UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Valutazione del tasso d'errore di Kraken 2 per sequenze metagenomiche di lunghezze variabili

Relatore:

prof. Matteo Comin

Laurenado:

Trincanato Marco

ANNO ACCADEMICO: 2021/2022

Data di laurea: 23 Settembre 2022

Abstract

La metagenomica consiste nello studio di comunità microbiche direttamente nel loro ambiente. In quest'ambito, un'operazione fondamentale è la classificazione tassonomica di sequenze metagenomiche. Questo processo richiede lo sviluppo di strumenti sempre più performanti in termini di tempo e affidabilità. Kraken 2 è un software che permette l'assegnamento di etichette tassonomiche alle letture di sequenziamento. Tuttavia, per rendere il software più veloce, sono state adottate delle tecniche che possono introdurre un errore nella classificazione. In questa tesi, si valuterà il tasso d'errore di kraken 2 per sequenze di lunghezze diverse.

Abstract

Metagenomics is the study of microbial communities directly in their environment. In this context, a fundamental operation is the taxonomic classification of metagenomic sequences. This process requires the development of increasingly performing tools in terms of time and reliability. Kraken 2 is a software that allows the assignment of taxonomic labels to sequencing reads. However, to make the software faster, have been adopted techniques that may produce some errors in the classification. In this thesis, we will evaluate the kraken 2 error rate for sequences of different lengths.

Indice

1 Introduzione

La metagenomica mira a quantificare la composizione del microbioma attraverso il recupero e il sequenziamento di tutto il DNA proveniente dai campioni estratti direttamente dall'ambiente. I primi studi furono condotti da Staley e Konopka, nel 1985. Negli anni '90, diverse ricerche ampliarono la conoscenza in questo campo e, nel 1998, il termine "metagenomica" è stato proposto per la prima volta [2]. Dagli anni 2000, si è assistito ad un notevole progresso nella metagenomica grazie all'aumento della disponibilità di tecnologie di sequenziamento ad alto rendimento che hanno ridotto costi e tempi [1].

Un tipico studio metagenomico comprende diverse fasi. Si inizia con la raccolta e il sequenziamento dei campioni e si termina con l'analisi computazionale dei risultati ed un'eventuale validazione [7]. Esistono principalmente due approcci per eseguire il sequenziamento dei campioni: lo studio di alcuni geni marcatori, solitamente regioni di un gene rRNA (16S per i batteri) e il sequenziamento "shotgun" di tutto il DNA microbico, senza la selezione di nessun gene specifico. Infatti, il termine "approccio metagenomico" si riferisce ad un ampio spettro di metodi per l'analisi di informazioni genetiche delle comunità microbiche recuperate dall'ambiente. Tuttavia, il termine "metatassonomica" è più adatto ad indicare il sequenziamento del gene marcatore, poiché non consente di indagare sull'intero genoma [3]. Nel sequenziamento "shotgun", i genomi vengono suddivisi casualmente in numerosi piccoli frammenti, che vengono poi sequenziati e assemblati in contigs, cioè gruppi di lunghi frammenti di DNA sovrapposti. Invece, la metatassonomica utilizza un approccio metodologico chiamato metabarcoding. Il metabarcoding[6] consiste nell'identificare la composizione tassonomica di una comunità attraverso l'amplificazione e il sequenziamento dei geni marcatori quali il 16S per i procarioti, ITS per i funghi e 18S per la maggior parte degli eucarioti.

L'approccio metagenomico ha il vantaggio di non amplificare il campione di interesse delle regioni genomiche, rendendo la classificazione immune ai bias che possono verificarsi durante questa fase. Tuttavia, ha un limite importante rispetto al metabarcoding: si concentra soprattutto sulle comunità microbiche, a causa di diversi fattori. Tra questi, quello più significativo è la relativa scarsità di sequenze genomiche per organismi multicellulari nei database [4]. Il divario tra i database metagenomici e di metabarcoding si sta, però, colmando rapidamente. Ad esempio, il progetto Earth BioGenome mira a sequenziare i genomi di oltre un milione specie eucariotiche entro il prossimo decennio. In questo modo, la metagenomica potrà essere usata per valutare l'appartenenza alle comunità di piante e

animali, soprattutto nei casi in cui gli organismi sono difficili da osservare o degradati.

1.1 Tecnologie di lettura

La lunghezza di lettura delle tecnologie di sequenziamento si riferisce al numero di coppie di basi (bp) sequenziate da un frammento di DNA. Dopo il sequenziamento, è possibile ricostruire l'intera sequenza di DNA utilizzando le regioni di sovrapposizione tra le letture per assemblare e allineare le letture a un genoma di riferimento. Le letture si dividono generalmente in lunghe e corte.

Tra le tecnologie che generano letture corte, si evidenziano Sanger, 454/Roche ed Illumina [8]. Sanger, tra le tre, è stata la prima tecnologia sviluppata, e può generare letture anche superiori a 700bp, con un basso tasso di errore. 454/Roche ed Illumina sono, invece, tecnologie di nuova generazione, più precisamente di seconda generazione. La denominazione di "letture di nuova generazione" (NGS) è data dalla capacità di elaborare una maggior qunatità di letture di sequenze in parallelo.

Il sistema 454/Roche applica la reazione a catena della polimerasi in emulsione (eP-CR) per clonare frammenti di DNA casuali, che sono attaccati a microsfere. Le microsfere vengono depositate in pozzetti di una piastra da microtitolazione e quindi pirosequenziate singolarmente e in parallelo. Il processo di pirosequenziamento prevede l'aggiunta sequenziale di tutti e quattro i deososinucleosidici trifosfati, che, se complementari al filamento stampo, sono incorporati da una DNA polimerasi. Questa reazione di polimerizzazione rilascia pirofosfato, che viene convertito per produrre luce. In questa fase avvengon circa 1,2 milioni di reazioni. La produzione di luce viene rilevata e convertita nella sequenza effettiva di DNA [9].

La tecnologia Illumina immobilizza frammenti di DNA casuali su una superficie e quindi esegue l'amplificazione PCR su superficie solida, ottenendo gruppi di frammenti di DNA identici. Questi vengono quindi sequenziati con terminatori reversibili in un processo di sequenziamento per sintesi. La lunghezza della lettura si avvicina a 150 bp e i frammenti raggruppati possono essere sequenziati da entrambe le estremità. È possibile ottenere informazioni su sequenze di quasi 300 bp accoppiando due letture da 150 bp.

Negli ultimi anni, sia il sequenziamento a lettura lunga che l'analisi metagenomica sono avanzati significativamente. L'avvento delle tecnologie di "terza generazione" a lettura lunga a singola molecola (PacBio e Oxford Nanopore) ha avuto un impatto significativo sulle analisi metagenomiche, in particolare per l'assemblaggio del genoma [4]. Queste

tecnologie consentono lunghezze di lettura di coppie di oltre 10 kilobase (Kbp), in forte contrasto con Illumina. Tuttavia, sia la tecnologia PacBio che Nanopore hanno tassi di errore più elevati (precisione del 88–94% per Nanopore e del 85–87% per PacBio). Studi recenti che hanno impiegato dati metagenomici di una finta comunità batterica basati su Nanopore hanno mostrato un'accuratezza relativamente alta nelle classificazioni a livello di genere, di circa il 93% [5]. Le letture più lunghe, quindi, nonostante il loro più alto tasso di errore, possono migliorare considerevolmente l'accuratezza della classificazione rispetto alle letture più brevi, e questo è particolarmente vero per specifici taxa, cioè gruppi ordinati di organismi a qualsiasi livello gerarchico.

1.2 Analisi computazionale

Una componente essenziale dello studio metagenomico consiste nel riconoscimento degli organismi presenti in un campione. Se la maggior parte dei genomi presenti nel campione sono sconosciuti, vengono utilizzati i metodi di assemblaggio senza sequenze di riferimento, chiamati strumenti di binning. In caso contrario, è possibile confrontare i dati sequenziati con un database di riferimento che memorizza le informazioni genomiche relative ai vari taxa. Quest'ultimo caso è noto anche come classificazione tassonomica o binning tassonomico [3]. A questo scopo, sono stati sviluppati numerosi software che associano delle etichette tassonomiche alle letture di sequenziamento, sfruttando diversi metodi. La maggior parte dei sistemi attuali sono ottimizzati per funzionare con letture brevi e accurate tecnologie di sequenziamento di seconda generazione. Tuttavia, a causa di un aumento della loro precisione e throughput, le tecnologie di sequenziamento a lettura lunga stanno guadagnando popolarità [11].

Inizialmente, lo strumento più utilizzato era BLAST (Basic Local Alignment Search Tool) ed è considerato il gold standard [10]. Questo tool approssima direttamente gli allineamenti che soddisfano dei criteri di somiglianza locale, cioè il maximal segment pair score (MSP). Tuttavia, con l'arricchimento dei database e l'aumento della lunghezza delle letture, BLAST è risultato inadatto dal punto di vista computazionale. Pertanto, negli ultimi anni sono stati sviluppati degli algoritmi per la classificazione metagenomica veloce che utilizzano tecniche diverse. Alcuni sono basati su k-mer, cioè delle particolari sottostringhe delle letture genomiche. Tra questi si annoverano CLARK [14], Kraken [15] o Centrifuge [16]. Ad esempio, Kraken, per la tassonomia, costruisce una struttura ad albero, in cui un elenco di k-mer significativi è associato a ciascun nodo, foglia e nodo in-

terno. Dato un nodo dell'albero tassonomico, il suo elenco di k-mer rappresenta l'etichetta tassonomica e verrà utilizzato per la classificazione delle letture metagenomiche. Nella fase di classificazione, ogni lettura viene scomposta nei suoi l-mer e si cerca, nell'albero tassonomico, il k-mer corrispondente avente il percorso con il peso più alto [3]. Successivamente, si analizzerà in dettaglio un software basato su Kraken, ma più performante, Kraken2. Esistono, tuttavia, altri approcci, come quelli basati su amminoacidi alfabetici ridotti, come DIAMOND [18] o sulla mappatura, come MetaMap [19] e MEGAN-LR [20].

1.3 Contenuto dell'opera

In questo lavoro, si vuole indagare la variazione dell'output del software Kraken 2 al mutare delle lunghezze delle letture, del numero di letture generate e degli errori presenti in esse.

Nelle prossime pagine, si procede, quindi, a presentare Kraken 2 e i programmi di generazione di letture, quali Simlord e Mason 2. Attraverso questi due software, da 40 specie di riferimento, si generano diversi set di letture che saranno dati come input a Kraken 2. Infine, si analizzano gli output ottenuti e si discutono i risultati.

