

Dry Lab



Davide La Rocca,
UOSD Ricerca e controllo degli organismi geneticamente modificati
Valeria Russini,
UOC Microbiologia degli Alimenti



Dry Lab

Scaletta

Si parlerà di:

1. Unix, un po' di storia
2. GNU-Linux , concetti introduttivi
3. Esempio di analisi di amplicon sequencing



Unix

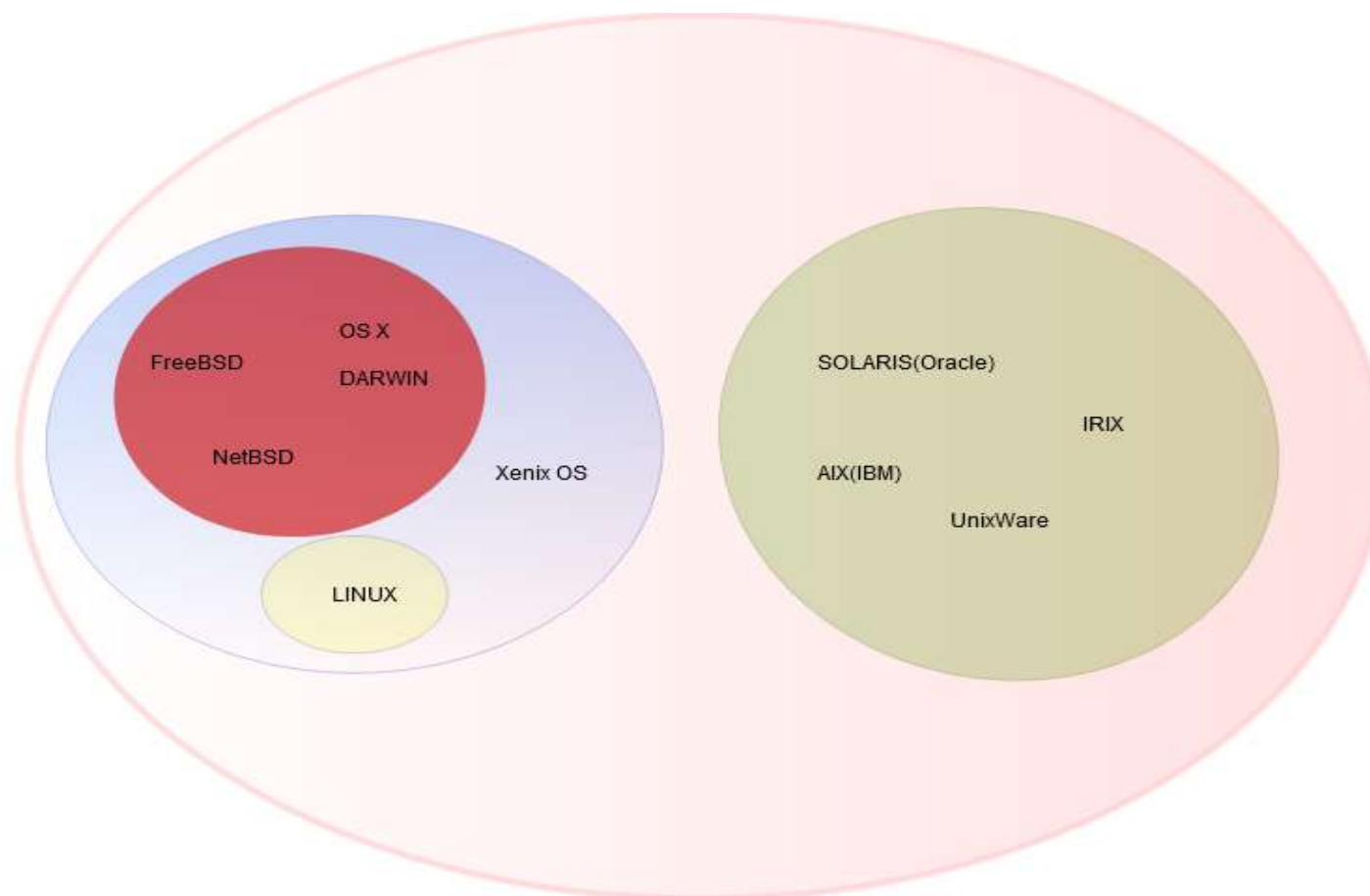
Introduzione e concetti più importanti

Unix è stato progettato nei Bell Laboratories (AT&T Corp.). Il primo sistema operativo che può definirsi a tutti gli effetti come "Unix" fu sviluppato Ken Thompson nel 1969 per poter eseguire un programma chiamato "Space Travel" che simulava i movimenti del sole e dei pianeti, così come il movimento di una navicella spaziale che poteva atterrare in diversi luoghi



Unix

Introduzione e concetti più importanti



GNU

Introduzione e concetti più importanti

GNU appartiene alla famiglia UNIX ed ideato nel 1984 da Richard Stallman ed è promosso dalla Free Software Foundation allo scopo di ottenere un sistema operativo completo utilizzando esclusivamente software libero.



«Il mio lavoro sul software libero è motivato da un obiettivo idealistico: diffondere libertà e cooperazione. Voglio incoraggiare la diffusione del software libero, rimpiazzando i programmi proprietari che proibiscono la cooperazione, e quindi rendere la nostra società migliore. Questa è la ragione fondamentale per cui la GNU General Public License è stata scritta così com'è - come copyleft»

(Richard M. Stallman)



GNU Linux

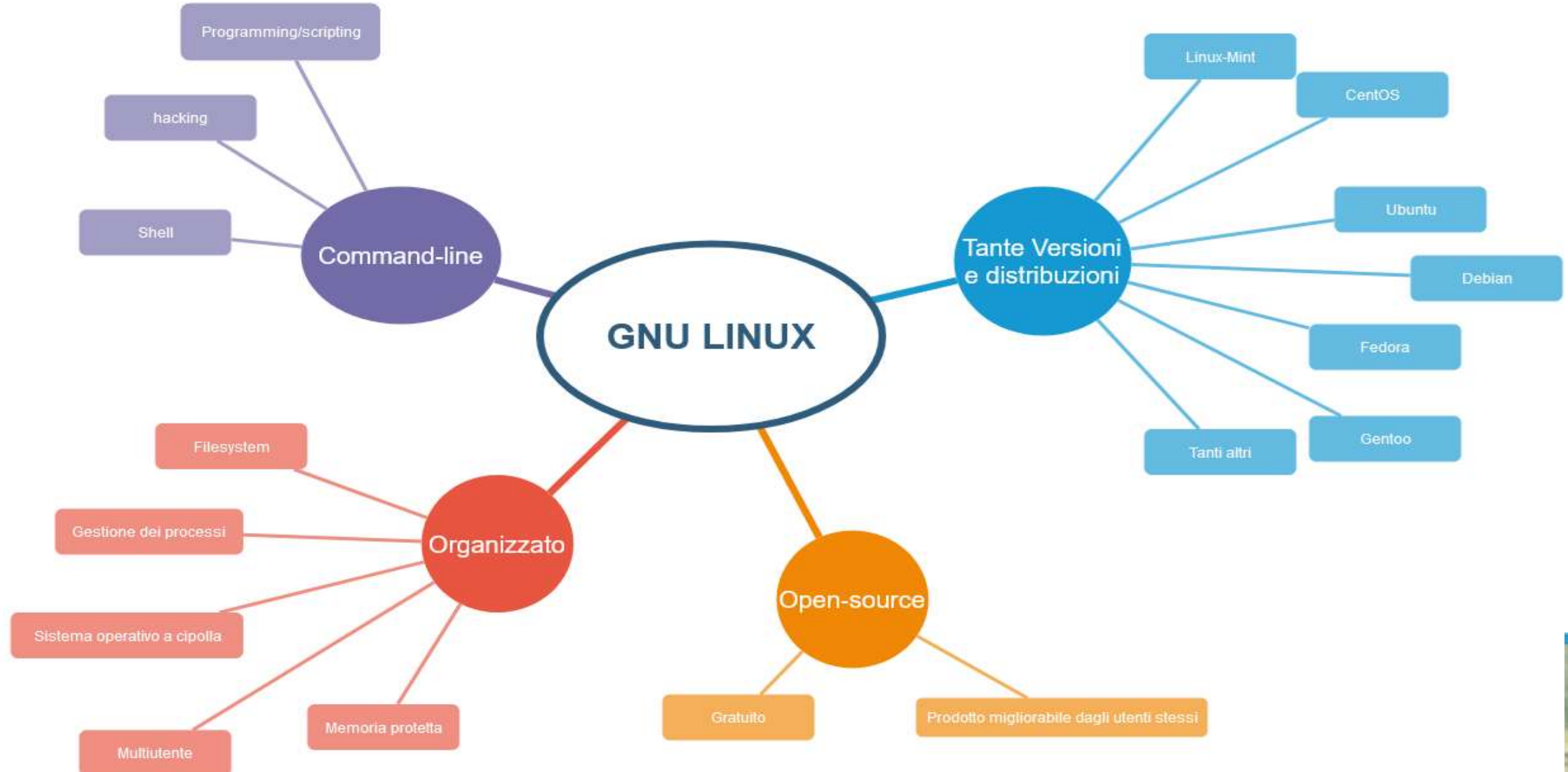
Introduzione e concetti più importanti

Quello che ancora mancava era il kernel. Nel 1990, i membri del progetto GNU cominciarono lo sviluppo di un kernel chiamato GNU hurd, che deve ancora raggiungere il livello di maturità richiesto per l'uso diffuso. Nel 1991, Linus Torvalds, uno studente finlandese, usò gli strumenti di sviluppo GNU per produrre il kernel Linux.



