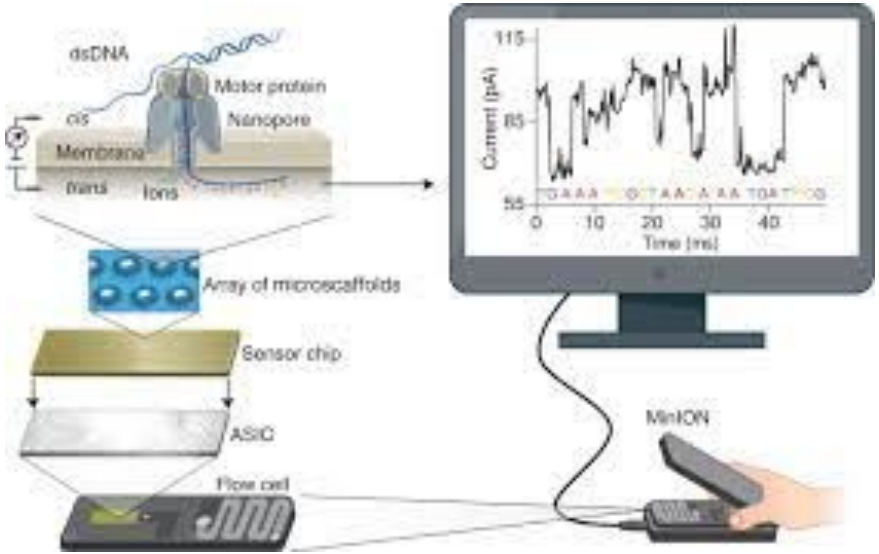
Controllo long reads nanopore: NANOPLOT



Nanoplot è uno strumento che viene 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.

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	

I moduli includono:

Statistiche riassuntive

- Media/Mediana/N50 reads length
- Media/Mediana/N50 reads quality
- Numero di reads
- Totale delle basi generate
- Q-scores

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	

- **" mean read length"** è una misura statistica che indica la lunghezza media delle read prodotte dal sequenziatore. Questo valore è espresso solitamente in numero di basi o in kilobasi (kb).

" mean read length"

Cosa influenza questo parametro?

Il valore della lunghezza media delle read può essere influenzato da vari fattori, come il tipo di sequenziatore utilizzato, la qualità del campione sequenziato e le condizioni di sequenziamento. In generale, una maggiore lunghezza media delle read può indicare una maggiore 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. Inoltre, a seconda dell'applicazione specifica, la lunghezza media delle read potrebbe non essere il parametro più importante da considerare. Ad esempio, 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.

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	

"Mean Read Quality" nel report di sequenziamento genomico Nanopore rappresenta la media dei valori di qualità di tutte le basi nelle read prodotte dal sequenziatore.

"Mean Read Quality"

· Questo valore fornisce un'indicazione generale della qualità complessiva dei dati di sequenziamento, in quanto riflette la precisione media del sequenziatore nella lettura delle basi.

Un valore di "Mean Read Quality" più elevato indica una maggiore precisione e affidabilità dei dati di sequenziamento, mentre un valore più basso indica una maggiore presenza di errori nella lettura delle basi. Tuttavia, è importante notare 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.

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	

"Median Read Length" in un report di sequenziamento genomico nanopore rappresenta la lunghezza mediana delle read prodotte dalla tecnologia di sequenziamento nanopore.

"Median Read Length"

- Questo valore indica la lunghezza delle read che divide l'insieme delle read in due parti uguali, in modo che la metà delle read sia più corta della mediana e l'altra metà sia più lunga.

La mediana è un valore statistico più robusto rispetto alla media, in quanto è meno influenzata da valori anomali o outliner. Questo significa che la "Median Read Length" può fornire un'indicazione più accurata della lunghezza delle read tipiche prodotte dalla tecnologia di sequenziamento nanopore.

La "Median Read Length" può essere 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 maggiore "Median Read Length" indica una maggiore copertura e una maggiore precisione nell'assemblaggio dei dati di sequenziamento.

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	

"Median Read Quality" in un report di sequenziamento genomica nanopore indica la qualità mediana delle read prodotte dalla tecnologia di sequenziamento nanopore.

"Median Read Quality"

- Questo valore rappresenta la qualità delle basi nella posizione mediana di tutte le read sequenziate.

Una "Median Read Quality" più alta indica che la maggior parte delle basi nelle read hanno una maggiore precisione nella chiamata della base. Tuttavia, è importante notare che la qualità delle basi può variare all'interno della stessa read e tra le diverse read sequenziate.

