



QUAST

Chiara Coscarelli
Marco Santoriello

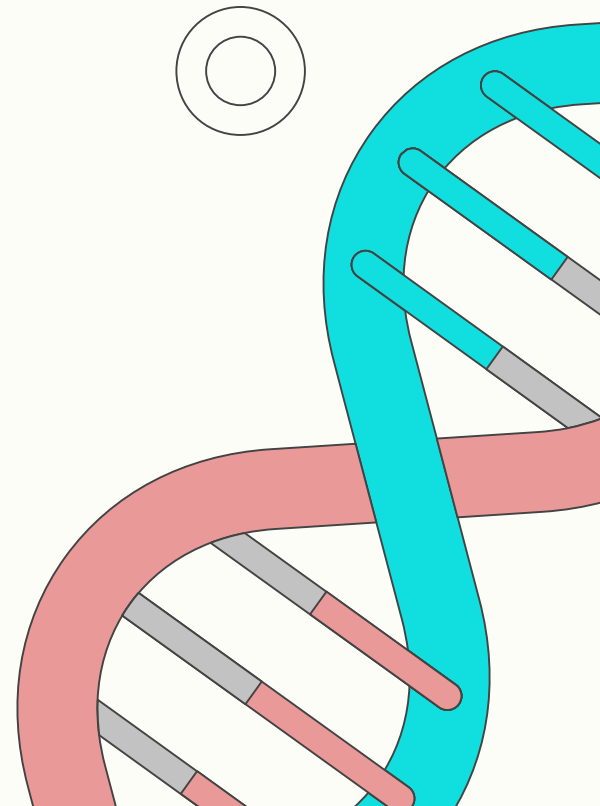




Table of contents

01

Assemblaggio

02

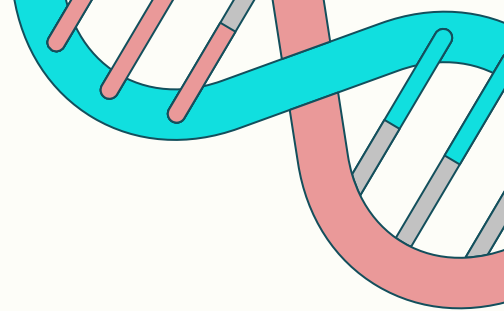
QUAST

03

Analisi del codice

04

Galaxy workflow





01

Assemblaggio



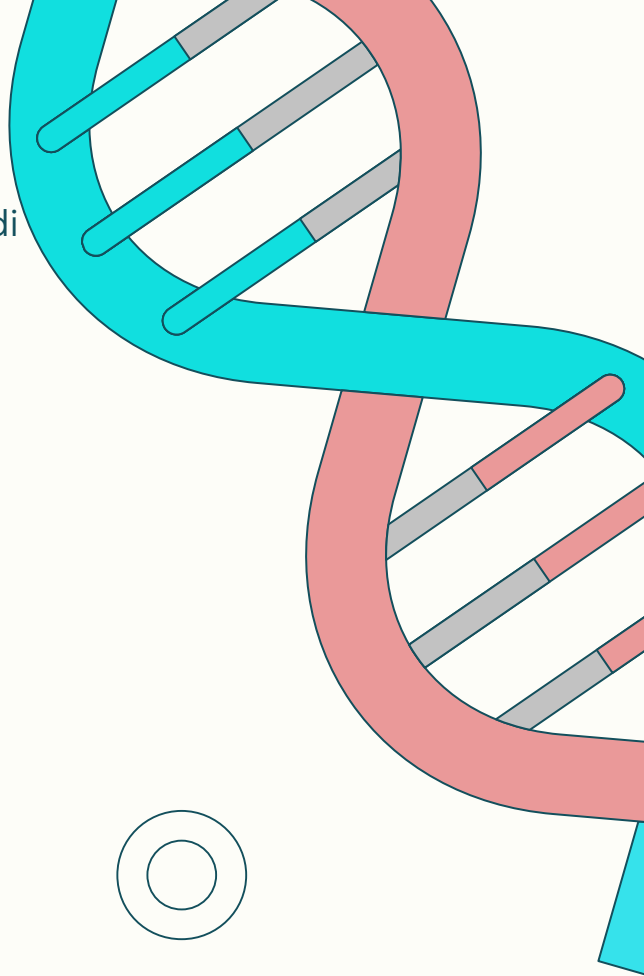
Assemblaggio

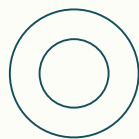
L'**assemblaggio del genoma** è il processo di unione di frammenti di DNA (*reads*) per creare sequenze più lunghe (**contig**).

È essenziale per analizzare la struttura e la funzionalità del genoma, ma la sua qualità è difficile da misurare.

