# BIOINFORMATICA
# II

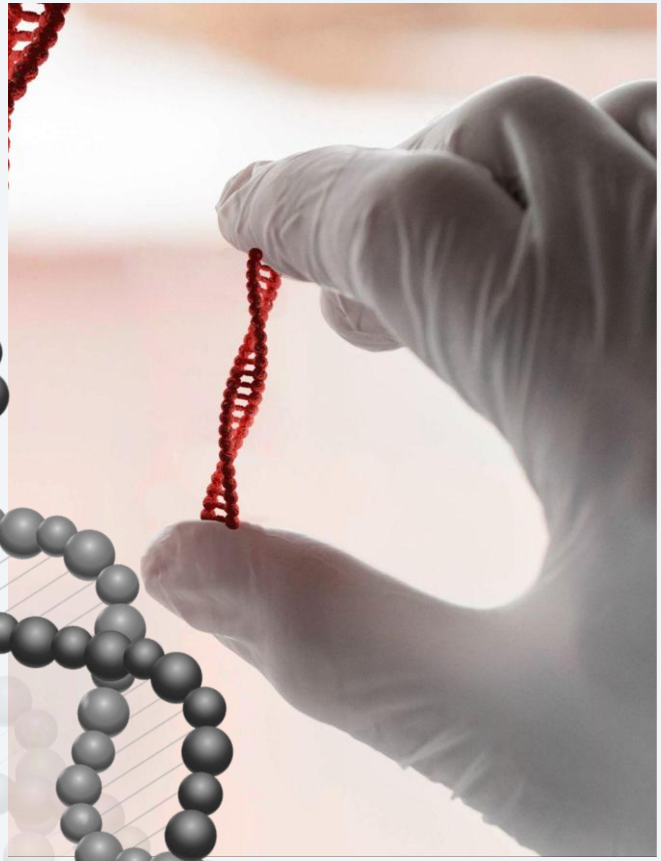
## *IL PROCESSO NGS*
## *CONTROLLO QUALITÁ*

# ARGOMENTI

01 PROCESSO NSG

02 CONTROLLO DI QUALITÁ DELLE READS

03 CONTROLLO DI QUALITA NGS III GENERAZIONE



# 01

# PROCESSO NSG

# PROCESSO NGS

Il **Next-Generation Sequencing** (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza.
Ecco i passaggi comuni di un processo NGS:

1. **Preparazione del campione**

***Isolamento degli acidi nucleici:*** *Si estraggono DNA o RNA dalla matrice biologica (es. sangue, tessuti). Se si lavora con RNA, può essere necessario convertire l'RNA in cDNA tramite trascrizione inversa.*
***Quantificazione e qualità:*** *Si valuta la qualità e la quantità degli acidi nucleici utilizzando strumenti come Nanodrop, Qubit o Bioanalyzer.*

# PROCESSO NGS

Il Next-Generation Sequencing (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza. Ecco i passaggi comuni di un processo NGS:

**2. Preparazione della libreria**

***Frammentazione:*** *gli acidi nucleici sono frammentati in segmenti di dimensioni specifiche (tipicamente 200-500 bp). Si possono usare metodi chimici, enzimatici o meccanici (es.: sonicazione).*

***Adattatori:*** *ai frammenti sono aggiunti adattatori specifici, che includono sequenze per il riconoscimento da parte della piattaforma NGS e, talvolta, codici a barre per la multiplexing.*

***Amplificazione:*** *si amplifica la libreria con* PCR *per ottenere una quantità sufficiente di DNA/RNA.*

# PROCESSO NGS

Il Next-Generation Sequencing (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza. Ecco i passaggi comuni di un processo NGS:

**3. Quantificazione e normalizzazione della libreria**
*Si verifica la qualità e la concentrazione della libreria finale tramite strumenti come il Qubit, il Bioanalyzer o il qPCR.*
*Le librerie sono normalizzate per garantire un caricamento uniforme nel sequenziatore*

# PROCESSO NGS

Il Next-Generation Sequencing (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza. Ecco i passaggi comuni di un processo NGS:

**4. Clustering (per piattaforme specifiche)**
*Illumina: I frammenti vengono immobilizzati su una flow cell e amplificati per formare cluster di molecole identiche.*
*Altre piattaforme: In tecnologie come quella di Ion Torrent o PacBio, il processo varia, ma il principio è quello di preparare i frammenti per il sequenziamento.*

# PROCESSO NGS

Il Next-Generation Sequencing (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza. Ecco i passaggi comuni di un processo NGS:

**5. Sequenziamento**

*La sequenza di basi viene determinata attraverso metodi specifici della piattaforma:*

*Illumina: Sequenziamento per sintesi con fluorescenza.*

*Ion Torrent: Misurazione dei cambiamenti nel pH.*

*PacBio/ONT: Rilevamento in tempo reale delle modifiche durante la replicazione*

# PROCESSO NGS

Il Next-Generation Sequencing (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza. Ecco i passaggi comuni di un processo NGS:

**6. Analisi dei dati grezzi**

***Conversione:*** *I segnali grezzi (fluorescenza o elettrici) sono convertiti in sequenze di basi (file FASTQ).*

***Qualità:*** *Si valuta la qualità delle sequenze (es. con strumenti come FASTQC).*

**7. Allineamento e analisi bioinformatica**

***Allineamento:*** *Le sequenze sono confrontate con un genoma di riferimento usando software come Bowtie, BWA o STAR.*

*Chiamata di varianti: Si identificano mutazioni, SNPs, INDELs, o altre variazioni genomiche.*

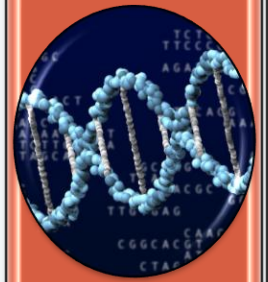
***Analisi personalizzate:*** *A seconda dell'obiettivo (es. espressione genica, analisi di microbiomi, studio di mutazioni), si utilizzano tool specifici.*

# PROCESSO NGS

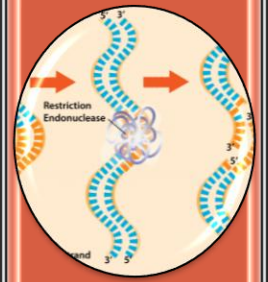
Il Next-Generation Sequencing (NGS) è una tecnologia avanzata che consente di sequenziare DNA o RNA in modo rapido e ad alta efficienza. Ecco i passaggi comuni di un processo NGS:

**8. Interpretazione dei risultati**
*I dati elaborati sono interpretati per rispondere alla domanda biologica o clinica primordiale.*
*Possono essere utilizzati strumenti come annotatori di varianti o analisi statistica per derivare conclusioni.*
*Il processo può variare leggermente in relazione alla piattaforma utilizzata (Illumina, Ion Torrent, PacBio, Oxford Nanopore, ecc.) e al tipo di applicazione (es. RNA-Seq, WGS, target sequencing).*

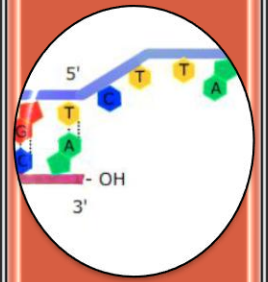
# PROCESSO NGS



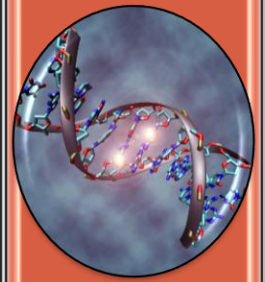
Estrazione del DNA dai campioni



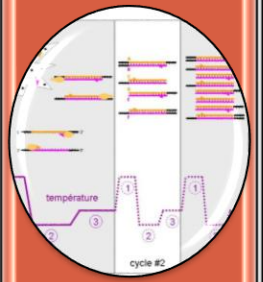

Taglio del DNA (metodi: nebulizzazione, sonificazione,…)



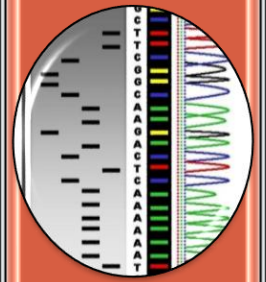

Aggiunta di adattatori ai frammenti



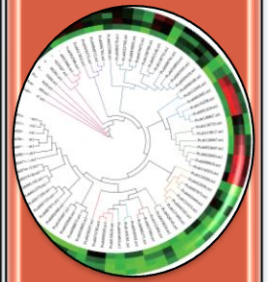
Creazione di una libreria (selezione dei frammenti)




Amplificazione (PCR ad emulsione – Roche e Ion Torrent; PCR a ponte -Illumina)