GNU-Linux

Introduzione e concetti più importanti



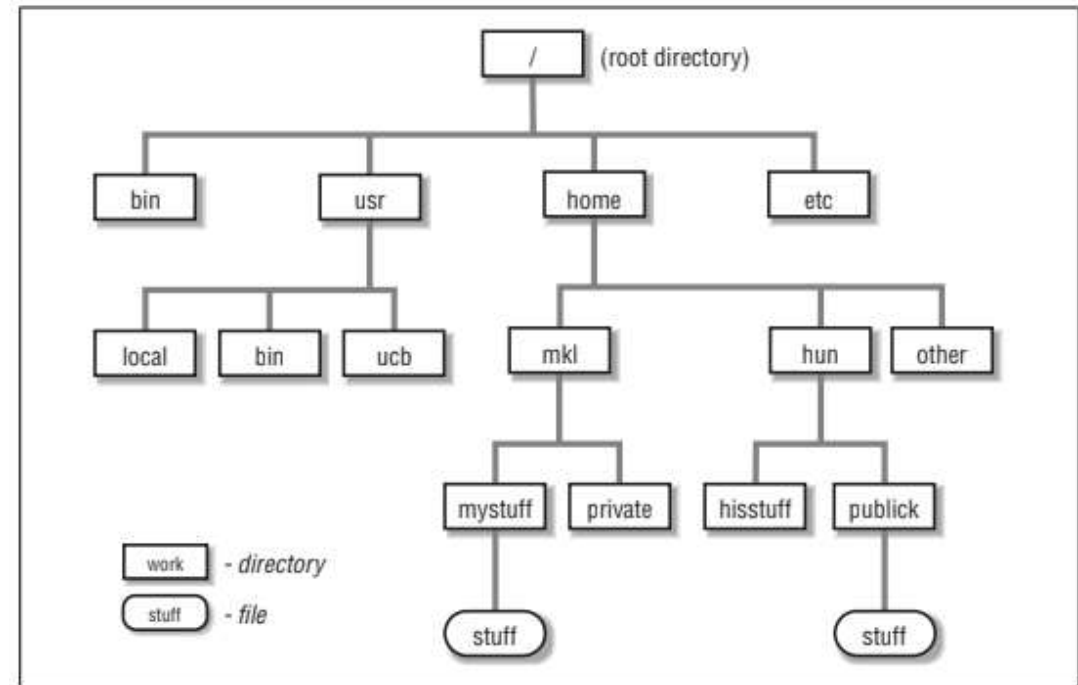
GNU Linux

Introduzione e concetti più importanti

Cos'è filesystem?

In informatica è il meccanismo con cui i file sono posizionati e organizzati sui dispositivi informatici per l'archiviazione dei dati.

Linux come altri sistemi operativi possiede una struttura ad albero o gerarchico.



GNU Linux

Introduzione e concetti più importanti

Una shell è un interprete dei comandi. Il suo principale compito è **interpretare i comandi che digita l'utente** ed eseguire i programmi specificare nelle righe di comando. Per impostazione predefinita, la shell legge i comandi dalla tua tastiera e fa in modo che altri programmi possano scrivere i loro risultati lì. **La shell protegge Unix/GNU Linux dall'utente**

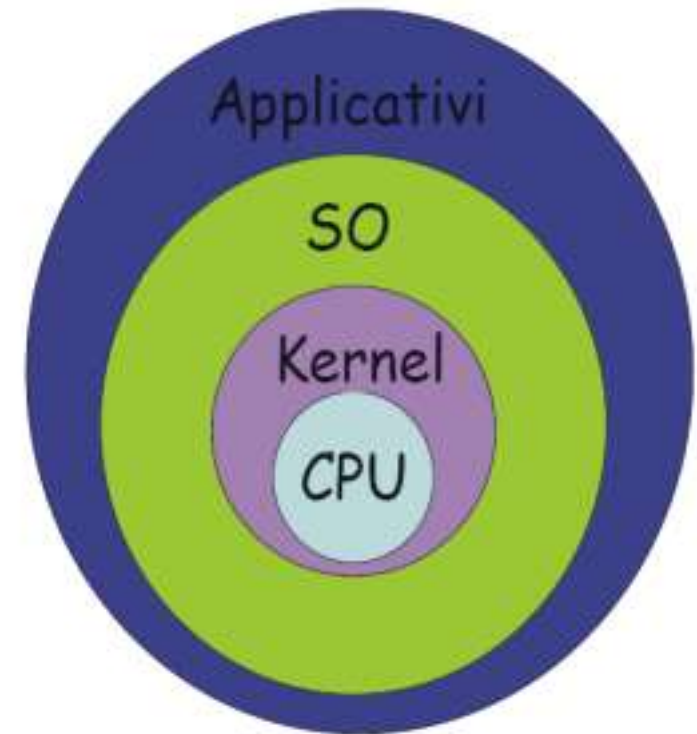
```
crogm@crogm-VirtualBox: ~  
File Modifica Visualizza Cerca Terminale Aiuto  
crogm@crogm-VirtualBox:~$ for i in {1..10};do echo "CIAO AL PARTECIPANTE NUM$i"  
; done; echo "SONO CONTENTO DI VEDERVI, QUI NON CI SONO VESPE";  
CIAO AL PARTECIPANTE NUM1  
CIAO AL PARTECIPANTE NUM2  
CIAO AL PARTECIPANTE NUM3  
CIAO AL PARTECIPANTE NUM4  
CIAO AL PARTECIPANTE NUM5  
CIAO AL PARTECIPANTE NUM6  
CIAO AL PARTECIPANTE NUM7  
CIAO AL PARTECIPANTE NUM8  
CIAO AL PARTECIPANTE NUM9  
CIAO AL PARTECIPANTE NUM10  
SONO CONTENTO DI VEDERVI, QUI NON CI SONO VESPE  
crogm@crogm-VirtualBox:~$
```



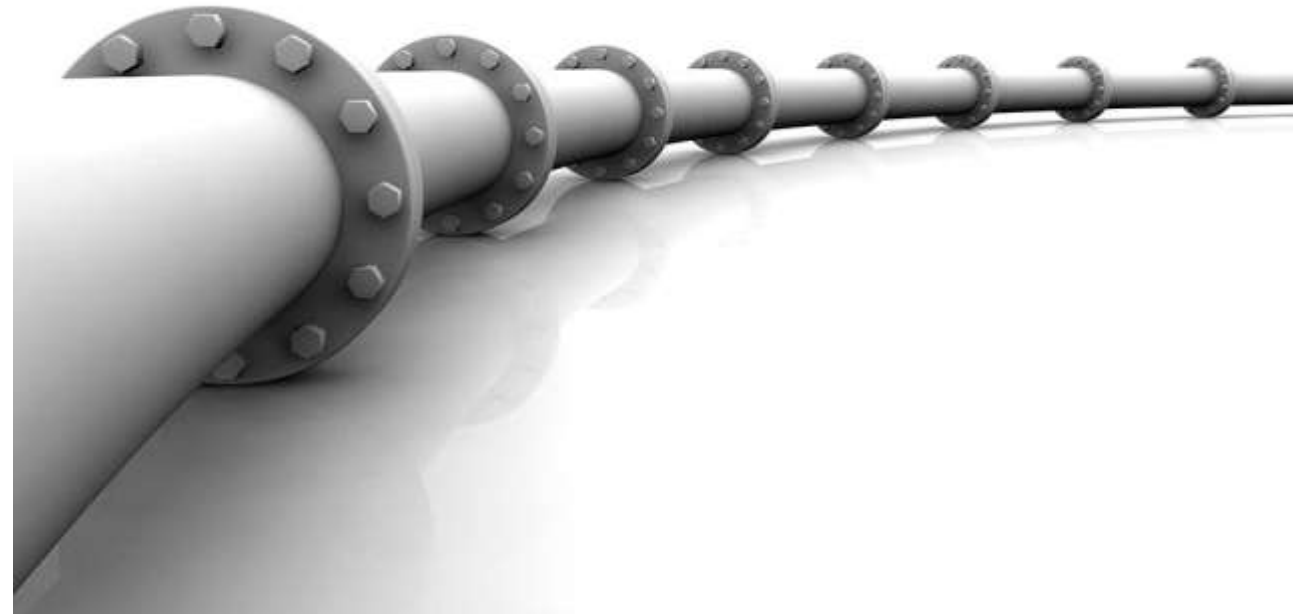
GNU Linux

Introduzione e concetti più importanti

Il kernel (monolitico) è il cuore del sistema operativo Unix/GNU Linux stesso. Il kernel assegna memoria a ciascuno dei programmi in esecuzione, partiziona il tempo in modo equo in modo che ogni programma possa svolgere il proprio lavoro, gestisce tutte le operazioni di I / O (input / output).



Come posso collegare tra loro
tutti i programmi che fanno
parte di un determinato
workflow? ? ?



Unix/GNU Linux

Che cos'è una pipeline?

