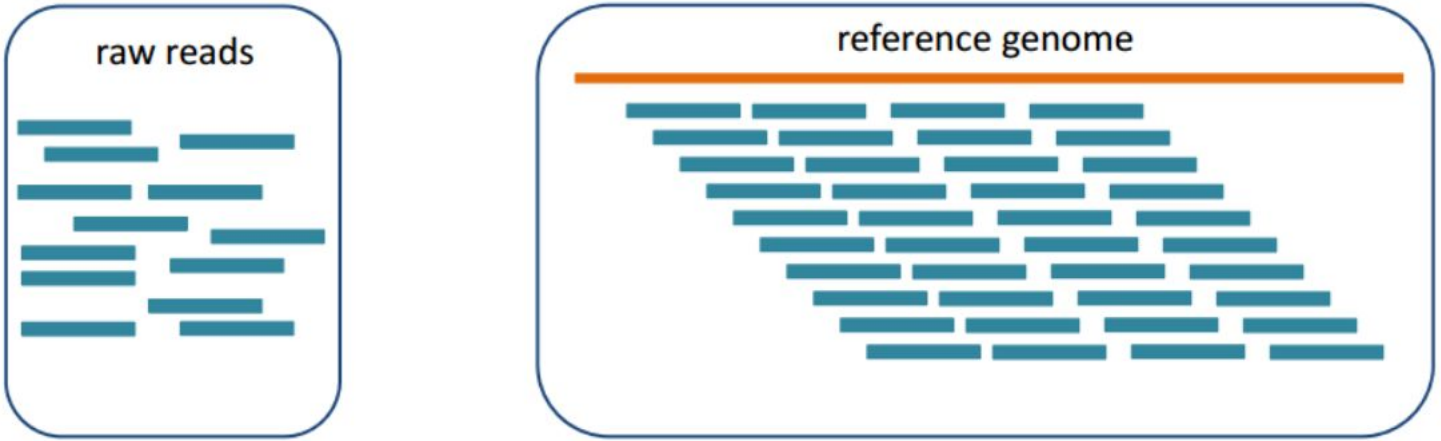
Algoritmi di allineamento sul genoma

- Allineamento: identificare la posizione delle reads rispetto al genoma di riferimento
- Processo nel quale si determina la posizione di provenienza più probabile di una read all'interno del genoma



Qualità dell'allineamento

Allineamento a un genoma di riferimento - qualità di mappatura (MQ)

Cosa succede se ci sono diversi posti possibili per allineare la ***read*** del sequenziamento? Ciò potrebbe essere dovuto a:

- Elementi ripetuti nel genoma
- Sequenze a bassa complessità
- Errori e lacune di riferimento

MQ è un punteggio Phred della qualità dell'allineamento

Con letture paired-end: la qualità del mapping è determinata sulla coppia, quindi anche se una lettura può essere mappata in più punti, la mappatura della sua coppia può aiutare a localizzarla correttamente.

raw reads

reference genome

low MQ: the probability of mapping to different locations is high, but no perfect multiple matches

high MQ: a single match

MQ0: a perfect multiple match

Metodi di allineamento

- Un buon metodo di allineamento dovrebbe:

– Gestire grandi **quantità** di reads: allineare centinaia di milioni di reads velocemente e senza richiedere enormi risorse computazionali

– Gestire sequenze di DNA **corte**: identificare la posizione genomica di piccole sequenze di DNA (35-300bp) con sicurezza

– Buon rapporto tra **sensibilità e velocità** di ricerca: identificare velocemente la giusta posizione genomica

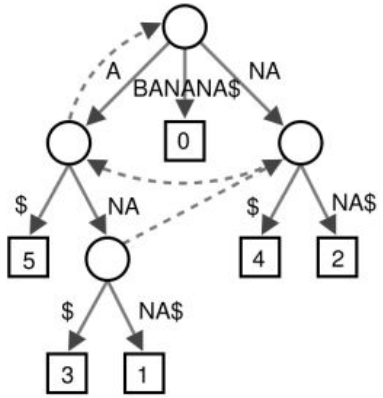
– Gestire **genomi** di riferimento di **grandi dimensioni**: ricercare in maniera veloce le posizioni delle reads in genomi lunghi alcune GB senza richiedere enormi risorse computazionali

Metodi di allineamento

- **Spaced-seed indexing:** divisione del genoma e delle reads in sottoinsiemi (seed) e ricerca delle corrispondenze tramite l'uso di indici
- **Trasformata di Burrows-Wheeler:** compressione del genoma di riferimento per una ricerca più veloce

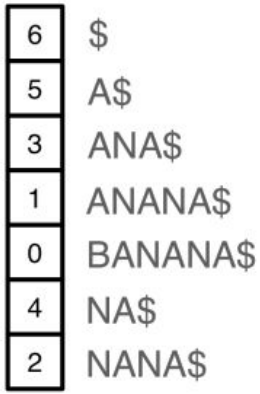
Indicizzazione del genoma

- Genomi e reads sono troppo grandi per approcci diretti (ad es. programmazione dinamica)
- É necessario creare un indice del genoma



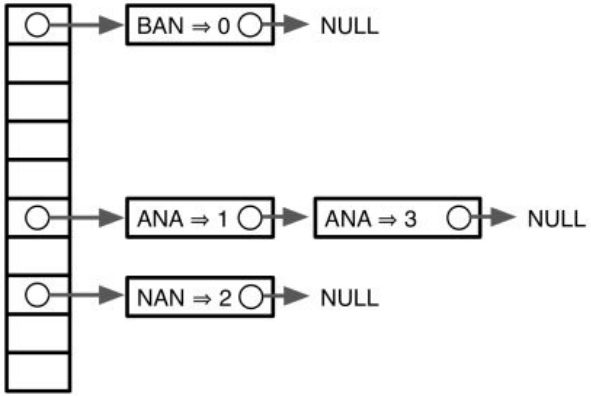
Suffix tree

> 35 GBs



Suffix array

> 12 GBs



Seed hash tables

Many variants, incl. spaced seeds

> 12 GBs

La scelta dell'indice è critica per le performance della mappatura.

Spaced-seed indexing

è un metodo utilizzato nell'allineamento di sequenze genomiche o nell'analisi di similitudine tra sequenze. Questo metodo è particolarmente utile quando si tratta di allineare sequenze con molte mutazioni o inserimenti/eliminazioni (indel) rispetto a un riferimento.

come funziona:

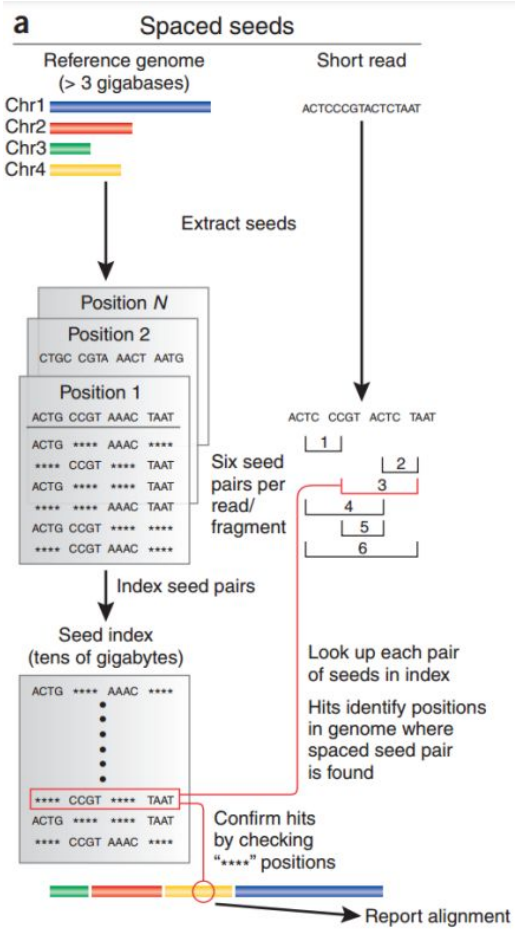
1. **Creazione dei seed:** Inizialmente, vengono generati dei “seed” di lunghezza fissata dalla sequenza di riferimento. Un seed è una sottosequenza di nucleotidi o basi (A, C, G, T) estratti a intervalli regolari dalla sequenza di riferimento. Questi seed sono progettati in modo da catturare le regioni conservate e permettere la rilevazione di sequenze simili nonostante le variazioni.

2.Costruzione dell'indice: I seed vengono quindi utilizzati per costruire un indice del riferimento. Questo indice consente di recuperare rapidamente le posizioni dei semi all'interno della sequenza di riferimento.

3.Allineamento: Quando si desidera allineare una sequenza di query (ad esempio, una sequenza di DNA sconosciuta) con il riferimento, vengono generati anche semi dalla sequenza di query. L'indice viene utilizzato per cercare corrispondenze tra i seed della query e quelli del riferimento.