2 Strumenti

2.1 Kraken 2

Kraken 2 [13] è stato sviluppato da Derrick E. Wood, Jennifer Lu and Ben Langmead sulla base di Kraken. Kraken è un software che associa brevi sequenze di sottostringhe genomiche, chiamate k-mer, indicizzate da minimizzatori, con l'unità tassonomica avente il lowest common ancestor (LCA). Kraken ha dimostrato di avere un'alta efficienza ed accuratezza in molte analisi, ma la quantità di memoria richiesta costringe ad adottare delle soluzioni per limitarne l'uso, come, ad esempio, ridurre la dimensione del database di riferimento. In Karken 2, sono state apportate delle modifiche che permettono di ridurre l'utilizzo di memoria di circa l'85%. E' stato introdotto, inoltre, lo schema di ricerca spaced seed che migliora l'abilità di classificazione delle letture. In questa sezione, sono mostrate le caratteristiche che lo distinguono da Kraken e che permettono di ottenere queste performance.

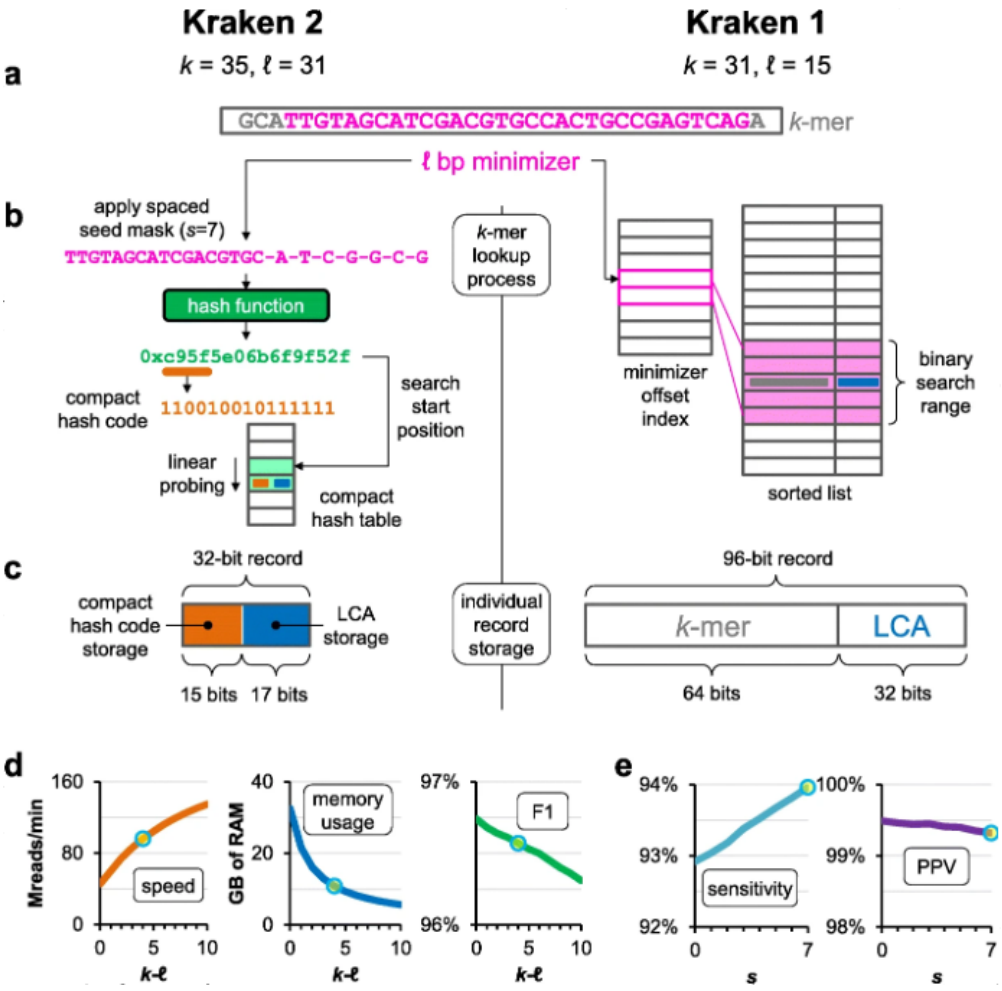

Figura 1: *Differenze tra Kraken 1 e Kraken 2 e grafici delle performance*

2.1.1 Database interno

Kraken 2, al contrario di Karaken 1, apporta alcune modifiche alla rappresentazione interna del database inserito dall'utente. Per prima cosa, Kraken 2 trova un set minimale di nodi che consiste di tutti i nodi ai quali è assegnata una sequenza di riferimento e di tutti gli antenati dei nodi. I vertici tra i nodi rimangono gli stessi della tassonomia fornita dall'utente, mantenendo la struttura ad albero nella rappresentazione interna. Successivamente, Kraken 2 assegna degli ID numerici in ordine crescente ai nodi del set minimale usando una ricerca breadth-first search (BFS), iniziando dalla radice a cui viene assegnato il numero 1. La BFS garantisce che i nodi antenati abbiano valori numerici minori dei loro discendenti.

Per rendere i risultati facilmente interpretabili e mostrare l'output utilizzando la tassonomia esterna, Kraken 2 memorizza una mappatura della tassonomia interna degli ID alla tassonomia esterna. L'uso della tassonomia interna permette una computazione più semplice degli LCA di due nodi, perché gli ID forniscono informazioni sulla loro posizione all'interno dell'albero. Inoltre, l'uso degli ID consente di utilizzare il numero minimo di bit per salvare i dati, dando il massimo spazio agli hash code compatti e riducendo la probabilità di collisioni nella hash table compatta (CHT), che verrà presentata in seguito.

2.1.2 Sottocampionamento basato su minimizzatori

Kraken 2 sottocampiona il set delle sottostringhe genomiche e inserisce nel database solo minimizzatori diversi. Il minimizzatore l bp di un k-mer (l<k) è l'l-mer canonico lessicograficamente più piccolo trovato tra i k-mer. Un l-mer si dice canonico se è lessicograficamente minore o uguale al suo complemento inverso. Se k = l, non avviene alcun campionamento e Kraken 2 inserisce la sottostringa nella struttura dati senza variazioni. Inoltre, al crescere della differenza tra l e k, sono inserite meno sottostringhe nella CHT, riducendo la dimensione, la memoria richiesta e il runtime. I valori di default per Kraken 2 sono k=35, l=31. Nella Figura 1.d si possono vedere i grafici che raffigurano i valori di velocità, memoria utiilizzata e accuratezza delle classificazioni al variare del valore $k - l$.

Kraken 2 determina quali l-mer sono minimizzatori usando un algoritmo sliding window minimum calculation che usa una coda doppia (deque). Questo permette una determinazione più veloce dei minimizzatori: usa un tempo medio O(1) per calcolare un nuovo minimizzatore. L'algoritmo sliding window include anche l'operazione di shuffling xor (or esclusivo). Quest'operazione serve a permutare l'ordinamento degli l-mer nel calcolo dei

minimizzatori e aiuta a evitare una preferenza verso gli l-mer a bassa complessità quando si seleziona il minimizzatore di un k-mer.

2.1.3 Hash table compatta

kraken 2 introduce una hash table compatta (CHT) e probabilistica per mappare i minimizzatori agli LCA. Questa tabella utilizza un terzo della memoria di una hash table standard, riducendo la specificità e l'accuratezza. Inoltre, kraken 2 conserva, nella sua struttura dati, solo minimizzatori (di lunghezza l, $l \leq k$) , invece che tutti i k-mer. Solo i minimizzatori distinti dalla query (read) attivano l'accesso alla hash table.

L'hash table usata in Kraken 2 è molto simile ad una hash table tradizionale che usa la metodologia di ricerca lineare per risolvere le collisioni, ma presenta alcune modifiche. Viene usato un array di dimensione fissa con celle di hash di 32 bit per immagazzinare le coppie chiave-valore. In una cella, il numero di bit usati per una coppia chiave-valore varia in base al numero di bit necessari a rappresentare tutti gli ID tassonomici che si trovano nel database di riferimento. Il valore è conservato nei bit meno significativi della cella e deve essere un intero positivo. Nei bit più significativi rimanenti, vengono salvati i bit più significativi dell'hash code della chiave (hash code compatto). La ricerca di una chiave K nella CHT viene svolta calcolando l'hash code della chiave (h(K)) e successivamente scandendo linearmente l'array della tabella partendo dalla posizione h(k) mod (T), dove T è il numero di celle dell'array. Il funzionamento della CHT è schematizzato nella Figura 1.b, c.

Questa organizzazione delle coppie introduce, però, un nuovo modo in cui le chiavi possono "collidere". Due chiavi distinte possono essere considerate identiche se hanno lo stesso hash code compatto e le loro posizioni di partenza sono abbastanza vicine da causare l'incontro di un uguale hash code compatto prima di una cella vuota da parte della ricerca lineare. Questa caratteristica impatta sull'accuratezza delle query e dà alla CHT una natura probabilistica, dove sono possibili 2 tipi di query falso positive:

- una chiave che non è stata inserita può essere riportata come presente nella tabella;
- il valore di 2 chiavi può essere confuso tra loro.

Il secondo errore non è propriamente un falso positivo, ma si traduce in un LCA meno specifico assegnato al minimizzatore. La probabilità di uno di questi errori è minore dell'1% con l'impostazione predefinita di Kraken 2 e un fattore di carico del 70%, anch'esso

fissato come valore di default. L'effetto dannoso della classificazione a livello di lettura è ridotto dall'algoritmo che Kraken 2 usa per combinare informazioni provenienti da tutte le letture che è invariato rispetto a Kraken1. Quest'algoritmo utilizza le informazioni di tuti i k-mer di una sequenza per contrastare i valori LCA errati.

Le chiavi possono essere uguali per 2 motivi:

- 2 diversi k-mer condividono lo stesso minimizzatore (collisione di minimizzatori);
- 2 distinti minimizzatori sono indistinguibili nella CHT.

Le collisioni tra minimizzatori non sono sempre dannose: tra k-mer di genomi strettamente collegati, una collisione di questo tipo potrebbe rilevare una vera omologia anche di fronte a singoli polimorfismi nucleotidici e/o errori di sequenziamento. La collisione di minimizzatori tra k-mer di genomi molto diversi può produrre o valori elevati di LCA, se entrambi i genomi sono nel database di riferimento, o classificazioni incorrette di k-mer, se un genoma non è nel database di riferimento. Questi errori sono tutti a livello di singoli k-mer e possono non compromettere la classificazione di una sequenza di query, poichè per determinare un'etichetta tassonomica sono necessari molti k-mer.