- Questo termine (in inglese *tubatura* — composta da più elementi collegati — o *condotto*) viene utilizzato per indicare un insieme di componenti software collegati tra loro in cascata, in modo che il risultato prodotto da uno degli elementi (output) sia l'ingresso di quello immediatamente successivo (input).
- L'accezione più comune della parola *pipeline* indica un comando di shell composito, in cui un programma *sorgente* genera un flusso di dati testuali che si propagano attraverso le pipe ("|") tramite una sequenza di filtri, fino ai *destinatari* (spesso file o terminale). Questi programmi sono collegati tra loro tramite l'operatore *pipe*, che in una riga di comando significa che lo standard output del programma a sinistra dell'operatore va passato allo standard input del programma alla sua destra.

```
echo $(wc -l $1 | grep -o '[0-9]\+' | sed 's/.$//') >> $3
```



Esempio Amplicon Sequencing

[illegible]

Analisi dei dati di sequenziamento

Cos'è il formato file fastq?

- Un tipo di file che contiene le reads di sequenziamento
- Ogni read è descritta da 4 righe

1. **Header**
2. **Prodotto di sequenziamento (la nostra sequenza)**
3. **Il carattere '+'**
4. **Una serie di caratteri che descrivono il Phred score**

@M03865:13:000000000-J3DTD:1:1104:22979:3134

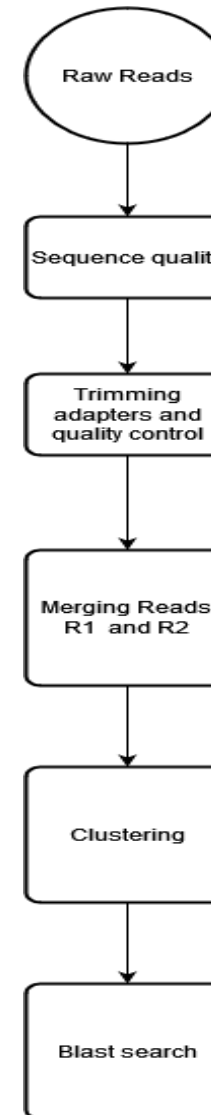
GTATGTGAGTACACATGTTGACAAACCGTTCCTGTGCCCCATGAAGACGATTTGCTCACAGTGGCTTGGAAACGGCGCTT
TAACTTTTGATTGCAAAAAGCCGTCTGTTCCGGTTGCAGTGCCGCTTTGGCCCTTGGCATACCCACTTACCATTGT
ACCGTAA

+

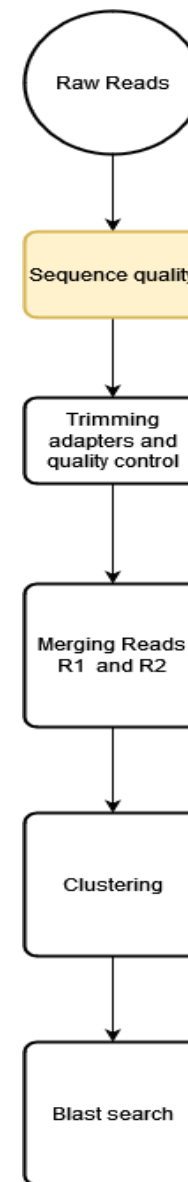
C-B@CCG9CFGFGGFGGEGGFFFGF<D:CFG@CEE@FFEFGFFA<9E@FCCCE9F,@F<FCFG?EEC7:DCECFGG+8:
DFE9E9E<EFD<FDGFGF<=<CFFGG7FF,C=EGGC@FDA=FGFGGGGF,EDF<84?A5EE@E8FF8?F<,,??AFFDC@
D?BCF<:



Esempio Analisi dei dati di sequenziamento

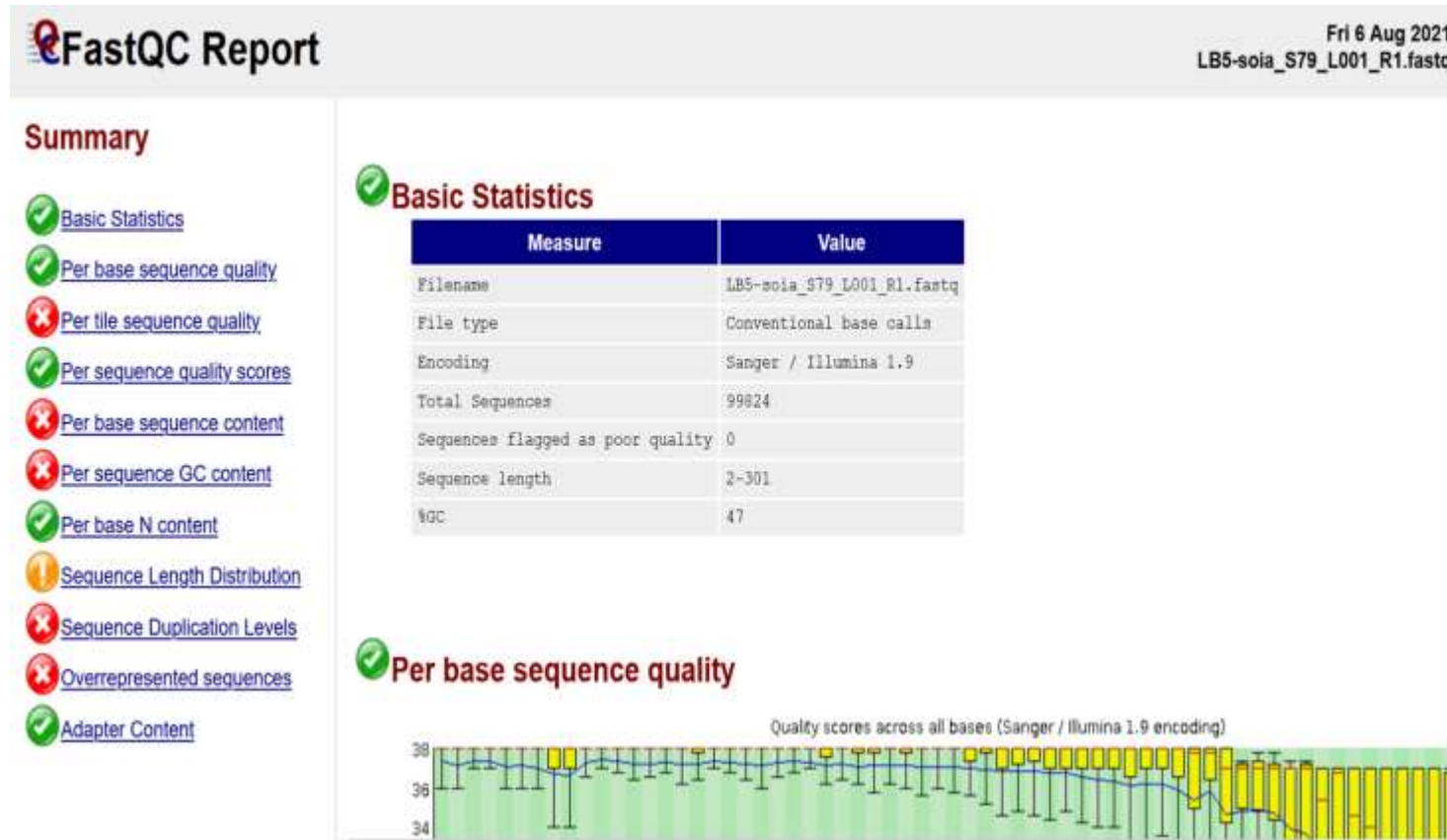


Qualità di sequenza

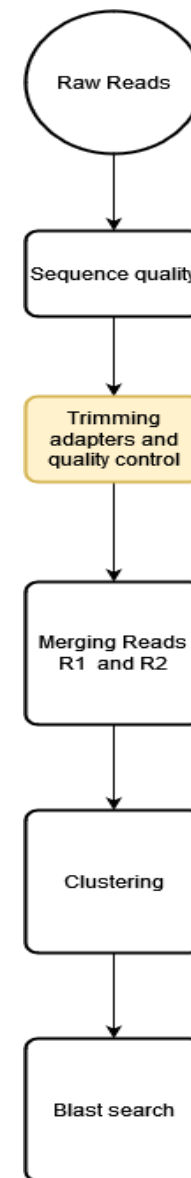


Analisi dei dati di sequenziamento:

Qualità di sequenza



Trimming adattatori e quality check



Analisi dei dati di sequenziamento: Trimming adattatori e quality check

Esempio :Adattatore Nextera
AGATGTGTATAAGAGACAG

```
egm@bioinfogn: ~/cartella_lavoro/amplicon_project
egm@bioinfogn:~/cartella_lavoro/amplicon_project$ zcat /RawData/OGM_2_2021/Pool-4a_S5_L001_R1_001.fastq.gz | grep -B 1 "AGATGTGTATAAGAGACAG"
@M03865:43:000000000-D9NV9:1:1101:8435:9796 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGCTCGGCTGCGTCAGATGTGTATAAGAGACAGTTTTTATTCGGTTTTTCGCTATCGACGATGATGGTCCCTTTTGTTCATTCTCA
--
@M03865:43:000000000-D9NV9:1:1101:25288:11053 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTCGTCGGCAGCGTAGATGTGTATAAGAGACAGGGGATGACGTTAATTGGCTCTGAGCTTCGTCCTCTTAAGTTCATGCTTCTGTTTCCACGGCGTGATGCTTACGGTCAAGCAGCCGTCAGCAACTGCTCGTAAGTCTCTGGTCTTTCTGGAACCGTCCGTAATCCAGGTGACAAGTCTATCTCCACAGGTGCTTCATGTT
--
@M03865:43:000000000-D9NV9:1:1101:4616:12503 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGTCAGCAACAGATGTGTATAAGAGACAGTTATTATTGGTTTTGGCCTTTTGATCGTTCTCA
--
@M03865:43:000000000-D9NV9:1:1101:7118:15809 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGTCAGCGTCAGATGTGTATAAGAGACAGTTTTTATTCGGTTTTTCGCTATCGATATGGTCCCTTTTGTTCATTCTCA
--
@M03865:43:000000000-D9NV9:1:1101:28233:16347 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTATTCGGTTTTTCGCTATTGTTTCTTCACGGTCTGTATTACACATATCCGATCCACGAGACGGCCTCTATCTCGTATGCCGCTTCTGCTTGAAAAAAAAAAAAACAACAAAGGACTCCTATCTCGTATGCCGCTTATGCTTGAAAAAAAAAAAAATAAAACTTAATGACATT
GCAATATCGTATA
--
@M03865:43:000000000-D9NV9:1:1101:24501:16519 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTATTCGGTTTTTCGCGATCAGGAGACGGCCTCTTGTCTTATGACTGCCCTTATTACATCTCCGAGCCACGAGACGGACTCCTATCTCGTATGCCGCTTCTGCTTGAAAAAAAAAAAAACAACAAAGATAGCAAAAAAAAAATATTGAATAAAGTGAATCTAATGAAAG
AAAAACCTTTTAG
--
@M03865:43:000000000-D9NV9:1:1101:13145:19246 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTATTCGGTTTTTCGCTATCGACGATGCGGCTTTGTTTATTCTCACTTTTCTTATACAAATCTCCGAGCCACGAGACGGACTCCTATCTCGTATGCCGCTTCTGCTTGAAAAAAAAAAAAACTACAGTCTTATACAACTACACTTCCCAACACAAACACCCACTCCTACTCACTC
TCAAACTCTCGT
--
@M03865:43:000000000-D9NV9:1:1101:19724:19637 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGTCAGCGTCAGATGTGTATAAGAGACAGTTTTTATTCGGTTTTTCGCTATCTGTCTGTTACACGGTCTGATTACGCATCGTAATGCCACGTGTCGCTCTTTGTTCTCTTCC
--
@M03865:43:000000000-D9NV9:1:1101:22169:21321 1:N:0:GGACTCCT+CTCTCTAT
TTTTTATTCGGTTTTTCGCTATCGTCGGCAGCATCAGATGTGTATAAGAGACAGTTTTTATTCGGACATCTACCGCGCACGAGTCGGCTTCTATCCGTATCCCTGCTTCTTATTACATCTCCGAGCCACGAGACGGACTCCTATCTCGTATGCCGCTTCTGCTTGAAAAAAAAAACTCCACCTCTACTATCAAACCTCCCACTCTATCTACCCCTCTCACTCAATCCCT
CTAATCTCCTT
```



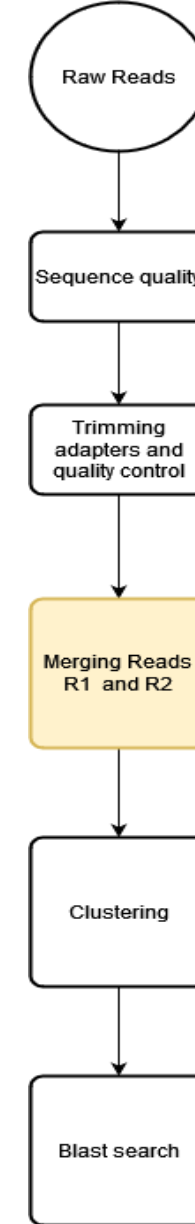
Analisi dei dati di sequenziamento: Trimming adattatori e quality check

Confronto prima e dopo trimming qualità:

```
ogm@bioinfogm:~/iaetella_labseq/results_cab/trim/trimm$ zcat /RawData/OGM_2_2021/Pool-4a_S5_L001_R1_001.fastq.gz | grep -A 1 "@M03865:43:000000000-D9NV9:1:1101:14562:2235 1:N:0:"
@M03865:43:000000000-D9NV9:1:1101:14562:2235 1:N:0:GGACTCCT+CTCTCTAT
CCACGGCGTGCATGCTTCACGGTGCAAGCAGCCGTCAGCAACTGCTCGTAAGTCCTCTGGTCTGTCTCTATACACATCTCCGAGCCACGAGACGGACTCCTATCTCGTATGCCGCTTCTGCTTGAAAAAAAATTCTCTTTTAACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATTATTTTTTTTTTTTATTTTTTTTATTTTATCTTTTTTTTTTTTTTC
TTTTTTTTTTTTTT
ogm@bioinfogm:~/iaetella_labseq/results_cab/trim/trimm$ grep -A 1 "@M03865:43:000000000-D9NV9:1:1101:14562:2235 1:N:0:" Pool-4a_S5_L001_1P
@M03865:43:000000000-D9NV9:1:1101:14562:2235 1:N:0:GGACTCCT+CTCTCTAT
CCACGGCGTGCATGCTTCACGGTGCAAGCAGCCGTCAGCAACTGCTCGTAAGTCCTCTGG
ogm@bioinfogm:~/iaetella_labseq/results_cab/trim/trimm$ _
```

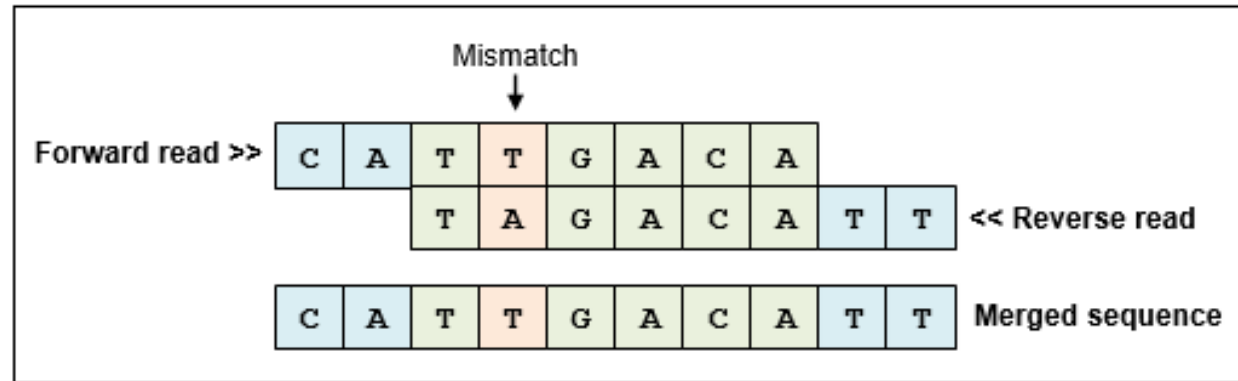


Merging Reads R1 R2



Analisi dei dati di sequenziamento:

Merging Reads R1 R2

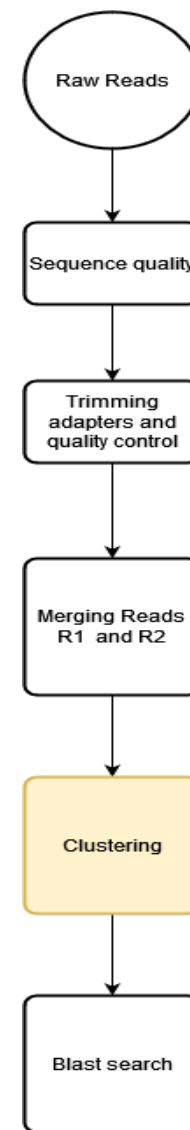


```

ogm@bioinfogm:~/lavoro/progetti_cad/trimmomatic$ grep -A 1 "@M03865:43:000000000-D9NV9:1:1101:15877:2286" Pool-4a_S5_L001_1P
@M03865:43:000000000-D9NV9:1:1101:15877:2286 1:N:0:GGACTCCT+CTCTCTAT
GGGATGACGTTAATTGGCTCTGAGCTTCGTCTCTTAAGGTCATGCTTCTGTTTCCACGGCGTGTCATGCTTCACGGTGCAAGCAGCCGTCCAGCAACTGCTCGTAAGTCCTCTGGTCTTTCTGGAACCGTCCGTATTCCAGGTGACAAGTCTATCTCCACCGGTCCTTCATGTT
ogm@bioinfogm:~/lavoro/progetti_cad/trimmomatic$ grep -A 1 "@M03865:43:000000000-D9NV9:1:1101:15877:2286" Pool-4a_S5_L001_2P
@M03865:43:000000000-D9NV9:1:1101:15877:2286 2:N:0:GGACTCCT+CTCTCTAT
ACATGAAGGACCGGTGGGAGATAGACTTGTACCTGGAATACGGACGGTTCAGAAAGACCAGAGGACTTACGAGCAGTTGCTGGACGGCTGCTTGACCCGTGAAGCATGCACGCCGTGGAACAGAGAAGACATGACCTTAAGAGGACGAAGCTCAGAGCCAATTAACGTCATCCC
ogm@bioinfogm:~/lavoro/progetti_cad/trimmomatic$ echo "GGATGACGTTAATTGGCTCTGAGCTTCGTCTCTTAAGGTCATGCTTCTGTTTCCACGGCGTGTCATGCTTCACGGTGCAAGCAGCCGTCCAGCAACTGCTCGTAAGTCCTCTGGTCTTTCTGGAACCGTCCGTATTCCAGGTGACAAGTCTATCTCCACCGGTCCTTCATGTT" | tr "ATCG" "TAGC" | rev
AACATGAAGGACCGGTGGGAGATAGACTTGTACCTGGAATACGGACGGTTCAGAAAGACCAGAGGACTTACGAGCAGTTGCTGGACGGCTGCTTGACCCGTGAAGCATGCACGCCGTGGAACAGAGAAGACATGACCTTAAGAGGACGAAGCTCAGAGCCAATTAACGTCATCCC
ogm@bioinfogm:~/lavoro/progetti_cad/trimmomatic$
  
```



Clustering



Analisi dei dati di sequenziamento: Clustering

E' una tecnica **descrittiva** del processo di Data Analytics (estrazione di informazioni utili da dati); fa parte del cosiddetto "*unsupervised learning*", ovvero , ci sono dati input ma nessun output risultante(a differenza del "*supervised learning*"), (dati non etichettati), perciò impariamo le etichette "latenti".