4.Estensione dell'allineamento: Una volta trovate le corrispondenze iniziali tra i seed, viene eseguita un'estensione per identificare la migliore corrispondenza tra le due sequenze. Questa estensione può includere il calcolo della lunghezza dell'allineamento, la gestione degli indel e il calcolo della similarità.

5.Valutazione dell'allineamento: L'allineamento risultante viene valutato in base a vari criteri, come la similarità, la lunghezza dell'allineamento e la presenza di indel. L'allineamento migliore è solitamente selezionato come risultato finale.



Spaced seeds

- Spezzare il genoma in una serie di 4 sottosequenze di lunghezza uguale, chiamati “seed”
- Creare le possibili sei combinazioni di due coppie di seed lungo il genoma
- Creare un indice delle possibili combinazioni (molto voluminoso)
- Ripetere il processo di creazione dei seed per ogni read
- Identificare il match tra i seed sulla read e quelli nell'indice
- SNPs e INDELs saranno tollerati grazie agli “spazi” (****) tra i seed

Svantaggi

- Tollerante a polimorfismi che colpiscono al massimo due seed
- La creazione dell'indice del genoma di riferimento è un passaggio dispendioso in termini di tempo e risorse (circa 50GB di memoria per indicizzare il genoma umano)
- Velocità dipendente dalle risorse computazionali a disposizione

Trasformata di Burrows-Wheeler:

è una tecnica di compressione dei dati utilizzata anche nell'allineamento di sequenze genomiche. La BWT viene spesso utilizzata per preparare il testo di riferimento in modo da semplificare l'allineamento. Il principale algoritmo di allineamento basato sulla BWT è noto come "FM-index" (Index di Ferragina e Manzini), che è utilizzato in molti software di allineamento genetico come BWA e Bowtie.

come funziona:

1. **Creazione della BWT:** La BWT viene applicata alla sequenza di riferimento. Questa trasformazione riorienta il testo in modo da raggruppare caratteri simili vicino l'uno all'altro, rendendo più facile trovare corrispondenze tra la sequenza di query e il testo di riferimento.

2.Creazione dell'FM-index: Dalla BWT, viene creato l'FM-index, che è un'opportuna struttura dati che consente di cercare rapidamente corrispondenze tra la sequenza di query e il testo di riferimento. L'FM-index include informazioni su come i caratteri sono ordinati nella BWT e consente di eseguire ricerche di corrispondenze in modo efficiente.

3.Allineamento: Per allineare una sequenza di query al genoma di riferimento, viene utilizzato l'FM-index per cercare corrispondenze iniziali tra la sequenza di query e il testo di riferimento. Questa ricerca può essere fatta in modo molto efficiente grazie all'FM-index.

4.Estensione dell'allineamento: Una volta trovate le corrispondenze iniziali, viene eseguita un'estensione dell'allineamento simile a quanto avviene in altri metodi di allineamento. L'obiettivo è estendere l'allineamento in modo da coprire l'intera regione di interesse e gestire eventuali indel (inserimenti/eliminazioni) o mutazioni.

5.Valutazione dell'allineamento: L'allineamento risultante viene quindi valutato in base a criteri come la similarità, la lunghezza dell'allineamento e la qualità delle corrispondenze. L'allineamento migliore è selezionato come risultato finale.

Trasformata di Burrows-Wheeler

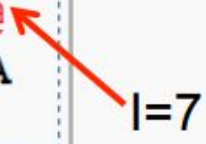
- Data una stringa T, un carattere speciale (e precedente a tutti i caratteri di T in ordine lessicografico) viene appeso in coda a T;

- Tutte le rotazioni cicliche di T formano la matrice di Burrows-Wheeler;

- Le righe della matrice vengono ordinate in ordine lessicografico;

- I caratteri nell'ultima colonna della matrice ordinata costituiscono la trasformazione di Burrows-Wheeler (tutto il resto è scartato); viene riportata anche la posizione nella matrice della stringa originale.

Transformation			
Input	All Rotations	Sort the Rows	Output
^BANANA@	^BANANA@ @^BANANA A@^BANAN NA@^BANA ANA@^BAN NANA@^BA ANANA@^B BANANA@^	ANANA@^**B** ANA@^BA**N** A@^BANA**N** BANANA@**^** NANA@^B**A** NA@^BAN**A** ^BANANA@ @^BANAN**A**	BNN^AA@A



- BWT(T) è reversibile, cioè è possibile ricostruire da essa la stringa di partenza;

- E' comprimibile, perchè caratteri identici finiscono spesso adiacenti uno all'altro in sottostringhe (è usata dal compressore bzip2);

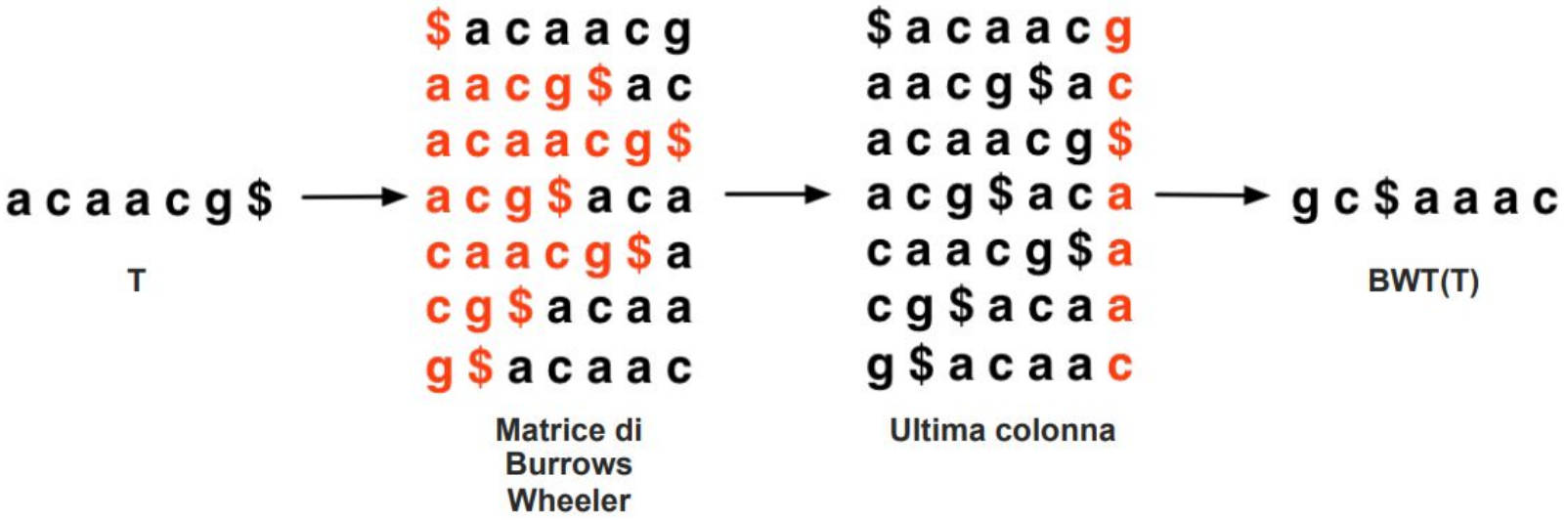
- Permette la ricerca veloce di sottostringhe (ad es. una read);

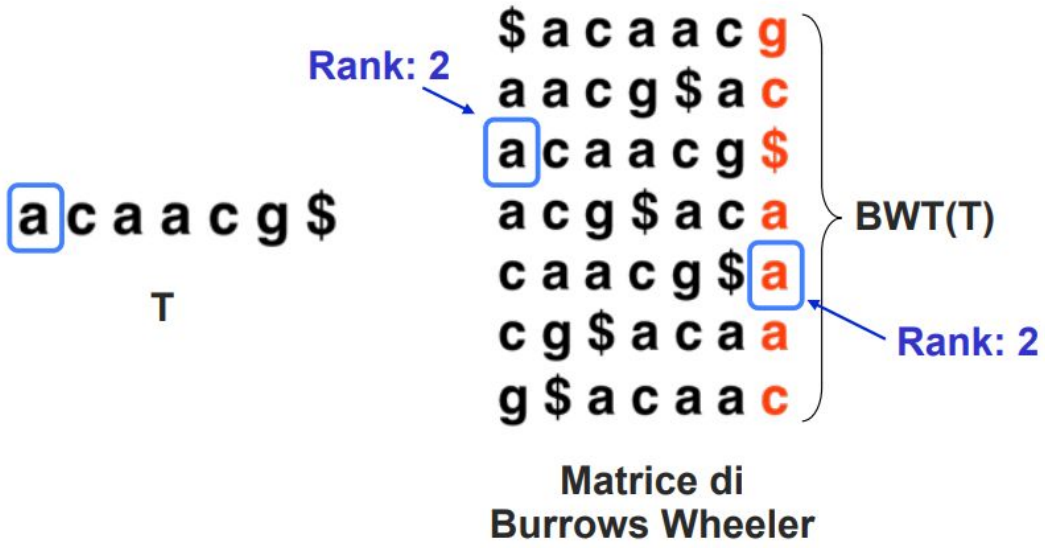
- Una volta trovato un buon match, se ne vogliono conoscere le coordinate nella stringa di partenza (cioè nel genoma).