Il tasso di collisioni tra minimizzatori è influenzato dalla lunghezza dei minimizzatori usati, poiché la lunghezza è direttamente collegata al numero di possibili minimizzatori. E' stato verificato sperimentalmente che maggiore è l, cioè la lunghezza dei minimizzatori, minore è la probabilità che si verifichino collisioni tra minimizzatori.

2.1.4 Spaced-seed

Kraken 2 usa un semplice approccio spaced seed dove l'utente specifica, quando costruisce il database, un intero s che indica quante posizioni saranno nascoste nel minimizzatore, cioè le posizioni che non verranno considerate durante la ricerca. Iniziando dalla posizione più a destra, ogni altra posizione viene mascherata fino a quando non saranno state nascoste s posizioni. Con Kraken 2 si possono ottenere risultati molto simili a Kraken 1 settando k = l = 31 e s=0, poiché questi valori evitano i sottocampionamenti basati su minimizzatori e l'uso di spaced seed. Il valore di default di s è 7.

Come si vede in Figura 1.b, i canonici l-mer candidati minimizzatori sono nascosti con la maschera spaced seed prima del loro inserimento nella deque attraverso l'algoritmo sliding window. Canonizzando i candidati minimizzatori prima di applicare la maschera spaced seed assicura che il risultato sia lo stesso sia che lo spaced-seed venga applicato

al l-mer che al suo complemento inverso. L'uso di spaced seed e del sottocampionamento dei minimizzatori candidati rende la sensibilità di Kraken 2 governata da l-s (il numero di basi confrontate nelle sottostringhe cercate). In Figura 1.e si può vedere come cambia la sensitività e il positive predictive value (PPV), cioè $\frac{TP}{TP+FP}$, al variare del parametro s.

2.1.5 Sottocampionamento basato su hash

Per stimare la capacità richiesta dalla tabella di hash, Kraken 2 utilizza i valori k, l, s e il database di riferimento scelti dall'utente. È possibile specificare una dimensione massima del database. Se la stima della capacità richiesta è più grande della dimensione massima specificata, i minimizzatori verranno nuovamente sottocampionati usando una funzione di hash. Data una capacità stimata S' e una massima capacità S (S<S') si può calcolare il valore f = S/S' che è la frazione dei minimizzatori che l'utente può immagazzinare nel database. Può anche essere calcolato un valore minimo di hash v = (1-f)M, dove M è il massimo valore di output della funzione di hash h. Ogni minimizzatore della libreria di riferimento con un hash code minore di v non sarà inserito nella tabella di hash. Questo valore è dato anche al classificatore affinché solo i minimizzatori con hash code uguali o maggiori di v sono usati per la ricerca nell'hash table, evitando che vengano cercati minimizzatori di cui è certa l'assenza.

2.2 Simlord

SimLoRD [22], Simulation of Long Read Data, è un simulatore di letture di sequenziamento di terza generazione focalizzato sul modello d'errore SMRT Pacific Biosiences (PacBio). È un tool a linea di comando, implementato in Python, che può generare letture sia da filamenti di sequenze forniti dall'utente che da filamenti di sequenze generati randomicamente. Produce anche gli allineamenti delle letture simulate alle reference, salvati in formato SAM. Simlord ha a disposizione un numero elevato di parametri che permettono di generare letture specifiche. Tra questi, si menzionano:

- Il parametro -rr (read reference) usato per indicare il file da cui generare le letture;
- Il parametro -n (num reads) che determina il numero di letture da simulare;
- Il parametro -fl (fixed readlength) che determina la lunghezza delle letture da simulare;

- Il parametro -c (coverage) che viene usato da simlord per calcolare il numero di letture da simulare;
- I parametri -ps, -pi e -pd che rappresentano, la percentuale di errori di, rispettivamente, sostituzioni, inserzioni e cancellazioni che si presenteranno nelle letture.

. Per utilizzare tali parametri, a linea di comando, si scrive il nome del parametro seguito dal volore desiderato.

2.3 Mason2

Mason 2 [21] è un software di simulazione di letture per letture Illumina, 454 e Sanger implementato in C ++ utilizzando la libreria SeqAn, estensibile e liberamente disponibile sotto GPL/LGPL. È stato scritto pensando alle prestazioni e può campionare letture da grandi genomi. E' possibile inserire tassi d'errore specifici e valori di qualità di base. Per le letture Illumina, è fornita un'analisi completa con dati empirici per l'errore e il modello di qualità. Per le altre tecnologie vengono utilizzati i modelli della letteratura. Mason è una raccolta di strumenti per la simulazione di sequenze biologiche. La raccolta è composta dai seguenti strumenti:

- mason_frag_sequencing, per la simulazione del campionamento di frammenti da un genoma;
- mason_genome, per la simulazione di sequenze genomiche casuali;
- mason_materializer, per applicare la variazione da un file VCF a un genoma in un file FASTA;
- mason_methylation che simula i livelli di metilazione per un genoma dipendente dal contesto per ogni possibile sito;
- mason_simulator, per la simulazione di letturedi seconda generazione dato un genoma ed, eventualmente, anche un file VCF con le varianti, dato un filamento di sequenze fornito dall'utente;
- mason_splicing che calcola il trascrittoma (la totalità degli RNA trascritti a partire da un genoma) da un file FASTA del genoma e un file GFF con i geni;
- mason_variator che simula SNP, piccoli indel e varianti strutturali per i dati genomici scrivendo il risultato come file VCF.

Nella parte sperimentale, verrà utilizzato lo strumento mason_simulator. Sono disponibili diversi parametri che peremettono di generare letture specifiche. Tra questi, sono presenti:

- Il parametro -ir (input reference), usato per indicare il file da cui generare le letture;
- Il parametro -n (num reads) che determina il numero di letture da simulare;
- Il parametro –ilumina-read-length che determina la lunghezza delle letture da simulare e, automaticamente, utilizza i valori d'errore di default della tecnologia Illumina;
- Il parametro -o (output), usato per specificare il percorso e il nome del file generato da Mason 2 ;

Per utilizzare tali parametri, a linea di comando, si scrive il nome del parametro seguito dal volore desiderato.

3 Metodi

Per valutare gli effetti della lunghezza della lettura sull'accuratezza della classificazione, si utilizzano delle letture simulate con Mason2 e Simlord. Sono state utilizzate le letture della sequenza genomica simulata da 40 specie di riferimento selezionate (visibili in Tabella 14) ed è stata valutata la correttezza della classificazione delle letture del DNA a livello di specie. Gli ID tassonomici e i ranghi sono stati estratti dal file nodes.dmp scaricato dal sito web dell'NCBI. Attraverso un programma sviluppato in C è stata calcolata la sensibilità, la prcisione e il valore F1 delle classificazioni di Kraken 2.

3.1 Generazione delle letture

Per generare le letture corte abbiamo utilizzato Mason2. La focalizzazione sulla tecnologia Illumina e la velocità di calcolo hanno favorito la scelta di tale strumento per questi test.

Per le letture lunghe, invece, è stato privilegiato Simlord, dotato di diversi parametri, tra cui il parametro coverage che abbiamo utilizzato anche per la generazione di alcune letture brevi con i tassi d'errore di PacBio, Illumina e Nanopore.

Tutte le letture sono state generate utilizzando le 40 specie di riferimento della Tabella 14.

3.1.1 Letture corte con lunghezze di lettura diverse, numero fisso di letture generate e tecnologia Illumina

È stato scelto di generare un numero costante di 2000 letture per ogni esecuzione. Si è variata la lunghezza delle letture: da 100bp a 1000bp, con intervalli di 100bp. Per ogni specie è stato eseguito Mason2 con tali intervalli di lunghezza. Le letture generate appartengono alla tecnologia Illumina, i cui tassi d'errore sono inseriti automaticamente da Mason2. Il comando usato si presenta in questo modo:

```
mason_simulator -ir fileInput -n 2000 --illumina-read-length len --
    fragment-mean-size 1000 -o fileOutput.fastq
```

Dove, `fileInput` rappresenta il file fna che contiene la specie da cui si genereranno le letture, `len` si riferisce alla lunghezza delle letture che si vogliono ottenere e `fileOutput` è il nome assegnato al file fastq che verrà generato. Il parametro `--fragment-mean-size` è stato introdotto per allungare la dimensione media dei frammenti, evitando, così, un errore che si presentava con alcune specie.

Si possono raggruppare le letture generate in 10 set, ognuno dei quali avente 80 000 (2000x40) letture con lunghezze di letture crescenti da 100bp nel primo set a 1000bp nell'ultimo set. Nell'Appendice, alla Tabella 1, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

3.1.2 Letture corte con lunghezze di lettura diverse, numero di letture costante, tecnologia PacBio e Nanopore

Per indagare al meglio le potenzialità delle letture lunghe, si è deciso di generare, prima, delle letture corte utilizzando i tassi d'errore specifici delle letture lunghe, cioè quelli di PacBio e Oxford Nanopore. Attraverso Simlord sono state, quindi, generate delle letture aventi lunghezza variabile da 100bp a 1000bp, con intervalli di 100bp. Il comando che genera le letture secondo i tassi d'errore PacBio si presenta in questo modo:

```
simlord --read-reference fileInput -n 2000 -fl len -pi 0.11 -pd 0.04
    -ps 0.01 --no-sam fileOutput
```

Il parametro`--no-sam` è stato utilizzato per evitare la generazione dei file di allineamento che non erano necessari ai test.