Esistono diversi tipi di approcci per il clustering:

- k means
- clustering gerarchico
- analisi relazionale
- ecc..



Analisi dei dati di sequenziamento: Clustering

In questo esempio è stato utilizzato un programma che usa un clustering *centroid based*(che riprende un algoritmo molto simile ai k means , ovvero l'algoritmo greedy) basato sulla similarità di sequenza (alle sequenze di input si possono applicare diversi pre-sorting, tra cui quelli sulle lunghezze, sulle abbondanze o sull'ordine definito dall'utente)

```
ogm@bioinfogm:~/cartella_lavoro/results_CA1/modifiche_Alessandro/ultimi_risultati/multiple_aln$ awk -F"\t" '{if($1 != "C") print $0}' cotone_pr_r1_clus_65.uc | head -n 10
S      0      44      *      *      *      *      *      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795      *
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:12833:17567;size=427      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:11152:17954;size=400      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:11106:6784;size=392      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:12917:9382;size=363      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      45      100.0    +      0      0      D44M     M05888:24:000000000-C94BD:1:1101:11383:18887;size=335      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:10793:10499;size=310      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:10544:16944;size=300      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      48      100.0    +      0      0      4D44M    M05888:24:000000000-C94BD:1:1101:11128:9314;size=288      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
H      0      47      100.0    +      0      0      3D44M    M05888:24:000000000-C94BD:1:1101:24417:4486;size=254      M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
ogm@bioinfogm:~/cartella_lavoro/results_CA1/modifiche_Alessandro/ultimi_risultati/multiple_aln$
```



Analisi dei dati di sequenziamento: Clustering

centroide

>M05888:24:000000000-C94BD:1:1101:10124:10793;size=795
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTCTCAATCGGACCATCACATCAATCCACTTGCTTTGAAGACGTGGTTGGAACGTCTTCTTTTCCACGATGCTCCTCGTGGGTGGGGTCCATCTTTGGGACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:12833:17567;size=427
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTCTCAATCGGACCATCACATCAATCCACTTGCTTTGAAGACGTGGTTGGAACGTCTTCTTTTCCACGATGCTCCTCGTGGGTGGGGTCCATCTTTGGGACCAGTGTGGCAGAGGCATCTACCAGCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:11152:17954;size=400
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTCTCAATACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:11106:6784;size=392
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:12917:9382;size=363
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:11383:18887;size=335
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:10793:10499;size=310
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTCTCAATCGGACCATCACATCAATCCACTTGCTTTGAAGACGTGGTTGGAACGTCTTCTTTTCCACGATGCTCCTCGTGGGTACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:10544:16944;size=300
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTCTCAATCGGACCATCACATCAATCCACTTGCTTTGAAGACGTGACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

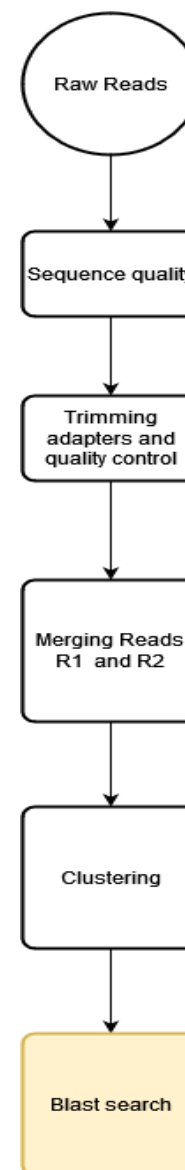
>M05888:24:000000000-C94BD:1:1101:11128:9314;size=288
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGATATTACCCTTTGTTGAAAAGTCTCAATCGGACCATCACATCAATCCACTTGCTTTGAAGACGTGGTTGGAACGTCTTCTTTTCCAACCAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

>M05888:24:000000000-C94BD:1:1101:24417:4486;size=254
GCTGGGCAATGGAATCCGAGGAGGTTTCCGGACAGCCCGGGCCGTCGACCACGCGTGCCCTATAGTGAGTCGTATTAC

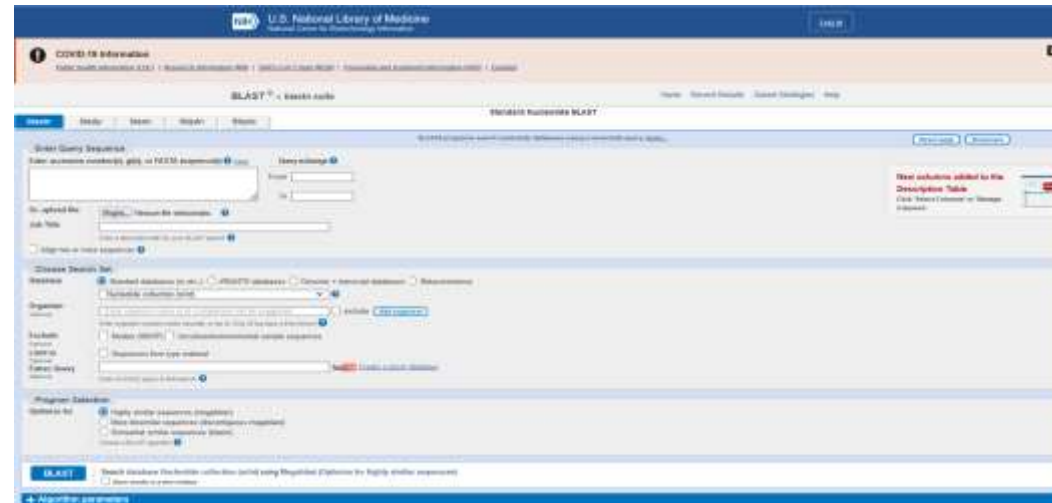
sequenze del
cluster



Blast



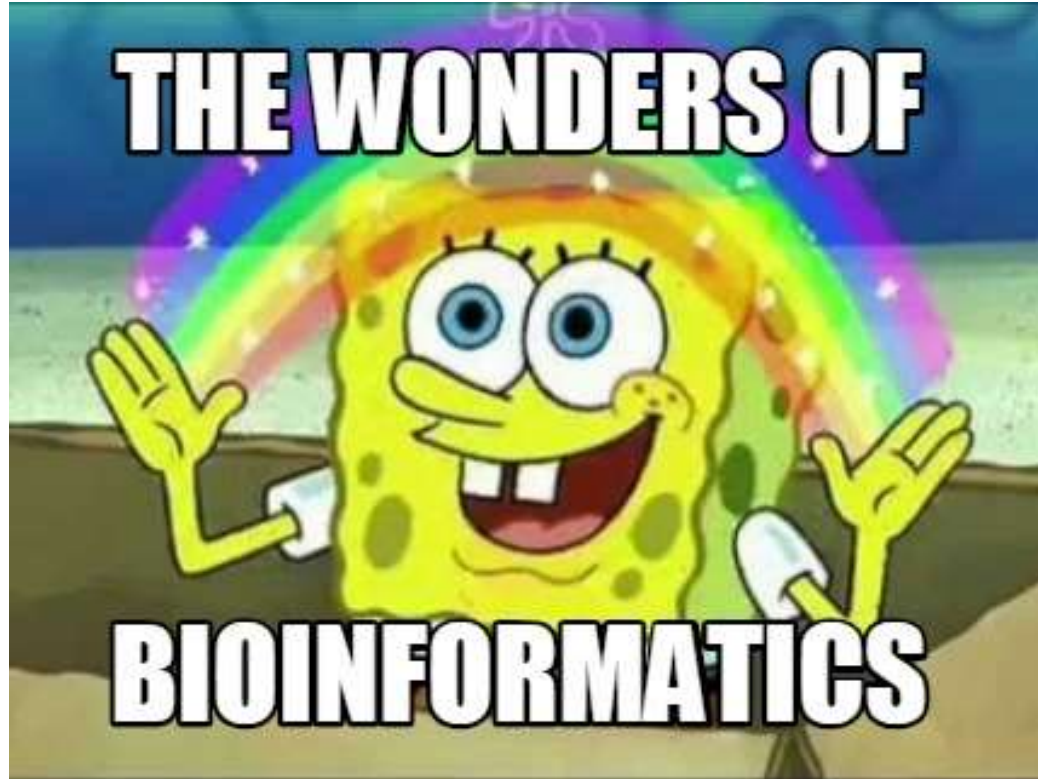
Analisi dei dati di sequenziamento: Blast search



```
ogm@bioinform: $ blastn -query LB5-cotone_S81_L001_r2_centri.fasta -db nt -outfmt '6 qaccver saccver pident length mismatch gapopen qstart qend qlen qseq sstart send eval evalue bitscore' -out output.csv
```