- tenere traccia in un vettore delle coordinate di ogni nucleotide del genoma nella BWT(T) (occupa molta memoria);

- tenere traccia in un vettore solo di alcune posizioni ad intervalli fissi, trovare quello più vicino alla read in input, e ricostruire da lì.

La trasformazione BWT(T), una volta costruita, è l'unica cosa che viene mantenuta, la matrice non è più necessaria





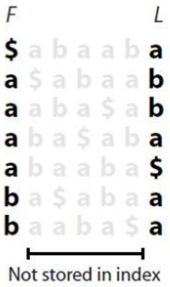
La proprietà che rende BWT(T) reversibile è detta “LF Mapping” : L’occorrenza i esima di un carattere nell’ultima (Last) colonna corrisponde alla occorrenza i esima dello stesso tipo di carattere nella prima (First) colonna.

FM Index

FM Index: an index combining the BWT *with a few small auxilliary data structures*

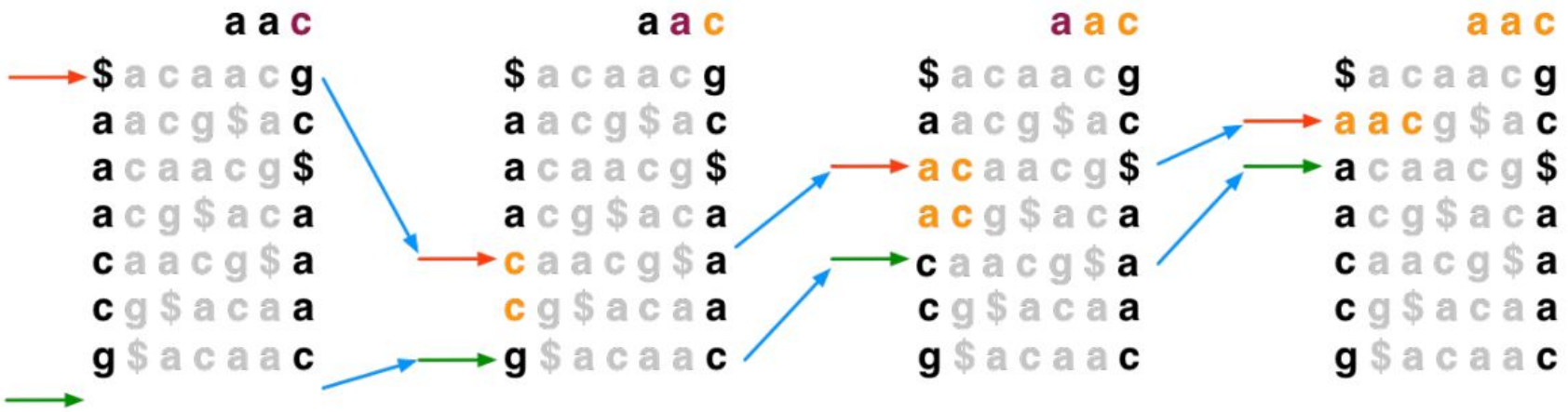
"FM" supposedly stands for "Full-text Minute-space." (But inventors are named Ferragina and Manzini)

Core of index consists of *F* and *L* from BWM:
F can be represented very simply (1 integer per alphabet character)
And *L* is compressible
Potentially very space-economical!



Paolo Ferragina, and Giovanni Manzini. "Opportunistic data structures with applications." *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on.* IEEE, 2000.

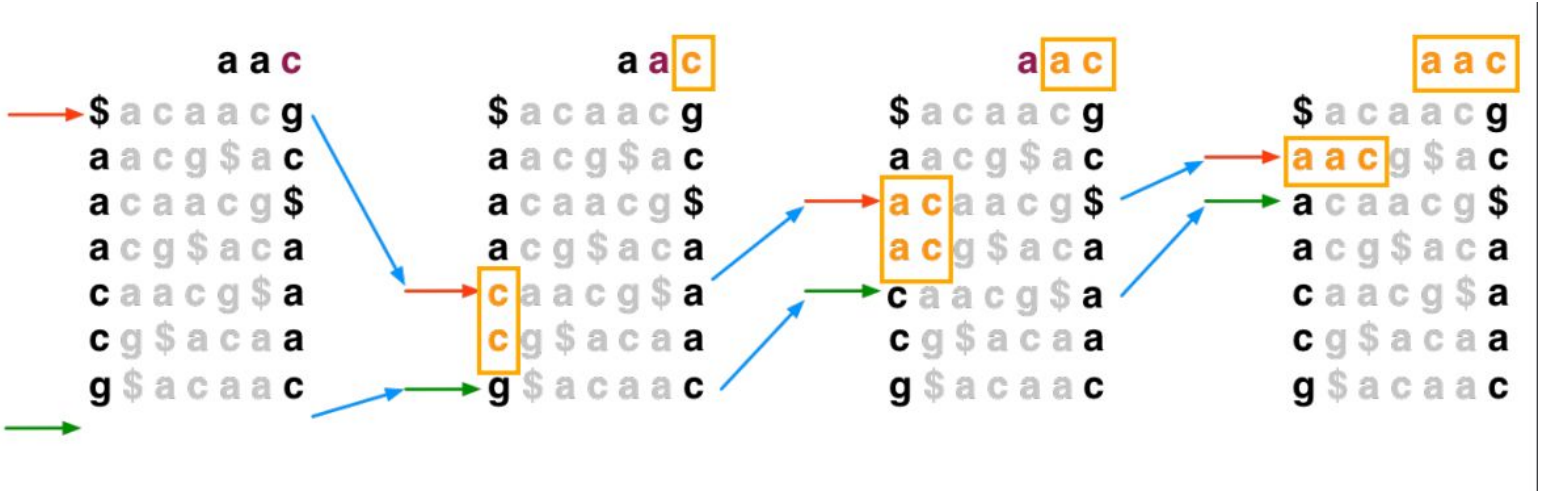
La proprietà che rende BWT(T) reversibile è detta "LF Mapping" : L'occorrenza i esima di un carattere nell'ultima (Last) colonna corrisponde alla occorrenza i esima dello stesso tipo di carattere nella prima (First) colonna.



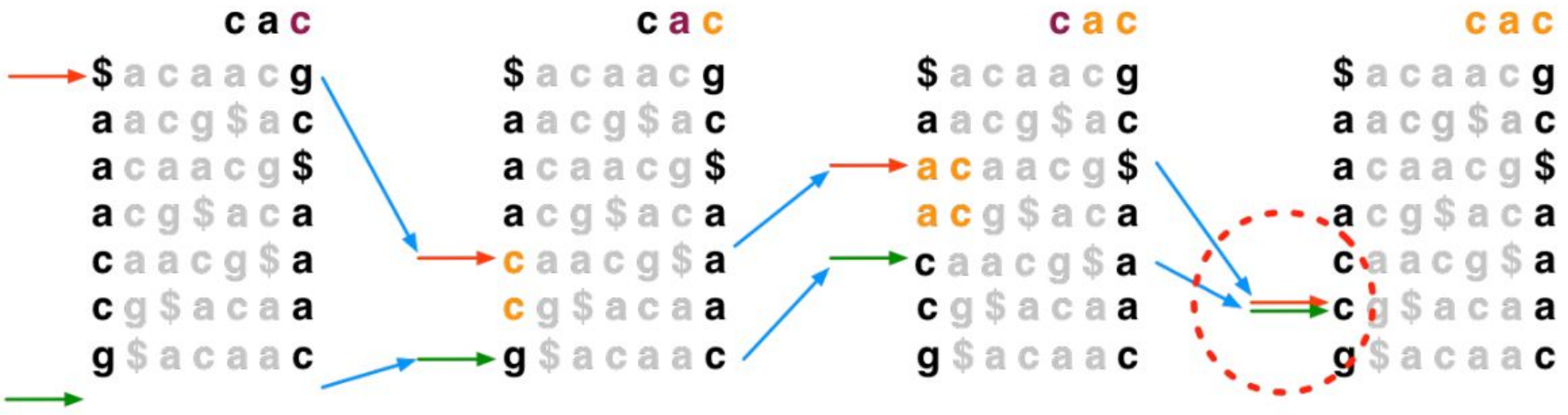
Per trovare un match perfetto di una stringa Q in BWT(T) si deve applicare ricorsivamente la regola:

top = LF(top, qc); bottom = LF(bottom, qc)

In cui qc è il prossimo carattere in Q (da destra a sinistra) e LF(i, qc) mappa la riga i alla riga il cui primo carattere corrisponde all'ultimo carattere di i