Metriche di Valutazione

- ◆ **Contiguità**: lunghezza e numero di contig. Un buon assemblaggio ha pochi contig molto lunghi.
- ◆ **Completezza**: misura quanto l'assemblaggio copra il genoma di riferimento.
- ◆ **Correttezza**: verifica l'ordine e la posizione dei contig rispetto al genoma reale.





Metodi esistenti per il confronto



Plantagora

Piattaforma che aiuta i ricercatori a confrontare diverse strategie di sequenziamento e assemblaggio del genoma delle piante.

LIMITI:

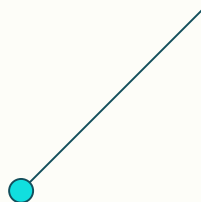
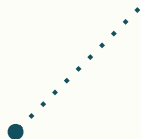
- Dipendenza da un genoma di riferimento noto.
- Non adatto per specie mai sequenziate.

GAGE

Utilizzato per confrontare diversi assemblatori di genomi su 4 dataset, valutando metriche come errori di assemblaggio.

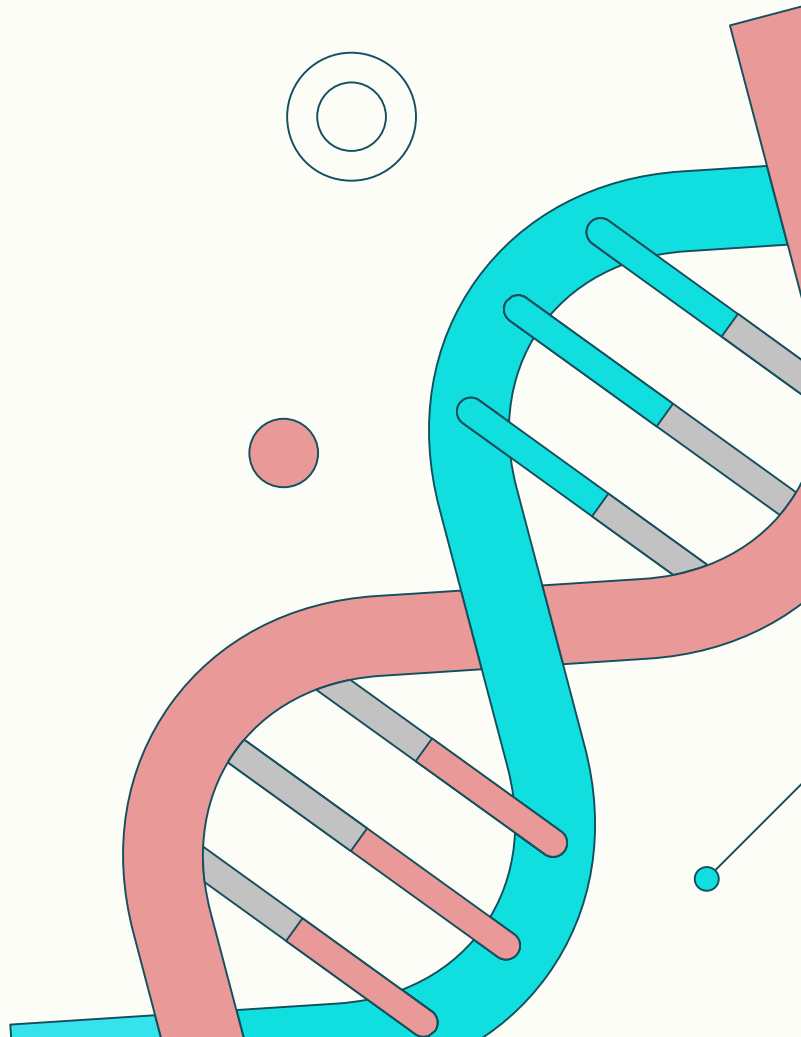
LIMITI:

- Dipendenza da un genoma di riferimento noto.
- Analizza un solo dataset alla volta.



02

QUAST



• QUAST: un nuovo approccio alla valutazione

è uno strumento avanzato per la valutazione degli assemblaggi genomici.
Si distingue per le seguenti caratteristiche innovative:

Ampia gamma di metriche

Include tutte le metriche rilevanti, bilanciando completezza e chiarezza.

Facilità d'uso

Interfaccia intuitiva e visualizzazioni immediate.

Valutazione senza riferimento

Ideale per specie nuove o non ancora sequenziate.