Grazie per l'attenzione





Istituto Zooprofilattico Sperimentale
del Lazio e della Toscana *M. Aleandri*

Next Generation Sequencing (NGS) e bioinformatica: applicazioni nel campo della sicurezza alimentare 26-27 ottobre 2021

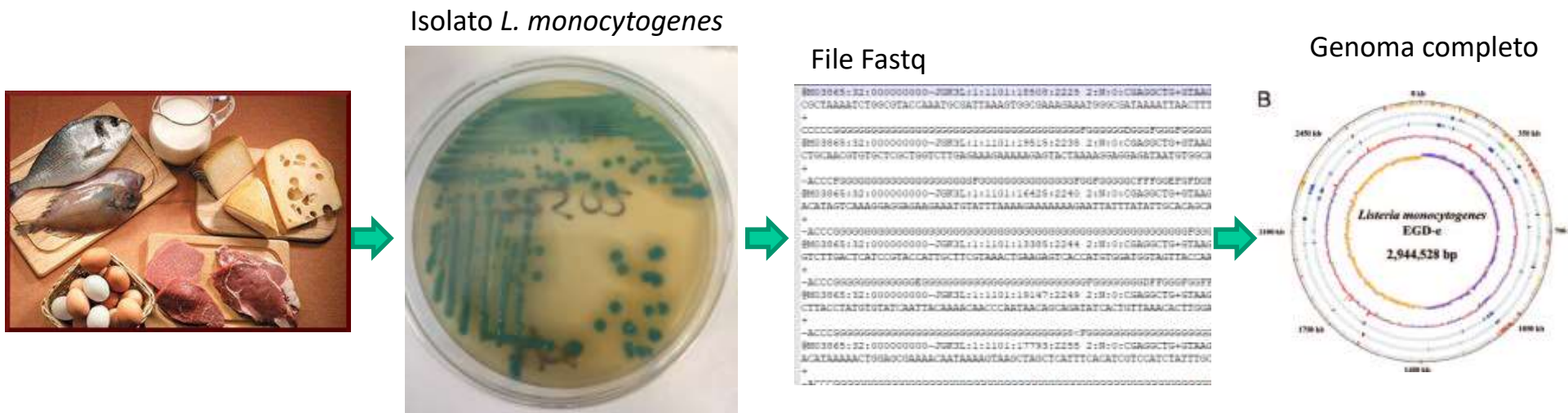
Analisi bioinformatiche per Whole Genome
Sequencing (WGS)



Che cos'è il WGS

Il Whole Genome Sequencing è il metodo utilizzato per l'analisi dell'intero DNA genomico di un organismo.

Si parte da **isolato** batterico patogeno da **alimento** o da matrice di origine **umana**



Consente la caratterizzazione dell'intero genoma



Istituto Zooprofilattico Sperimentale
del Lazio e della Toscana M. Aleandri

Applicazioni del WGS

Diagnosi:

Identificazione del batterio a livello di ceppo

Riconoscimento geni antibiotico resistenza e virulenza *in silico*

Sorveglianza:

Genotipizzazione *in silico* (MLST, cgMLST)

Quali genotipi/ceppi sono presenti nel territorio



Epidemiologia molecolare:

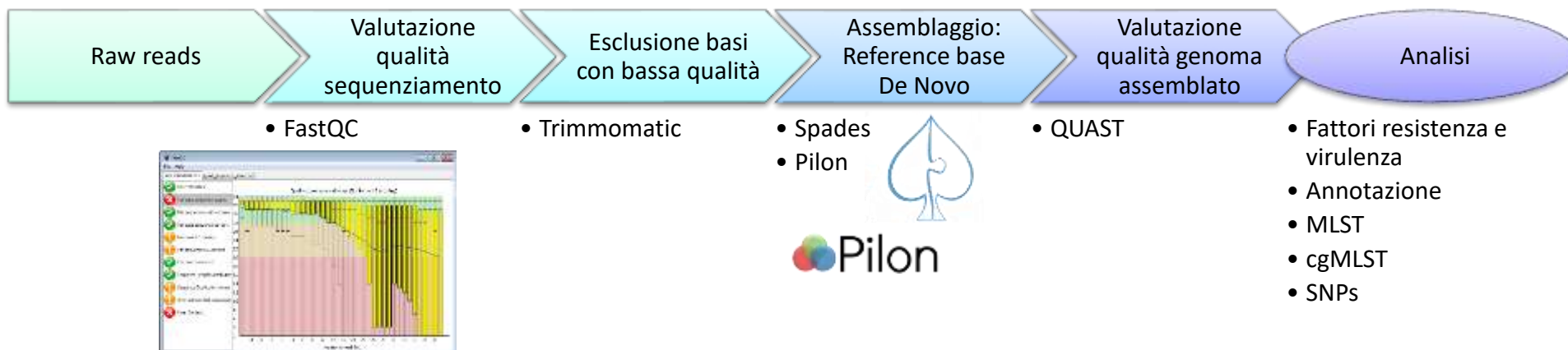
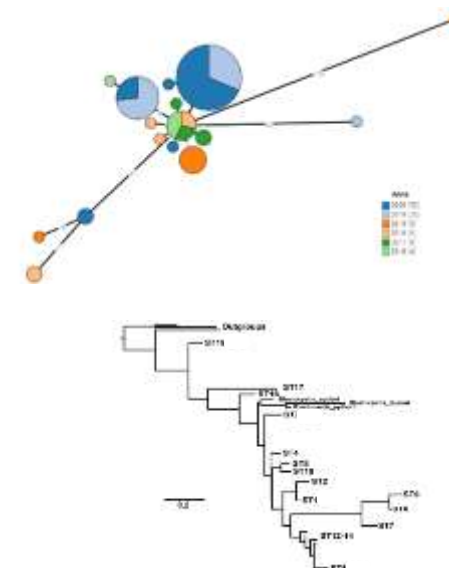
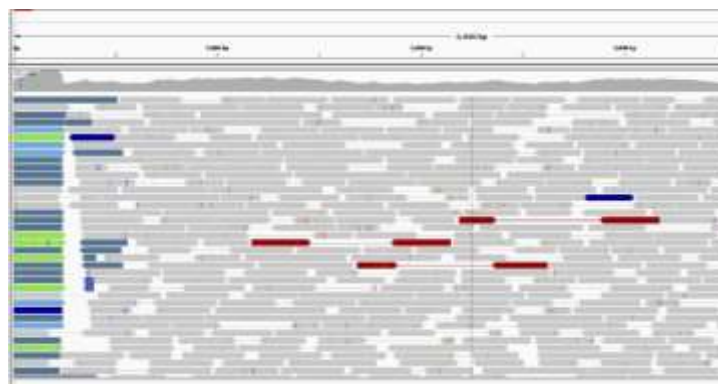
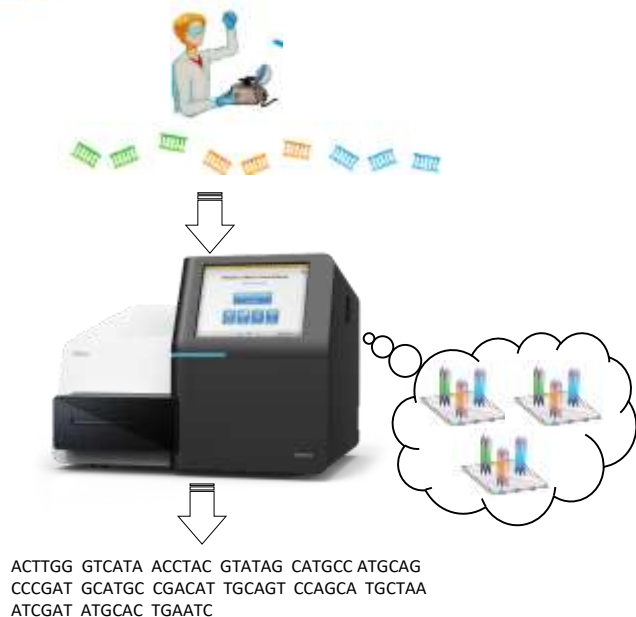
Riconoscimento tempestivo di focolai

Individuazione dell'origine della contaminazione (relazioni filogenetiche, SNPs)

Conferma ipotesi e fornisce prove solide



Pipeline per l'analisi di ogni campione





Qualità reads prodotte

Quality	Chance it's wrong	Accuracy	Description
10	1 in 10	90%	Bad
20	1 in 100	99%	Maybe
30	1 in 1000	99.9%	OK
40	1 in 10,000	99.99%	Very good
50	1 in 100,000	99.999%	Excellent

$$Q = -10 \log_{10} P \quad \Leftrightarrow \quad P = 10^{-Q/10}$$

Q = Phred quality score P = probability of base being incorrect

[illegible]