In round successivi, le posizioni top & bottom delimitano il range di righe che iniziano con suffissi progressivamente più lunghi di Q

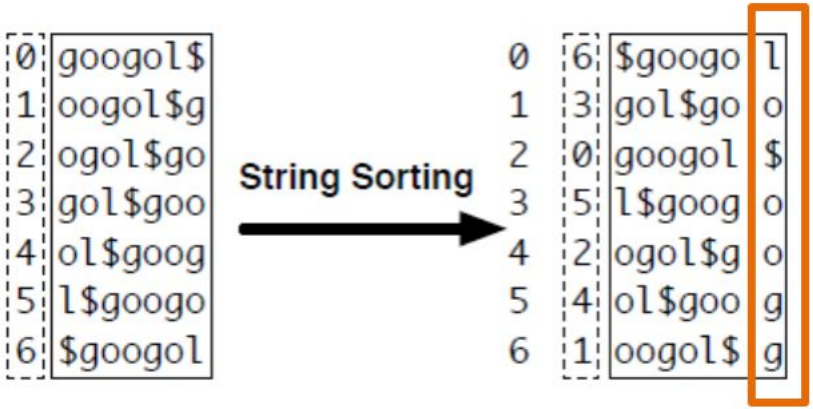


Se il range diventa vuoto (top = bottom) il suffisso di Q, e quindi la sottostringa Q, non è presente nel testo (cioè nel genoma nel nostro caso).

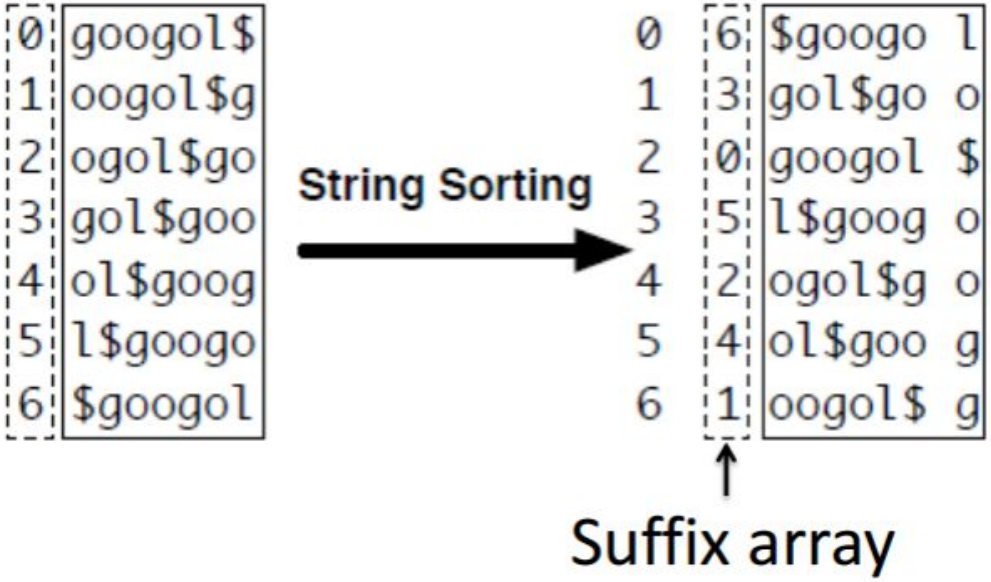
Per contemplare casi di errori di sequenziamento, o di polimorfismi rispetto al genoma di riferimento, esistono varianti di questo algoritmo di base.

Trasformata di Burrows-Wheeler

1. Testo originale = “googol”
2. Aggiungere ‘$’ alla fine = X = “googol$”
3. Mettere in ordine lessicografico tutte le rotazioni del testo X
4. Prendere l'ultima colonna



BW & suffix arrays



un suffix array per una stringa data è un array di interi che rappresenta l'ordine lessicografico di tutti i suffissi di quella stringa. In altre parole, è una lista ordinata di tutti i possibili suffissi di una stringa.

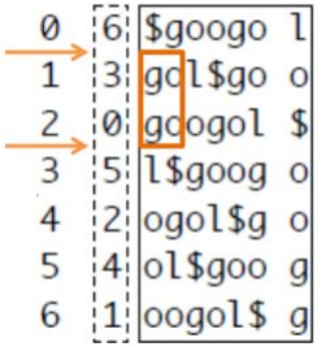
BW & suffix arrays interval

- Tutte le sottosequenze con un suffisso comune W appariranno vicine tra loro, definendo un intervallo nel suffix array

- Intervallo suffix array di

  “go” = [1,2]

W = “go”



X = googol$

0 3

Vantaggi di utilizzo della trasformata di Burrows-Wheeler (BWT) e dell'FM-index :

- **Efficienza computazionale:** La BWT e l'FM-index permettono di effettuare ricerche di allineamento in modo molto efficiente. Questi algoritmi sono progettati per ridurre il tempo di calcolo necessario per trovare corrispondenze tra la sequenza di query e il genoma di riferimento, il che è particolarmente importante quando si lavora con sequenze genomiche lunghe su genomi di riferimento di grandi dimensioni.

- **Gestione di mutazioni e indel:** L'FM-index è in grado di gestire efficacemente variazioni genomiche come mutazioni e indels. Questo lo rende adatto per l'allineamento di sequenze con mutazioni o variazioni rispetto al riferimento, come sequenze di tumori o di popolazioni genetiche diverse.

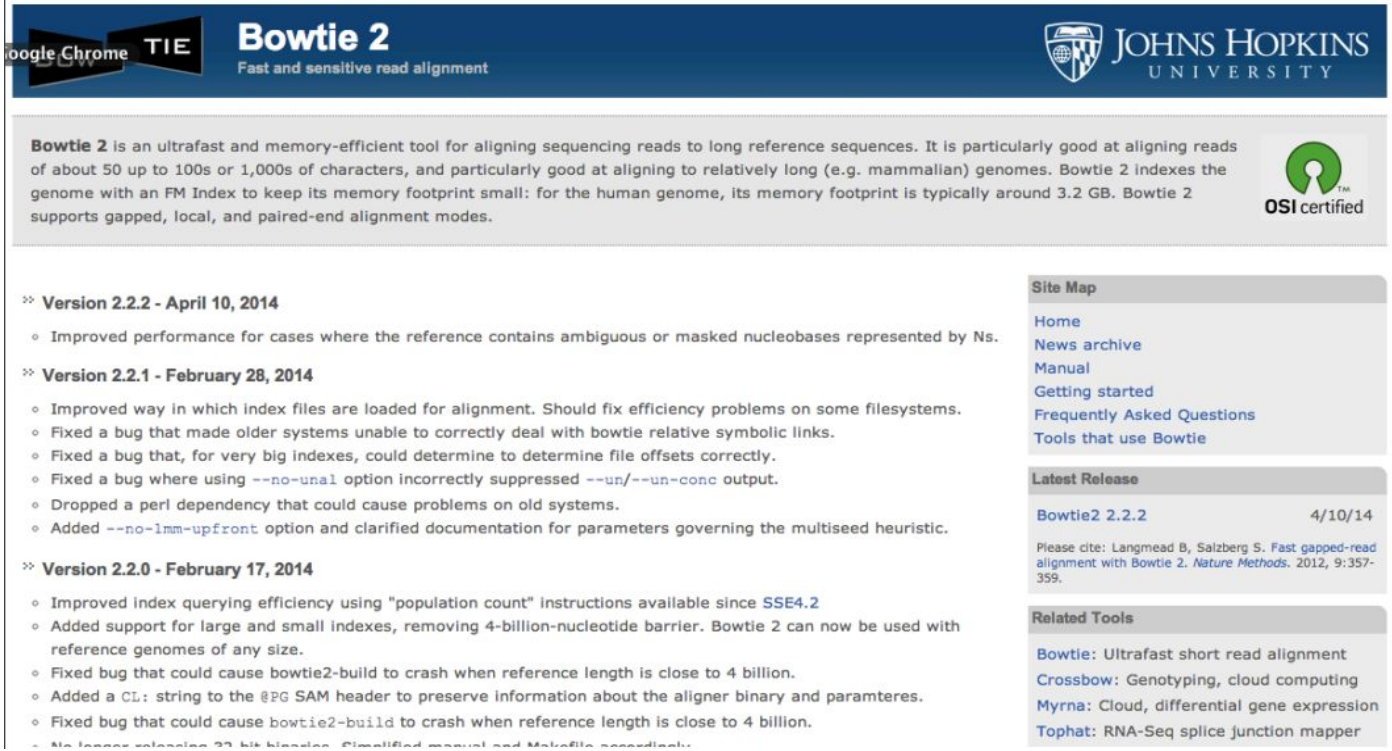
- **Spazio di memoria ridotto:** L'FM-index richiede meno spazio di memoria rispetto ad alcune altre strutture dati utilizzate nell'allineamento genetico. Questo significa che è possibile mantenere l'indice in memoria anche su hardware con risorse limitate.