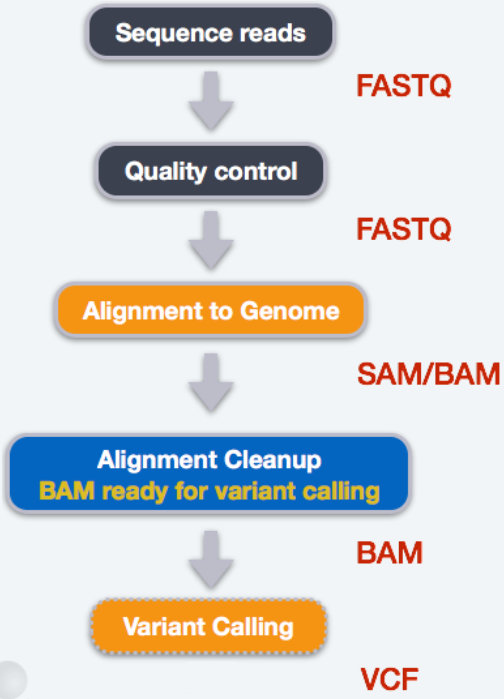
Sequenziamento



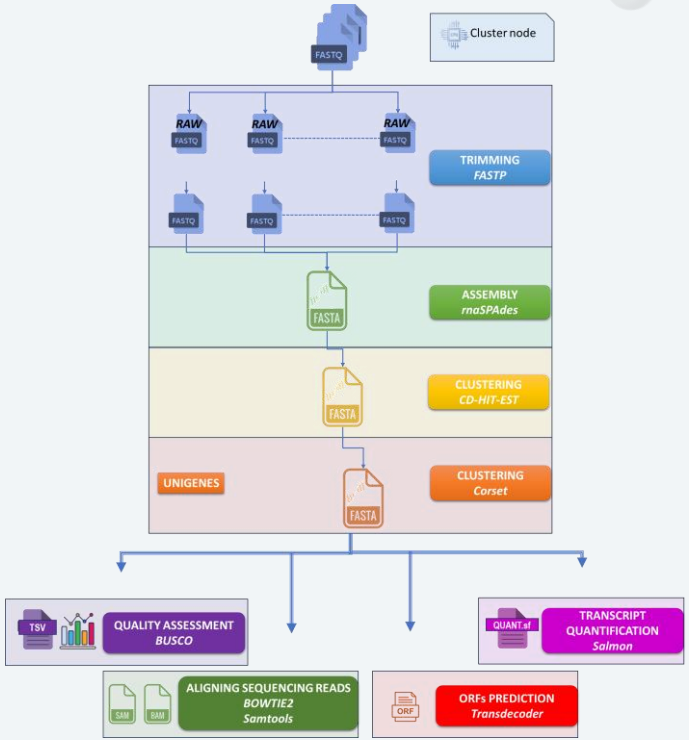
Assemblamento e analisi bioinformatiche

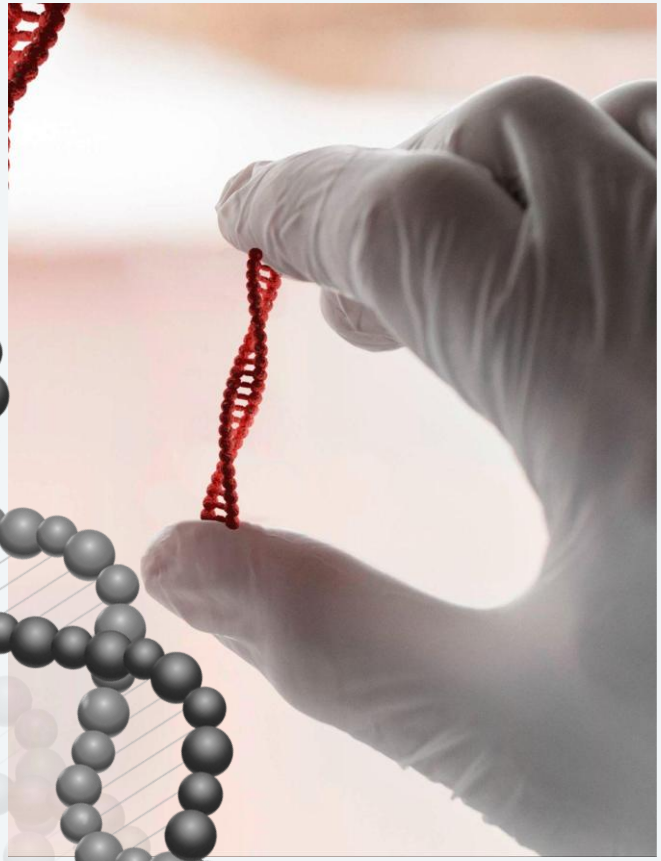
# PROCESSI POST NGS

## VARIANT CALLING




## TRASCRITTOMA *DE NOVO*





# 02

# CONTROLLO DI QUALITÁ DELLE READS

# CONTROLLO QUALITÁ

Dopo il sequenziamento il **controllo di qualità** è una fase fondamentale per garantire che i dati ottenuti siano accurati, affidabili e interpretabili.
Ci sono diverse ragioni per cui il controllo di qualità va svolto:

**1. Identificazione di errori di sequenziamento**
Durante il sequenziamento, possono verificarsi errori, come inserzioni, delezioni o errori di lettura delle basi. Il controllo di qualità aiuta a identificare questi errori e a determinarne la frequenza, così da ridurre la possibilità di falsi positivi o falsi negativi.

# CONTROLLO QUALITÁ

## 2. Verifica della qualità delle letture

Il NGS produce milioni o miliardi di letture (reads) che devono avere una qualità sufficiente per essere utilizzabili. Gli strumenti di controllo di qualità analizzano i punteggi di qualità di ciascuna read per assicurarsi che siano sufficientemente alti, in modo da garantire che le basi siano state lette correttamente e non ci siano troppi errori.

## 3. Rimozione di sequenze di bassa qualità e adattatori

Le read possono contenere regioni di bassa qualità o sequenze di adattatori (aggiunte artificialmente durante la preparazione della libreria) che devono essere rimosse. Questo passaggio migliora l'accuratezza delle analisi successive, evitando che regioni non informative o rumorose interferiscano con i risultati

# CONTROLLO QUALITÁ

**4. Controllo della contaminazione**
Durante la fase di preparazione e di sequenziamento, può verificarsi contaminazione con DNA estraneo o proveniente da altre fonti (ad esempio, altri campioni o microrganismi ambientali). Il controllo di qualità aiuta a identificare potenziali contaminazioni, che potrebbero alterare i risultati finali.
**5. Ottimizzazione dei parametri di analisi**
La qualità delle letture influenza direttamente i parametri delle analisi bioinformatiche successive. Conoscere la qualità dei dati aiuta a scegliere i migliori parametri per l'allineamento e l'analisi delle varianti, riducendo il rischio di errori.

# CONTROLLO QUALITÁ

**6. Riduzione dei costi e risparmio di tempo**
Identificare e risolvere eventuali problemi già nelle fasi iniziali di analisi permette di risparmiare tempo e risorse. Se i dati di sequenziamento sono di bassa qualità, è spesso preferibile ripetere il sequenziamento invece di proseguire con un'analisi compromessa, che potrebbe portare a risultati sbagliati.

In sintesi, il controllo di qualità post-NGS è essenziale per assicurare che i dati di sequenza siano utilizzabili per l'analisi e l'interpretazione e per evitare errori che potrebbero avere un impatto sui risultati finali, soprattutto in ambiti come la ricerca genetica e la diagnostica medica.

# CONTROLLO QUALITÁ: Interventi

Non tutte le basi che compongono le READS hanno lo stesso livello di qualità: la qualità tende a degenerare mentre ci si avvicina alla fine 3’ e muta al variare delle della posizione all’interno delle diverse reads

***Basi con qualità<20 sono generalmente ritenute non qualificanti ai fini dell’analisi***

Nel sequenziamento Illumina si uniscono degli adattatori (P5 e P7, sequenze di oligonucleotidi) utili per l'amplificazione di frammenti di DNA e il loro ancoraggio alle celle di flusso (flow cell) durante il processo di sequenziamento.

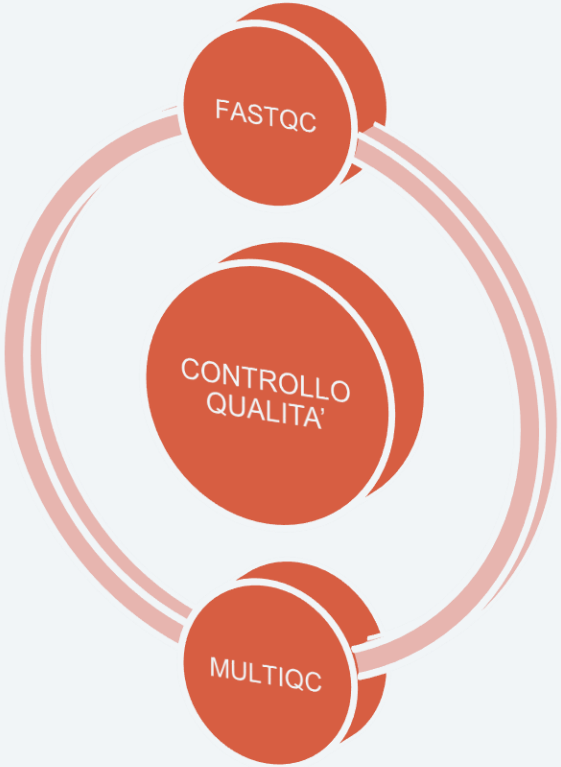
***In alcuni casi le reads possono mantenere completamente o parzialmente queste adattatori che devono essere rimossi ai fini dell'analisi***

# CONTROLLO QUALITÁ

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| TGAGGTAGTAGATTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC | 10865 | 4.346 | Illumina Multiplexing PCR Primer 2.01 (100% over 28bp) |
| TAGCTTATCAGACTGATGTTGACAGATCGGAAGAGCACACGTCTGAACTC | 10845 | 4.338 | Illumina Multiplexing PCR Primer 2.01 (100% over 27bp) |
| TCTTTGGTTATCTAGCTGTATGAGATCGGAAGAGCACACGTCTGAACTCC | 7062 | 2.8247999999999998 | Illumina Multiplexing PCR Primer 2.01 (100% over 28bp) |
| TCTTTGGTTATCTAGCTGTATGAAGATCGGAAGAGCACACGTCTGAACTC | 4056 | 1.6223999999999998 | Illumina Multiplexing PCR Primer 2.01 (100% over 27bp) |
| TGAGGTAGTAGTTTGTGCTGTTAGATCGGAAGAGCACACGTCTGAACTCC | 3737 | 1.4948 | Illumina Multiplexing PCR Primer 2.01 (100% over 28bp) |
| TGAGGTAGTAGTTTGTACAGTTAGATCGGAAGAGCACACGTCTGAACTCC | 3549 | 1.4196 | Illumina Multiplexing PCR Primer 2.01 (100% over 28bp) |
| TAGCTTATCAGACTGATGTTGACCAGATCGGAAGAGCACACGTCTGAACT | 277 | 0.11080000000000001 | Illumina Multiplexing PCR Primer 2.01 (100% over 26bp) |
| TGAGGTAGTAGTTTGTGCTGTTTAGATCGGAAGAGCACACGTCTGAACTC | 276 | 0.1104 | Illumina Multiplexing PCR Primer 2.01 (100% over 27bp) |
| TCAGTGCACTACAGAACTTTGTAGATCGGAAGAGCACACGTCTGAACTCC | 268 | 0.1072 | Illumina Multiplexing PCR Primer 2.01 (100% over 28bp) |
| AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGCTGTGATCTCGTA | 267 | 0.10679999999999999 | TruSeq Adapter, Index 10 (97% over 38bp) |
| TCTTTGGTTATCTAGCTGTATGATAGATCGGAAGAGCACACGTCTGAACT | 264 | 0.10560000000000001 | Illumina Multiplexing PCR Primer 2.01 (100% over 26bp) |
| CTAGACTGAGGCTCCTTGAGGAGATCGGAAGAGCACACGTCTGAACTCCA | 264 | 0.10560000000000001 | Illumina Multiplexing PCR Primer 2.01 (100% over 29bp) |
| CATGCCTTGAGTGTAGGACTGTAGATCGGAAGAGCACACGTCTGAACTCC | 264 | 0.10560000000000001 | Illumina Multiplexing PCR Primer 2.01 (100% over 28bp) |