Si possono raggruppare le letture ottenute in 10 set, ognuno dei quali avente 80 000 (2000x40) letture con lunghezze di letture crescenti da 1000bp nel primo set a 10 000bp nell'ultimo set. Nell'Appendice, alla Tabella 2, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

Invece, il comando usato per generare le letture secondo i tassi d'errore Nanopore si presenta in questo modo:

```
simlord --read-reference fileInput -n 2000 -fl len -pi 0.04 -pd 0.05
    -ps 0.03 --no-sam fileOutput
```

Nell'Appendice, alla Tabella 3, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

3.1.3 Letture lunghe con lunghezze di lettura diverse, numero fisso di letture generate e tecnologia PacBio e Nanopore

Anche per le letture lunghe, rimane costante il numero di letture ad ogni esecuzione (2000). Si è variata la lunghezza delle letture: da 1000bp a 10 000bp, con intervalli di 1000bp. Per ogni specie è stato eseguito Simlord con questi intervalli di lunghezza. Per

la generazione delle letture sono stati utilizzati i tassi d'errore di Pacific Biosience. Il comando Simlord usato si presenta in questo modo:

```
simlord --read-reference fileInput -n 2000 -fl len -pi 0.11 -pd 0.04
    -ps 0.01 --no-sam fileOutput
```

Si possono raggruppare i risultati ottenuti in, 10 set, ognuno avente 80 000 (2000x40) letture con lunghezze di letture crescenti da 1000bp nel primo set a 10 000bp nell'ultimo set. Nell'Appendice, alla Tabella 4, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

Si è, in seguito, ripetuta la generazione di letture utilizzando, però, i tassi d'errore Oxford Nanopore, attraverso tale comando:

```
simlord --read-reference fileInput -n 2000 -fl len -pi 0.04 -pd 0.05
    -ps 0.03  --no-sam fileOutput
```

Nell'Appendice, alla Tabella 5, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

3.1.4 Letture corte con uguale lunghezza di lettura, numero diverso di letture generate e tecnologia Illumina

Nei test successivi si è deciso di mantenere costante la lunghezza di lettura, ma variare il numero di letture generate ad ogni esecuzione. È stata scelta quindi la lunghezza di 500bp per la grenerazione di tutti i file fastq. Trattandosi di letture corte, è stato utilizzato Mason2 e sono stati mantenuti i tassi d'errore standard di Illumina. Il numero di letture generate è stato fatto variare da 1000 a 10 000, con intervalli di 1000. Il comando usato si presenta in questo modo:

```
mason_simulator -ir fileInput -n numReads --illumina-read-length 500
    --fragment-mean-size 1000 \ -o fileOutput
```

Dove, `numReads` rappresenta il numero di letture che varia da 1000 a 10 000.

Si possono raggruppare i risultati ottenuti in 10 set, ognuno avente uguale lunghezza di lettura, ma un numero crescente di letture, da 40 000 (1000x40) letture contenute nel primo set a 400 000 (10 000x40) letture contenute nell'ultimo set. Nell'Appendice, alla Tabella 6, è possibile trovare i risultati in dettaglio delle classificazioni svoltei con questa tipologia di letture.

3.1.5 Letture lunghe con uguale lunghezza di lettura, numero diverso di letture generate, tecnologia PacBio e Nanopore

È stato ripetuto lo stesso procedimento con letture lunghe: è stata fissata la lunghezza di lettura a 5000bp e si è variato il numero di letture generate da 1000 a 10 000, con intervalli di 1000. Per simulare le letture è stato utilizzato Simlord con i tassi d'errore PacBio. Il comando usato si presenta in questo modo:

```
simlord --read-reference fileInput -n numReads -fl 5000 -pi 0.11 -pd
    0.04 -ps 0.01 --no-sam fileOutput
```

Dove, `numReads` rappresenta il numero di letture che varia da 1000 a 10 000.

Si possono raggruppare le letture ottenute in 10 set, ognuno avente uguale lunghezza di lettura, ma un numero crescente di letture, da 40 000 (1000x40) letture contenute nel primo set a 400 000 (10 000x40) letture contenute nell'ultimo set. Nell'Appendice, alla Tabella 7, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

Si esegue lo stesso metodo utilizzando i tassi d'errore specifici della tecnologia Oxford Nanopore:

```
simlord --read-reference fileInput -n numReads -fl 5000 -pi 0.04 -pd
    0.05 -ps 0.03 --no-sam fileOutput
```

Nell'Appendice, alla Tabella 8, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

3.1.6 Letture lunghe con lunghezze di lettura diverse, numero di letture determinato dal parametro coverage, tecnologia PacBio e Nanopore

Per questo test, si è scelto di lasciare a Simlord il compito di determinare il numero di letture da generare utilizzando il parametro coverage. Coverage (profondità di lettura o profondità) è il numero medio di letture che rappresentano un dato nucleotide nella sequenza ricostruita [23]. Può essere calcolato attraverso la lunghezza del genoma originale (O), il numero di letture (N) e la lunghezza media di lettura (M), come $\frac{N \times M}{O}$. In questi test, è stata variata anche la lunghezza delle letture da 1000bp a 10 000bp, con intervalli di 1000bp. Sono stati utilizzati i tassi d'errore PacBio. Il comando usato si presenta in questo modo:

```
simlord --read-reference fileInput -c 20 -fl len -pi 0.11 -pd 0.04 -
    ps 0.01 --no-sam fileOutput
```

Dove `-c` è il parametro coverage.

Si possono raggruppare le letture ottenute in 10 set, ognuno avente un numero diverso di letture con lunghezze di letture crescenti da 1000bp nel primo set a 10 000bp nell'ultimo set. Nell'Appendice, alla Tabella 9, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

Si ripete lo stesso procedimento utilizzando i tassi d'errore specifici della tecnologia Oxford Nanopore:

```
simlord --read-reference fileInput -c 20 -fl len -pi 0.04 -pd 0.05 -
    ps 0.03 --no-sam fileOutput
```

Nell'Appendice, alla Tabella 10, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

3.1.7 Letture corte con lunghezze di lettura diverse, numero di letture determinato dal parametro coverage e tecnologia Illumina, PacBio, Nanopore

Si utilizza il parametro coverage di Simlord anche per generare letture corte da 100bp a 1000bp, con intervalli di 100bp. Ora, però, vengono utilizzati i tassi d'errore di Illumina. Il comando usato si presenta in questo modo:

```
simlord --read-reference fileInput -c 20 -fl len -pi 0.00001 -pd
    0.00001 -ps 0.01 --no-sam fileOutput
```

Dove, sono cambiati i valori dei parametri dei tassi d'errore per rispettare lo standard Illumina. Si possono raggruppare le letture ottenute in 10 set, ognuno avente un numero diverso di letture con lunghezze di letture crescenti da 100bp nel primo set a 1000bp nell'ultimo set. Nell'Appendice, alla Tabella 11, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

Si utilizza lo stesso parametro coverage anche per generare letture corte con tassi d'errore PacBio:

```
simlord --read-reference fileInput -c 20 -fl len -pi 0.11 -pd 0.04 -
    ps 0.01 --no-sam fileOutput
```

Anche queste letture si possono raggruppare in 10 set. Nell'Appendice, alla Tabella 12, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

Infine, con tassi d'errore Nanopore:

```
simlord --read-reference fileInput -c 20 -fl len -pi 0.04 -pd 0.05 -
    ps 0.03 --no-sam fileOutput
```

Nell'Appendice, alla Tabella 13, è possibile trovare i risultati in dettaglio delle classificazioni svolte con questa tipologia di letture.

3.2 Utilizzo dei dati e analisi dei risultati

Precedentemente, abbiamo virtualmente raggruppato i file generati in set. Ogni set contiene 40 file fastq che hanno delle caratteristiche in comune (uguale lunghezza di lettura o uguale numero di letture). Si utilizza, quindi, un software sviliuppato in python per raggruppare i file contenuti in ogni set in un singolo file fasta. Avendo 70 set, otteniamo 70 file fasta. Ogni file viene caricato nel cluster del Dipartimento di Ingegneria dell'Informazione (DEI) ed è usato come input per l'esecuzione di Kraken 2. Kraken 2 viene utilizzato con i parametri di default (l=31, k=35, s=7 e fattore di carico della CHT del 70%). Il database di riferimento è lo standard database generato da Kraken 2 attraverso il comando:

```
kraken2-build --standard --db $DBNAME
```

Per valutare il tasso d'errore di Kraken 2 si utilizza il file avente estensione .out generato dallo stesso. Si modifica il file attraverso il comandoo:

```
cut -f 2-3 file.out > file.res
```

A partire dai file fastq di ogni set, attraverso un programma scritto in python sono stati creati i ground truth necessari per valutare le performance di Kraken 2.

La classificazione di Kraken 2 è stata confrontata con il ground truth attraverso un programma, chiamato evaluate_calls, scritto in C e contenuto nel server del DEI. Tale programma accetta come input il file nodes.dmp, il livello al quale si vuole verificare la correttezza della classificazione di Kraken 2, impostato a specie, il ground truth e il file risultante da Kraken 2 adeguatamente modificato. Il programma restituisce i valori d'interesse per l'analisi dell'output di Kraken 2.

3.2.1 Metriche di valutazione

Per valutare la qualità della classificazione, sono stati definiti cinque gruppi di base:

- true positive (TP): il numero di letture che sono state classificate correttamente. Solitamente, i taxa che sono ben rappresentati nel database e che hanno pochi taxa strettamente correlati avranno tassi elevati di corrispondenze vere;
- false positive (FP): il numero di letture che sono state classificate in modo errato. In questo gruppo si trovano taxa con molti parenti stretti nel database che generano molte false corrispondenze;
- true negative (TN): il numero di letture rimaste non classificate che appartengono ad un organismo non presente nel database;
- false negative (FN): il numero di letture rimaste non classificate, ma appartenenti ad un organismo presente nel database;
- letture corrette, ma al di sopra del rango (OK).

Con il graduale ampliamento dei database genomici, si prevede che la frazione di query non riuscite (FN) diminuirà. Allo stesso tempo ci si aspetta che la frazione di FP possa aumentare. L'esatta natura di questo compromesso non è ben esplorata, ma si stanno già studiando nuovi approcci statistici che possano affrontare il problema [4].

Inoltre, il gruppo TN non rappresenta una situazione biologicamente realistica, in quanto tutte le sequenze derivano da un taxon e, con il progressivo aggiornamento dei database genomici, anche la frazione TN diminuirà. Tuttavia, questo aspetto è utile quando si vogliono valutare le prestazioni dei classificatori.

Per quanto riguarda i risultati attesi, si possono fare delle ipotesi. Ad esempio, se il taxon A domina la comunità, non può avere un'alta frazione di falsi positivi rispetto ai veri positivi perché la stragrande maggioranza delle query lette dalla comunità proverrà dal taxon A e quindi da veri positivi. Al contrario, se il taxon B è estremamente raro, ci sarà un gran numero di falsi positivi rispetto ai veri positivi, poiché pochissime query di lettura proverranno dal taxon B, risultando in una frazione molto piccola di veri positivi.

Questi quattro valori vengono combinati per calcolare metriche di valutazione più complesse. La prima metrica utilizzata è la sensitivity: la percentuale di letture corrette classificate:

$$Sensitivity(SENS) = \frac{TP}{TP + FN + FP + OK}$$

La sensitivity della classificazione non è del tutto appropriata per valutare set di dati sbilanciati. Pertanto, come seconda metrica, si usa il punteggio F1, calcolato usando anche il valore di precisione:

$$Precision(PR) = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times PR \times SENS}{PR + SENS}$$

Il valore di precisione fornisce un indicazione della correttezza della classificazione considerando solo le letture che Kraken 2 riesce a classficare ed è più indicato per il calcolo del tasso d'errore di Kraken2.

Nelle analisi delle performance di Kraken 2 verranno evidenziate le precentuali di sensitivity, precisione e i punteggi F1 al fine di determinare le differenze tra le diverse tecnologie e lunghezze delle letture.