In generale, quindi, una "Median Read Quality" elevata è un indicatore positivo della qualità dei dati di sequenziamento prodotti dalla tecnologia di sequenziamento nanopore. 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.

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	

. **"Number of Reads"** in un report di sequenziamento genomica nanopore indica il numero totale di read prodotte dalla tecnologia di sequenziamento per il campione sequenziato.

I numero di read è un parametro importante per valutare la copertura del campione sequenziato e la quantità di informazioni che possono essere ottenute dal sequenziamento.

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	

"Read Length 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.

"Read Length N50"

Il valore N50 è un indice della distribuzione delle lunghezze delle read prodotte dalla tecnologia di sequenziamento genomica nanopore 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.

Ad esempio, un valore di "Read length N50" pari a 10.000 indicherebbe 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.

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	19,343,583,808.0
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	

"stdev read length" (deviazione standard delle lunghezze delle read) questo valore fornisce informazioni sulla variabilità delle lunghezze delle read all'interno dei dati di sequenziamento ottenuti.

"stdev read length"

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. 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.

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.

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.

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	

- **"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.

"Total bases"

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.

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	

"Number, percentage and megabases of reads above quality cutoffs"

- rappresenta le informazioni relative alla qualità delle read prodotte dal sequenziatore.

"Number, percentage and megabases of reads above quality cutoffs"

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.

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.

Il report può anche fornire la quantità totale di megabasi delle read che superano i cut off di qualità specificati. 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. 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.

"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.

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

untitled

File Edit Format Run Options Window Help

```python
import sys
import gzip

# Funzione per calcolare il numero di basi con punteggio di qualità Q30 in una stringa di qualità
def calculate_q30(quality_string):
    # La stringa di qualità in un file FASTQ è codificata in ASCII con un offset di 33.
    # Per esempio, il carattere "I" rappresenta un punteggio Phred di 40.
    return sum(1 for char in quality_string if ord(char) - 33 >= 30)

def main():
    # Verifica che il numero di argomenti sia corretto
    if len(sys.argv) != 3:
        print("Utilizzo: python calculate_q30.py <file_fastq_sample1> <file_fastq_sample2>")
        sys.exit(1)

    total_bases = 0  # Contatore per il totale delle basi
    q30_bases = 0    # Contatore per le basi con qualità Q30

    # Analizza entrambi i file FASTQ forniti come argomenti
    for fastq_file in sys.argv[1:3]:
        # Apri il file FASTQ, potrebbe essere compresso (gzip) o no
        with gzip.open(fastq_file, 'rt') if fastq_file.endswith(".gz") else open(fastq_file, 'r') as f:
            # Ogni sequenza nel file FASTQ è rappresentata da 4 righe
            for line in f:
                sequence = next(f).strip()  # Sequenza di basi
                _ = next(f)  # Linea del separatore (+)
                quality = next(f).strip()   # Stringa di qualità

                # Aggiorna i contatori
                total_bases += len(quality)
                q30_bases += calculate_q30(quality)

    # Calcola la percentuale di basi con qualità Q30
    q30_percentage = (q30_bases / total_bases) * 100
    print(f"Percentuale di basi con Q30: {q30_percentage:.2f}%")

# Esegui la funzione main() quando lo script viene avviato
if __name__ == "__main__":
    main()
```

Esempio di codice python per calcolare la quality Q30 di due campioni paired end

"Scala Phred"

La scala Phred (o sistema Phred) è una scala logaritmica utilizzata per esprimere la qualità delle chiamate delle basi nucleotidiche nei dati di sequenziamento del DNA o RNA. Questa scala è ampiamente utilizzata nell'ambito della genomica e della bioinformatica per quantificare e rappresentare la precisione delle chiamate delle basi.

La scala Phred è definita dalla seguente equazione:

$Q = -10 * \log10(P)$

Dove:

- "Q" rappresenta il punteggio Phred.
- "P" rappresenta la probabilità di errore nella chiamata della base (generalmente compresa tra 0 e 1).

Quindi, per calcolare il punteggio Phred, si prende il logaritmo base 10 della probabilità di errore, lo si moltiplica per -10 e si ottiene il punteggio Phred risultante.

I punteggi Phred vengono espressi come numeri interi positivi. Maggiore è il punteggio Phred, maggiore è la qualità e minore è la probabilità di errore nella chiamata della base.