# BIOINFORMATICA

# CONTROLLO QUALITÁ



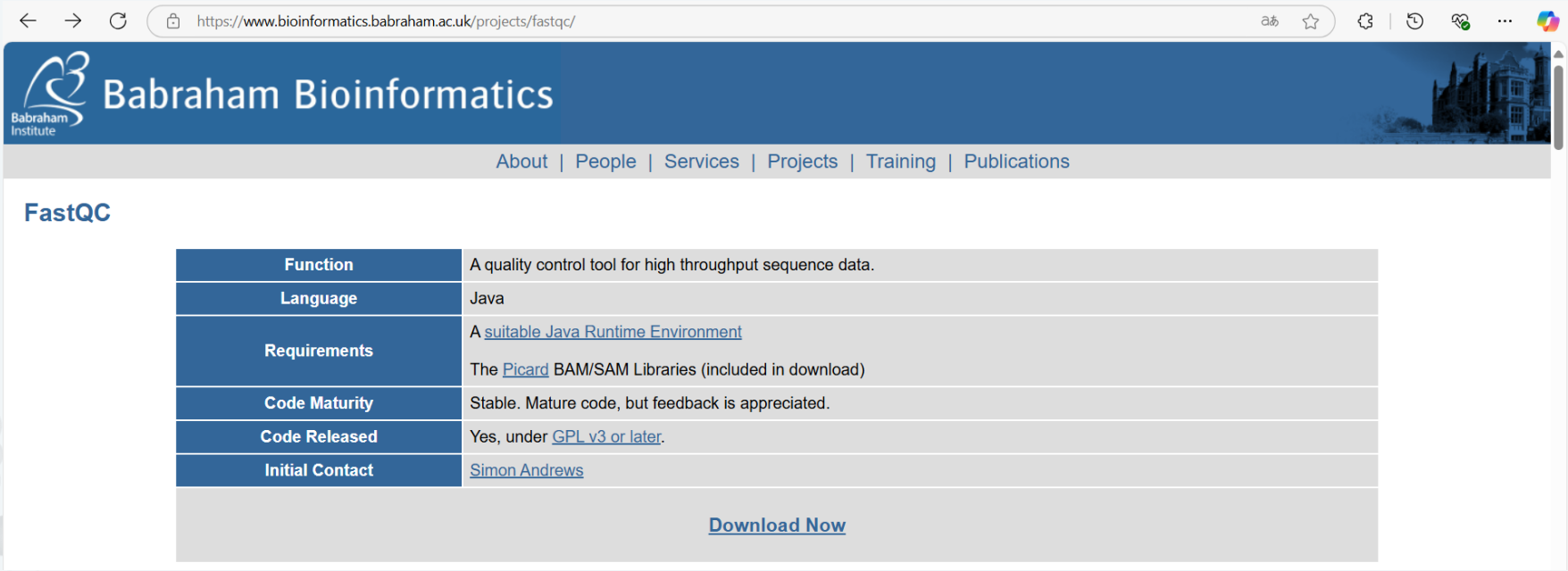
# FASTQC

Il software più accreditato per avere informazioni sulla qualità dei dati è **FastQC** sviluppato dal Babraham Institute

FastQC esegue un'analisi approfondita delle statistiche di qualità per ciascuna base in ogni sequenza. Esamina vari aspetti, tra cui la distribuzione dei punteggi di qualità Phred (Phred Quality Scores), la presenza degli adattatori (adapters) o regioni di basi con bassa qualità, la distribuzione delle lunghezze delle sequenze e altro ancora.

# FASTQC: Sito

**https://www.bioinformatics.babraham.ac.uk/projects/fastqc/**



# FASTQC: Installazione

È sufficiente il download e la decompressione del file

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip  
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.dmg 

Dipendenze:
Richiede la Java Virtual Machine
https://www.java.com/it/download/manual.jsp

# FASTQC: Comando

Il comando per eseguire FASTQC su un file è:

**fastqc [OPZIONI] <file1> <file2>**

| Opzione | Significato |
|---|---|
| **-o** | Specifica la directory di output dove salvare i risultati |
| **-f** | Specifica il formato del file di input (ad esempio, fastq, bam, ecc.) |
| **-t** | Specifica il numero di thread da utilizzare per l'elaborazione parallela |
| **-k** | Lunghezza massima delle sequenze k-mer da analizzare |
| **--noextract** | Evita l'estrazione dei file zip generati (mantiene solo i file compressi). |
| **--extract** | Forza l'estrazione del file zip generato. Questa opzione è di solito abilitata per impostazione predefinita |
| **--nogroup** | Disabilita il raggruppamento delle sequenze per lunghezza durante il calcolo delle statistiche (utile per sequenze molto lunghe) |

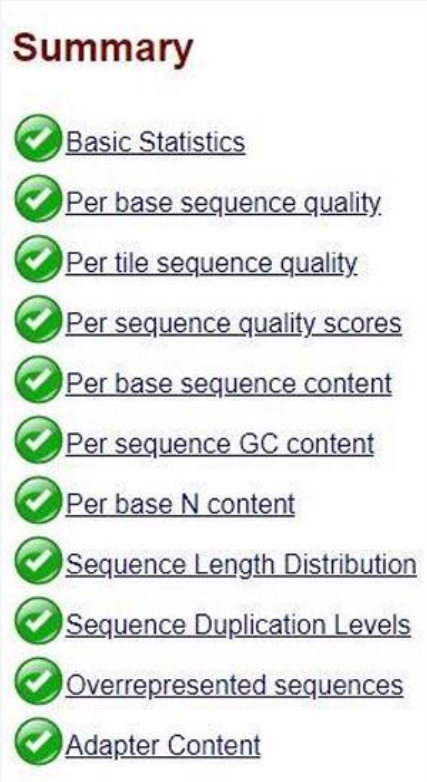
# FASTQC: Comando (esempio)

Ad esempio:

**fastqc -o results/ -t 4 -f fastq --nogroup sample1.fastq sample2.fastq**

I risultati sono salvati nella directory *results/*
Sono utilizzati 4 thread
I file in input sono in formato fastq
Non è effettuato il raggruppamento per lunghezza delle sequenze

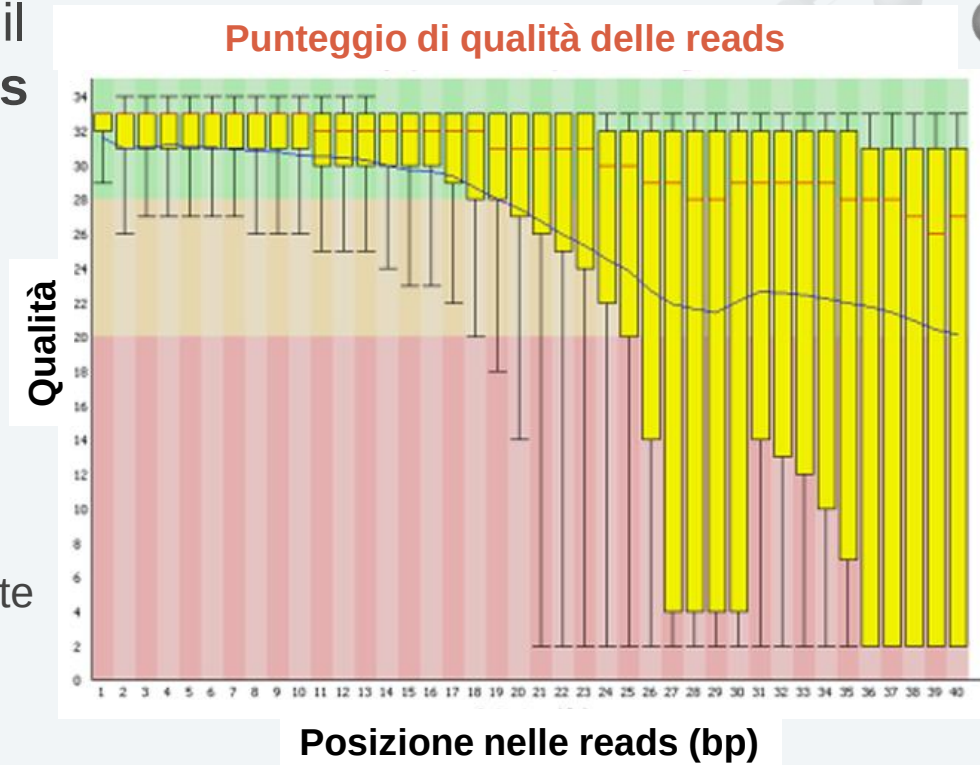
# FASTQC: Risultato principale




# FASTQC: Analisi

Il primo grafico che si analizza è il **punteggio di qualità delle reads** che offre una panoramica della qualità delle basi in ciascuna posizione nel file FASTQ.

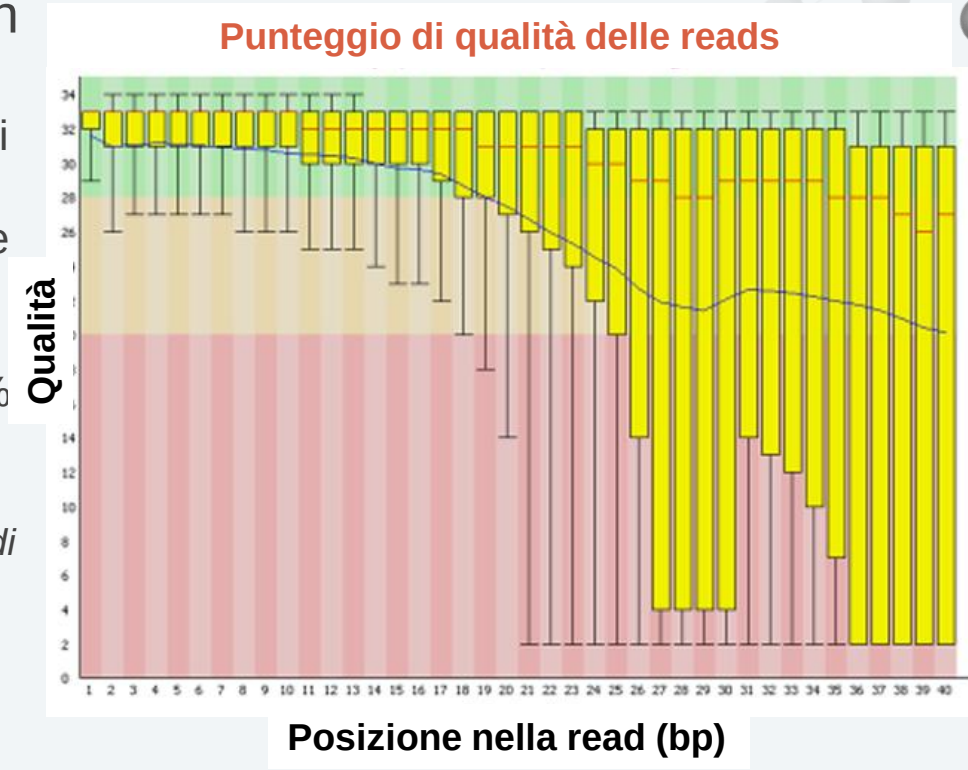
L'asse Y del grafico mostra i punteggi di qualità. Più alto è il punteggio, migliore è l'identificazione della base. Lo sfondo del grafico divide l'asse y in chiamate di qualità molto buona (verde), chiamate di qualità ragionevole (arancione) e chiamate di scarsa qualità (rosso).