Symbol	Q-Score
=	28
>	29
?	30
@	31
A	32
B	33
C	34
D	35
E	36
F	37
G	38
H	39
I	40

Probabilità di errore di chiamata di una base

Scala logaritmica

Idealmente $Q > 30$

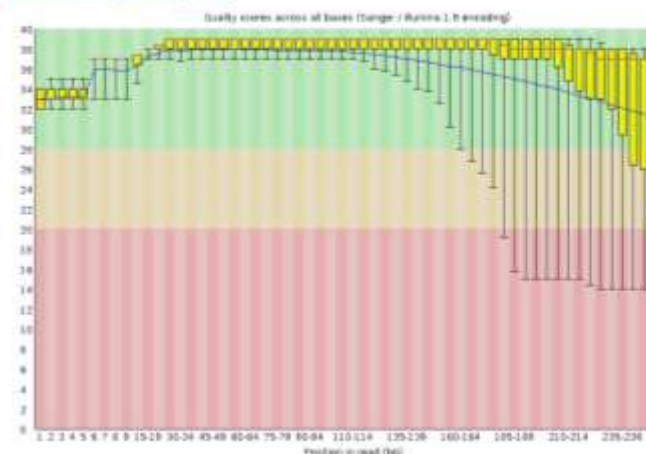


Qualità reads prodotte: FASTQC

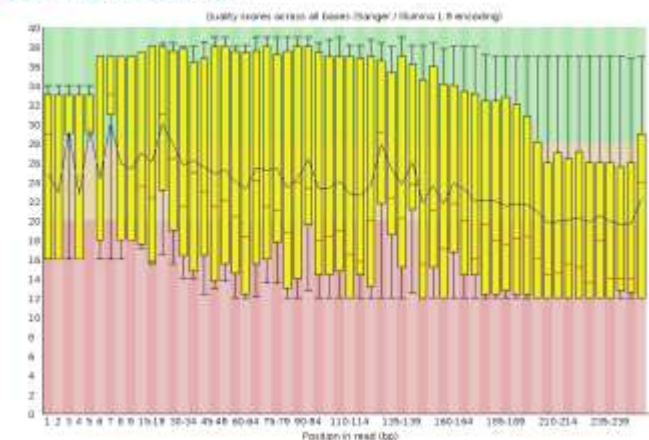


- La linea rossa centrale è il valore mediano
- Il riquadro giallo rappresenta l'intervallo interquartile (25-75%)
- I baffi superiore e inferiore rappresentano i punti 10% e 90%
- La linea blu rappresenta la qualità media

✓ Per base sequence quality

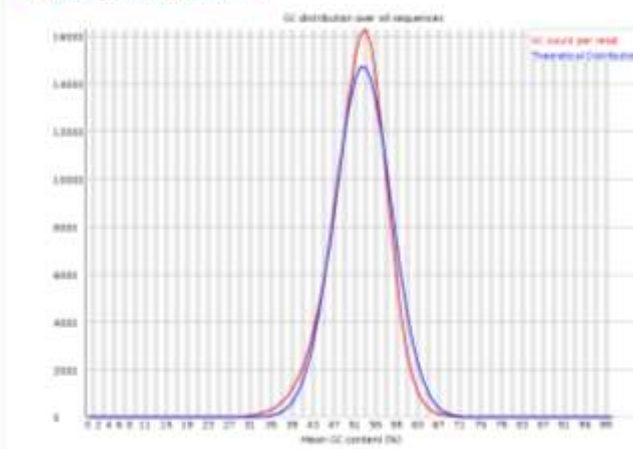


✗ Per base sequence quality

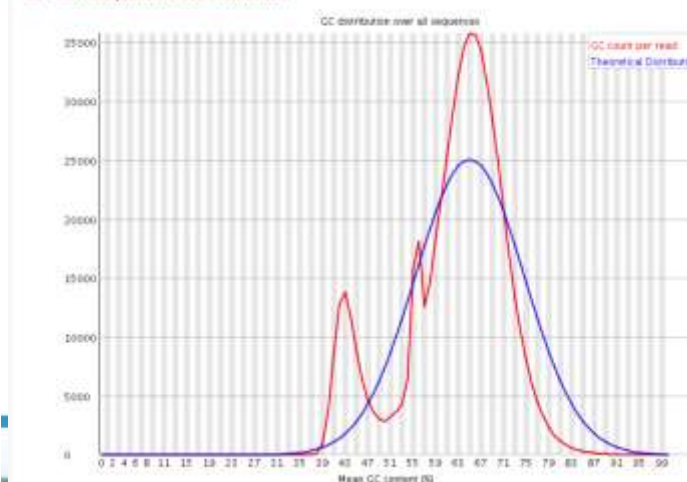


Il contenuto in GC del genoma è specie specifico, utile per capire se ci sono contaminazioni nel nostro sequenziamento

✓ Per sequence GC content



✗ Per sequence GC content





È molto comune che non tutte le metriche di qualità non siano ottimali.

Bisogna rimuovere alcune delle sequenze di bassa qualità per ridurre la probabilità di errori di sequenziamento

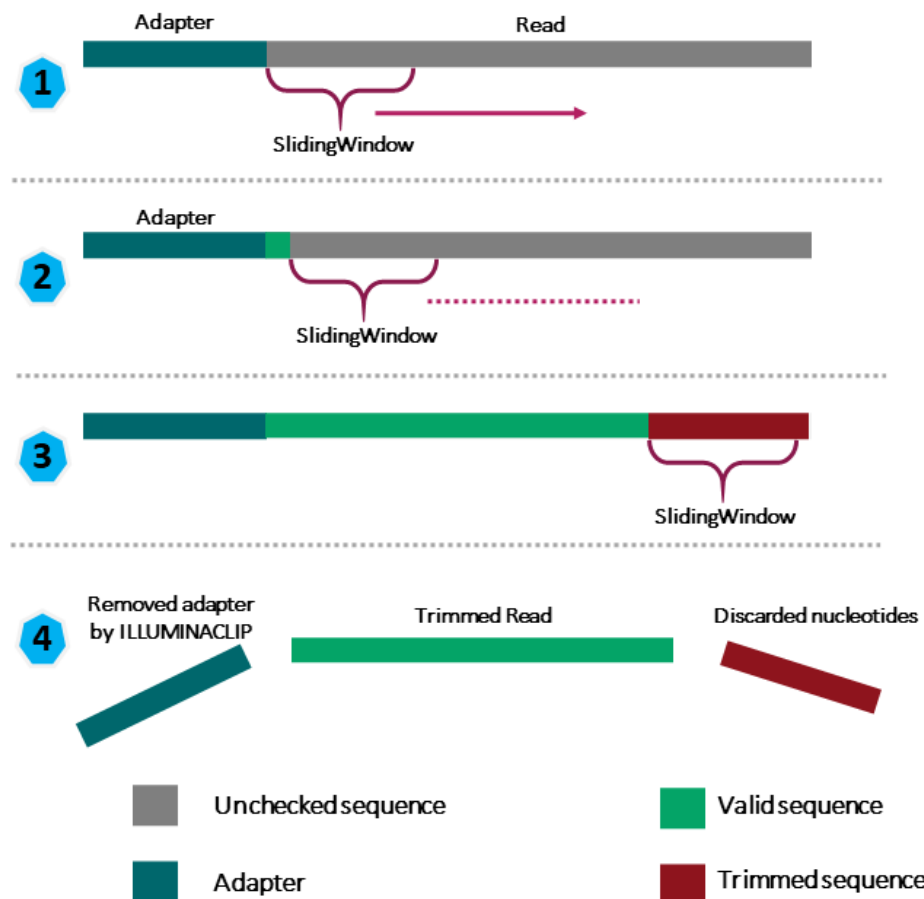
Si utilizza Trimmomatic per filtrare le reads e tagliare le basi di scarsa qualità dai nostri campioni

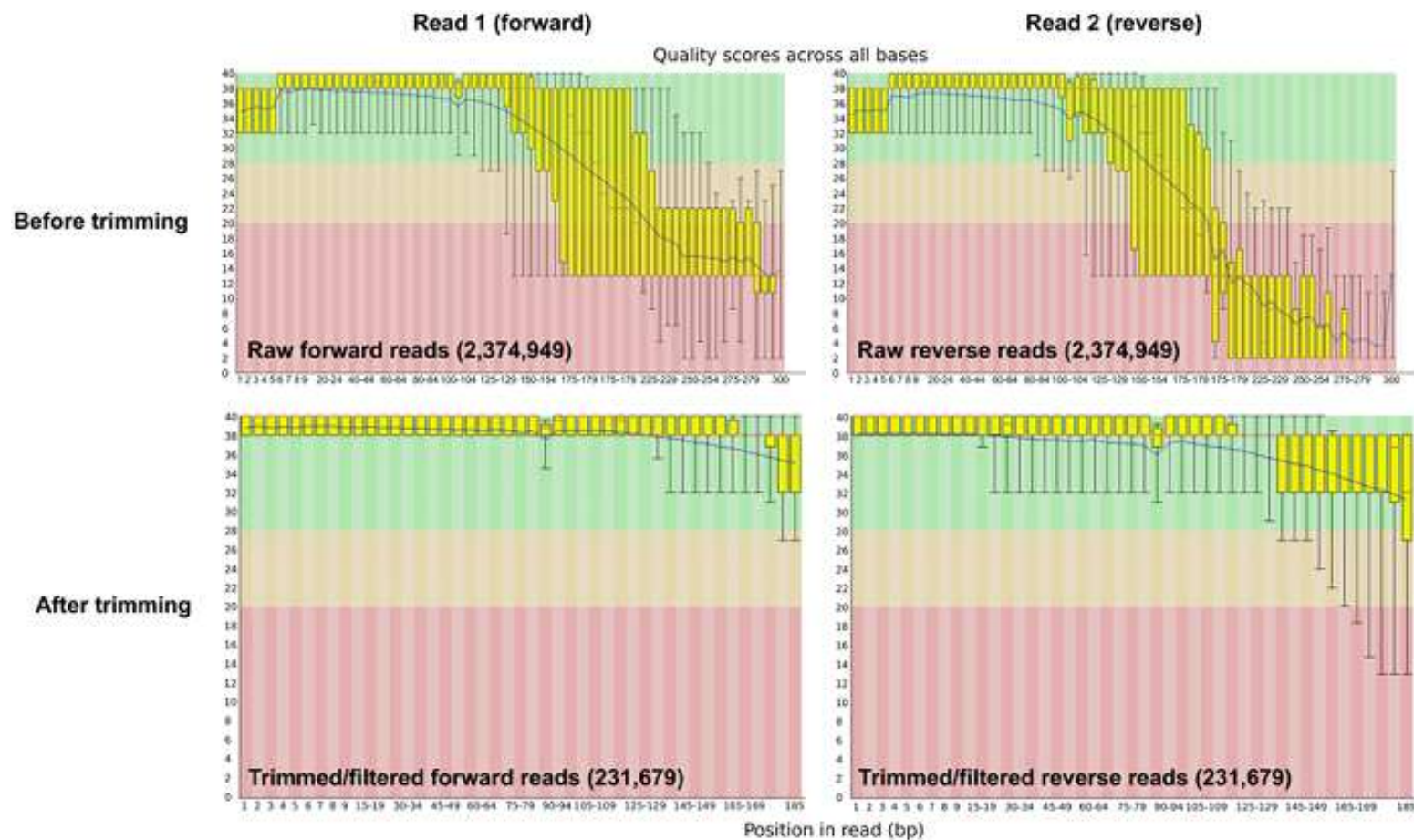
In questo passaggio vengono eliminati anche gli adattatori usati durante il sequenziamento