- **Adattabilità a diverse applicazioni:** L'FM-index può essere utilizzato per una vasta gamma di applicazioni nell'analisi genomica, tra cui l'allineamento di sequenze DNA e RNA, la ricerca di varianti genetiche, la determinazione della similarità tra sequenze, la ricerca di regioni funzionali e altro ancora.

- **Software ampiamente disponibili:** Esistono diversi software ben sviluppati che sfruttano l'FM-index, come BWA, Bowtie, Bowtie2 e HISAT2. Questi strumenti consentono di eseguire facilmente analisi genomiche avanzate senza la necessità di scrivere da zero un proprio algoritmo di allineamento.

- **Allineamento accurato:** L'FM-index è in grado di produrre allineamenti molto accurati, il che è essenziale per molte applicazioni nell'analisi genomica, come la scoperta di varianti genetiche rilevanti per la salute umana.

Bowtie (Langmead, 2009)



Google Chrome TIE

Bowtie 2

Fast and sensitive read alignment

JOHNS HOPKINS UNIVERSITY

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

OSI certified

Version 2.2.2 - April 10, 2014

- Improved performance for cases where the reference contains ambiguous or masked nucleobases represented by Ns.

Version 2.2.1 - February 28, 2014

- Improved way in which index files are loaded for alignment. Should fix efficiency problems on some filesystems.
- Fixed a bug that made older systems unable to correctly deal with bowtie relative symbolic links.
- Fixed a bug that, for very big indexes, could determine to determine file offsets correctly.
- Fixed a bug where using `--no-unal` option incorrectly suppressed `--un`/`--un-conc` output.
- Dropped a perl dependency that could cause problems on old systems.
- Added `--no-lmm-upfront` option and clarified documentation for parameters governing the multiseed heuristic.

Version 2.2.0 - February 17, 2014

- Improved index querying efficiency using "population count" instructions available since SSE4.2
- Added support for large and small indexes, removing 4-billion-nucleotide barrier. Bowtie 2 can now be used with reference genomes of any size.
- Fixed bug that could cause bowtie2-build to crash when reference length is close to 4 billion.
- Added a `CL:` string to the `@PG` SAM header to preserve information about the aligner binary and paramteres.
- Fixed bug that could cause `bowtie2-build` to crash when reference length is close to 4 billion.

Site Map

Home
News archive
Manual
Getting started
Frequently Asked Questions
Tools that use Bowtie

Latest Release

Bowtie2 2.2.2 4/10/14

Please cite: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

Related Tools

Bowtie: Ultrafast short read alignment
Crossbow: Genotyping, cloud computing
Myrna: Cloud, differential gene expression
Tophat: RNA-Seq splice junction mapper

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

https://github.com/BenLangmead/bowtie2

Bowtie2 è un popolare strumento di allineamento genomico utilizzato per mappare rapidamente le sequenze di lettura su un genoma di riferimento. È ampiamente utilizzato nella bioinformatica e nella genomica per diverse applicazioni, come il sequenziamento del DNA, il sequenziamento dell'RNA e l'identificazione di varianti genetiche.

Caratteristiche e vantaggi

- **Algoritmo di allineamento efficiente:** utilizza un algoritmo di allineamento basato sulla Trasformata di Burrows-Wheeler (BWT) e sfrutta l'FM-index per ottenere risultati di allineamento molto veloci, anche su genomi di riferimento di grandi dimensioni.

- **Allineamenti multipli:** supporta l'allineamento di sequenze di lettura multiple e permette di trovare tutte le posizioni di allineamento possibili per ciascuna sequenza di query, fornendo una panoramica completa delle corrispondenze tra le sequenze di lettura e il riferimento.

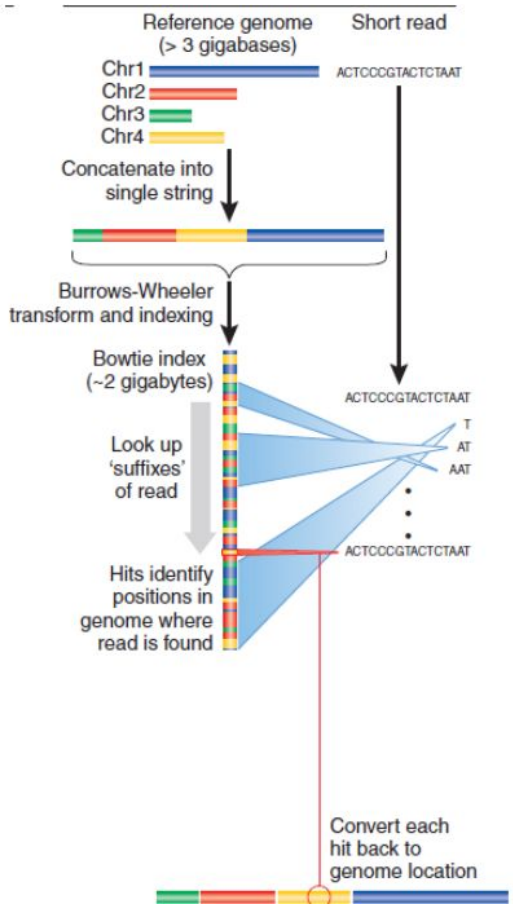
- **Gestione di indel:** è in grado di gestire sequenze di lettura con indel (inserimenti/eliminazioni) rispetto al genoma di riferimento. Ciò lo rende adatto per l'allineamento di dati di sequenziamento che possono contenere variazioni genomiche.

- **Configurabilità:** L'utente può configurare diversi parametri di allineamento per adattare l'analisi alle specifiche esigenze del progetto, inclusi i parametri di mismatch e gap, il tipo di output desiderato e altro ancora.

- **Compatibilità con formati standard:** supporta numerosi formati di input e output standard dell'industria, tra cui il formato FASTQ per le sequenze di lettura e il formato SAM/BAM per i risultati di allineamento, consentendo una facile integrazione con altri strumenti bioinformatici.

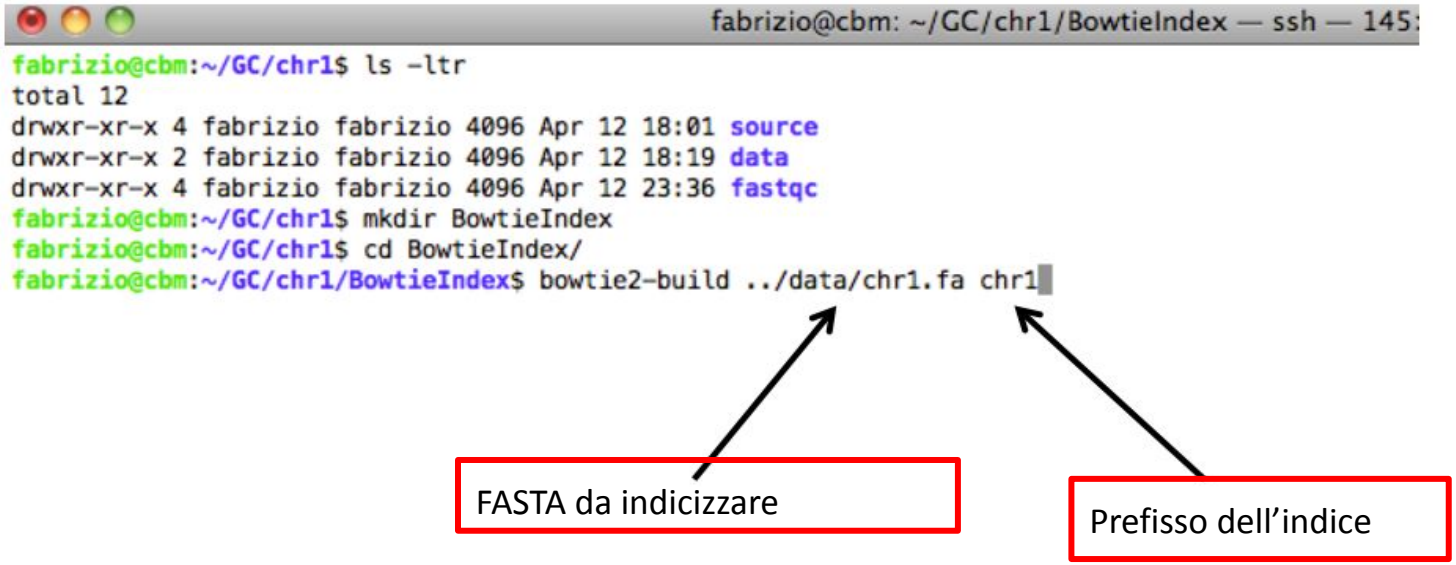
- **Ampia documentazione e comunità di supporto:** è ben documentato, con un manuale utente completo e una comunità attiva di sviluppatori e utenti che forniscono supporto e risorse online.

- **Open source e gratuito:** è un software open-source distribuito sotto la licenza GPLv3, il che significa che è gratuito da utilizzare e può essere personalizzato o esteso per adattarsi alle specifiche esigenze.