# FASTQC: analisi

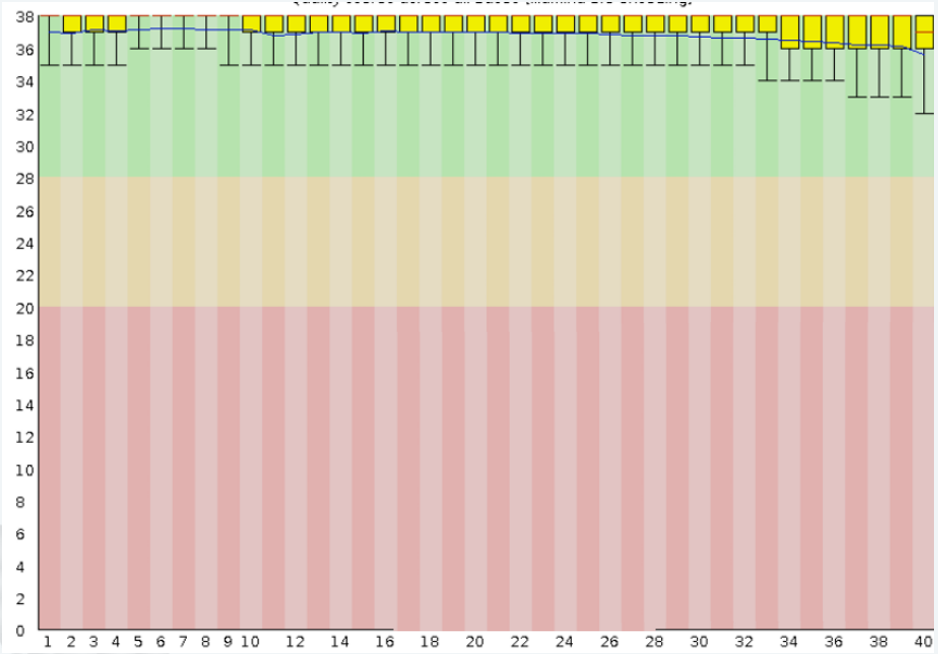
Per ogni posizione è disegnato un grafico *BoxWhisker*.
Il box giallo rappresenta la distribuzione di qualità delle read tra il 25% e il 75%
La riga rossa nel box rappresenta il valore mediano
I whiskers (le stanghette) sopra e sotto il box rappresentano il limite del 10% e 90%
La riga blu è la qualità media

*Di solito la qualità peggiora mentre il processo di sequenziamento procede, quindi è comune vedere la qualità delle basi cadere nell'area arancione verso la fine di una lettura*

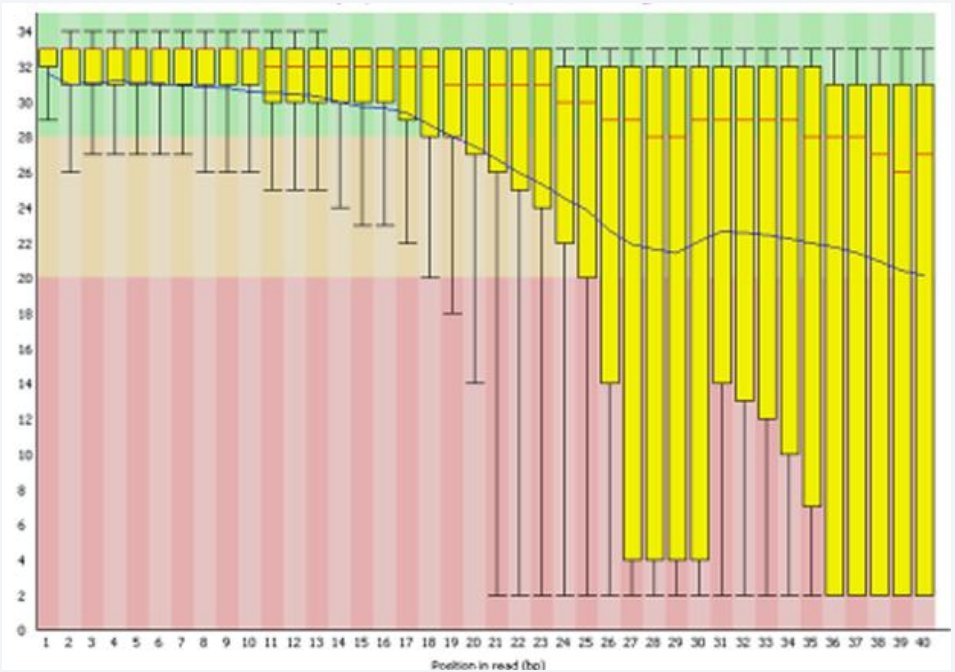


# FASTQC: Analisi

## QUALITÁ BUONA




## QUALITÁ SCARSA

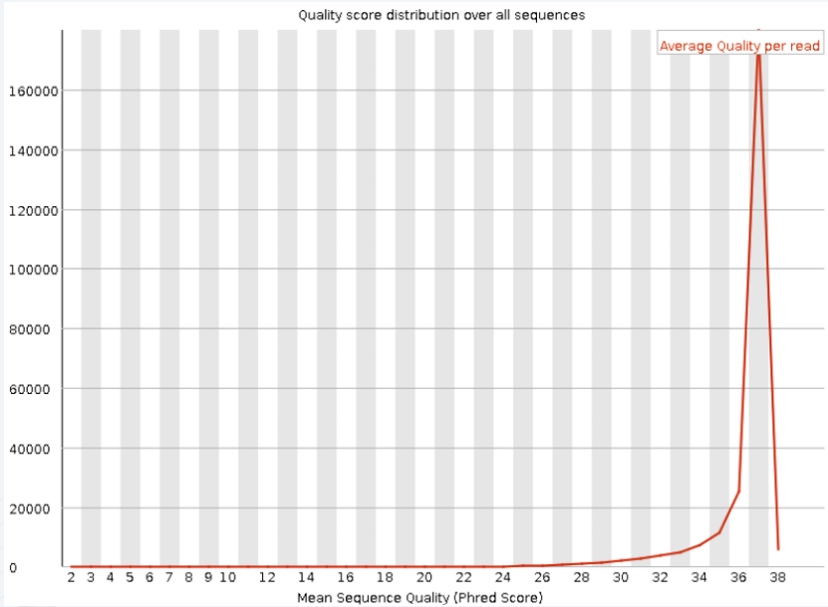


# FASTQC: Analisi

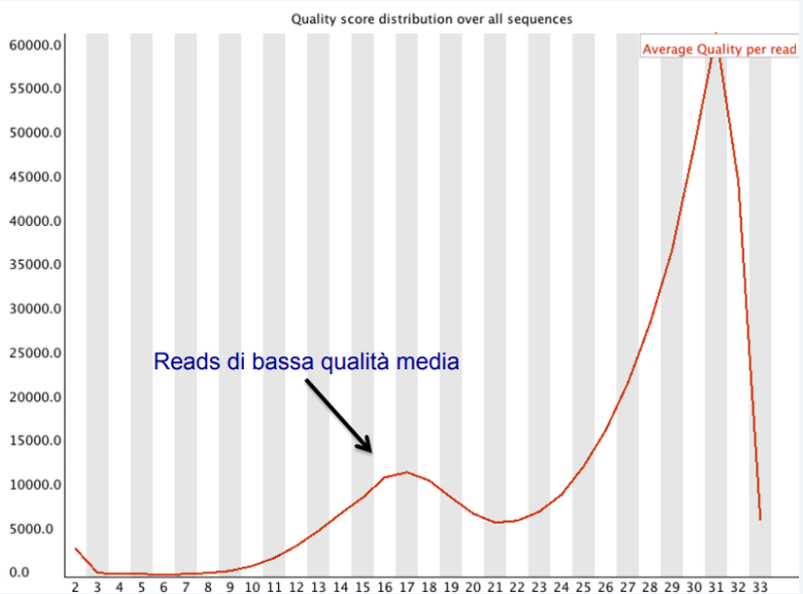
La **distribuzione della quantità delle reads** rappresenta il numero totale di reads (asse delle y) rispetto al punteggio di qualità medio (punteggio Phred asse delle x) dell'intera sequenza
Se si ha un picco verso valori alti significa che molte reads hanno una ottima qualità

# FASTQC: Analisi

## QUALITÁ BUONA

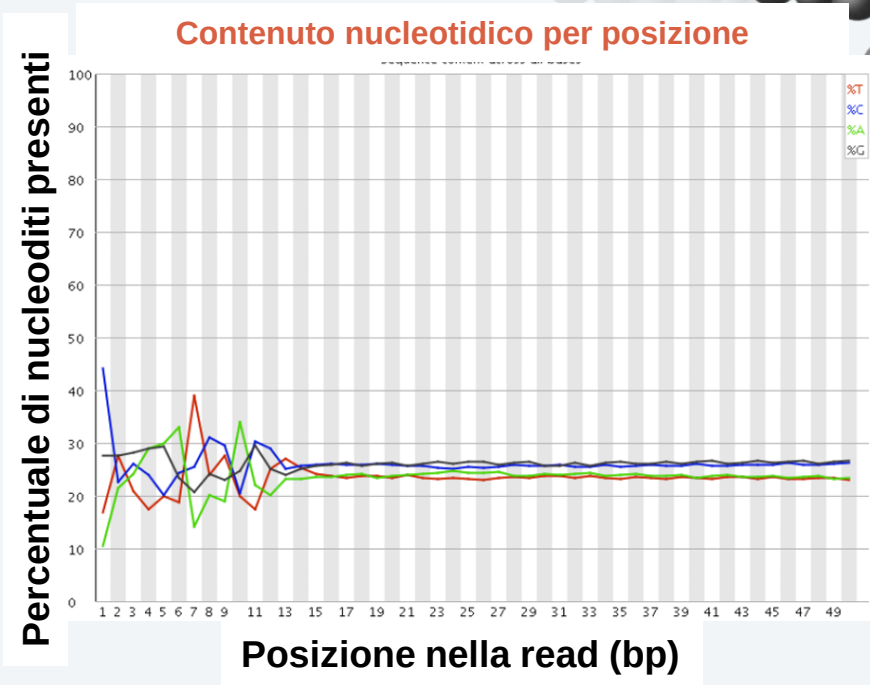



## QUALITÁ SCARSA



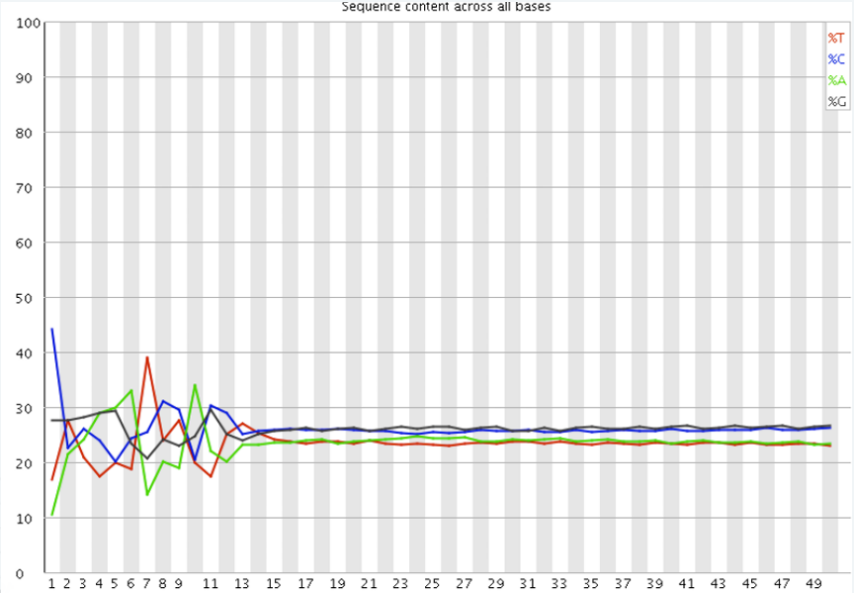
# FASTQC: Analisi

Il grafico di **contenuto nucleotidico per posizione** riporta la percentuale di basi presenti (per ciascuno dei quattro nucleotidi) nell'i-esima posizione di tutte le reads