4 Risultati

4.0.1 Letture con diversa lunghezza di lettura e stesso numero di letture generate

Nel seguente gruppo di grafici è possibile osservare i valori sensitivity, precisione ed F1 per le letture con lunghezza di lettura variabile delle tecnologie Illumina, Oxford Nanopore e PacBio.

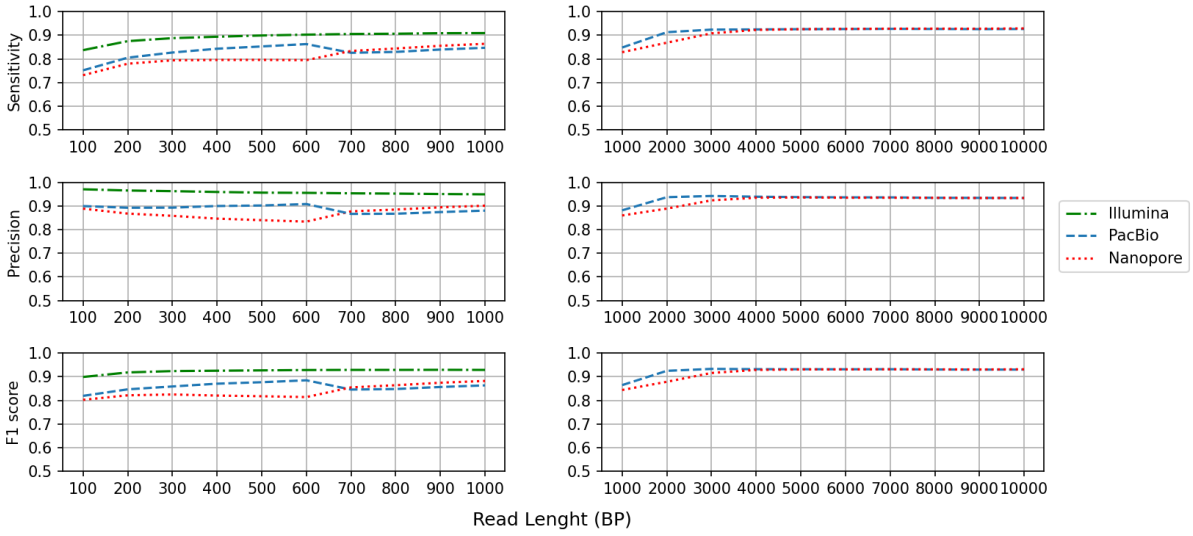

Figura 2: *Grafico contenente i valori sensitivity, precision ed F1 per le letture lunghe e corte, con tassi d'errore Illumina, Oxford Nanopore e PacBio*

I risultati della classificazione di Kraken 2 delle letture generate da Mason 2 con i tassi d'errore Illumina sono, in generale, molto accurati. Per ogni set di letture, si è trovata una senstivity superiore all'83%. La media delle sensitivity delle letture è dell'89,2%. Tale risultato è dovuto in gran parte al basso tasso d'errore delle letture Illumina. Si nota una tendenza crescente all'aumentare delle lunghezze di lettura: il valore peggiore si registra a 100bp, con un incremento significativo a 200bp e una crescita meno marcata, ma continua, dopo i 300bp. I valori di precisione, come ci si aspetta, sono più elevati rispetto ai valori di sensitivity, in quanto questa metrica non considera le letture che Kraken 2 non ha classificato. Tuttavia, se i valori di sensitivity avevano un andamento crescente, i valori di precisione presentano misure progressivamente più basse che, però, non vanno al di sotto del 94% e la media risulta del 95,7%. Ciò si traduce in un tasso d'errore di circa il 5%. L'F1 score presenta, invece, valori più simili alla sensitivity. I valori si aggirano attorno al 90%, con un solo valore all'89% a 100bp e una media del 92,3%. La curva si presenta crescente, anche se non si apprezza nessuna crescita evidente dopo i 400bp.

La classifcazione delle letture corte aventi tecnologia PacBio ottiene dei risultati peggiori e presenta un andamento discontinuo. Infatti, si può osservare un picco a 600bp con sensitivity dell'86% per poi retrocedere all'82% a 700bp. La media delle sensitivity è dell'82,9%, inferiore del 6,5% alla media delle sensitivity della tecnologia Illumina. I valori di precisione, invece, presentano misure pressochè costanti, con valori di circa il 90% fino a 600bp, per poi subire una brusca decrescita che abbassa la media dei valori all'82,8%. Ciò si traduce in un tasso d'errore di circa l'8%. Anche l'F1 score segue lo stesso andamento con il picco a 600bp dell'88% . La media è, invece, dell'86,2%, inferiore dell 6,3% rispetto alla media F1 delle letture corte con tecnologia Illumina.

Anche le letture corte con tecnologia Nanopore ottengono risultati inferiori. È interessante notare che fino a 700bp si hanno valori di sensitivity più bassi rispetto a quelli ottenuti con tecnologia PacBio, ma dagli 800bp i valori si allineno e Nanopore ottiene risultati di poco superiori. Al contrario delle letture corte generate da PacBio, la curva della sensitivity delle letture di Nanopore è crescente. La media dei valori di sensitivity è del 80,8%, comunque inferiore alla media delle sensitivity registrate con PacBio. I valori di precisione, invece, presentano valori discendenti fino a 600bp, con un valore dell'83%, ma poi risalgono e si allineano ai valori di precisione di Nanopore. La media dei valori di precisione è dell'86,9%. Anche la curva dei valori F1 non si presenta crescente, al contrario di quella delle sensitivity. Infatti, si trovano valori progressivamente più bassi da 200bp a 600bp, per poi risalire all'83% a 700bp. La media dei valori F1 è dell'83,7%, anch'essa più bassa rispetto alla media ottenuta con PacBio.

Le letture lunghe con tecnologia PacBio hanno, invece, riportato risultati migliori. Mentre a 1000bp la sensibility è solamente di circa l'85%, già a 2000bp passa al 91%, per poi stabilizzarsi intorno al 91,6%. La media delle sensitivity è dell'91,6%, superiore dell'8,7% alla media delle sensitivity delle letture corte con stessa tecnologia e del 2,4% rispetto alla media calcolata sulle letture corte con tecnologia Illumina. La curva dei valori di precisione si presenta simile ma scalata in alto: la media dei valori è del 93,1%. Ciò si traduce in un tasso d'errore di circa il 7%. Anche la curva contenente i valori F1 al variare della lunghezza di lettura ha un comportamento simile. A 1000bp si usserva circa 86% che cresce e si attesta a vaslori intorno al 93%. La media è del 92,3%, superiore sia a quella risultante dalle letture corte della stessa tecnologia che alle letture corte aventi tassi d'errore Illumina.

Le letture lunghe con tecnologia Oxford Nanopore hanno una media delle sensitivity molto simile a quella delle letture lunghe PacBio, cioè del 90%. Tuttavia, a 1000bp e

2000bp hanno valori più bassi, rispettivamente del 82% e 86%, rispetto all'84% e al 91% di PacBio. Da 4000bp, invece, i valori di sensitivity Nanopore e PacBio risultano molto simili (a 10000bp: 92,8% per Nonopore e 92,6% per PacBio). Anche i valori di precisione e F1 rispettano la curva dei valori di sensitivity, con una media rispettivamente del 92,1% e del 91,5%.

4.0.2 Letture corte con uguale lunghezza di lettura e numero diverso di letture generate

Nei seguenti grafici si possono osservare i valori sensitivity, precisione e F1 per numeri diversi di letture generate, avendo fissato a 500bp la lunghezza di lettura per le letture generate utilizzando la tecnologia Illumina e a 5000bp la lunghezza di lettura avente tassi d'errore PacBio e Nanopore.

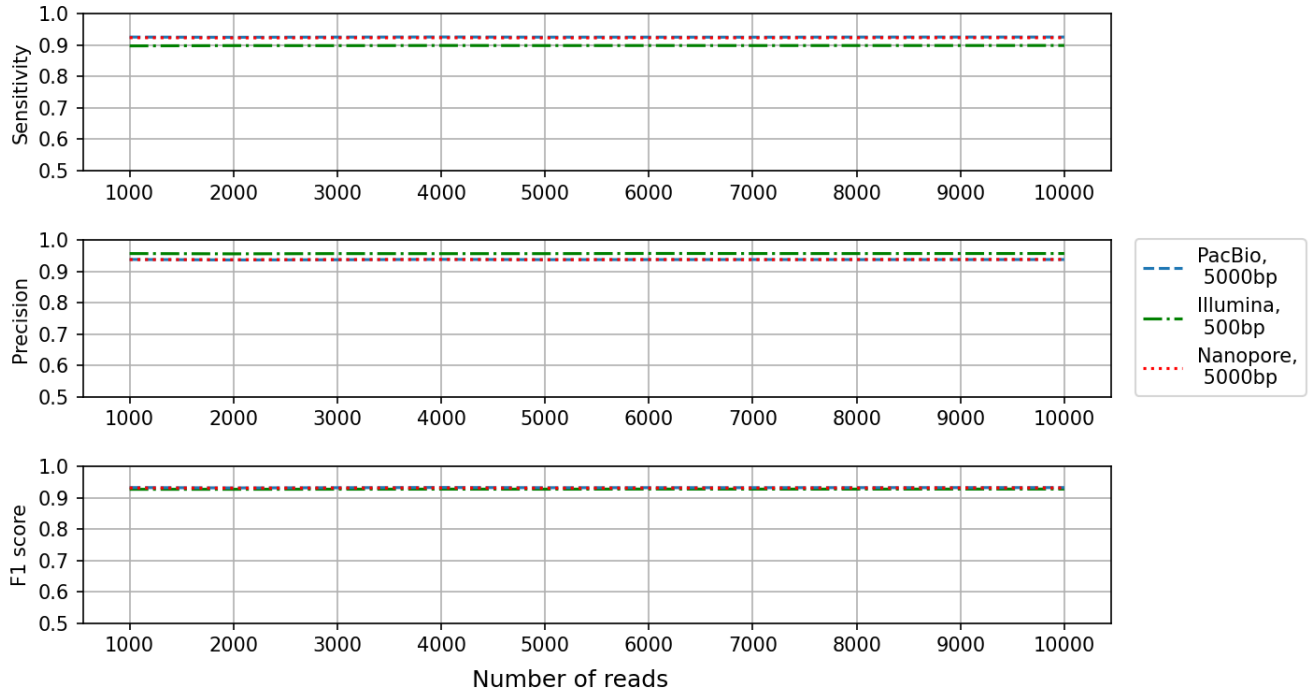

Figura 3: *Grafico contenente i valori sensitivity, precision ed F1 per un numero di letture variabile con tassi d'errore Illumina, PacBio e Oxford Nanopore*

Si può subito notare che l'intervallo scelto per la generazione di letture porta a risultati molto simili nella classificazione. Le letture generate attraverso Mason2 hanno un incremento di 0,2 punti percentuale per la sensitivity: dall'89,7% con 1000 letture generate al 92,7% con la generazione di 10000 letture, anche se é già possibile vedere questo risultato a 4000bp. La media delle sensitivity è dell'89,9%. Anche i valori di precisione e F1 score si dispongono su una retta, con la media rispettivamente del 95,6% e 92,7%.