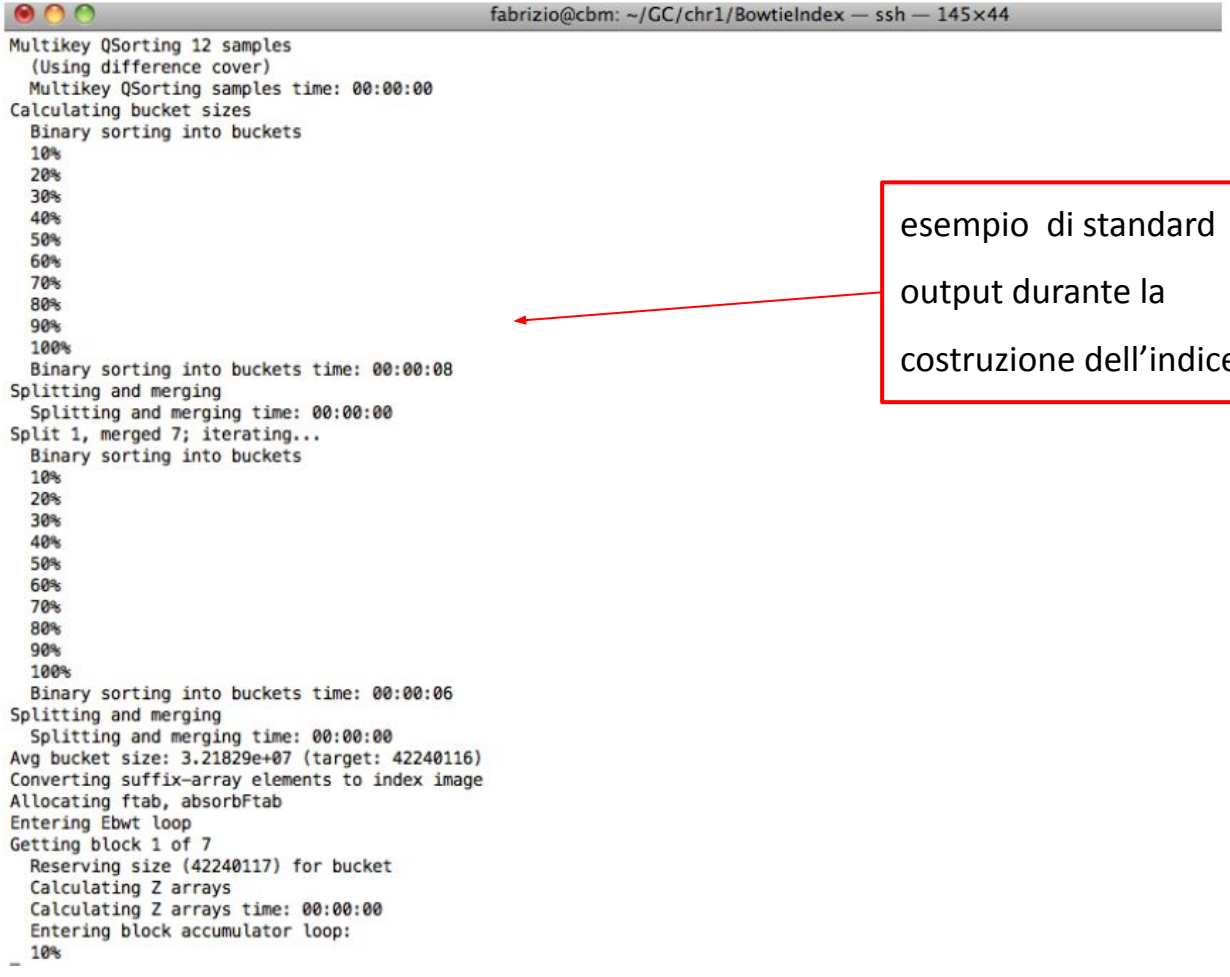
L'algoritmo bowtie2 implementa una trasformazione di Burrows-Wheeler per comprimere e indicizzare il genoma, in modo da occupare meno memoria e permettere la ricerca di sottostringhe (nel nostro caso le reads) in maniera veloce.

Come si lancia Bowtie2-build



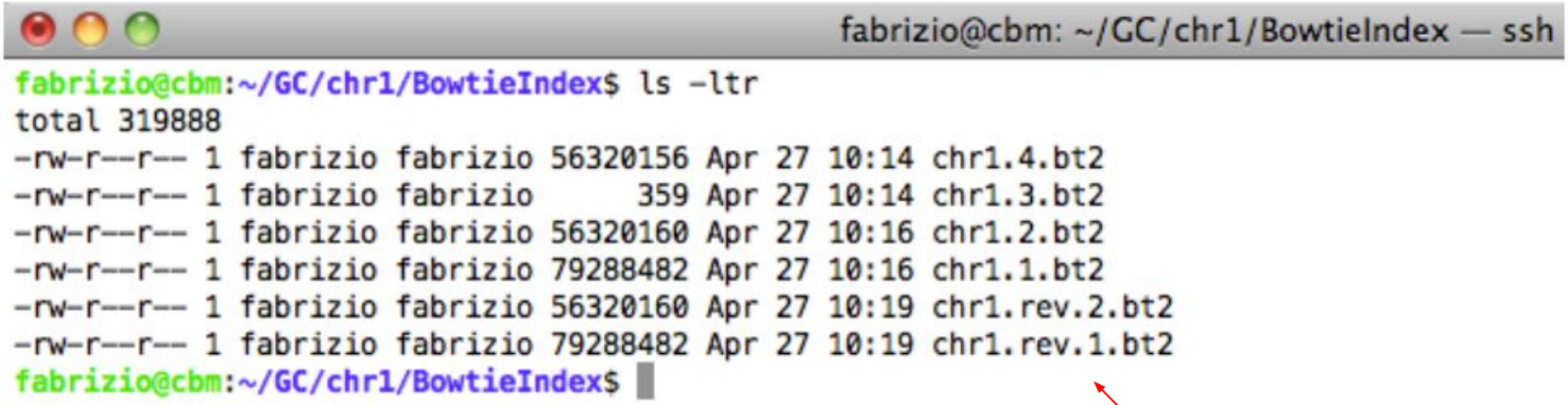
bowtie2-build crea vari files i cui nomi iniziano tutti con il prefisso prescelto

Il prefisso è poi utilizzato dagli altri programmi della procedura per specificare l'indice da usare



esempio di standard output durante la costruzione dell'indice

Output di bowtie2-build:



file prodotti durante l'indicizzazione del Chr1

Burrows-Wheeler Alignment (BWA) Tool on Biowulf

BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence, except for disallowing gaps close to the end of the query. It can also be tuned to find a fraction of longer gaps at the cost of speed and of more false alignments.

Quick Links

- Documentation
- Notes
- Interactive job
- Batch job
- Swarm of jobs

BWA excels in its speed. Mapping 2 million high-quality 35bp short reads against the human genome can be done in 20 minutes. Usually the speed is gained at the cost of huge memory, disallowing gaps and/or the hard limits on the maximum read length and the maximum mismatches. BWA does not. It is still relatively light-weighted (2.3GB memory for human alignment), performs gapped alignment, and does not set a hard limit on read length or maximum mismatches.

Given a database file in FASTA format, BWA first builds BWT index with the 'index' command. The alignments in suffix array (SA) coordinates are then generated with the 'aln' command. The resulting file contains ALL the alignments found by BWA. The 'samse/sampe' command converts SA coordinates to chromosomal coordinates. For single-end reads, most of computing time is spent on finding the SA coordinates (the aln command). For paired-end reads, half of computing time may be spent on pairing (the sampe command) given 32bp reads. Using longer reads would reduce the fraction of time spent on pairing because each end in a pair would be mapped to fewer places.

References:

- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
- If you use BWA-SW, please cite:
  Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub.

Documentation

https://hpc.nih.gov/apps/bwa.html

https://bio-bwa.sourceforge.net/

https://github.com/lh3/bwa

BWA è ampiamente utilizzato nella bioinformatica e nella genomica per mappare rapidamente sequenze di lettura su un genoma di riferimento

caratteristiche e vantaggi:

- **Efficienza computazionale:** BWA utilizza un algoritmo di allineamento basato sulla Trasformata di Burrows-Wheeler (BWT) e sfrutta un indice dell'FM-index per ottenere risultati di allineamento molto veloci, specialmente su genomi di riferimento di grandi dimensioni.

- **Allineamenti multipli:** BWA supporta l'allineamento di sequenze di lettura multiple e permette di trovare tutte le posizioni di allineamento possibili per ciascuna sequenza di query, fornendo una panoramica completa delle corrispondenze tra le sequenze di lettura e il riferimento.

- **Gestione di indel:** BWA è in grado di gestire sequenze di lettura con indel (inserimenti/eliminazioni) rispetto al genoma di riferimento. Ciò lo rende adatto per l'allineamento di dati di sequenziamento che possono contenere variazioni genomiche.