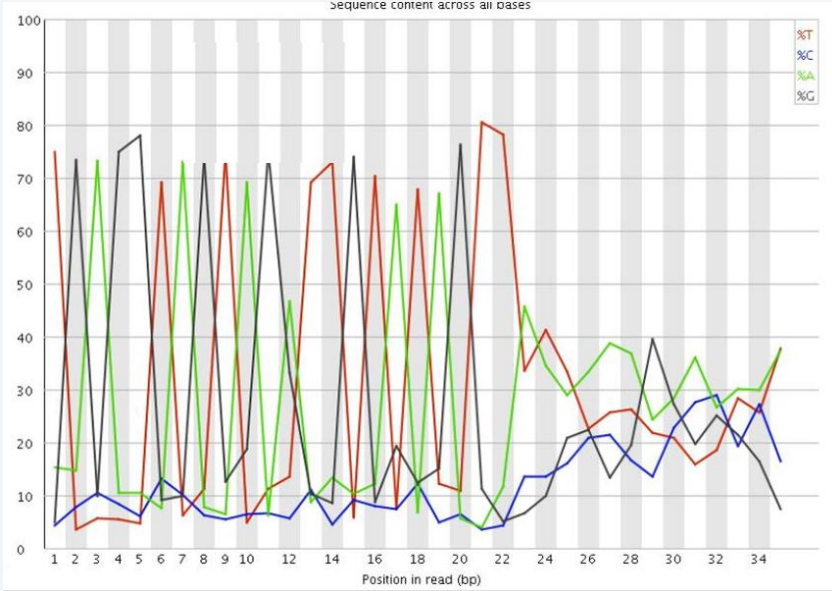


# FASTQC: Analisi

## QUALITÁ BUONA



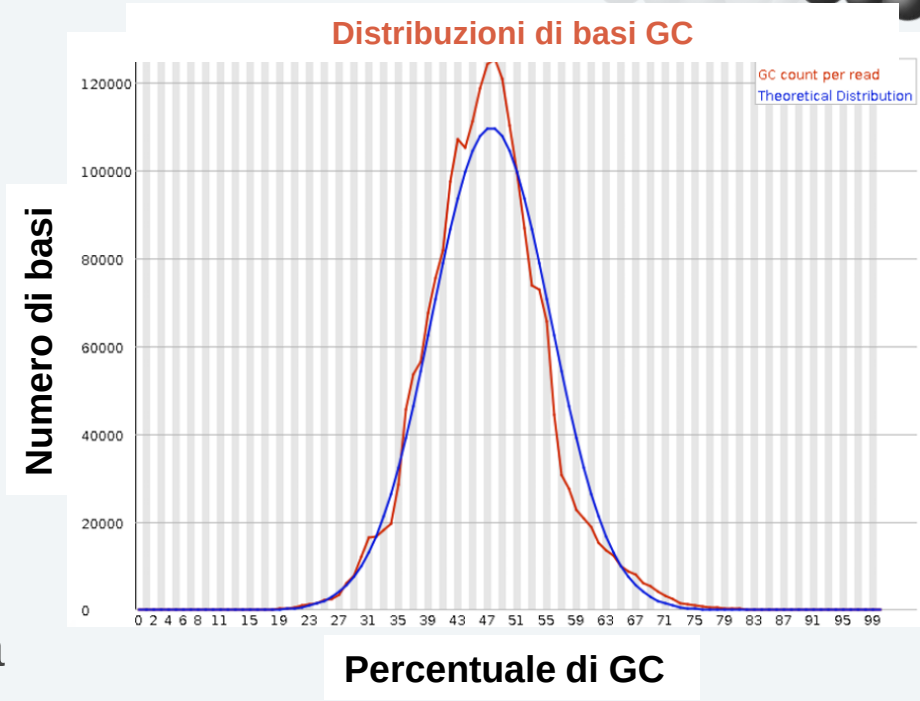

## QUALITÁ SCARSA



# FASTQC: Analisi

Il grafico della **distribuzione di basi GC nella sequenza** mostra l'andamento in relazione alla distribuzione teorica nel caso di un contenuto uniforme

In altre parole la linea rossa (GC *count per read*) indica la presenza delle basi GC per read mentre la linea blu (la distribuzione teorica) la presenza delle basi GC per read con una distribuzione normale

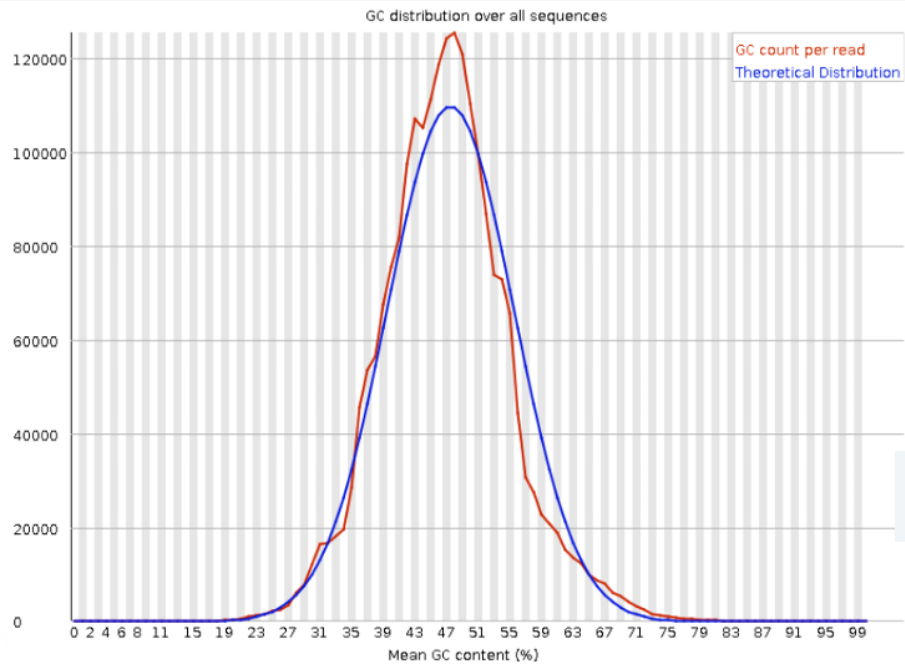


# FASTQC: Analisi (perché GC)

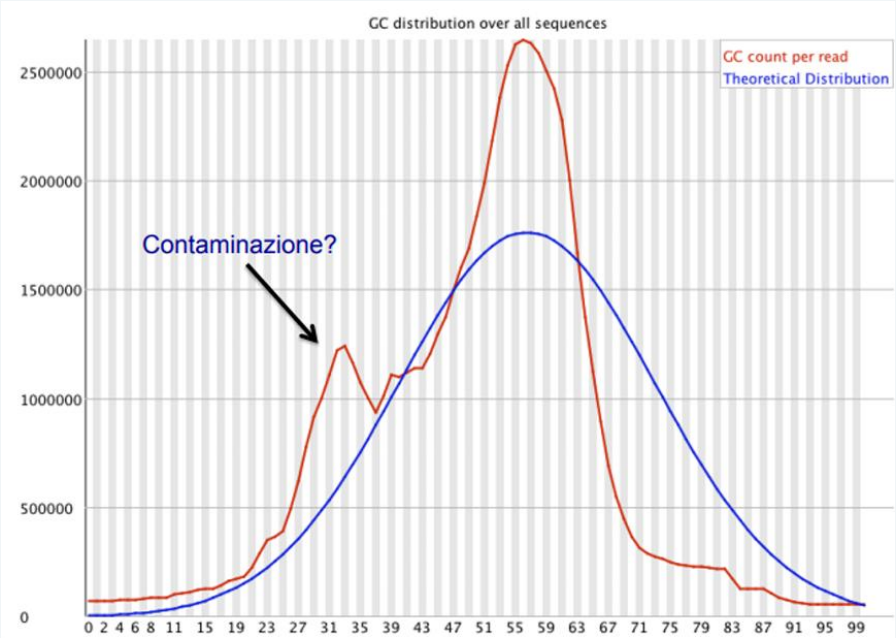
- Le coppie di basi GC sono legate da tre legami idrogeno, rispetto ai due legami idrogeno delle coppie AT (adenina e timina). Questo rende le regioni ricche di GC più stabili e difficili da denaturare (separare i filam1enti di DNA) a temperature elevate
- Regioni ricche di GC: Queste aree possono essere associate a promotori e altre regioni regolatorie del genoma. La loro stabilità intrinseca può influenzare l'interazione con proteine e altri fattori di trascrizione.
  Regioni povere di GC: Sono spesso associate a sequenze introniche o inter-geniche e possono svolgere ruoli specifici in funzione del contesto genomico
- In bioinformatica, la distribuzione di GC è spesso analizzata per:
  - Individuare regioni anomale nel genoma, come isole CpG (aree ricche di CG spesso correlate a promotori nei mammiferi)
  - Valutare bias nelle sequenze prodotte da tecniche di sequenziamento o amplificazione.

# FASTQC: Analisi

## QUALITÁ BUONA




## QUALITÁ SCARSA



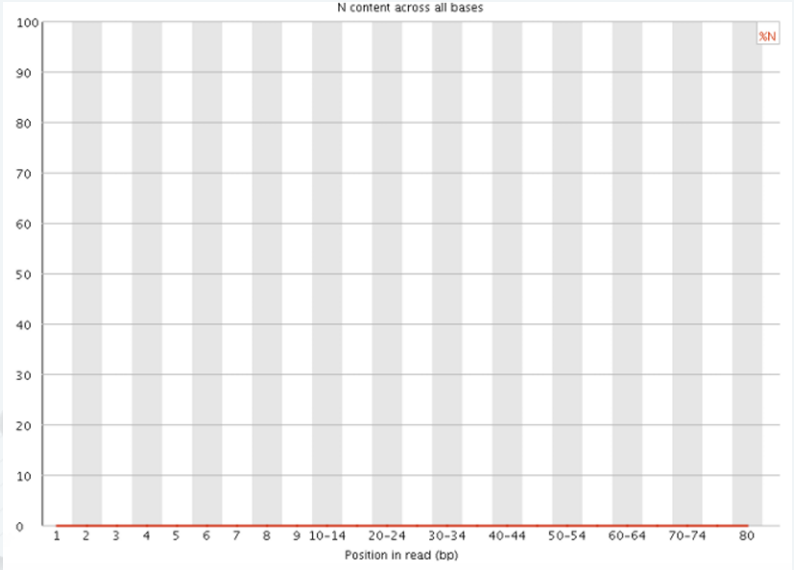
# FASTQC: Analisi

Il grafico **basi non assegnate per posizione** aiuta a identificare le posizioni nelle sequenze in cui sono presenti basi non identificabili ('N').