Efficienza elevata

Utilizza elaborazione parallela per analisi rapide su grandi dataset



Metriche valutate da QUASt

Metriche usate da QUASt per valutare la qualità degli assemblaggi

01

**Metriche legate al
numero di contigs**

02

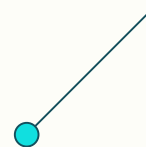
**Errori di
misassemblaggio e
variazioni strutturali**

03

**Rappresentazione
genomica ed elementi
funzionali**

04

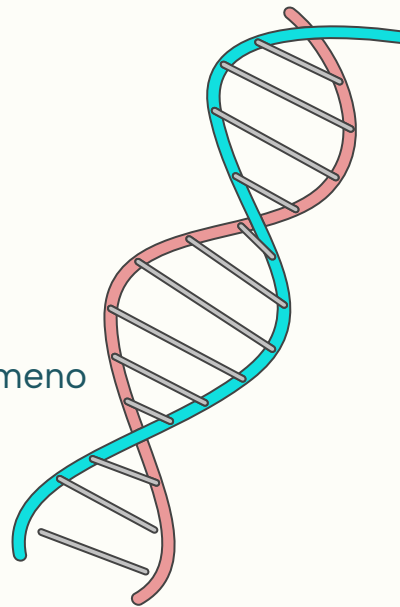
Metriche basate su N50





Metriche legate al numero di contigs

- **Numero di Contigs:** totale dei contigs nell'assemblaggio.
- **Contig più lungo:** lunghezza del contig più grande ricostruito.
- **Lunghezza totale:** numero totale di basi nell'assemblaggio.
- **Nx:** lunghezza del contig più corto, che sommato ad altri più lunghi, copre almeno l' $x\%$ della lunghezza totale dell'assemblaggio.
- **NGx:** simile a Nx, ma riferito alla lunghezza del genoma di riferimento.