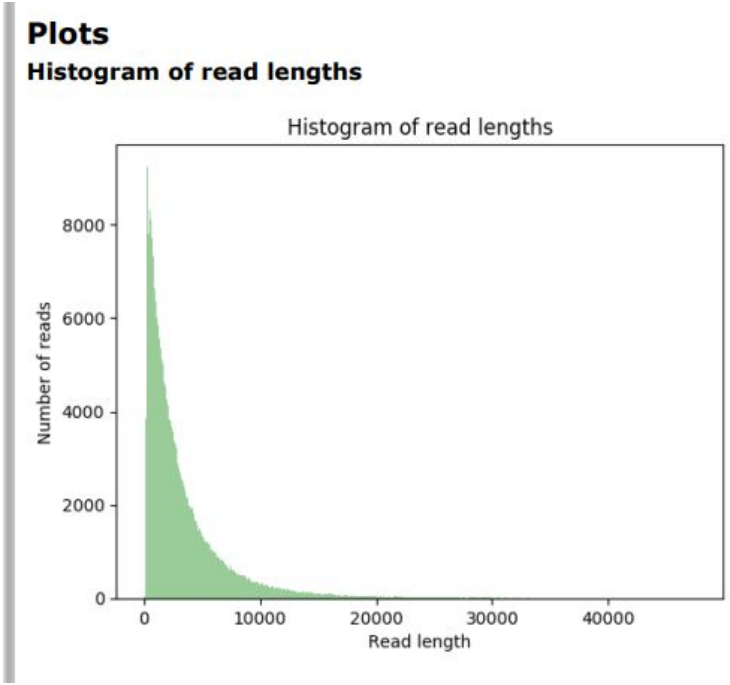
GRAFICI REPORT NANOPLOT

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"

"Histogram of read lengths"



Il grafico rappresenta la distribuzione delle lunghezze delle read in un dataset di sequenziamento. Questo tipo di grafico è comunemente utilizzato per visualizzare la variabilità delle lunghezze delle read all'interno di un campione e per valutare la qualità dei dati di sequenziamento.

- **Asse delle X:** rappresenta le lunghezze delle read. Ogni bin o intervallo sull'asse delle X corrisponde a un intervallo specifico di lunghezza delle read.
- **Asse delle Y:** rappresenta il numero di read che cadono in ciascun intervallo di lunghezza. L'asse delle Y mostra quindi la frequenza con cui le read hanno lunghezze comprese in ciascun intervallo.

La distribuzione può fornire informazioni sulle caratteristiche del campione sequenziato, come la presenza di contaminanti, la qualità del DNA o la presenza di regioni ripetitive.

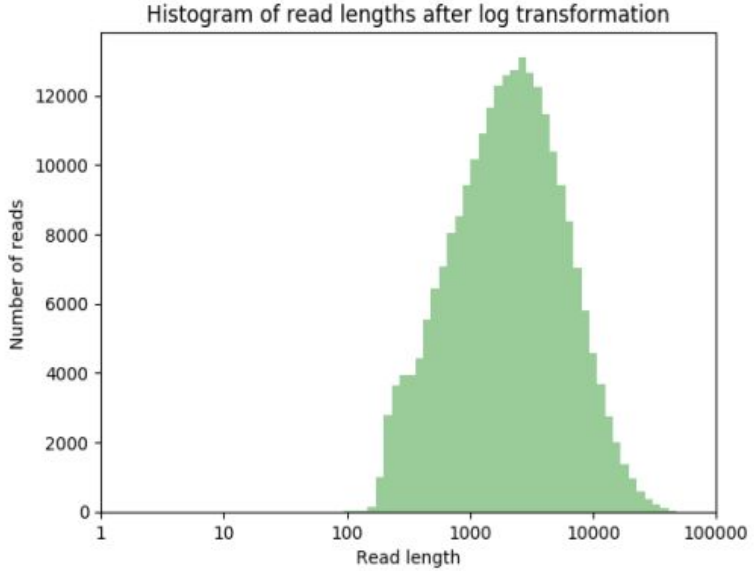
"Histogram of read lengths"

cosa si può apprendere da questo tipo di istogramma:

- **Lunghezza dominante delle read:** L'istogramma mostrerà un picco o uno o più picchi. La lunghezza corrispondente al picco più alto rappresenta la lunghezza dominante delle read nel campione. Questa informazione è importante per la pianificazione delle analisi successive.
- **Variabilità delle lunghezze:** Guardando la larghezza dell'istogramma e la diffusione delle lunghezze delle read, puoi valutare la variabilità delle lunghezze delle read nel campione. Una distribuzione stretta indica lunghezze uniformi, mentre una distribuzione ampia suggerisce maggiore variabilità.
- **Presenza di read anormali:** Le code dell'istogramma possono indicare la presenza di read anomale. Ad esempio, code lunghe potrebbero essere indicative di read estremamente lunghe o contaminazioni, mentre code corte potrebbero suggerire errori di sequenziamento o problemi nel campione.
- **Qualità dei dati:** La forma dell'istogramma può rivelare informazioni sulla qualità dei dati di sequenziamento. Ad esempio, una distribuzione regolare e ben definita è spesso indicativa di dati di alta qualità, mentre una distribuzione irregolare o multi-modale può indicare problemi.
- **Adattamento agli obiettivi dell'analisi:** Comprendere la distribuzione delle lunghezze delle read è essenziale per determinare se i dati soddisfano gli obiettivi specifici dell'analisi. Ad esempio, se stai cercando di assemblare un genoma, è importante che le read abbiano lunghezze adeguate per coprire le regioni genomiche più lunghe.

"Histogram of read lengths after log transformation"

Histogram of read lengths after log transformation



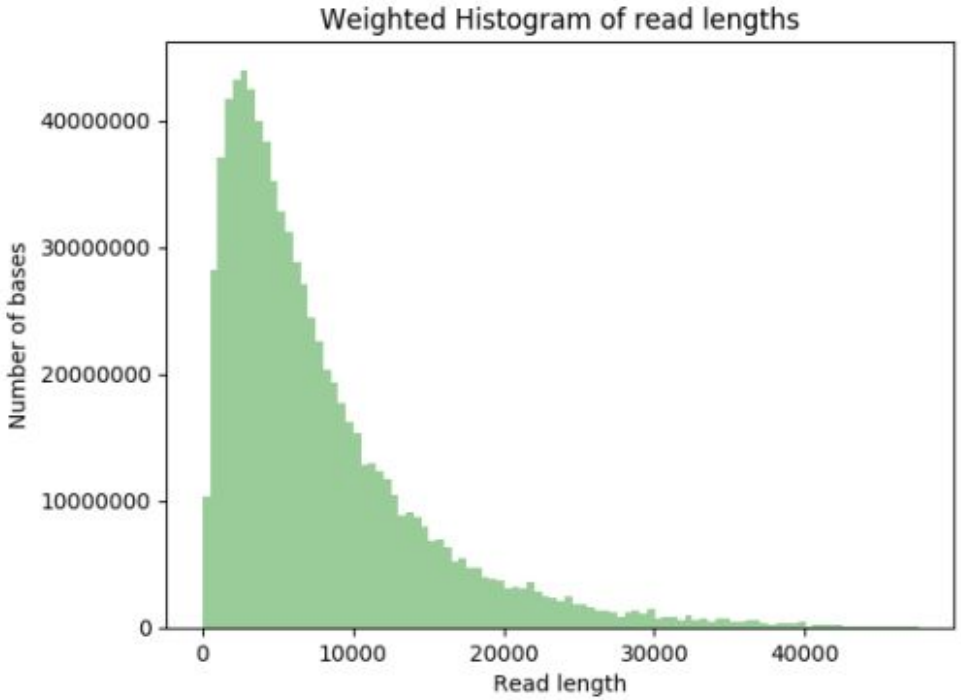
Questo viene spesso fatto per visualizzare meglio i dati che hanno un'ampia gamma di valori, poiché la scala logaritmica può aiutare a evidenziare le differenze nei dati che potrebbero non essere evidenti con la scala lineare.

"Histogram of read lengths after log transformation"

cosa si può apprendere da un istogramma delle lunghezze delle read dopo la trasformazione logaritmica:

- **Visualizzazione delle code:** La trasformazione logaritmica può "espandere" le code dell'istogramma, rendendo più evidenti le differenze nelle lunghezze delle read più lunghe o più corte. Questo può essere utile per identificare eventuali read estremamente lunghe o corte nel dataset.
- **Distribuzione delle lunghezze delle read:** Dopo la trasformazione logaritmica, l'istogramma potrebbe mostrare una distribuzione più simile a una distribuzione normale o gaussiana. Questo può rendere più facile valutare la simmetria e la forma generale della distribuzione delle lunghezze delle read.
- **Rivelazione di dettagli:** La trasformazione logaritmica può mettere in evidenza dettagli che potrebbero essere compressi in un istogramma standard. Ad esempio, se ci sono variazioni significative nelle lunghezze delle read in una parte specifica del dataset, queste variazioni potrebbero essere più evidenti dopo la trasformazione logaritmica.