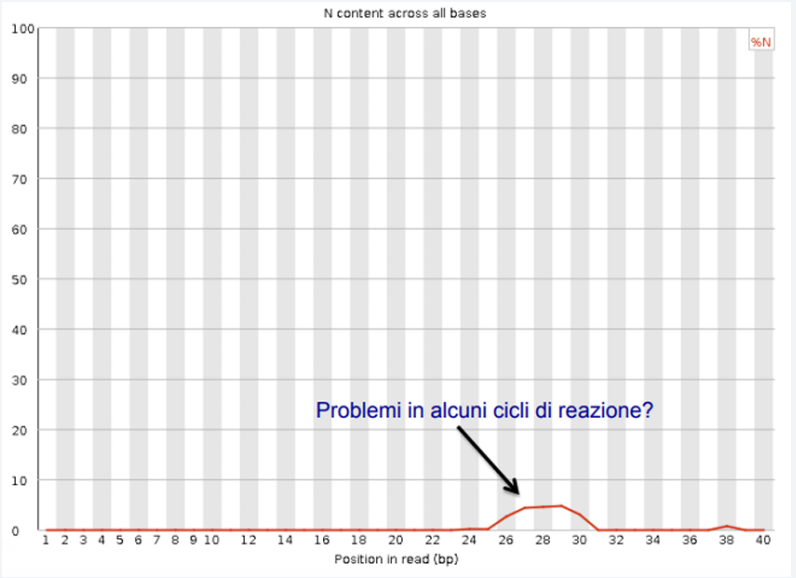
Una percentuale elevata di 'N' in alcune posizioni suggerisce che quelle parti delle sequenze non sono state sequenziate correttamente

# FASTQC: Analisi

## QUALITÁ BUONA

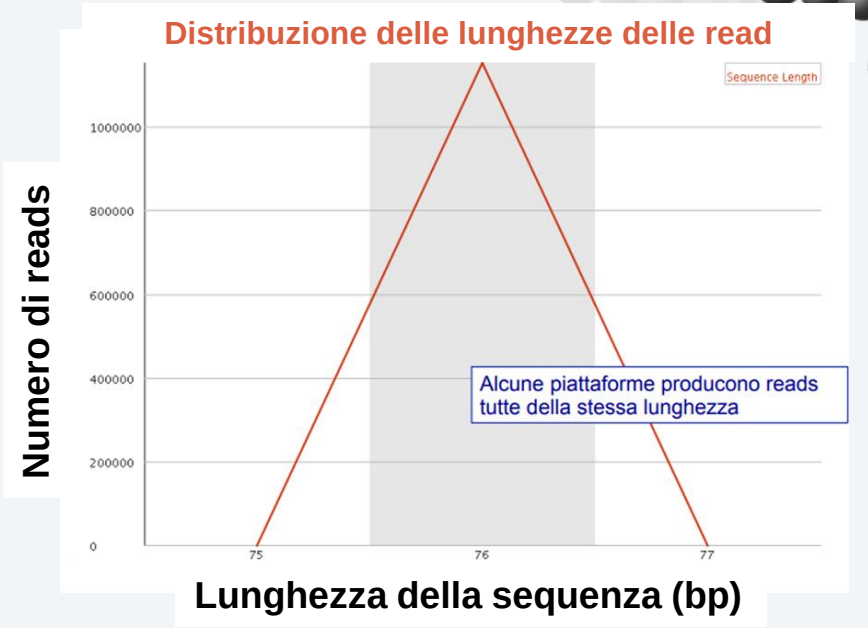



## QUALITÁ SCARSA



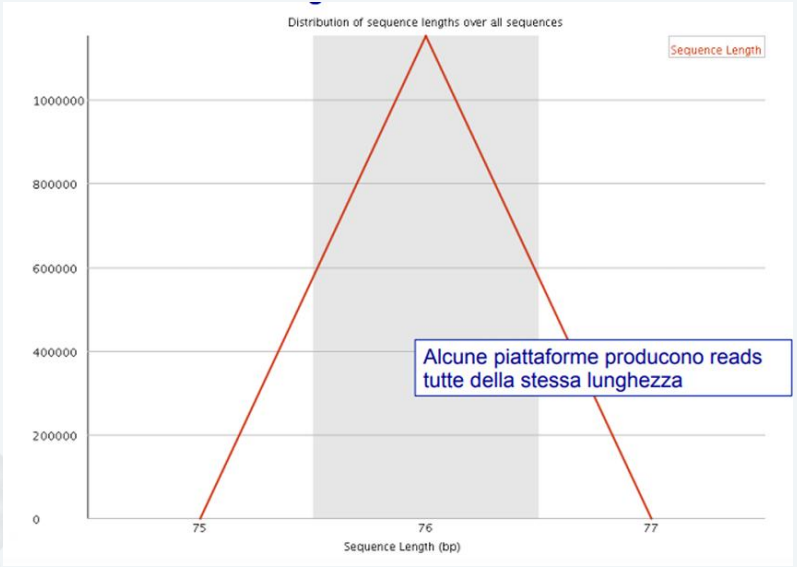
# FASTQC: Analisi

Il grafico **distribuzione delle lunghezze delle read** mostra la distribuzione delle lunghezze delle reads su tutte le sequenze.
Avere reads della stessa lunghezza è ottimale per le analisi

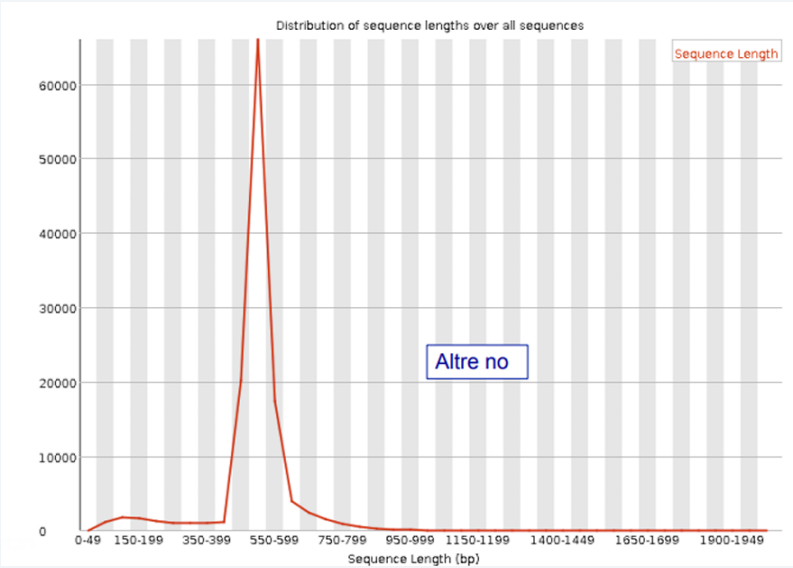


# FASTQC: Analisi

## QUALITÁ BUONA




## QUALITÁ SCARSA



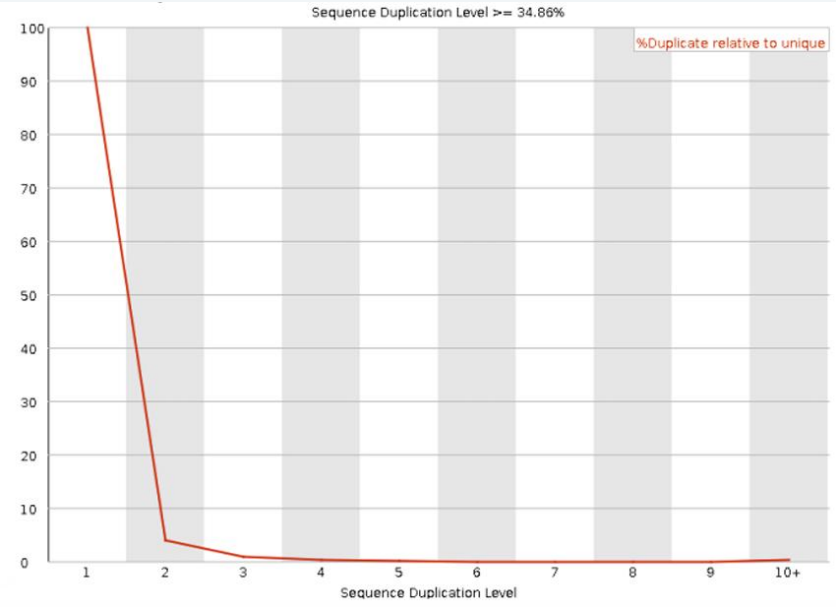
# FASTQC: Analisi

Il grafico del **livello di duplicazione delle reads** mostra quante volte una specifica read appare in una serie di dati sequenziati, in relazione al numero di copie di quella sequenza nel file.

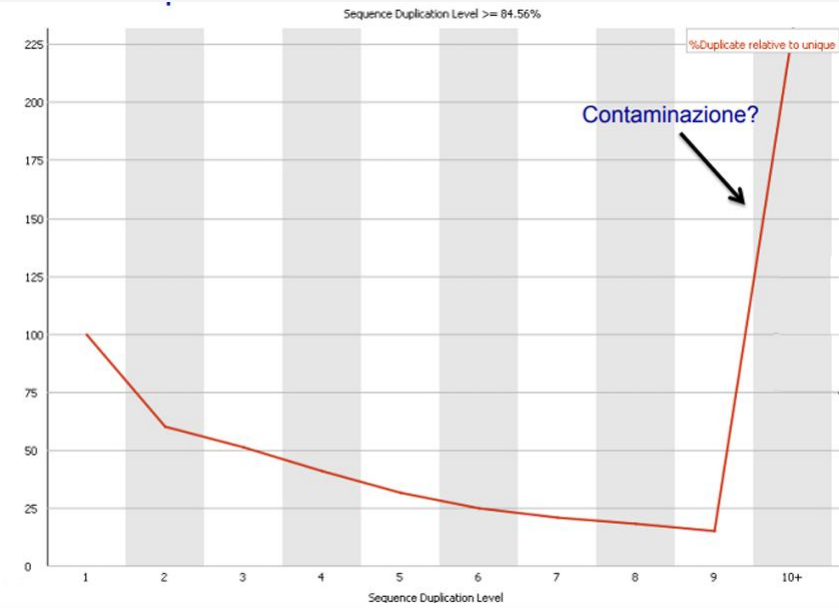
Questo tipo di grafico è spesso utilizzato per valutare la diversità delle sequenze all'interno di un campione sequenziato. Se una sequenza appare molte volte nel campione e costituisce una grande percentuale delle letture, potrebbe essere una sequenza dominante o molto comune (contaminazione)