Le letture generate attraverso Simlord utilizzando i tassi d'errore tipici di PacBio hanno tutti valori di sensitivity di poco superiori al 92,6%, tranne quelli compresi tra le 5000 e le 7000 letture che hanno valori di poco inferiori. La media dei valori di sensibility, come ci si aspetta, è del 92,6%. Invece, la media dei valori di precisione è del 93,7%. Ciò si traduce in un tasso d'errore di circa il 7%, come già si era verificato con la tecnologia PacBio. La media degli F1 score è invece del 93,2% .

Le letture generate secondo i tassi d'errore Nanopore presentano risultati molto simili a PacBio nella classificazione. A partire dalle 2000 reads si registrano valori di sensitivity superiori al 92,3% e la media si attesta sullo stesso valore. I valori di F1 differiscono maggiormente: il valore piùà basso, 80%, si registra a 1000bp, mentre si trova il valore 88% a 10000bp. La media si attesta all'83,7%.

4.0.3 Letture con lunghezze di lettura diverse e numero di letture generate attraverso il parametro coverage

I seguenti grafici mostrano i valori di sensitivity, precisione ed F1 risultanti dalla classificazione delle letture generate usando una lunghezza di lettura variabile e un numero di letture anch'esso variabile determinato dal parametro coverage con i tassi d'errore delle tecnologia Illumina, PacBio e Nanopore.

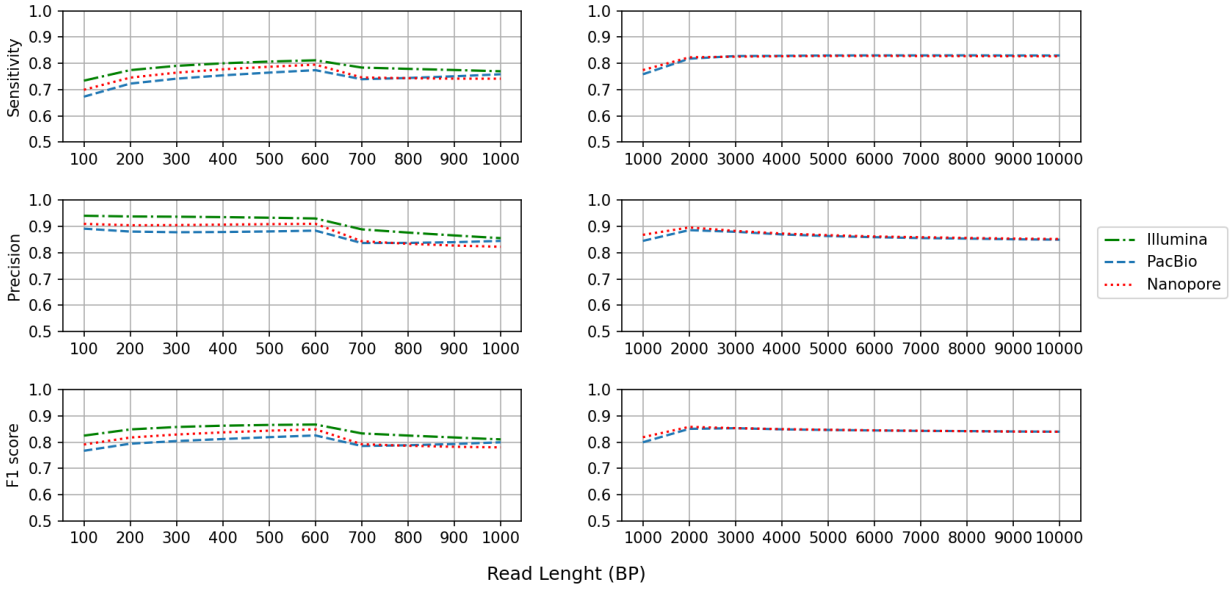

Figura 4: *Grafico contenente i valori sensitivity, precision ed F1 per le letture con numero di letture determinato attraverso il parametro coverage, con tassi d'errore Illumina e PacBio*

Si evidenzia immediatamente una sensibile dimunizione del valore di sensitivity per tutte le lunghezze di lettura generate. Le letture corte con tecnologia Illumina hanno la

performance peggiore, del 73%, a 100bp e il risultato migliore a 600bp con sensitivity dell'81%. La media dei valori di sensitivity è del 78,2%. I valori di precisione, invece, presentano un andamento decrescente con una media del 91%. Ciò si traduce in un tasso d'errore medio di circa il 9%. I valori F1 non hanno risultati molto superiori alla sensitivity: con una media del 84,1%, si trova il valore peggiore a 1000bp (81%) e il valore migliore a 600bp (86%). In questo caso non è possibile notare una correlazione tra miglioramento delle performance di classificazione e aumento della lunghezza di lettura: se da 100bp a 600bp i valori di sensitivity e F1 diventano progressivamente più elevati, dai 700bp si osserva una brusca caduta e un andamento decresente.

Anche le letture corte con tecnologia PacBio presentano un picco a 600 bp per sensitivity ed F1: la sensitivity presenta il valore più alto del 77% e l'F1 score dell'82% . Il valore di precisione più alto si registra invece a 100bp con un valore dell'89%.

Gli ultimi set di letture corte, con i tassi d'errore Nanopore presentano anch'essi un andamento simile a quello delle letture eseguite con gli altri tassi. Possiamo trovare una media delle sensitivity dell' 80,8%, una media dei valori di precisione del 86,9% e l'83,7% per l'F1 score.

Le letture lunghe con tecnologia PacBio, invece, sono più stabili. Dopo una sensitivity relativamente bassa del 75%, si nota un miglioramento e i valori successivi oscillano intorno all'83%. La media delle sensitivity è del 82,1%. Ciò si traduce in un tasso d'errore medio di circa il 18%, inferiore a quello ottenuto con letture corte di tecnologia Illumina. L'F1 score segue lo stesso andamento con una media dell'84%.

Le letture lunghe con tecnologia Nanopore hanno risultati simili a PacBio. Infatti, le due curve sono quasi sovarpponibili e non sorprende, quindi, che la media sia anch'essa dell'82,1%. Le stesse osservazioni si rilevano per i valori F1, dove la media si attesta all'84,3%, di poco superiore alla media dei valori F1 per la tecnologia PacBio.

5 Conclusioni

Le classificazioni tassonomiche di Kraken 2 permettono di elaborare alcune considerazioni. In generale, si può osservare una correlazione tra lunghezza delle letture e classificazioni corrette di Kraken 2. All'aumentare della lunghezza di lettura, la classificazione migliora fino a raggiungere un picco oppure un valore di soglia. I bassi tassi d'errore della tecnologia Illumina rendono la classificazione di kraken 2 accurata già a 300bp, con una sensitivity dell'89%, mentre, utilizzando i tassi d'errore più elevati, per raggiungere valori simili è necessario avere lunghezze di lettura maggiori o uguali a 2000bp per PacBio e 3000bp per Nanopore. Tuttavia, anche a brevi lunghezze di lettura si trovano valori relativamente elevati (F1 score della tecnologia Nanopore di circa l'80% a 100bp) . A 4000bp raggiungono un valore che rimane pressochè costante (F1 score di circa il 93% sia per PacBio che per Nanopore). In tal senso, sia la tecnologia Pacific Bioscience che la tecnologia Oxford Nanopore portano a risultati simili nelle classificazioni di Kraken 2 per letture lunghe ($\geq 4000bp$). Inoltre, la sensitivity e l'F1 score delle tecnologie di terza generazione presentano valori più elevati delle misurazioni rilevate con Illumina, di circa 2 punti percentuale. Invece, si registra l'esatto opposto per quanto riguarda i valori di precisione.

Considerando i risultati delle classificazioni delle letture con il parametro coverage, si nota una sensibile riduzione dei valori di tutte le misurazioni effettuate. Inoltre, per le letture corte, se da 100bp a 600bp le performance migliorano, da 600bp a 1000bp le performance peggiorano progressivamente, riportandosi ai livelli di 100bp circa. Ciò non si verifica invece per le letture lunghe che dopo i 2000bp si mantengono circa agli stessi valori, anche se l'F1 score presenta un lieve peggioramento. Si possono formulare delle ipotesi sul motivo di queste performance. Infatti, viene generato un numero di letture diverso per ogni specie che porta ad uno squilibrio nella quantità di letture per specie. Inoltre, le dimensioni dei file generati aumentano il fattore di carico della CHT di Kraken 2.

Per quanto riguarda, invece, il numero di letture, non si registra alcuna correlazione particolare. Infatti, con una lunghezza fissa di 5000bp per letture PacBio e Nanopore e 500bp per Illumina, il grafico dei valori sensitivity, precisione ed F1 score è comparabile ad una linea retta. Si può evincere, però, che le letture aventi tecnologia Illumina ottengono sensitivity e precisione inferiori rispetto a PacBio e Nanopore (anche in questo caso molto simili). Tuttavia, i valori dell'F1 score si sovrappongono.

Il tasso d'errore delle classificazioni di Kraken2 che viene calcolato utilizzando la me-

trica precision, risulta, quindi accettabile per la maggior parte dei set di letture, con un valore di circa il 5% per le letture corte con tecnologia Illumina e del 7% letture lunghe con tecnologia Oxford Nanopore e Pacific Bioscience. Tuttavia, le letture generate utilizzando il parametro coverage portano a tassi d'errore molto più elevati, di circa il 13% per PacBio e Nanopore e dell'8% per Illumina.