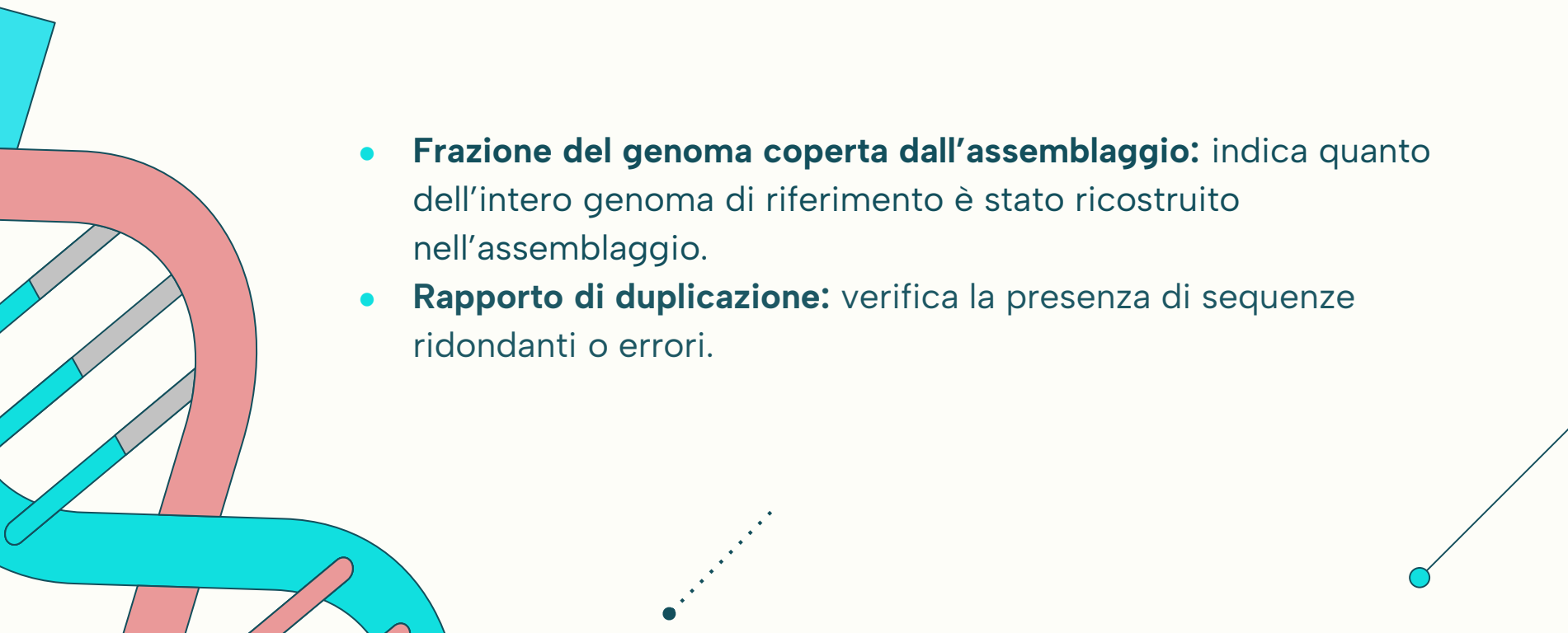


Errori di misasseblaggio e variazioni strutturali

- **Numero di misassebliaggi:** indica quanti errori sono stati trovati nell'asseblaggio confrontandolo con il genoma di riferimento.
- **Numero di contigs misasseblati:** indica quanti contigs contengono almeno un errore di misassemblaggio.
- **Lunghezza dei contigs misassemblati:** rappresenta il numero totale di basi presenti nei contig che contengono errori.
- **Numero dei contigs non allineati:** numero dei contigs che non riescono a trovare alcuna corrispondenza nel genoma di riferimento.
- **Numero di contigs mappati ambigamente:** contigs che si allineano bene a più punti nel genoma di riferimento.



Rappresentazione genomica ed elementi funzionali

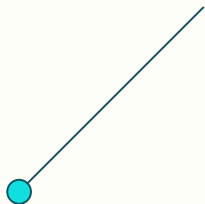
- 
- **Frazione del genoma coperta dall'assemblaggio:** indica quanto dell'intero genoma di riferimento è stato ricostruito nell'assemblaggio.
 - **Rapporto di duplicazione:** verifica la presenza di sequenze ridondanti o errori.

Metriche basate sull'N50



N50 non riflette accuratamente la qualità di un assemblaggio:

- **Non tiene conto della lunghezza reale del genoma:** N50 si basa solo sulla distribuzione dei contig, ma se l'assemblaggio è più corto o più lungo del genoma vero non se ne accorge.
 - **Non verifica se i contig sono corretti:** anche un assemblaggio con errori strutturali può avere un N50.
 - **Non considera l'accuratezza del posizionamento dei contig:** un assemblaggio può avere contig lunghi ma posizionati nel punto sbagliato.
-
- **NAx:** corregge il primo problema → usa la lunghezza del genoma di riferimento invece della lunghezza dell'assemblaggio per calcolare le soglie.
 - **NGAx:** corregge il secondo e il terzo problema → considera solo i contig che si allineano correttamente al genoma di riferimento.



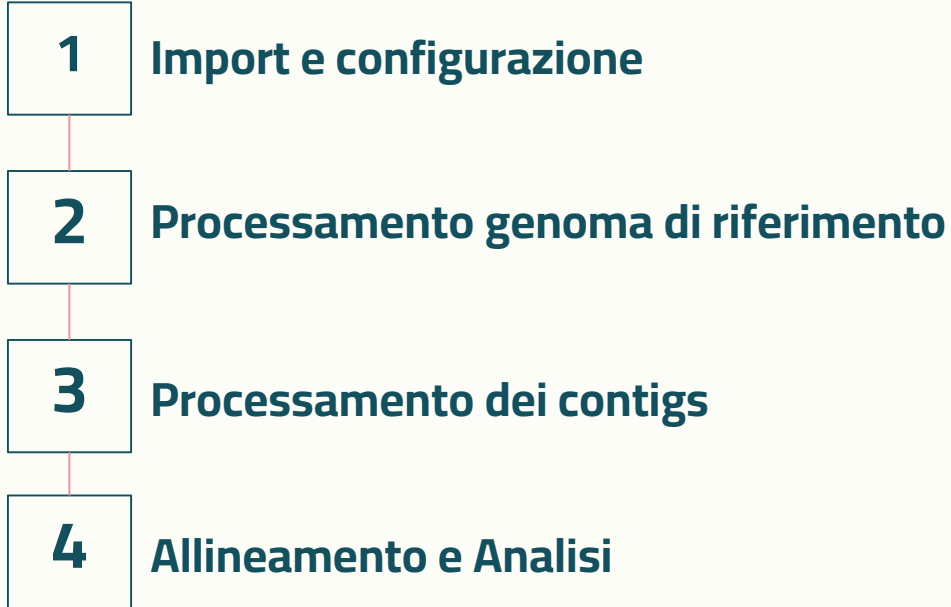


03

Analisi del codice

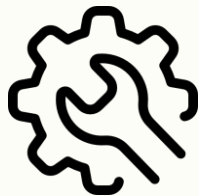


Flusso Operativo





Fase di Import e Configurazione

- Importazione librerie esterne e native
- Controllo della versione di Python
- Processing degli argomenti e conversione in comandi





Processamento del genoma di riferimento

- 
- **Opzionale**
 - **Correzione del genoma di riferimento**
 - Rimozione di sequenze più corte di una certa **soglia** (specificata) e di sequenze che contengono caratteri diversi da A, C, G, T (**basi azotate** del DNA) ed N (rappresenta aree incerte nella sequenza)
 - **Generazione dell'assemblaggio ottimale (UBA)**
 - Opzionale
 - Simulazione di come sarebbe un assemblaggio perfetto
 - **Aggiornamento della lista dei Contigs**
 - Viene aggiunto il percorso all'assemblaggio ottimale
- 