"Weighted Histogram of read lengths"



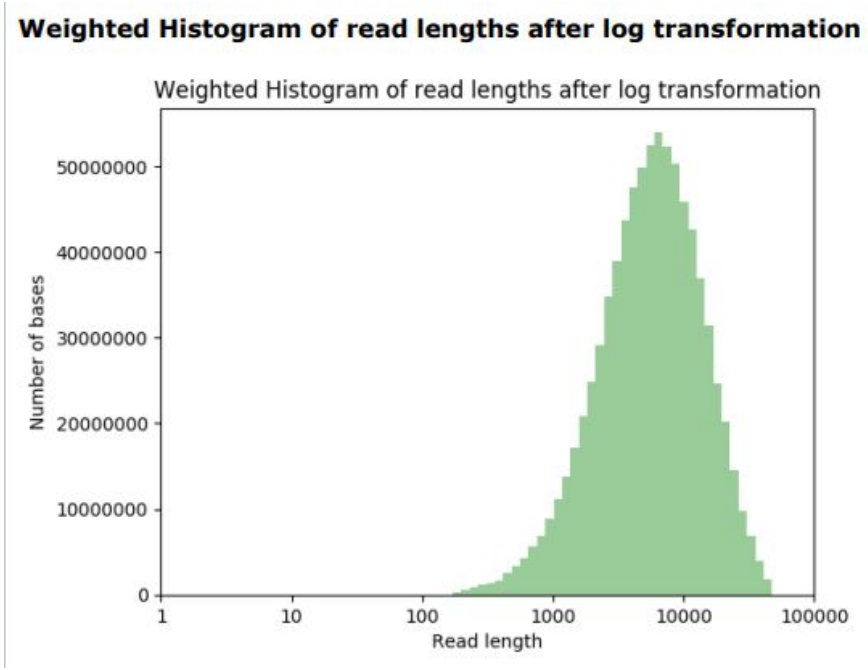
A "Weighted Histogram of read lengths" è un tipo di grafico che rappresenta la distribuzione delle lunghezze delle read in un dataset di sequenziamento, ma con l'aggiunta di un peso associato a ciascuna read. Questo tipo di grafico è utilizzato per enfatizzare l'importanza di alcune read rispetto ad altre in base a un criterio specifico, come la qualità, la copertura o altre metriche di interesse.

"Weighted Histogram of read lengths"

cosa si può apprendere da un istogramma ponderato delle lunghezze delle read:

- **Distribuzione delle lunghezze delle read pesate:** L'istogramma mostra la distribuzione delle lunghezze delle read, ma le read sono ponderate in base a un criterio specifico. Ad esempio, le read di alta qualità possono essere pesate più pesantemente rispetto a quelle di bassa qualità.
- **Focus su subset specifici:** Questo tipo di grafico consente di concentrarsi sulle read che soddisfano un certo requisito o criterio di selezione. Ad esempio, se stai cercando di analizzare solo le read di alta qualità, puoi assegnare loro un peso maggiore e visualizzare solo quelle nel grafico.
- **Valutazione dell'impatto delle read pesate:** Puoi determinare quanto influiscono le read con un peso maggiore sulla distribuzione complessiva delle lunghezze delle read. Questo è particolarmente utile quando si prendono decisioni sull'elaborazione o l'analisi dei dati in cui alcune read sono considerate più importanti di altre.
- **Rilevare pattern o tendenze specifiche:** La ponderazione può evidenziare pattern o tendenze specifiche nelle lunghezze delle read che potrebbero non essere immediatamente evidenti in un istogramma standard.