Riferimenti bibliografici

[1] Chistoserdova, L. Recent progress and new challenges in metagenomics for biotechnology. Biotechnol Lett 32, 1351–1359 (2010). https://doi.org/10.1007/s10529-010-0306-9

[2] Semenov, M.V. Metabarcoding and Metagenomics in Soil Ecology Research: Achievements, Challenges, and Prospects. Biol Bull Rev 11, 40–53 (2021). https://doi.org/10.1134/S2079086421010084

[3] Matteo Comin, Barbara Di Camillo, Cinzia Pizzi, Fabio Vandin, Comparison of microbiome samples: methods and computational challenges, Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 88–95, https://doi.org/10.1093/bib/bbaa121

[4] Pearman, W.S., Freed, N.E. and Silander, O.K. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. BMC Bioinformatics 21, 220 (2020). https://doi.org/10.1186/s12859-020-3528-4

[5] Bonnie L. Brown, Mick Watson, Samuel S. Minot, Maria C. Rivera, Rima B. Franklin, MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach, GigaScience, Volume 6, Issue 3, March 2017, gix007, https://doi.org/10.1093/gigascience/gix007

[6] Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, et al. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett. 2013;16(10):1245–57. https://doi.org/10.1111/ele.12162

[7] Quince, C., Walker, A., Simpson, J. et al. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35, 833–844 (2017). https://doi.org/10.1038/nbt.3935

[8] Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008 Mar;24(3):133-41. doi: 10.1016/j.tig.2007.12.007. Epub 2008 Feb 11. PMID: 18262675

[9] Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012 Feb 9;2(1):3. doi: 10.1186/2042-5783-2-3. PMID: 22587947; PMCID: PMC3351745

[10] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

[11] Josip Marić, Krešimir Križanović, Sylvain Riondet, Niranjan Nagarajan, Mile Šikić bioRxiv 2020.11.25.397729; doi: https://doi.org/10.1101/2020.11.25.397729

[12] Pearman, W.S., Freed, N.E. & Silander, O.K. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. BMC Bioinformatics 21, 220 (2020). https://doi.org/10.1186/s12859-020-3528-4

[13] Wood, D.E., Lu, J. and Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257 (2019). https://doi.org/10.1186/s13059-019-1891-0

[14] Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236. https://doi.org/10.1186/s12864-015-1419-2

[15] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46

[16] Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721–9. https://doi.org/10.1101/gr.210641.116

[17] Luciani Mattia, Metagenomic classification of long reads with overlap graphs [tesi di laurea magistrale]. Padova: Università degli studi di Padova, 2022.

[18] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176

[19] A. T. Dilthey, C. Jain, S. Koren, and A. M. Phillippy, "Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps," Nat. Commun., vol. 10, no. 1, p. 3066, Jul. 2019

[20] Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, et al. (2016) MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Computational Biology 12(6): e1004957. https://doi.org/10.1371/journal.pcbi.1004957

[21] M. Holtgrewe, “Mason – A Read Simulator for Second Generation Sequencing Data”, 2010-10

[22] Bianca K. Stöcker, Johannes Köster, Sven Rahmann, SimLoRD: Simulation of Long Read Data, Bioinformatics, Volume 32, Issue 17, 1 September 2016, Pages 2704–2706, https://doi.org/10.1093/bioinformatics/btw286

[23] Sims, D., Sudbery, I., Ilott, N. et al. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121–132 (2014). https://doi.org/10.1038/nrg3642

Appendice

In questa Appendice sono contenenti tutti i risultati derivanti dai test descritti nella sezione Metodi.

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
100bp	66953	2045	6140	4861	0	79999	0,836923	0,970361	0,898716	0,985974
200bp	69998	2516	4695	2789	0	79998	0,874997	0,965303	0,917934	0,993129
300bp	71037	2766	4286	1937	0	80026	0,887674	0,962522	0,923584	0,995806
400bp	71535	3055	3989	1472	0	80051	0,893618	0,959043	0,925175	0,997072
500bp	71912	3292	3671	1124	0	79999	0,898911	0,956226	0,926683	0,997936
600bp	72209	3376	3521	915	0	80021	0,902376	0,955335	0,9281	0,99851
700bp	72427	3539	3361	689	0	80016	0,905156	0,953413	0,928658	0,998758
800bp	72533	3643	3224	637	0	80037	0,906243	0,952177	0,928642	0,999011
900bp	72694	3795	3007	514	0	80010	0,908561	0,950385	0,929003	0,99915
1000bp	72756	3909	2901	468	0	80034	0,909064	0,949012	0,928608	0,999269
mean	\	\	\	\	\	\	0,8923523	0,9573777	0,9235103	0,9964615

Tabella 1: *Output del progrmma evaluate_calls per letture corte con tecnologia Illumina.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
100bp	64987	7270	9461	4743	0	86461	0,751634	0,899387	0,818899	0,986321
200bp	69392	8362	5502	2941	0	86197	0,80504	0,892456	0,846497	0,992541
300bp	70828	8497	4393	1940	0	85658	0,82687	0,892884	0,85861	0,995571
400bp	71497	7971	3861	1491	0	84820	0,842926	0,899695	0,870386	0,997102
500bp	71890	7822	3435	1170	0	84317	0,852616	0,901872	0,876552	0,99762
600bp	72113	7303	3193	998	0	83607	0,862523	0,908041	0,884697	0,99831
700bp	72182	11134	3218	887	0	87421	0,825683	0,866364	0,845534	0,998577
800bp	72324	11090	3015	770	0	87199	0,829413	0,867049	0,847813	0,998703
900bp	72528	10443	2804	622	0	86397	0,839474	0,874137	0,856455	0,998926
1000bp	72635	9889	2700	543	0	85767	0,846887	0,880168	0,863207	0,999005
mean	\	\	\	\	\	\	0,8283066	0,8882053	0,856865	0,9962676

Tabella 2: *Output del progrmma evaluate_calls per letture corte con tecnologia PacBio.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
100bp	63665	7991	10664	4769	0	87089	0,731034	0,888481	0,802104	0,986465
200bp	68827	10501	6058	2897	0	88283	0,779618	0,867626	0,821271	0,992677
300bp	70632	11666	4665	2002	0	88965	0,79393	0,858247	0,824837	0,995751
400bp	71431	12966	3910	1488	0	89795	0,79549	0,846369	0,820141	0,997291
500bp	71873	13700	3578	1169	0	90320	0,79576	0,839903	0,817235	0,997684
600bp	72171	14384	3357	908	0	90820	0,79466	0,833817	0,813767	0,998415
700bp	71976	10086	3333	960	0	86355	0,83349	0,877093	0,854736	0,998368
800bp	72139	9394	3168	787	0	85488	0,843849	0,884783	0,863831	0,998582
900bp	72263	8546	2934	750	0	84493	0,855254	0,894244	0,874315	0,998783
1000bp	72323	7940	2809	665	0	83737	0,863692	0,901075	0,881988	0,998781
mean	\	\	\	\	\	\	0,8086777	0,8691638	0,8374225	0,9962797

Tabella 3: *Output del progrmma evaluate_calls per letture corte con tecnologia Oxford Nanopore.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000bp	72643	9763	2731	531	0	85668	0,84796	0,881526	0,864417	0,999114
2000bp	73050	4912	1768	305	0	80035	0,912726	0,936995	0,924701	0,999443
3000bp	73243	4497	1302	251	0	79293	0,923701	0,942153	0,932836	0,999633
4000bp	73369	4766	1005	222	0	79362	0,924485	0,939003	0,931688	0,999748
5000bp	73443	4912	779	219	0	79353	0,925523	0,937311	0,93138	0,999788
6000bp	73496	5003	632	226	0	79357	0,926144	0,936267	0,931178	0,99983
7000bp	73527	5010	527	263	0	79327	0,926885	0,936208	0,931523	0,999865
8000bp	73529	5164	473	211	0	79377	0,926326	0,934378	0,930335	0,999874
9000bp	73464	5183	432	276	0	79355	0,925764	0,934098	0,929912	0,999813
10000bp	73522	5223	327	299	0	79371	0,926308	0,933672	0,929975	0,999867
mean	\	\	\	\	\	\	0,9165822	0,9311611	0,9237945	0,9996975

Tabella 4: *Output del progrmma evaluate_calls per letture lunghe con tecnologia PacBio.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000bp	72659	11850	2697	521	0	87727	0,82824	0,859778	0,843714	0,999056
2000bp	73200	9165	1680	250	0	84295	0,868379	0,888727	0,878435	0,999576
3000bp	73382	6068	1169	209	0	80828	0,907878	0,923625	0,915684	0,999745
4000bp	73441	5132	833	208	0	79614	0,922463	0,934685	0,928534	0,999769
5000bp	73483	5042	656	204	0	79385	0,925653	0,935791	0,930695	0,999782
6000bp	73554	5128	506	188	0	79376	0,926653	0,934826	0,930722	0,999846
7000bp	73590	5136	404	207	0	79337	0,927562	0,934761	0,931148	0,999858
8000bp	73595	5193	326	214	0	79328	0,92773	0,934089	0,930899	0,999881
9000bp	73614	5249	284	227	0	79374	0,927432	0,933442	0,930427	0,999894
10000bp	73672	5236	215	203	0	79326	0,928725	0,933644	0,931178	0,999927
mean	\	\	\	\	\	\	0,9090715	0,9213368	0,9151436	0,9997334

Tabella 5: *Output del progrmma evaluate_calls per letture lunghe con tecnologia Oxford Nanopore.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000	35928	1611	1894	576	0	40009	0,897998	0,957085	0,9266	0,997776
2000	71912	3292	3671	1124	0	79999	0,898911	0,956226	0,926683	0,997936
3000	107877	4854	5599	1697	0	120027	0,898773	0,956942	0,926946	0,997835
4000	143859	6503	7328	2283	0	159973	0,899271	0,956751	0,927121	0,997872
5000	179814	8114	9238	2855	0	200021	0,898976	0,956824	0,926998	0,99789
6000	215802	9599	11153	3454	0	240008	0,899145	0,957414	0,927365	0,997869
7000	251732	11263	12973	4041	0	280009	0,899014	0,957174	0,927183	0,997824
8000	287749	12913	14788	4580	0	320030	0,899131	0,957051	0,927188	0,997862
9000	323718	14468	16648	5191	0	360025	0,899154	0,957219	0,927278	0,997843
10000	359721	16078	18454	5726	0	399979	0,89935	0,957216	0,927381	0,997867
mean	\	\	\	\	\	\	0,8989723	0,9569902	0,9270743	0,9978574

Tabella 6: *Output del progrmma evaluate_calls al variare del numero di letture con tecnologia Illumina.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000	36728	2415	393	120	0	39656	0,926165	0,938303	0,932195	0,999842
2000	73444	4964	730	220	0	79358	0,925477	0,93669	0,93105	0,999795
3000	110247	7365	1129	312	0	119053	0,926033	0,937379	0,931671	0,999832
4000	146936	9620	1611	442	0	158609	0,926404	0,938552	0,932439	0,999806
5000	183706	12218	1918	564	0	198406	0,925909	0,937639	0,931737	0,999814
6000	220388	14617	2300	717	0	238022	0,925914	0,937801	0,93182	0,999783
7000	257080	16990	2788	833	0	277691	0,925777	0,938009	0,931853	0,999778
8000	293912	19484	3060	896	0	317352	0,926139	0,937829	0,931947	0,999808
9000	330676	21912	3466	965	0	357019	0,926214	0,937854	0,931998	0,999806
10000	367490	24265	3935	1069	0	396759	0,92623	0,938061	0,932108	0,99982
mean	\	\	\	\	\	\	0,9260262	0,9378117	0,9318818	0,9998084