Processamento dei Contigs


- **Elaborazioni in parallelo**

- Possibilità di specificare il numero di **threads** da utilizzare per il processing, **ottimizzando** e **velocizzando** il processo

- **Correzione dei contigs**

- Rimozione caratteri speciali, come **+** o **-**, dai nomi dei contigs che possono causare errori durante l'elaborazione (es. con Nucmer)
- **Rimuove** contigs **corrotti** o problematici

- **Calcolo statistiche di copertura delle reads**

- **Allinea** le reads ai contigs e **calcola** le statistiche di copertura (es. Numero di basi coperte nel genoma di riferimento, numero di inserzioni e delezioni, ecc.)
- 

Allineamento e Analisi

- **Allineamento contigs al genoma di riferimento (opzionale)**
 - Controlla se il **genoma** è **ciclico**
 - **Effettua** l'allineamento
 - **Itera** sui contigs per determinare per ciascuno se l'**allineamento** è andato a buon fine o meno: se nessun contig è stato allineato con successo, l'analisi **termina**
- **Calcolo delle metriche, tra cui Nx, NGx, NAx**
- **Predizione genica (se specificata)**
 - Usa strumenti come Glimmer o GeneMark per **identificare** i **geni** nei contigs, Barnnap per trovare **rRNA**, BUSCO per verificare la presenza di **geni altamente conservati**.

Analisi NG50

- Rappresenta la **lunghezza** del contig più **corto** che, **sommato** ai contigs di lunghezza maggiore o uguale, copre **almeno** il 50% della lunghezza del genoma di riferimento
 - **numlist**: lista delle lunghezze dei contigs (ordine decrescente)
 - **s**: lunghezza totale del genoma di riferimento
 - **limit**: lunghezza necessaria per coprire almeno il 50% della lunghezza del genoma di riferimento
 - **lg50**: numero di contigs necessari per raggiungere NG50
 - **l**: la lunghezza del contig corrente
-
- Quando la lunghezza del genoma da coprire raggiunge/supera la soglia termina e restituisce ng50 e lg50
 - Se non si è raggiunta la soglia, restituisce None, None

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

return None, None
```

Esempio di Esecuzione

Esempio di esecuzione di QUASt utilizzando i dati di test offerti dagli sviluppatori

```
(myDefaultVenv) marcus@DESKTOP-G5UG03L:~/Workspace/quast$ quast.py test_data/contigs_1.fasta test_data/contigs_2.fasta -r test_data/reference.fasta.gz -o results/
```

File contigs

File contigs

Path genoma
di riferimento

Directory
risultati

```
Main parameters:
MODE: default, threads: 3, min contig length: 500, min alignment length: 65, min alignment IDV: 95.0, \
ambiguity: one, min local misassembly length: 200, min extensive misassembly length: 1000

Reference:
/home/marcus/Workspace/quast/test_data/reference.fasta.gz ==> reference

Contigs:
Pre-processing...
1 test_data/contigs_1.fasta ==> contigs_1
2 test_data/contigs_2.fasta ==> contigs_2

