocations	143
# translocations	0
# inversions	7
# misassembled contigs	74
Misassembled contigs length	3 759 719
# local misassemblies	73
# scaffold gap ext. mis.	0
# scaffold gap loc. mis.	0
# unaligned mis. contigs	22

Figure 7: Velvet

8.1.3 Rappresentazione genomica

Uno degli aspetti fondamentali per valutare la qualità di un assemblaggio genomico è la **sua capacità di ricostruire il genoma originale, misurata attraverso la percentuale del genoma ricostruito (Genome Fraction %)** e il **rapporto di duplicazione (Duplication Ratio)**. Queste metriche ci permettono di comprendere quanto dell'effettivo genoma è stato coperto dall'assemblaggio e se si sono verificate duplicazioni indesiderate, che potrebbero indicare errori nell'assemblaggio.

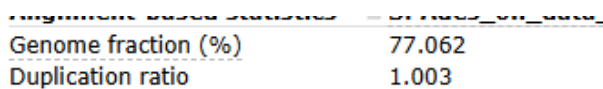
Il primo parametro da analizzare è la **percentuale del genoma ricostruito**, che rappresenta la frazione del DNA originale che è stata correttamente coperta dall'assemblaggio.

Dai risultati ottenuti emerge che **SPAdes ha ricostruito il 77,06% del genoma, mentre Velvet ha raggiunto il 76,73%**. Sebbene la differenza tra i due assemblatori sia minima, SPAdes mostra comunque un lieve vantaggio, riuscendo a coprire una porzione leggermente più ampia del genoma. Questo suggerisce che SPAdes ha avuto una maggiore capacità di recuperare sequenze utili e di assemblarle in modo efficace.

Un altro aspetto fondamentale da considerare è il **rapporto di duplicazione**, che ci dice se durante l'assemblaggio alcune regioni del genoma sono state ripetute erroneamente. Un valore vicino a 1 indica che l'assemblaggio non ha generato duplicazioni spurie, mentre valori più elevati possono segnalare problemi nel collegamento delle sequenze.

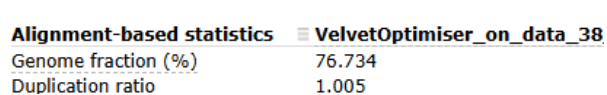
I dati mostrano che **SPAdes ha un rapporto di duplicazione di 1,003, mentre Velvet ha un valore di 1,005**. Questa differenza è trascurabile, il che significa che entrambi gli assemblatori hanno gestito molto bene il problema delle duplicazioni, evitando di introdurre errori strutturali significativi.

L'analisi della rappresentazione genomica suggerisce che SPAdes offre un leggero vantaggio rispetto a Velvet.



Alignment-based statistics	SPAdes_on_data_38
Genome fraction (%)	77.062
Duplication ratio	1.003

Figure 8: SPAdes



Alignment-based statistics	VelvetOptimiser_on_data_38
Genome fraction (%)	76.734
Duplication ratio	1.005

Figure 9: Velvet

8.1.4 Metriche basate su N50

Per valutare la qualità di un assemblaggio genomico, non basta considerare solo la lunghezza dei contig, ma è fondamentale capire quanto siano affidabili e quanto si avvicinino alla reale organizzazione del genoma. Due metriche particolarmente utili per questa valutazione sono NA50 e NGA50.

La metrica **NA50** è una versione più rigorosa del classico N50, perché considera solo i contig che non presentano errori evidenti e che rispettano criteri di qualità specifici.

Dai dati analizzati emerge che **SPAdes ha un NA50 di 21.734 bp mentre Velvet ha un NA50 di 14.711 bp**. SPAdes ha un valore significativamente più alto, il che significa che i contig prodotti sono mediamente più lunghi e meno frammentati rispetto a quelli di Velvet.