- **Configurabilità:** Gli utenti possono configurare vari parametri di allineamento per personalizzare l'analisi in base alle esigenze specifiche del progetto, inclusi i parametri di mismatch e gap, il tipo di output desiderato e altro ancora.

- **Compatibilità con formati standard:** BWA supporta numerosi formati di input e output standard dell'industria, come il formato FASTQ per le sequenze di lettura e il formato SAM/BAM per i risultati di allineamento, il che facilita l'integrazione con altri strumenti bioinformatici.

- **Ampia documentazione e comunità di supporto:** BWA è ben documentato, con un manuale utente completo e una comunità di sviluppatori e utenti attiva che fornisce supporto e risorse online.

- **Open source e gratuito:** BWA è un software open source distribuito sotto una licenza libera, il che significa che è gratuito da utilizzare e può essere personalizzato o esteso per adattarsi alle specifiche esigenze.

Differenze tra Bowtie2 e Bwa

Bowtie2 e BWA sono entrambi software popolari utilizzati per allineare reads di sequenziamento ad un genoma di riferimento, ma ci sono diverse differenze tra i due in termini di metodologia, funzionalità e prestazioni. Ecco alcune delle principali differenze:

1. **Metodologia di Indicizzazione**:
   - Bowtie2 utilizza un approccio basato sulla Burrows-Wheeler Transform (BWT) per costruire il suo indice. Questo approccio è ottimizzato per la velocità e l'efficienza.
   - BWA anch'esso utilizza la BWT, ma ha diversi algoritmi, tra cui BWA-backtrack, BWA-SW e BWA-MEM. BWA-MEM è il più recente e generalmente il più preferito per reads di lunghezza >70bp a causa della sua efficienza e accuratezza.

Differenze tra Bowtie2 e Bwa

2. **Tipi di Reads Supportate**:

- Bowtie2 supporta reads più lunghi rispetto alla sua versione precedente (Bowtie) e può gestire reads di qualsiasi lunghezza con una qualità di allineamento variabile.
- BWA è particolarmente efficiente con reads di lunghezza moderata (come 70-100bp da Illumina). Tuttavia, con BWA-MEM, può gestire reads molto lunghi come quelli prodotti da sequenziatori come Oxford Nanopore o PacBio.

3. **Scenari di Utilizzo**:

- Bowtie2 è spesso preferito per applicazioni che richiedono tempi di elaborazione rapidi, come applicazioni trascrittomiche dove è necessario mappare molte reads su trascritti noti.
- BWA, in particolare BWA-MEM, è spesso la scelta di riferimento per allineare reads su genomi di riferimento completi, come il genoma umano, a causa della sua accuratezza e efficienza.

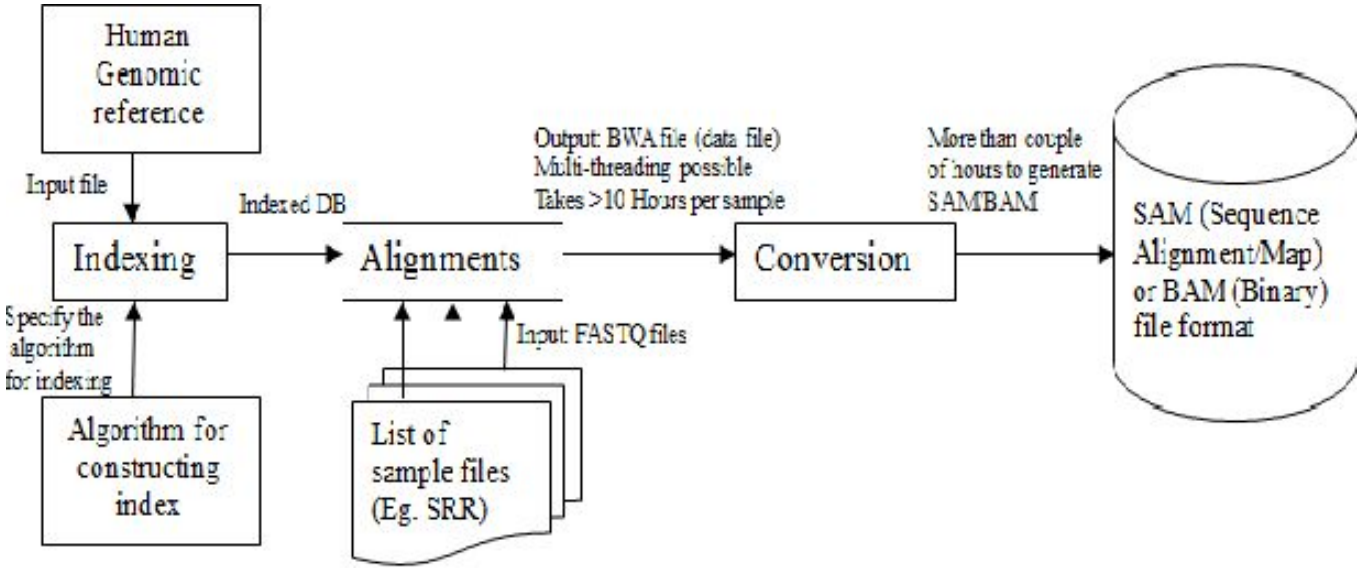
Differenze tra Bowtie2 e Bwa

4. **Funzionalità Aggiuntive**:
   - BWA-MEM, uno degli algoritmi di BWA, offre una migliore gestione delle regioni con alti tassi di variabilità o ripetizioni, e può identificare sub-optimal allineamenti, cosa che può essere utile per identificare varianti strutturali.

5. **Prestazioni**:
   - Sebbene sia Bowtie2 che BWA siano efficienti, le loro prestazioni possono variare in base alla natura dei dati e alle specifiche esigenze del progetto. Ad esempio, Bowtie2 potrebbe essere più veloce in certi contesti, mentre BWA potrebbe offrire un allineamento più accurato per reads complesse o problematiche.

Entrambi i tool sono ampiamente utilizzati nella comunità bioinformatica e hanno contribuito in modo significativo all'analisi dei dati di sequenziamento ad alto rendimento. La scelta tra i due spesso dipende dalla natura specifica dei dati e dagli obiettivi dell'analisi.

Tipico workflow bioinformatico per il mapping delle reads contro un reference genome

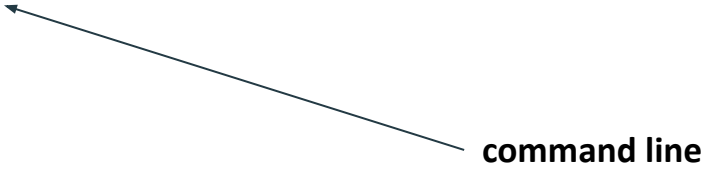


comandi BWA:

Indicizzazione del genoma di riferimento:

Per indicizzare il genoma di riferimento, è necessario eseguire il comando **bwa index**. Ad esempio, se hai un file FASTA del genoma di riferimento chiamato **reference.fasta**, puoi indicizzarlo con il seguente comando:

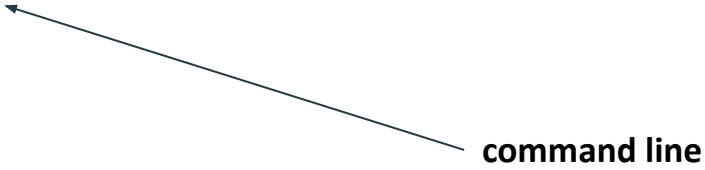
bwa index reference.fasta



Allineamento delle sequenze di lettura:

Per allineare le sequenze di lettura al genoma di riferimento, puoi utilizzare il comando **bwa mem**. Supponiamo di avere un file FASTQ contenente le sequenze di lettura chiamato **reads.fastq** e il genoma di riferimento indicizzato chiamato **reference.fasta**. Ecco come eseguire l'allineamento:

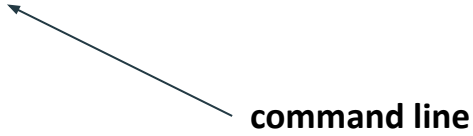
bwa mem reference.fasta reads.fastq > alignment.sam



Conversione in formato BAM:

Il file di output dell'allineamento è nel formato SAM (Sequence Alignment/Map). Puoi convertirlo in formato BAM (Binary Alignment/Map) utilizzando SAMtools, ad esempio:

samtools view -bS alignment.sam > alignment.bam



Formati per dati genomici

Dati su scala genomica (sequenze, annotazioni, variazioni, etc.) sono scambiati tramite files con specifici formati standard per facilitarne la diffusione e l'analisi bioinformatica.