Tabella 7: *Output del progrmma evaluate_calls al variare del numero di letture con tecnologia PacBio.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000	36651	2428	446	136	0	39661	0,924107	0,937869	0,930937	0,999755
2000	73300	4891	913	275	0	79379	0,923418	0,937448	0,93038	0,99974
3000	109935	7296	1356	420	0	119007	0,923769	0,937764	0,930714	0,999684
4000	146601	9655	1845	573	0	158674	0,923913	0,93821	0,931007	0,999694
5000	183209	12229	2220	742	0	198400	0,923432	0,937428	0,930377	0,999696
6000	219965	14546	2714	921	0	238146	0,923656	0,937973	0,93076	0,999717
7000	256575	17047	3201	944	0	277767	0,923706	0,937699	0,93065	0,999702
8000	293264	19372	3655	1128	0	317419	0,923902	0,938037	0,930916	0,999712
9000	329966	21737	4089	1312	0	357104	0,924005	0,938195	0,931046	0,999715
10000	366621	24129	4595	1445	0	396790	0,923967	0,93825	0,931054	0,999723
mean	\	\	\	\	\	\	0,9237875	0,9378873	0,9307841	0,9997138

Tabella 8: *Output del progrmma evaluate_calls al variare del numero di letture con tecnologia Oxford Nanopore.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000bp	2430616	445974	308958	20416	0	3205964	0,758154	0,844964	0,799209	0,998718
2000bp	1223441	157464	107840	7664	0	1496409	0,817585	0,88597	0,850405	0,999204
3000bp	817913	111954	52973	5402	0	988242	0,827644	0,879602	0,852833	0,999453
4000bp	614348	92073	30599	4613	0	741633	0,828372	0,869663	0,848515	0,999558
5000bp	492100	77887	19120	3905	0	593012	0,829831	0,863353	0,84626	0,999661
6000bp	410449	67348	12947	3701	0	494445	0,830121	0,859045	0,844335	0,999735
7000bp	351816	59346	9329	3336	0	423827	0,830093	0,855663	0,842684	0,999752
8000bp	307958	52779	7043	3096	0	370876	0,830353	0,853691	0,84186	0,999787
9000bp	273752	47866	5376	2912	0	329906	0,829788	0,851171	0,840344	0,999791
10000bp	246401	43611	4208	2712	0	296932	0,829823	0,849623	0,839607	0,999801
mean	\	\	\	\	\	\	0,8211764	0,8612745	0,8406052	0,999546

Tabella 9: *Output del progrmma evaluate_calls per letture lunghe con tecnologia PacBio e parametro coverage.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000bp	2422752	368266	315080	23114	0	3129212	0,774237	0,868053	0,818465	0,99859
2000bp	1220189	141215	112753	8883	0	1483040	0,822762	0,896273	0,857946	0,999078
3000bp	816043	108246	58257	6113	0	988659	0,825404	0,882887	0,853178	0,999336
4000bp	613199	89193	34391	4808	0	741591	0,82687	0,873015	0,849316	0,999483
5000bp	491323	75541	22414	4196	0	593474	0,827876	0,866739	0,846862	0,999599
6000bp	409604	65677	15579	3732	0	494592	0,828165	0,861814	0,844655	0,999648
7000bp	350977	57599	11821	3666	0	424063	0,827653	0,859025	0,843047	0,999644
8000bp	307028	51561	9085	3492	0	371166	0,827199	0,856211	0,841455	0,999645
9000bp	272913	46740	7131	3280	0	330064	0,826849	0,853779	0,840098	0,999656
10000bp	245618	42613	5707	3125	0	297063	0,826821	0,852157	0,839298	0,999675
mean	\	\	\	\	\	\	0,8213836	0,8669953	0,843432	0,9994354

Tabella 10: *Output del progrmma evaluate_calls per letture lunghe con tecnologia Oxford Nanopore e parametro coverage.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
1000bp	22543119	1420412	5355233	1397609	0	30716373	0,733912	0,940726	0,824549	0,989414
2000bp	11770685	777166	2264536	394324	0	15206711	0,774045	0,938064	0,848198	0,995155
3000bp	7952289	533199	1392167	184683	0	10062338	0,790302	0,937163	0,85749	0,996872
4000bp	6003718	413652	981590	106162	0	7505122	0,799949	0,935542	0,862449	0,997604
5000bp	4824607	344994	742756	68416	0	5980773	0,806686	0,933265	0,865371	0,998062
6000bp	4032210	301620	587318	47990	0	4969138	0,811451	0,930403	0,866865	0,998315
7000bp	3464149	433060	487028	35513	0	4419750	0,783788	0,888879	0,833033	0,998486
8000bp	3036007	427236	407121	27616	0	3897980	0,778867	0,876637	0,824865	0,998616
9000bp	2703035	419385	345359	22300	0	3490079	0,774491	0,865686	0,817553	0,998748
10000bp	2435179	411512	298776	18646	0	3164113	0,769625	0,855442	0,810267	0,998802
mean	\	\	\	\	\	\	0,7823116	0,9101807	0,841064	0,9970074

Tabella 11: *Output del progrmma evaluate_calls per letture corte con tecnologia Illumina e parametro coverage.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
100bp	21752823	2646410	6464919	1465489	0	32329641	0,672845	0,891537	0,766905	0,988503
200bp	11654599	1580929	2473233	418110	0	16126871	0,722682	0,880554	0,793845	0,994785
300bp	7923002	1101937	1468290	192846	0	10686075	0,741432	0,877901	0,803916	0,996683
400bp	5995869	827216	1018439	108858	0	7950382	0,754161	0,878762	0,811708	0,997544
500bp	4820663	650966	763202	69915	0	6304746	0,764609	0,881029	0,818701	0,997992
600bp	4030464	528569	600528	48806	0	5208367	0,773844	0,884061	0,825289	0,998273
700bp	3455925	673783	505186	38973	0	4673867	0,739414	0,836845	0,785119	0,998352
800bp	3029847	588812	422514	30412	0	4071585	0,744144	0,837284	0,787972	0,998491
900bp	2697357	513627	358580	24472	0	3594036	0,750509	0,840041	0,792755	0,998628
1000bp	2430843	446103	309047	20196	0	3206189	0,758172	0,844939	0,799207	0,99871
mean	\	\	\	\	\	\	0,7421812	0,8652953	0,7985417	0,9967961

Tabella 12: *Output del progrmma evaluate_calls per letture corte con tecnologia Pacbio e parametro coverage.*

N	TP	FP	FN	OK	TN	sum	sens	prec	F1	pearson
100bp	22208974	2199553	5922016	1437873	0	31768416	0,69909	0,909886	0,790679	0,988962
200bp	11731923	1238020	2363938	404088	0	15737969	0,745453	0,904547	0,81733	0,994974
300bp	7943995	832538	1429119	186966	0	10392618	0,764388	0,90514	0,828831	0,996791
400bp	6001567	615503	998346	107385	0	7722801	0,777123	0,906983	0,837046	0,997596
500bp	4822334	484082	752937	69272	0	6128625	0,786854	0,908774	0,843431	0,99802
600bp	4031247	398316	593554	48516	0	5071633	0,794862	0,910078	0,848577	0,998288
700bp	3461545	641178	500036	36470	0	4639229	0,746147	0,843719	0,791939	0,998441
800bp	3034353	604279	416388	28239	0	4083259	0,74312	0,833927	0,785909	0,998574
900bp	2701134	564123	354276	22850	0	3642383	0,741584	0,827235	0,782071	0,998692
1000bp	2434331	524929	305729	18482	0	3283471	0,74139	0,822615	0,779893	0,99878
mean	\	\	\	\	\	\	0,7540011	0,8772904	0,8105706	0,9969118

Tabella 13: *Output del progrmma evaluate_calls per letture corte con tecnologia Nanopore e parametro coverage.*

N	Specie	ID_NCBI
1	Amycolatopsis_mediterranei_Ref.fna	749927
2	Arthrobacter_arilaitensis_Ref.fna	861360
3	Azorhizobium_caulinodans_Ref.fna	438753
4	Bdellovibrio_bacteriovorus_Ref.fna	264462
5	Bifidobacterium_adolescentis.fna	367928
6	Bifidobacterium_animalis_Ref.fna	442563
7	Brachyspira_intermedia_Ref.fna	1045858
8	Candida_albicans.fna	237561
9	Clostridium_beijerinckii.fna	290402
10	Clostridium_tetani.fna	212717
11	Clostridium_thermocellum_Ref.fna	203119
12	Corynebacterium_ulcerans_Ref.fna	945711
13	Deinococcus_radiodurans.fna	243230
14	Ehrlichia_ruminantium.fna	254945
15	Erysipelothrix_rhusiopathiae_Ref.fna	525280
16	Faecalibacterium_prausnitzii.fna	411485
17	Fervidicoccus_fontis.fna	1163730
18	Fusobacterium_nucleatum.fna	393480
19	Lactobacillus_fermentum.fna	334390
20	Lactobacillus_gasseri.fna	324831
21	Lawsonia_intracellularis.fna	363253
22	Metallosphaera_cuprina.fna	1006006
23	Methanobrevibacter_smithii.fna	420247
24	Methanosarcina_barkeri.fna	1434109
25	Micrococcus_luteus.fna	465515
26	Mycoplasma_gallisepticum.fna	710128
27	Nitrosococcus_watsonii_Ref.fna	105559
28	Photobacterium_profundum_Ref.fna	314280
29	Porphyromonas_gingivalis.fna	242619
30	Rhodobacter_sphaeroides.fna	272943
31	Rickettsia_prowazekii_Ref.fna	1105097
32	Rickettsia_rickettsii.fna	452659
33	Roseburia_hominis.fna	585394
34	Saccharomyces_cerevisiae.fna	559292
35	Schaalia_odontolytica.fna	411466
36	Streptomyces_scabiei.fna	680198
37	Symbiobacterium_thermophilum_Ref.fna	292459
38	Thermococcus_sibiricus_Ref.fna	604354
39	Variovorax_paradoxus.fna	595537
40	Veillonella_rogosae.fna	1298595

Tabella 14: *Specie da cui sono stata generate le letture con relativi ID tassonomici forniti dall'NCBI.* `https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?`