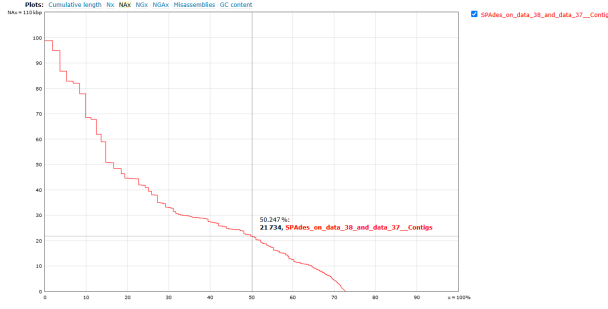


Figure 10: NA50 SPAdes

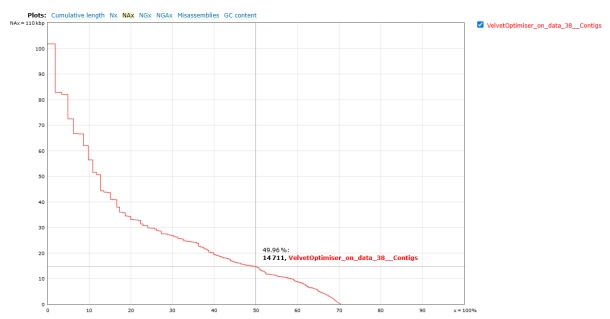


Figure 11: NA50 Velvet

Se NA50 misura la qualità dei contig in senso generale, **NGA50** offre un'indicazione ancora più accurata, perché tiene conto dell'allineamento delle sequenze rispetto al genoma di riferimento. I risultati mostrano che **SPAdes ha un NGA50 di 23.925 bp e Velvet ha un NGA50 di 16.119 bp**. Anche in questo caso, SPAdes si dimostra superiore, con un valore nettamente più alto.

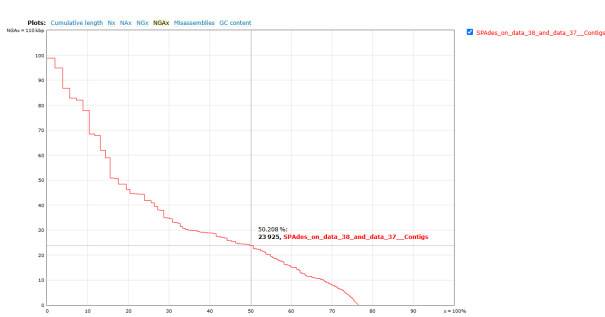


Figure 12: NGA50 SPAdes

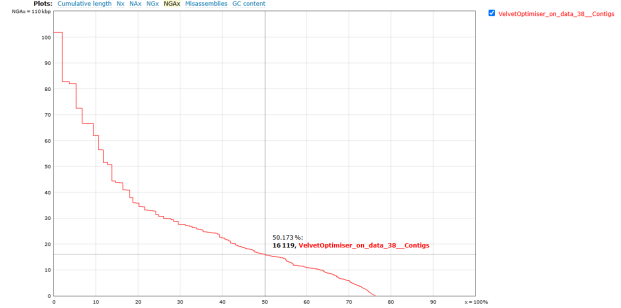


Figure 13: NGA50 Velvet

8.1.5 Conclusioni

Dall'analisi dettagliata delle diverse metriche di qualità dell'assemblaggio, emerge chiaramente che **SPAdes offre prestazioni superiori rispetto a Velvet** in diversi aspetti chiave dell'assemblaggio genomico. L'evidenza di questa superiorità è supportata da numerosi parametri che dimostrano una maggiore continuità delle sequenze, una minore frammentazione, una ridotta presenza di errori strutturali e una migliore rappresentazione del genoma di riferimento.

Uno degli aspetti più rilevanti è la **frammentazione dell'assemblaggio**, che risulta notevolmente inferiore in SPAdes rispetto a Velvet. Il numero totale di contig generati è significativamente più basso, suggerendo che l'assemblaggio di SPAdes riesce a ricostruire porzioni più ampie del genoma in sequenze lunghe e ben collegate, riducendo la dispersione delle informazioni genetiche. Contig più lunghi e meno frammentati permettono una migliore interpretazione del genoma e facilitano le analisi funzionali.