"Weighted Histogram of read lengths after log transformation"



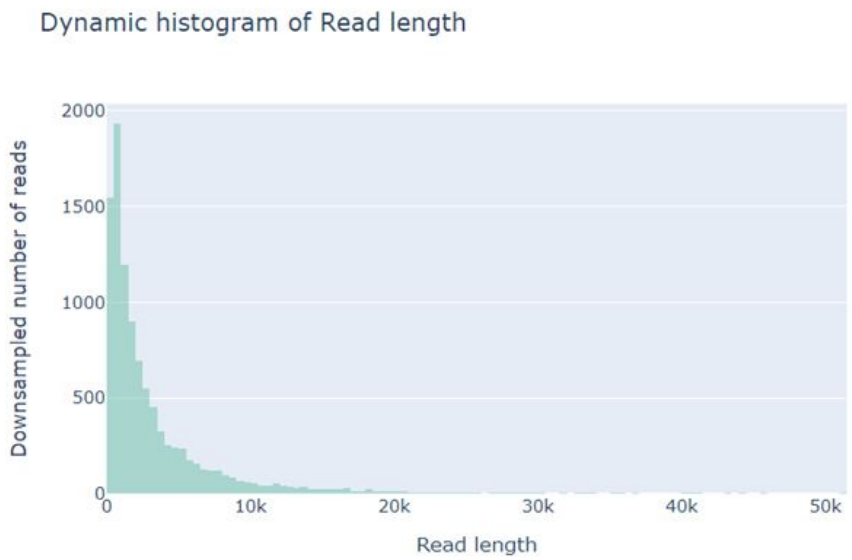
"Weighted Histogram of read lengths after log transformation" è un grafico che combina due aspetti importanti nell'analisi dei dati di sequenziamento: la distribuzione delle lunghezze delle read e l'effetto della trasformazione logaritmica, mentre si attribuiscono pesi alle read in base a criteri specifici.

"Weighted Histogram of read lengths after log transformation"

cosa si può apprendere da un istogramma di questo tipo:

- **Distribuzione delle lunghezze delle read:** L'istogramma mostra come le lunghezze delle read sono distribuite nel tuo dataset dopo la trasformazione logaritmica. Questa visualizzazione può aiutarti a valutare la distribuzione delle lunghezze delle read in modo più dettagliato rispetto a un istogramma non ponderato.
- **Effetto della trasformazione logaritmica:** La trasformazione logaritmica è spesso utilizzata per evidenziare dettagli nei dati e per gestire meglio la variabilità tra le read. L'istogramma mostra come questa trasformazione influisce sulla distribuzione delle lunghezze delle read e può rivelare eventuali pattern o tendenze emergenti dopo la trasformazione.
- **Ponderazione delle read:** L'aggiunta del peso alle read può evidenziare specifiche categorie di read o regioni che hanno un impatto maggiore sull'analisi. Questo consente di concentrarsi sulle read di maggiore interesse o di attribuire importanza a specifici aspetti dell'analisi dei dati.
- **Identificazione di outlier o gruppi distinti:** L'istogramma ponderato delle lunghezze delle read dopo la trasformazione logaritmica può rivelare gruppi distinti di read con lunghezze simili o outlier con lunghezze fuori dal comune. Questa informazione può essere utile per la segmentazione dei dati o per la selezione di read specifiche per ulteriori analisi.

"Dynamic Histogram of Reads Length"



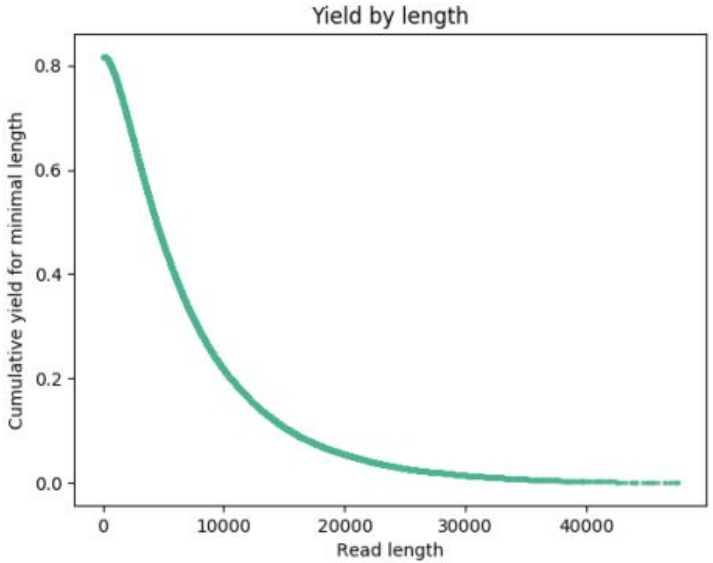
A "Dynamic Histogram of Reads Length" è un tipo di grafico che rappresenta la distribuzione delle lunghezze delle read in un dataset di sequenziamento in modo dinamico o in tempo reale. Questo tipo di grafico è spesso utilizzato per monitorare e visualizzare l'evoluzione delle lunghezze delle read durante il processo di sequenziamento o in un'applicazione di monitoraggio in tempo reale.