2025-03-05 19:25:35
Running Basic statistics processor...
Reference genome:
reference.fasta, length = 10000, num fragments = 1, GC % = 52.07
Contig files:
1 contigs_1
2 contigs_2
Calculating N50 and L50...
1 contigs_1, N50 = 3980, L50 = 1, auN = 2934.0, Total length = 6710, GC % = 51.28, # N's per 100 kbp = 0.00
2 contigs_2, N50 = 3360, L50 = 1, auN = 2875.4, Total length = 5460, GC % = 52.44, # N's per 100 kbp = 0.00
Drawing Nx plot...
saved to /home/marcus/Workspace/quast/results/basic_stats/Nx_plot.pdf
Drawing NGx plot...
saved to /home/marcus/Workspace/quast/results/basic_stats/NGx_plot.pdf
Drawing cumulative plot...
saved to /home/marcus/Workspace/quast/results/basic_stats/cumulative_plot.pdf
Drawing GC content plot...
saved to /home/marcus/Workspace/quast/results/basic_stats/GC_content_plot.pdf
Drawing contigs_1 GC content plot...
saved to /home/marcus/Workspace/quast/results/basic_stats/contigs_1_GC_content_plot.pdf
Drawing contigs_2 GC content plot...
saved to /home/marcus/Workspace/quast/results/basic_stats/contigs_2_GC_content_plot.pdf
Done.
```

- Processing del genoma di riferimento
- Calcolo delle metriche N50 e L50

Esempio di Esecuzione

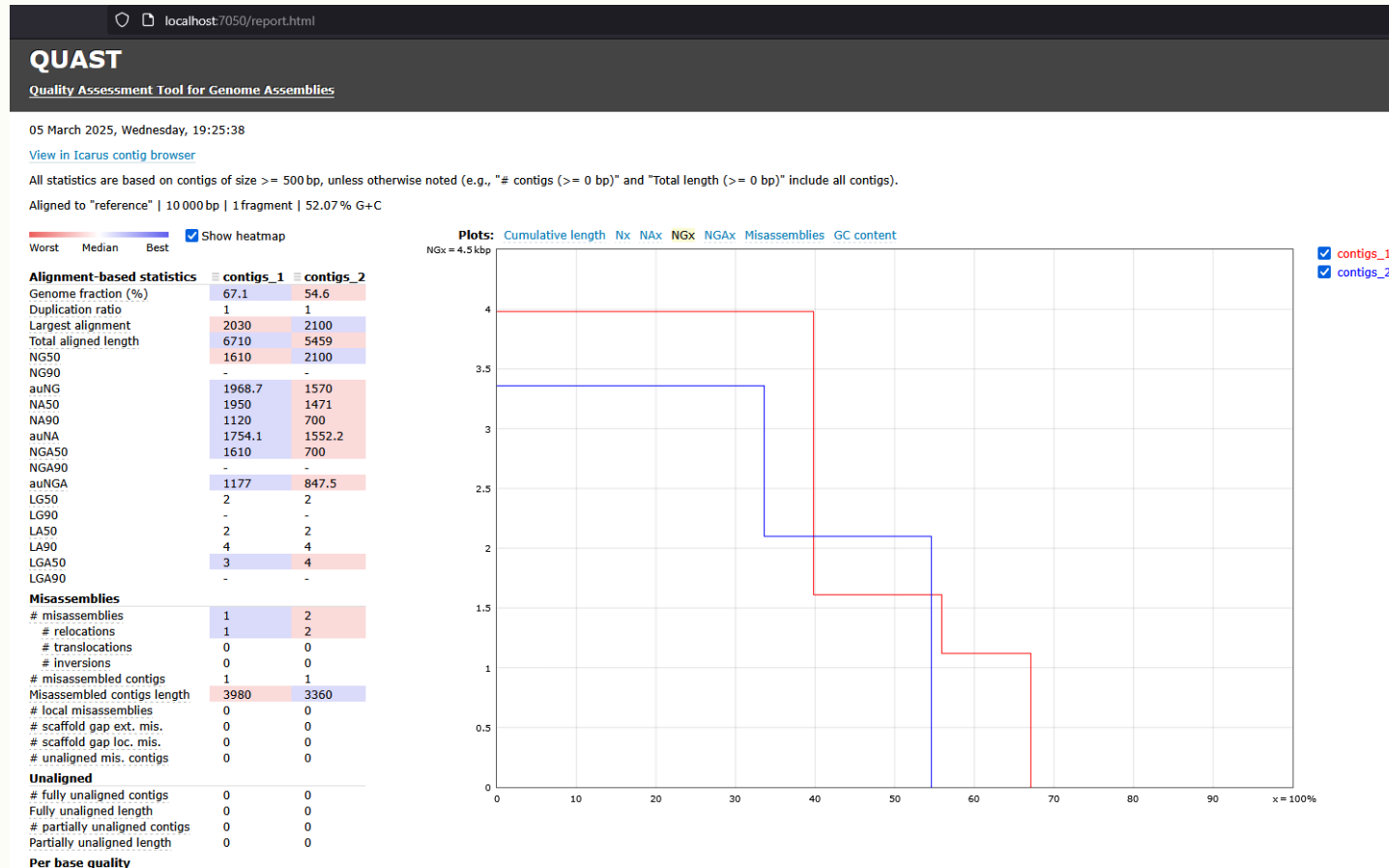
- Analisi dei contigs
- Allineamento
- Calcolo delle metriche NA-NGA

```
Running Contig analyzer...
1 contigs_1
2 contigs_2
2 Logging to files /home/marcus/Workspace/quast/results/contigs_reports/contigs_report_contigs_2.stdout and contigs_report_contigs_2.stderr...
1 Logging to files /home/marcus/Workspace/quast/results/contigs_reports/contigs_report_contigs_1.stdout and contigs_report_contigs_1.stderr...
1 Aligning contigs to the reference
2 Aligning contigs to the reference
2 Analysis is finished.
1 Analysis is finished.
Creating total report...
  saved to /home/marcus/Workspace/quast/results/contigs_reports/misassemblies_report.txt, misassemblies_report.tsv, and misassemblies_report.tex
Transposed version of total report...
  saved to /home/marcus/Workspace/quast/results/contigs_reports/transposed_report_misassemblies.txt, transposed_report_misassemblies.tsv, and transposed_report_misassemblies.tex
Creating total report...
  saved to /home/marcus/Workspace/quast/results/contigs_reports/unaligned_report.txt, unaligned_report.tsv, and unaligned_report.tex
Drawing misassemblies by types plot...
  saved to /home/marcus/Workspace/quast/results/contigs_reports/misassemblies_plot.pdf
Drawing misassemblies FRCurve plot...
  saved to /home/marcus/Workspace/quast/results/contigs_reports/misassemblies_frcurve_plot.pdf
Done.