La **continuità dell'assemblaggio** è un altro parametro in cui SPAdes dimostra un netto vantaggio. Metriche come N50, NG50 e NG90 confermano che i contig ottenuti da SPAdes sono mediamente più lunghi e

meglio organizzati, mentre in Velvet l'assemblaggio appare più frammentato e discontinuo. Questo significa che l'output di SPAdes fornisce una ricostruzione più affidabile e più utile per studi che richiedono una rappresentazione strutturale dettagliata del genoma.

Un elemento fondamentale nella valutazione dell'assemblaggio è anche la **correttezza strutturale delle sequenze ottenute**, misurata attraverso il numero di errori di misassemblaggio. Anche in questo caso, SPAdes dimostra una maggiore precisione rispetto a Velvet. Non solo genera meno errori globali, ma distribuisce gli eventuali misassemblaggi su un numero minore di contig, riducendo il rischio di distorsioni nella struttura genomica. Il numero di contig misassemblati e non allineati è nettamente inferiore, garantendo un'analisi più accurata e riducendo la necessità di correzioni post-assemblaggio.

Dal punto di vista della **rappresentazione genomica**, SPAdes offre una copertura leggermente superiore rispetto a Velvet, riuscendo a ricostruire una percentuale maggiore del genoma di riferimento. Sebbene la differenza tra i due assemblatori non sia enorme, ogni incremento nella genome fraction è significativo, soprattutto quando si lavora con organismi il cui genoma non è completamente noto. Inoltre, il rapporto di duplicazione è rimasto stabile e simile tra i due assemblatori, suggerendo che entrambi riescono a limitare la generazione di sequenze ripetute in modo spurio.

Infine, l'analisi delle metriche **NA50 e NGA50**, che valutano rispettivamente la qualità dell'assemblaggio rispetto ai contig validi e l'allineamento con il genoma di riferimento, conferma che SPAdes produce un output più accurato e meglio strutturato. I valori significativamente più alti rispetto a Velvet indicano che i contig generati non solo sono più lunghi, ma sono anche più fedeli alla reale organizzazione del genoma, riducendo la presenza di errori strutturali come riarrangiamenti e frammentazioni errate.

Riferimenti

- [1] Adam Thrash, Federico Hoffmann, and Andy Perkins. “Toward a more holistic method of genome assembly assessment”. In: *BMC bioinformatics* 21.Suppl 4 (2020), p. 249.
- [2] Alexey Gurevich et al. “QUAST: quality assessment tool for genome assemblies”. In: *Bioinformatics* 29.8 (2013), pp. 1072–1075.
- [3] Roger Barthelson et al. “Plantagora: modeling whole genome sequencing and assembly of plant genomes”. In: *PLoS One* 6.12 (2011), e28436.
- [4] Dent Earl et al. “Assemblathon 1: a competitive assessment of de novo short read assembly methods”. In: *Genome research* 21.12 (2011), pp. 2224–2241.
- [5] Steven L Salzberg et al. “GAGE: A critical evaluation of genome assemblies and assembly algorithms”. In: *Genome research* 22.3 (2012), pp. 557–567.
- [6] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (May 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/18/3094/48919122/bioinformatics_34_18_3094.pdf. URL: <https://doi.org/10.1093/bioinformatics/bty191>.
- [7] *GeneMark - Wikipedia* — *en.wikipedia.org*. <https://en.wikipedia.org/wiki/GeneMark>. [Accessed 02-01-2025].
- [8] Mosè Manni et al. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes”. In: *Molecular Biology and Evolution* 38.10 (July 2021), pp. 4647–4654. ISSN: 1537-1719. DOI: 10.1093/molbev/msab199. eprint: <https://academic.oup.com/mbe/article-pdf/38/10/4647/40449445/msab199.pdf>. URL: <https://doi.org/10.1093/molbev/msab199>.
- [9] *Glimmer* — *ccb.jhu.edu*. <https://ccb.jhu.edu/software/glimmer/index.shtml>. [Accessed 02-01-2025].
- [10] *Operone - Wikipedia* — *it.wikipedia.org*. <https://it.wikipedia.org/wiki/Operone>. [Accessed 02-01-2025].