"Dynamic Histogram of Reads Length"

cosa si può apprendere da un istogramma dinamico delle lunghezze delle read:

- **Monitoraggio in tempo reale:** Questo grafico ti consente di visualizzare le lunghezze delle read man mano che vengono generate durante il sequenziamento. Puoi osservare come le lunghezze delle read variano nel tempo.
- **Valutazione della stabilità:** Puoi utilizzare l'istogramma dinamico per verificare la stabilità delle lunghezze delle read durante il sequenziamento. Eventuali fluttuazioni o variazioni anomale possono essere immediatamente rilevate.
- **Controllo della qualità:** Se le lunghezze delle read sono importanti per la tua applicazione, puoi utilizzare questo tipo di grafico per rilevare tempestivamente eventuali problemi o errori che potrebbero influenzare la distribuzione delle lunghezze delle read.
- **Adattamento agli obiettivi:** Puoi regolare il grafico per visualizzare specifiche categorie di lunghezze delle read o specifici intervalli di interesse, a seconda degli obiettivi della tua analisi.
- **Identificazione di tendenze:** Puoi osservare tendenze emergenti nelle lunghezze delle read nel corso del tempo, il che può fornire insight sugli eventi o sui processi che stanno influenzando il sequenziamento.

"Yield by length"



Il grafico del "Yield by length" rappresenta la quantità totale di basi (yield) prodotta dal sequenziamento in funzione della lunghezza delle read.

Questo tipo di grafico è utile per valutare la resa del sequenziamento in relazione alle diverse lunghezze delle read.

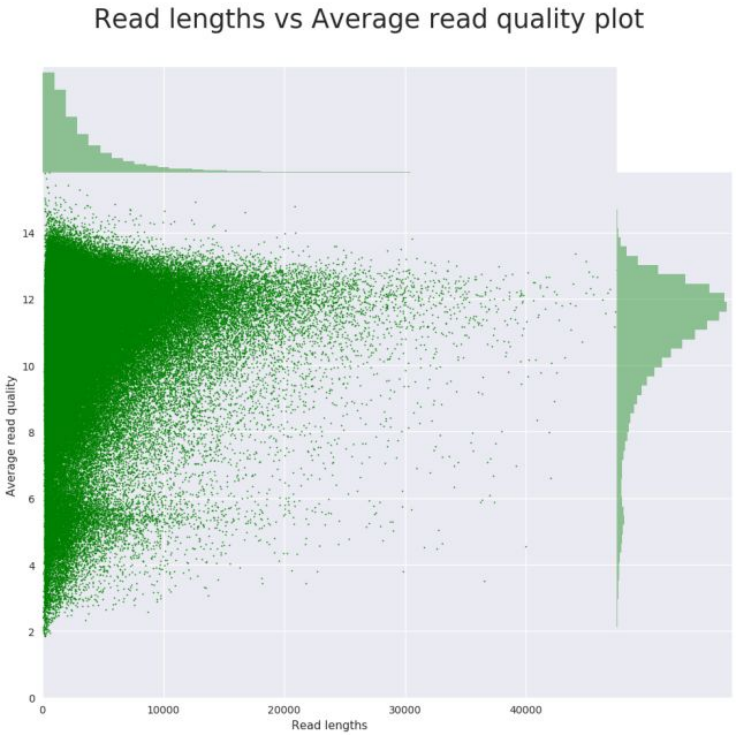
"Yield by length"

Questo tipo di grafico è utile per valutare la resa del sequenziamento in relazione alle diverse lunghezze delle read. L'interpretazione del grafico consiste nell'osservare se ci sono delle lunghezze di read che producono un yield maggiore rispetto ad altre.

In genere, è possibile notare un aumento dell'yield per le read di lunghezza medie e una diminuzione dell'yield per le read troppo corte o troppo lunghe. Inoltre, il grafico del "Yield by length" può fornire indicazioni sulla qualità del sequenziamento.

Se l'yield diminuisce bruscamente per le read di lunghezza maggiore, potrebbe essere un segnale di problemi nella qualità del sequenziamento in quella regione specifica del genoma o di altri fattori che influenzano la lettura delle basi. In generale, è importante valutare attentamente il grafico del "Yield by length" per ottimizzare il sequenziamento e ottenere i migliori risultati possibili.