Formati principali:

- FASTA
- FASTQ S
- SAM: allineamenti di reads al genoma
- BAM: formato binario di SAM
- VCF: varianti genomiche (SNPs, piccole delezioni/inserzioni)
- BED: mappatura di annotazioni sul genoma
- GFF/GTF: mappatura di geni

Formato SAM/BAM

- SAM – sequence alignment map
- BAM – binary alignment map Formato standard per conservare gli allineamenti (indipendente dal metodo utilizzato per allineare)

Il BAM è la versione binaria del SAM:

– dimensioni ridotte

– facilità di storage

– Accessibilità

– Non è un formato testuale → non è leggibile

Il formato SAM

Il formato SAM (Sequence Alignment/Map) è uno standard ampiamente adottato per rappresentare allineamenti di sequenze di reads di sequenziamento a genomi di riferimento. Esso permette di rappresentare in maniera efficace allineamenti complessi, includendo varie informazioni come la posizione di allineamento, la qualità dell'allineamento, e le sequenze stesse.

Il formato SAM

Composto da due parti:

– **Intestazione:** contiene informazioni sui campioni

– **Allineamenti:** contiene informazioni sulla posizione e sulla qualità di allineamento di tutte le reads

Il formato SAM

A

```
                    10        20        30        40
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGGCAT
```

B

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

QUAL (read quality; * meaning such information is not available)

Alignment section

QNAME (query template name, aka. read ID)

FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)

RNAME (reference sequence name, e.g. chromosome /transcript id)

POS (1-based position)

MAPQ (mapping quality)

CIGAR (summary of alignment, e.g. insertion, deletion)

RNEXT (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)

PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)

TLEN (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read

SEQ (read sequence)

Optional fields in the format of TAG:TYPE:VALUE

Il formato SAM

1. **Intestazione (Header)**: La sezione di intestazione del file SAM inizia con il carattere '@' e fornisce metadati sull'allineamento e sul genoma di riferimento. Ci sono diversi tipi di righe di intestazione, tra cui:

- `@HD`: Header. Fornisce informazioni sulla versione del formato SAM e altre informazioni globali.
- `@SQ`: Sequence. Descrive le sequenze del genoma di riferimento, fornendo informazioni come la lunghezza e il nome di ciascun cromosoma o contig.
- `@RG`: Read Group. Informazioni sul gruppo di reads.
- `@PG`: Program. Dettagli sul software o pipeline utilizzato per generare l'allineamento.
- Altre righe di intestazione possono includere `@CO` (commento), `@RP` (punto di rottura della ripetizione) e altre.

Il formato SAM

2. **Righe di Allineamento**: Ogni riga in questa sezione rappresenta l'allineamento di un singolo read o una coppia di reads (nel caso di sequenziamento paired-end). Ogni riga di allineamento ha vari campi, tra cui:
   - QNAME: Nome del read.
   - FLAG: Un valore intero che indica vari attributi dell'allineamento, come la direzione del read, se è appaiato, se è allineato, ecc.
   - RNAME: Nome del cromosoma o contig di riferimento.
   - POS: Posizione iniziale dell'allineamento sul riferimento.
   - MAPQ: Qualità dell'allineamento.
   - CIGAR: Rappresentazione compatta dell'allineamento, che mostra inserzioni, delezioni e altre varianti.
   - MRNM/RNEXT: Nome del compagno nel caso di sequenziamento paired-end.
   - MPOS/PNEXT: Posizione iniziale dell'allineamento del compagno.
   - ISIZE/TLEN: Lunghezza dell'inserto.
   - SEQ: Sequenza del read.
   - QUAL: Valori di qualità per ciascuna base del read.
   - Campi opzionali possono includere informazioni specifiche dell'allineamento o metadati addizionali.

CIGAR string

- Rappresentazione compatta delle caratteristiche di allineamento.

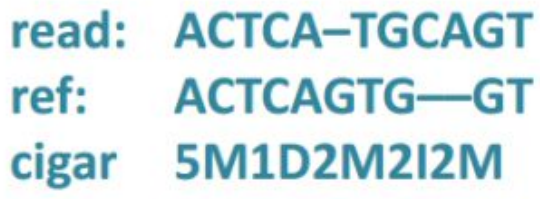
Include:

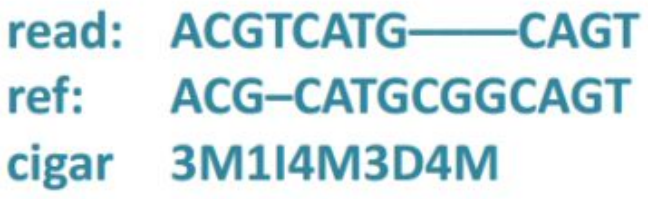
– M, Match, corrispondenza tra basi nella reference e nella read

– I, Insertion, inserzione di una o più basi nella read

– D, Deletion, delezione di una o più basi nella read

NB: i polimorfismi tra read e reference non sono indicati





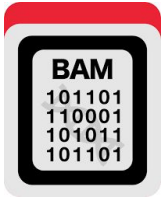
Il formato SAM

3. **Estensibilità**: Una delle forze del formato SAM è la sua estensibilità. Le righe di intestazione e i campi opzionali nelle righe di allineamento possono essere estesi per includere nuove informazioni o metadati specifici per particolari applicazioni o software.

4. **Formato BAM**: Per risparmiare spazio e migliorare l'efficienza, il formato SAM può essere compresso in un formato binario chiamato BAM. I file BAM conservano tutte le informazioni del file SAM ma sono molto più compatti e veloci da elaborare.

In sintesi, il formato SAM è uno standard fondamentale nella bioinformatica per rappresentare allineamenti di sequenze. Esso fornisce una rappresentazione dettagliata e flessibile degli allineamenti, permettendo agli ricercatori di analizzare e interpretare i dati di sequenziamento ad alto rendimento.

Il formato BAM

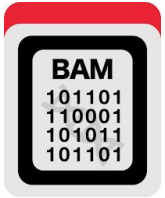


Il formato BAM (Binary Alignment/Map) è una rappresentazione binaria del formato SAM (Sequence Alignment/Map). Mentre il formato SAM è testuale e facilmente leggibile da un umano, il formato BAM è stato sviluppato per fornire una rappresentazione più compatta e efficiente per lo stoccaggio e l'analisi dei dati di allineamento.

Descrizione dettagliata del formato BAM:

1. **File Binario**: La caratteristica più evidente del formato BAM è che si tratta di un formato binario. Ciò significa che, a differenza dei file SAM, i file BAM non possono essere letti direttamente o visualizzati con semplici editor di testo. Invece, richiedono software specializzati per la visualizzazione e l'analisi.

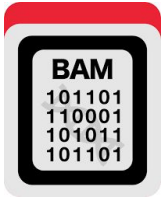
Il formato BAM



2. **Struttura e Contenuto**: Anche se BAM è in formato binario, contiene fondamentalmente le stesse informazioni di un file SAM. Ciò include l'intestazione (header) con metadati sull'allineamento e il genoma di riferimento, seguita dalle righe di allineamento che dettagliano come ciascun read si allinea al genoma di riferimento.

3. **Comprimibilità**: Una delle principali ragioni per utilizzare il formato BAM è la sua compressione. BAM utilizza una compressione a blocco che consente di ridurre notevolmente le dimensioni del file rispetto al corrispondente file SAM. Questa compressione è particolarmente vantaggiosa per progetti genomici di grandi dimensioni, in cui i dati di allineamento possono occupare una quantità significativa di spazio di archiviazione.

Il formato BAM

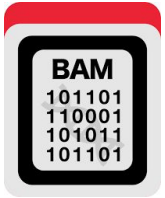


4. **Efficienza di Accesso**: Un altro vantaggio del formato BAM è l'efficienza con cui si possono recuperare specifiche regioni genomiche. Utilizzando un indice associato (tipicamente un file `.bai` o `.csi`), è possibile accedere rapidamente a regioni specifiche del genoma senza dover scorrere l'intero file.

5. **Software e Tool**: Vari strumenti e librerie, come Samtools e htslib, sono stati sviluppati per lavorare con file BAM. Questi strumenti permettono di visualizzare, indicizzare, ordinare e manipolare i file BAM con grande efficienza.

6. **Compatibilità**: Essendo una rappresentazione binaria del formato SAM, è possibile convertire facilmente tra SAM e BAM. Questa compatibilità garantisce che gli utenti possano passare tra una rappresentazione leggibile e una efficiente in base alle loro esigenze.

Il formato BAM



7. **Utilizzo Comune**: Il formato BAM è diventato uno standard de facto in bioinformatica, specialmente in progetti che coinvolgono allineamenti su larga scala come lo studio del genoma umano. Molti visualizzatori di genomi e piattaforme di analisi integrano la compatibilità con BAM proprio a causa della sua efficienza e della sua diffusione.

In sintesi, il formato BAM offre una soluzione efficiente e compatta per rappresentare allineamenti di sequenze genomiche. Nonostante la sua natura binaria, mantiene tutte le informazioni essenziali del formato SAM, garantendo allo stesso tempo velocità e flessibilità per analisi su larga scala.