# FASTQC: Analisi

## QUALITÁ BUONA




## QUALITÁ SCARSA



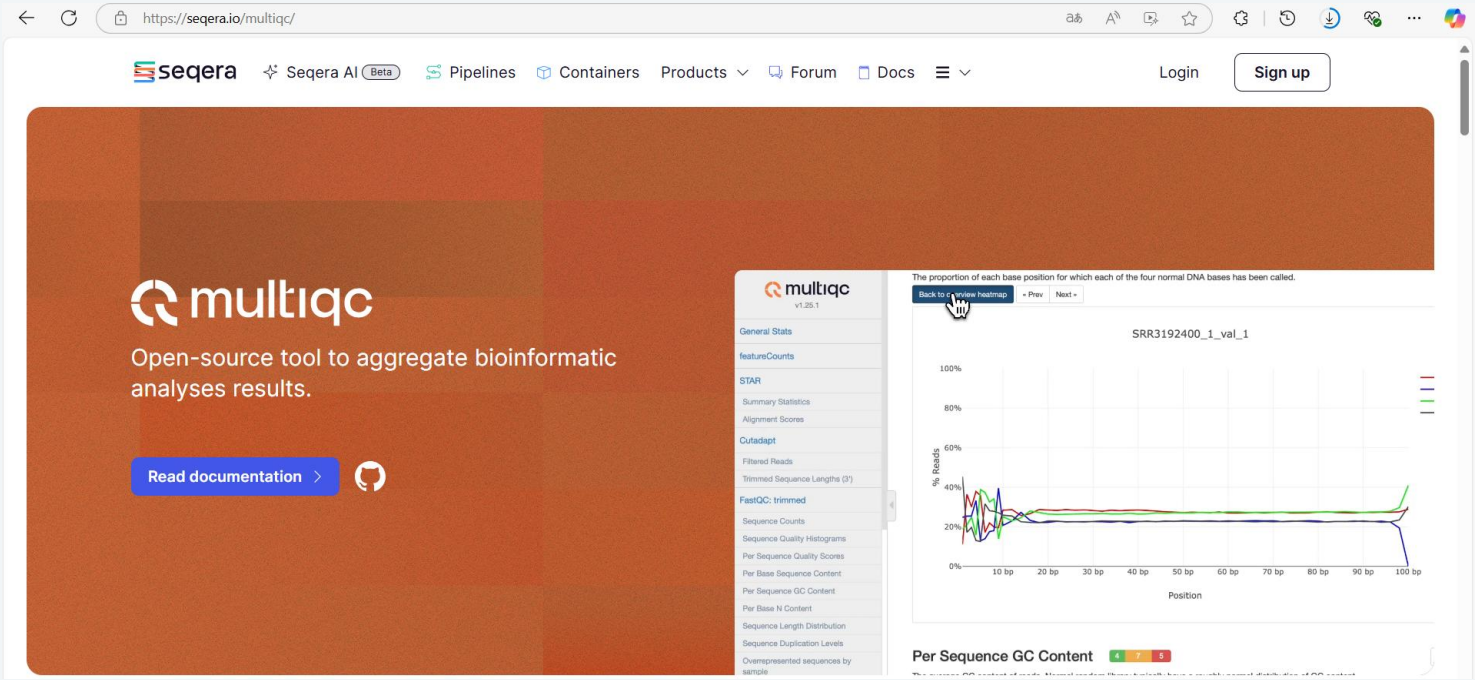
# MULTIQC

**MultiQC è un software che consente di generare report di qualità consolidati e visualizzazioni a partire dai risultati di più analisi FastQC o di altri strumenti di controllo della qualità**

- MultiQC unisce i risultati di più analisi in un unico report, semplificando la visualizzazione e la condivisione delle informazioni di qualità.
- Consente di confrontare facilmente le metriche di qualità tra più campioni o analisi.
- Aiuta i ricercatori a ottenere una visione d'insieme della qualità dei dati e a individuare rapidamente tendenze o anomalie tra i campioni o le analisi.
- Facilita anche la generazione di report chiari e comprensibili per la comunicazione dei risultati

# MULTIQC: Sito

## https://seqera.io/multiqc/



# MULTIQC: Installazione

E’ sufficiente il download e la decompressione del file

https://github.com/MultiQC/MultiQC 

Oppure procedere con il comando di installazione pip
**pip install multiqc**
In alternativa procedere con il comando di installazione conda
**conda install multiqc**

Dipendenze:
Richiede python

# MULTIQC: Comando

## multiqc [OPZIONI] -d *directory_contente_i_file_prodotti_da_FASTQC*

| Opzione | Significato |
|---|---|
| **-d o --dirs** | Scansiona anche le sottocartelle. Utile se i file di output sono organizzati in cartelle diverse |
| **-f o --force** | Sovrascrive i report esistenti senza chiedere conferma |
| **-n o --filename** | Specifica il nome del file di output per il report MultiQC (default: multiqc_report.html) |

# MULTIQC: Report

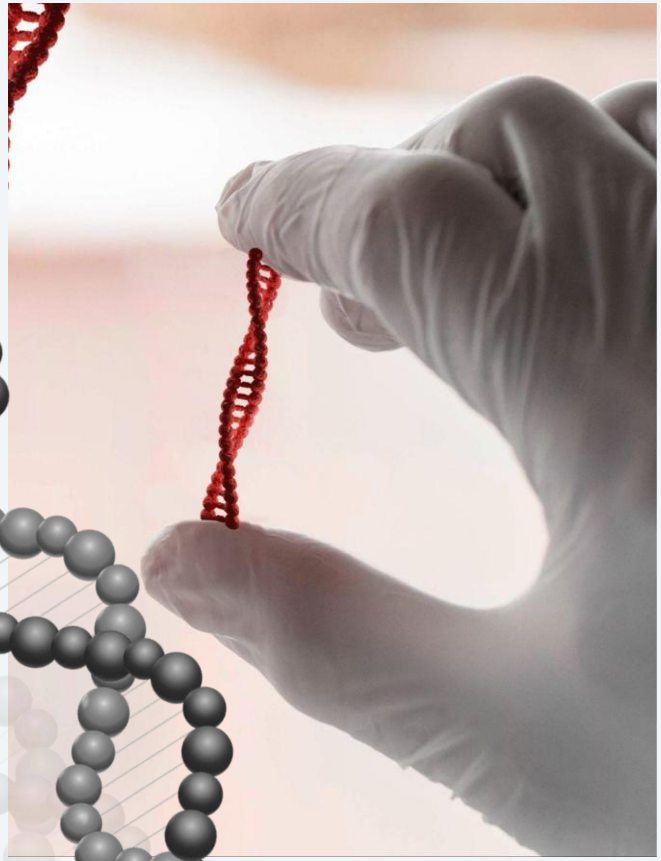
I report MultiQC hanno la tabella *Statistiche generali* che mostra una panoramica dei valori chiave.
Lo scopo della tabella è riunire le statistiche per ciascun campione dell'analisi in modo da poterle visualizzare in un unico posto.

General Statistics

Show Key

| Sample Name | % Assigned | M Assigned | % Mapped | M Mapped | Trimmed | % Dups | % GC | Length | M Seqs |
|---|---|---|---|---|---|---|---|---|---|
| SRR1067503_1 | 2.4% | 0.9 | 63.2% | 19.3 | 2.1% | 12.9% | 44% | 35 | 30.5 |
| SRR1067505_1 | 7.4% | 1.5 | 79.1% | 14.2 | 3.5% | 7.8% | 47% | 35 | 18.0 |
| SRR1067510_1 | 1.1% | 0.6 | 50.6% | 17.4 | 2.0% | 11.4% | 40% | 35 | 34.3 |
| SRR1067514_1 | 5.7% | 1.9 | 70.2% | 23.6 | 3.1% | 6.6% | 44% | 35 | 33.6 |
| SRR1067519_1 | 3.2% | 0.9 | 81.1% | 19.9 | 2.3% | 5.8% | 42% | 35 | 24.6 |
| SRR1067522_1 | 1.4% | 0.7 | 61.8% | 22.0 | 1.5% | 13.3% | 40% | 36 | 35.7 |

https://www.researchgate.net/publication/304033037_MultiQC_Summarize_analysis_results_for_multiple_tools_and_samples_in_a_single_report/figures?lo=1



# 03

# CONTROLLO DI QUALITA NGS III GENERAZIONE

# CONTROLLO QUALITÁ NANOPORE

**Nanoplot** è uno strumento che è utilizzato per l'analisi e la visualizzazione dei dati di sequenziamento generati da tecnologie come Oxford Nanopore Technologies (ONT).
Il report prodotto da Nanoplot contiene una serie di dati, grafici e metriche che forniscono informazioni dettagliate sulla qualità dei dati di sequenziamento e sulla loro distribuzione

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Il report include statistiche riassuntive:

- Media/Mediana/N50 reads lunghezza
- Media/Mediana/N50 reads qualità
- Numero di reads
- Totale delle basi generate
- Q-scores

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**La lunghezza media delle read prodotte dal sequenziatore**

Questo valore è espresso solitamente in numero di basi o in kilobasi (kb).

Il valore della lunghezza media delle read può essere influenzato da vari fattori, come il tipo di sequenziatore utilizzato, la qualità del campione e le condizioni di sequenziamento.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

In generale, una maggiore lunghezza media delle read può indicare una elevata qualità dei dati di sequenziamento, una migliore copertura del genoma sequenziato e una maggiore possibilità di identificare varianti genomiche.

Tuttavia, è importante considerare che una maggiore lunghezza media delle read potrebbe richiedere maggiori costi di sequenziamento e tempi di elaborazione dati più lunghi

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

In alcune applicazioni di assemblaggio del genoma, la presenza di read di corta lunghezza potrebbe essere utile per riempire lacune tra le regioni coperte dalle read di lunghezza maggiore

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

La **media dei valori di qualità di tutte le basi nelle read prodotte dal sequenziatore**
Questo valore fornisce un'indicazione generale della qualità complessiva dei dati di sequenziamento.
Un valore elevato di "Mean Read Quality" indica una maggiore precisione e affidabilità dei dati di sequenziamento, mentre un valore basso individua una maggiore presenza di errori nella lettura delle basi.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

È importante notare, però, che anche read di alta qualità possono contenere errori e che l'analisi successiva dei dati di sequenziamento può richiedere ulteriori controlli di qualità e correzioni degli errori

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**La lunghezza mediana delle read prodotte**

La mediana è un valore statistico più robusto rispetto alla media, in quanto è meno influenzata da valori anomali.

La "Median Read Length" fornisce un'indicazione più accurata della lunghezza delle read

# CONTROLLO QUALITÁ NANOPLOT

In statistica, data una distribuzione di un carattere quantitativo oppure qualitativo ordinabile (ovvero le cui modalità possano essere ordinate in base a qualche criterio), la mediana (o valore mediano) è il valore assunto nel mezzo della distribuzione

Per calcolare la mediana di  n dati

1. Si ordinano gli  n dati in ordine crescente;
2. Se il numero di dati è dispari la mediana corrisponde al valore centrale, ovvero al valore che occupa la posizione (n+1)/2.
   Se il numero n di dati è pari, la mediana è stimata utilizzando i due valori che occupano le posizioni n/2 e n/2+1 (generalmente si sceglie la loro media aritmetica)

| | | |
|---|---|---|
| Caso numero n dispari: | 2, 3, 5, 7, 10 | mediana=5 |
| Caso numero n pari : | 2, 3, 3, 5, 7, 10 | mediana=4 (cioè:(3+5)/2) |

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Il valore mediano della lunghezza delle reads è utile per valutare la distribuzione delle lunghezze delle read prodotte dal sequenziatore e per scegliere il parametro di taglio lunghezza da utilizzare durante l'analisi dei dati di sequenziamento. In generale, una valore alto della mediana indica una maggiore copertura e una maggiore precisione nell'assemblaggio dei dati di sequenziamento

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Tuttavia, è importante valutare anche altri parametri come la lunghezza delle read e la copertura del campione sequenziato per una valutazione completa della qualità dei dati di sequenziamento genomico

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**Il numero di reads indica il numero totale di read prodotte dal sequenziamento**

È un parametro importante per valutare la copertura del campione sequenziato e la quantità di informazioni che possono essere ottenute dal sequenziamento

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**Il valore N50 indica la lunghezza della read nella posizione mediana dell'intervallo di lunghezza delle read più lunghe**, ovvero la lunghezza a cui la metà delle basi sequenziate sono rappresentate da read di lunghezza uguale o superiore al valore N50

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Il valore N50 è un indice della distribuzione delle lunghezze delle read prodotte dal sequenziamento e rappresenta una stima della lunghezza mediana effettiva delle read.