"Read lengths vs Average read quality plot using dots"



Un grafico "Read lengths vs Average read quality plot using dots" è uno strumento utile per esplorare la relazione tra la lunghezza delle read e la loro qualità media all'interno di un dataset di sequenziamento.

In questo tipo di grafico, ogni punto rappresenta una read specifica, e la sua posizione sul grafico è determinata dalla sua lunghezza (asse delle X) e dalla sua qualità media (asse delle Y).

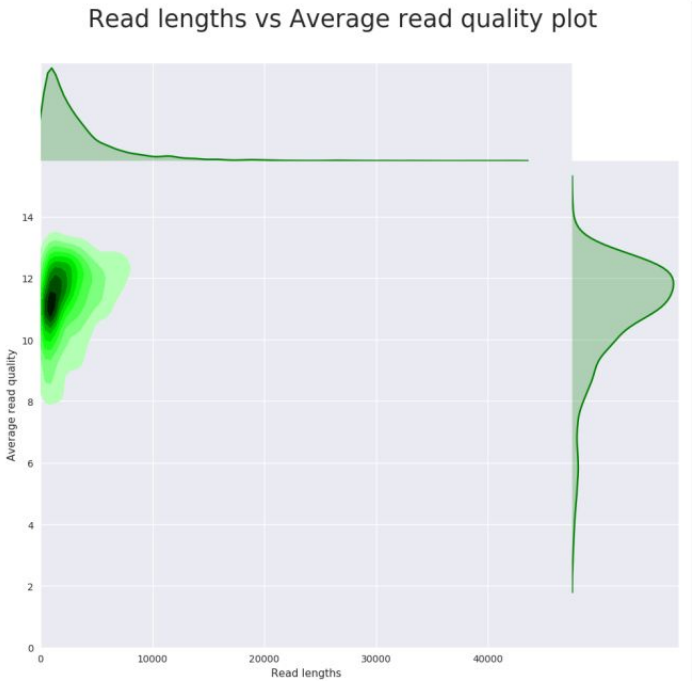
"Read lengths vs Average read quality plot using dots"

In questo grafico, ogni punto rappresenta una singola lettura e la sua posizione sul grafico è determinata dalla sua lunghezza e dalla sua qualità media delle basi. I punti sono disposti in modo da formare una nuvola di punti che può essere interpretata per comprendere la distribuzione di lunghezza e qualità delle letture.

L'interpretazione di questo grafico dipende dalla distribuzione dei punti. Se i punti sono distribuiti uniformemente lungo l'asse x e y, significa che ci sono letture di diverse lunghezze e qualità nel campione. Se, al contrario, i punti sono concentrati in una regione specifica del grafico, ciò indica che ci sono letture con una lunghezza e una qualità particolari che sono più comuni nel campione.

In generale, ci si aspetta che le letture più lunghe abbiano una qualità media inferiore rispetto alle letture più corte, poiché la probabilità di errori aumenta con la lunghezza della lettura. Tuttavia, questo può variare in base alla tecnologia di sequenziamento utilizzata e alle condizioni del campione.

"Read lengths vs Average read quality plot using a kernel density estimation"



· Il grafico "Read lengths vs Average read quality plot using a kernel density estimation" rappresenta la relazione tra la lunghezza delle letture e la qualità media delle basi, utilizzando un estimazione della densità del kernel.

In pratica, l'asse x rappresenta la lunghezza delle letture, mentre l'asse y rappresenta la qualità media delle basi. Il grafico mostra quindi come la qualità delle basi cambia al variare della lunghezza delle letture.

"Read lengths vs Average read quality plot using a kernel density estimation"

L'estimazione della densità del kernel è un metodo statistico per stimare la distribuzione di probabilità di una variabile casuale. In questo caso, viene utilizzata per stimare la densità di probabilità congiunta della lunghezza delle letture e della qualità media delle basi.

Interpretare questo grafico significa osservare come la qualità delle basi varia al variare della lunghezza delle letture. Ad esempio, se il grafico mostra una curva a campana con un picco intorno a una determinata lunghezza delle letture, potrebbe indicare che le letture di quella determinata lunghezza hanno una qualità media più alta rispetto ad altre lunghezze.

Invece, se il grafico mostra una distribuzione uniforme, potrebbe indicare che la qualità delle basi non dipende dalla lunghezza delle letture. Questo tipo di grafico può essere utile per valutare la qualità del sequenziamento e identificare eventuali problematiche come la presenza di errori sistematici nelle letture di determinate lunghezze.