2025-03-05 19:25:36
Running NA-NGA calculation...
1 contigs_1, Largest alignment = 2030, NA50 = 1950, NGA50 = 1610, LA50 = 2, LGA50 = 3
2 contigs_2, Largest alignment = 2100, NA50 = 1471, NGA50 = 700, LA50 = 2, LGA50 = 4
Drawing cumulative plot...
  saved to /home/marcus/Workspace/quast/results/aligned_stats/cumulative_plot.pdf
Drawing NAX plot...
  saved to /home/marcus/Workspace/quast/results/aligned_stats/NAX_plot.pdf
Drawing NGAX plot...
  saved to /home/marcus/Workspace/quast/results/aligned_stats/NGAX_plot.pdf
Done.
```

Esempio di Esecuzione

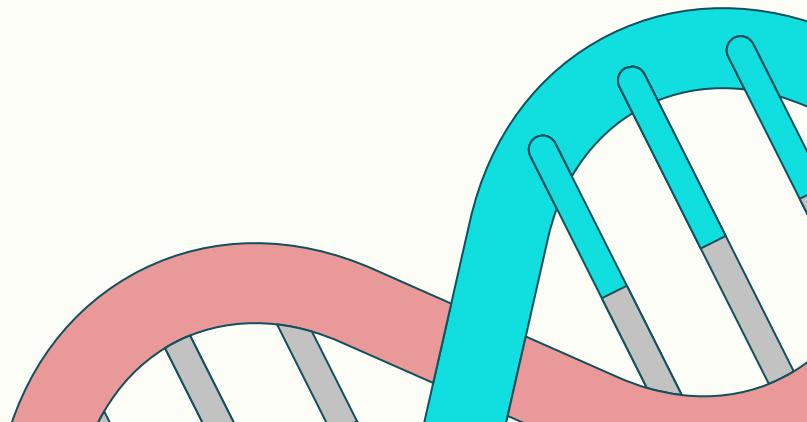
Visualizzazione dei risultati

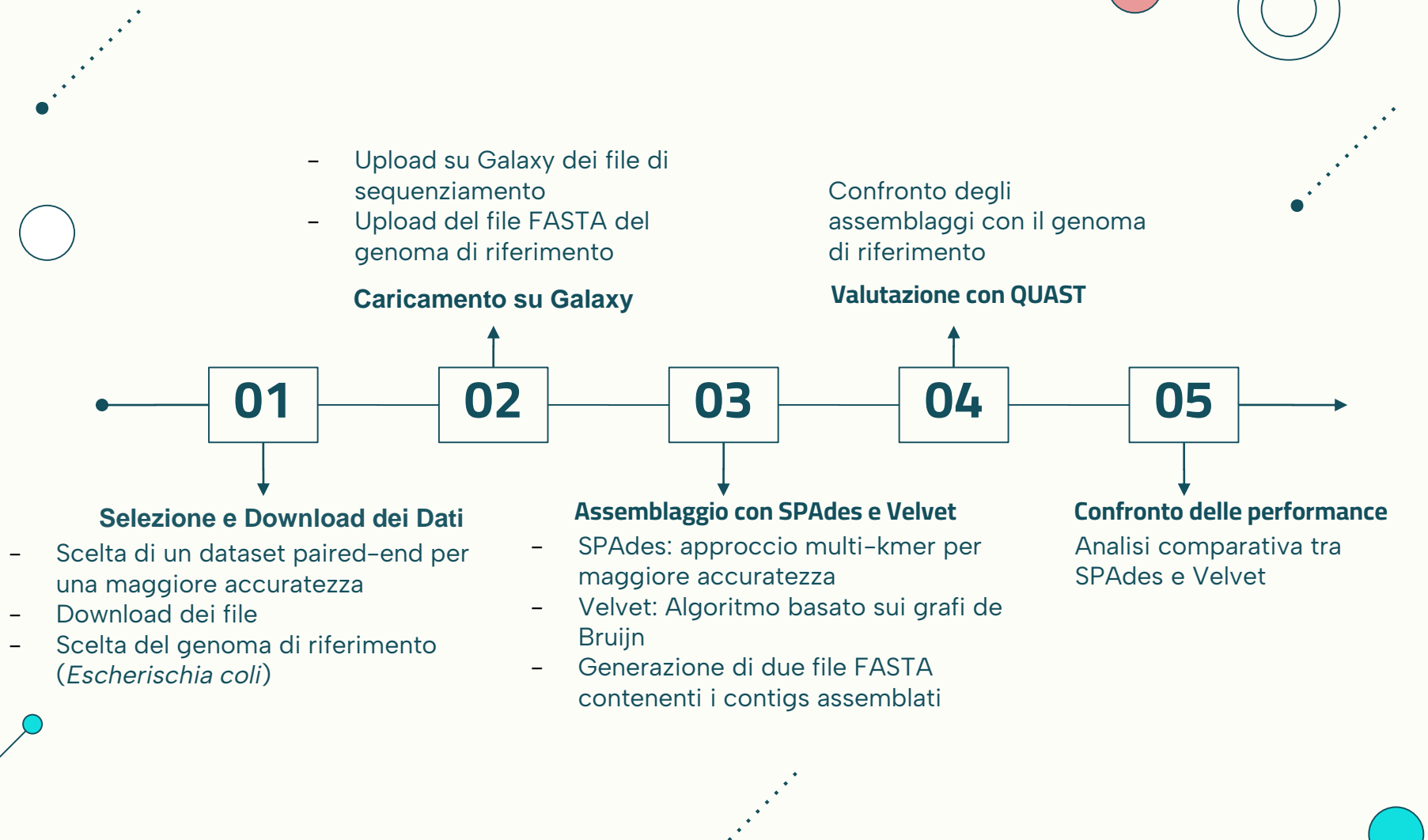




04

Galaxy workflow







Risultati

Metriche legate al numero di contig

- Numero totale di contig

- SPAdes: 190 contigs
- Velvet: 318 contigs

- Contig più lungo

- SPAdes: 313.898 bp
- Velvet: 289.097 bp

- Lunghezza totale

- SPAdes: 5.293.223 bp
- Velvet: 5.445.933 bp

- N50

- SPAdes: 124.687 bp
- Velvet: 59.880 bp

- NG50

- SPAdes: 127.550 bp
- Velvet: 64.016 bp

Statistics without reference

# contigs	190
# contigs (≥ 0 bp)	1113
# contigs (≥ 1000 bp)	134
Largest contig	313 898
Total length	5 293 223
Total length (≥ 0 bp)	5 414 009
Total length (≥ 1000 bp)	5 253 566
N50	124 687
N90	28 229

SPAdes

Statistics without reference

# contigs	318
# contigs (≥ 0 bp)	747
# contigs (≥ 1000 bp)	258
Largest contig	289 097
Total length	5 445 933
Total length (≥ 0 bp)	5 553 552
Total length (≥ 1000 bp)	5 402 552
N50	59 880
N90	9283