Un valore N50 elevato indica che una grande percentuale delle read prodotte ha lunghezze superiori o uguali al valore N50, il che può essere vantaggioso per l'assemblaggio del genoma e l'identificazione di mutazioni, varianti e altre caratteristiche genomiche.

# CONTROLLO QUALITÁ NANOPLOT

**Calcolo dello N50**

1. Si ordinano tutte le reads/contig (o scaffold) per lunghezza in ordine decrescente.
2. Si sommano le lunghezze delle reads/contig in questo ordine finché non si raggiunge almeno il 50% del totale delle basi assemblate.
3. La lunghezza della reads/contig che raggiunge o supera questa soglia è il valore N50

In altre parole, il N50 rappresenta la lunghezza della contig mediana ponderata dalla quantità di basi, piuttosto che dal numero delle contig.

# CONTROLLO QUALITÁ NANOPLOT

Assemblaggio con 5 contig di queste lunghezze (in base pari a nucleotidi, nt):
Contig 1: 5000 nt  Contig 2: 4000 nt  Contig 3: 3000 nt
Contig 4: 2000 nt  Contig 5: 1000 nt

Calcolo di N50:

1. Calcolo della lunghezza totale dell'assemblaggio:
   Totale basi=5000+4000+3000+2000+1000=15000
2. Determinare il 50% del totale:
   50% di 15000=7500 (nt50)
3. Sommare le lunghezze ordinate finché non si supera 7500:
   Contig 1: 5000 nt → somma parziale = 5000
   Contig 2: 4000 nt → somma parziale = 5000 + 4000 = 9000 (supera 7500!)
4. La contig che porta la somma a 9000 è Contig 2, lunga 4000 nt.

Risultato:
L'N50 di questo assemblaggio è 4000 nt.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Ad esempio, un valore di "Read length N50" pari a 10.000 indica che la lunghezza mediana delle read più lunghe è di 10.000 basi.

Tuttavia, è importante notare che il valore N50 non fornisce informazioni sulla qualità delle basi sequenziate e sulla presenza di errori nella sequenza.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

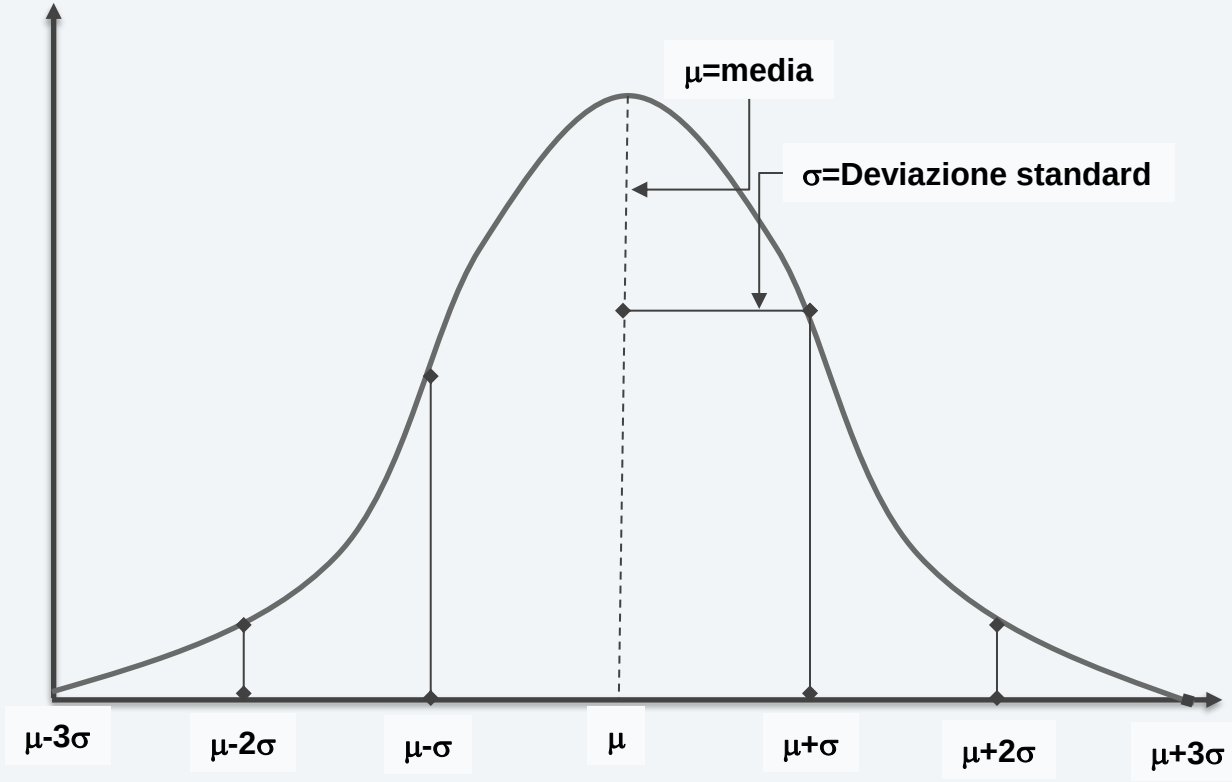
### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**La deviazione standard delle lunghezze delle read fornisce informazioni sulla variabilità delle lunghezze delle read all'interno dei dati di sequenziamento ottenuti**

Una deviazione standard delle lunghezze delle read maggiore indica una maggiore variabilità nelle lunghezze delle read, mentre una deviazione standard più piccola indica una minore variabilità e una maggiore uniformità nelle lunghezze delle read.

# CONTROLLO QUALITÁ NANOPLOT



# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

In generale, un valore più basso di stdev delle lunghezze delle read è preferibile poiché indica una maggiore uniformità nella lunghezza delle read sequenziate. Questa uniformità può semplificare l'analisi dei dati e migliorare la qualità complessiva dei risultati.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Un buon valore di stdev dipenderà anche dal tipo di campione e dall'obiettivo dell'analisi. Ad esempio, in alcuni casi, come nel sequenziamento di genomi, un valore di stdev delle lunghezze delle read molto basso potrebbe essere desiderabile per ridurre al minimo l'ambiguità nell'assegnazione delle read al genoma di riferimento.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Un valore di stdev delle lunghezze delle read che è significativamente più alto rispetto alla lunghezza media delle read può essere segnale di potenziali problemi durante la preparazione del campione o il processo di sequenziamento.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**La totalità delle basi (‘Total bases’) indica la somma di tutte le basi sequenziate per il campione considerato.**
Questo valore rappresenta la quantità totale di dati di sequenziamento generati per il campione, indipendentemente dalla lunghezza delle singole read prodotte

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Il valore ‘Total bases’ è un parametro importante per valutare la quantità di informazioni di sequenziamento prodotte per un campione e può essere utilizzato per valutare la copertura del genoma o il numero di regioni target coperte dal sequenziamento

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Il **numero**, la **percentuale**, e **megabasi** di reads sulle qualità rappresenta le informazioni relative alla qualità delle read prodotte dal sequenziatore

Il report può fornire il numero totale di read prodotte dal sequenziatore, nonché il numero e la percentuale di read che superano determinati cutoff di qualità, ovvero il valore minimo di qualità che deve essere soddisfatto per considerare una base come affidabile

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Una maggiore lunghezza media delle read prodotte dal sequenziatore Nanopore può offrire diversi vantaggi in termini di qualità dei dati di sequenziamento, come una migliore copertura del genoma sequenziato e una maggiore possibilità di identificare varianti genomiche.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Queste informazioni sono utili per valutare la qualità complessiva dei dati di sequenziamento prodotti e per determinare quali read sono più affidabili e quindi utilizzabili per l'analisi successiva.

In generale, il numero, la percentuale e le megabasi di read sopra i cut off di qualità sono importanti parametri di qualità dei dati di sequenziamento e rappresentano un indicatore della precisione e dell'affidabilità dei risultati dell'analisi successiva

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

Un elevato numero di read sopra i cut off di qualità indica una maggiore affidabilità dei dati di sequenziamento, mentre un basso numero di read sopra i cut off di qualità può indicare problemi di qualità dei dati o di preparazione del campione

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

**"Q-scores"**

I valori come Q5, Q7, Q10, Q12 e Q15 si riferiscono ai punteggi di qualità (Quality Scores) utilizzati per valutare la qualità dei dati di sequenziamento. Questi punteggi indicano la probabilità di errore nelle chiamate delle basi nucleotidiche in una sequenza di DNA o RNA e sono espressi sulla scala Phred.

# CONTROLLO QUALITÁ NANOPLOT

## NanoPlot report

### Summary statistics

| feature | |
|---|---|
| General summary | |
| Mean read length | 15,874.3 |
| Mean read quality | 31.4 |
| Median read length | 15,568.0 |
| Median read quality | 31.6 |
| Number of reads | 1,218,551.0 |
| Read length N50 | 16,093.0 |
| STDEV read length | 2,731.3 |
| Total bases | 19,343,585,808.0 |
| Number, percentage and megabases of reads above quality cutoffs | |
| >Q5 | 1218551 (100.0%) 19343.6Mb |
| >Q7 | 1218551 (100.0%) 19343.6Mb |
| >Q10 | 1218551 (100.0%) 19343.6Mb |
| >Q12 | 1218551 (100.0%) 19343.6Mb |
| >Q15 | 1218551 (100.0%) 19343.6Mb |
| Top 5 highest mean basecall quality scores and their read lengths | |

- Q5: indica che le basi sequenziate (in ciascuna reads) hanno una precisione del 67% o superiore, con una probabilità di errore del 33% o inferiore.
- Q7: indica che le basi sequenziate (in ciascuna reads) hanno una precisione dell'80% o superiore, con una probabilità di errore del 20% o inferiore.
- Q10: Indica che le basi sequenziate (in ciascuna reads) hanno una precisione del 90% o superiore, con una probabilità di errore del 10% o inferiore.
- Q15: Indica che le basi sequenziate (in ciascuna reads) hanno una precisione del 97% o superiore, con una probabilità di errore del 3% o inferiore.

# CONTROLLO QUALITÁ NANOPLOT

I grafici in un report generato da Nanoplot sono progettati per fornire una panoramica visuale delle metriche di qualità e delle caratteristiche dei dati di sequenziamento ottenuti.

I grafici in un report Nanoplot includono:

- Histogram of read lengths
- Histogram of read lengths after log transformation
- Weighted Histogram of read lengths
- Weighted Histogram of read lengths after log transformation
- Dynamic Histogram of Reads Length
- Il grafico del "Yield by length" rappresenta la quantità totale di basi (yield) prodotta dal sequenziamento in funzione della lunghezza delle read.
- Il grafico "Read lengths vs Average read quality plot using dots" rappresenta la relazione tra la lunghezza delle letture (asse x) e la qualità media delle basi nella lettura (asse y).
- "Read lengths vs Average read quality plot using a kernel density estimation"

# Grazie!

**Domande?**

franco.liberati@unitus.it
deb.scienceontheweb.com