Velvet

Errori di misassemblaggio

- **Numero di misasseblaggi**
 - **SPAdes:** 133 errori
 - **Velvet:** 150 errori
- **Numero di contig misassemblati**
 - **SPAdes:** 41 contigs con errori
 - **Velvet:** 74 contigs con errori
- **Numero dei contigs non allineati**
 - **SPAdes:** 12 contigs non allineati
 - **Velvet:** 22 contigs non allineati
- **Lunghezza dei contigs misasseblati**
 - **SPAdes:** 4 055 419
 - **Velvet:** 3 759 719

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

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

Velvet



Rappresentazione genomica

- **Percentuale genoma ricostruito (Genome Fraction %)**

- **SPAdes:** 77,06%
- **Velvet:** 76,73%

- **Rapporto di duplicazione**

- **SPAdes:** 1.003
- **Velvet:** 1.005

Genome fraction (%)	77.062
Duplication ratio	1.003

SPAdes

Genome fraction (%)	76.734
Duplication ratio	1.005

Velvet

● Metriche basate su N50

- **NA50**

- **SPAdes:** 21 734
- **Velvet:** 14 711

NA50	21 734
NA90	-
auNA	26 496
NGA50	23 925

SPAdes

- **NGA50**

- **SPAdes:** 23 925
- **Velvet:** 16 119

NA50	14 711
NA90	-
auNA	20 820
NGA50	16 119

Velvet

In the top right corner, there are several decorative elements: a short dotted line with a dark blue dot at its end, a small white circle with a thin blue outline, and a solid red circle.

FINE

In the bottom right corner, there are decorative elements: a thin blue circle with a concentric inner circle, and a thin blue line with a small cyan dot at its end.

Si ringrazia per l'attenzione