Es. di filtraggio:

«Finestra scorrevole» di una certa ampiezza che scorre nelle reads e la taglia in caso la media del Phred score sia minore al valore impostato

Trimmomatic





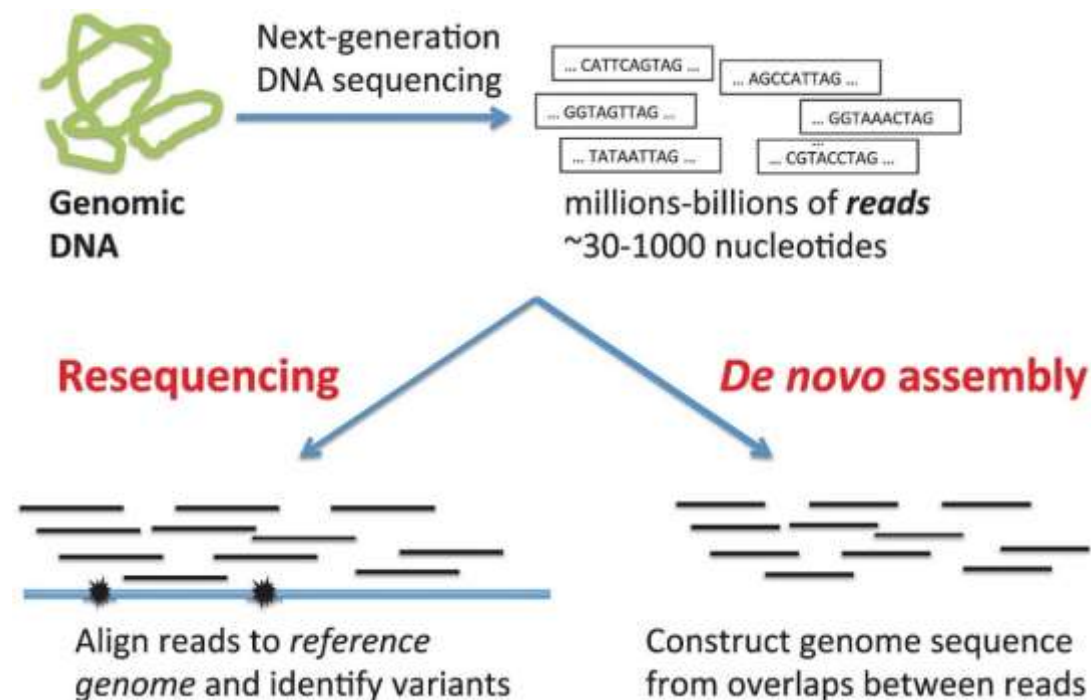


Dopo il filtraggio delle reads, è necessario assemblare le reads in sequenze più lunghe o allinearle a un genoma di riferimento.

Reference base: serve sequenza di riferimento “vicina”, metodo più accurato, utilizza solo geni già presenti in referenza. Come è più facile assemblare un puzzle se si conosce l'immagine, è più facile assemblare il genomi se si ha una buona idea dell'ordine delle sequenze

Assemblaggio *de novo*: non richiede sequenza di riferimento, si accede a tutto il genoma (possibilità di trovare nuovi plasmidi, nuovi geni di virulenza e resistenza), limitato dalla lunghezza delle reads, genera draft genomes

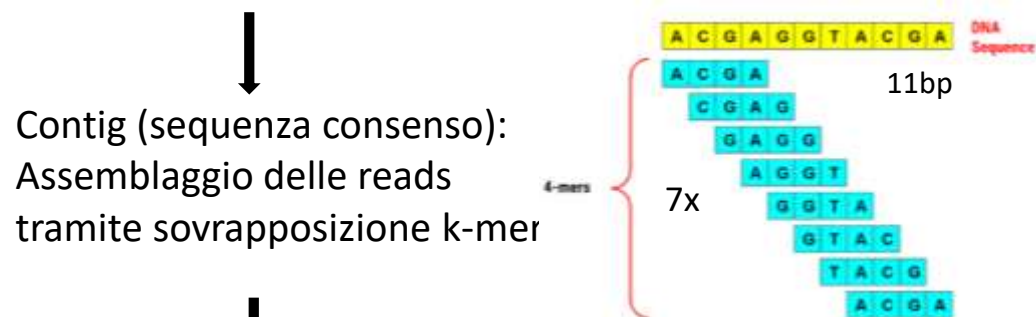
Assemblaggio del genoma



De novo assembly SPADES

A partire da reads trimmate

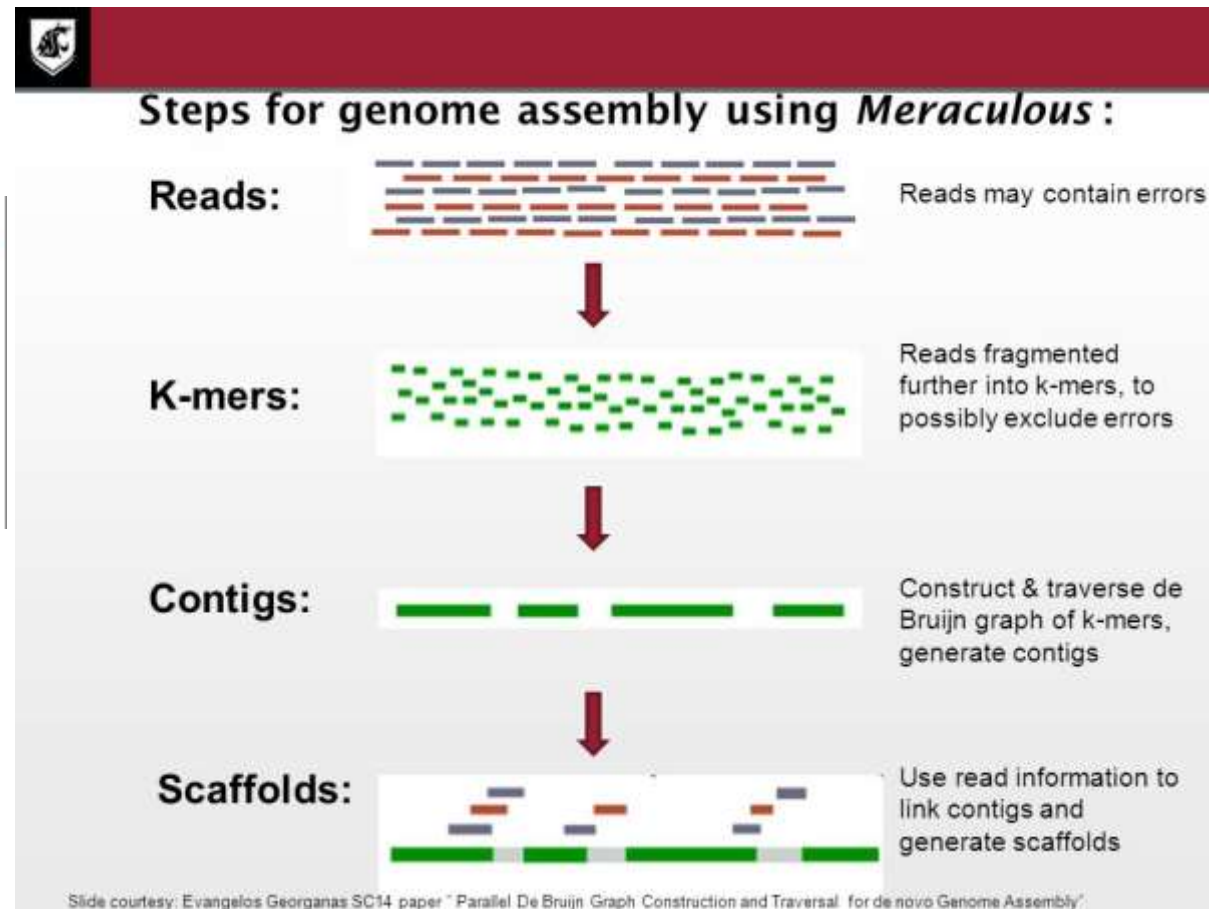
Utilizza i K-mer: reads frammentate
ulteriormente in sequenze da poche basi



↓

Scaffolds: Assemblaggio e ordinamento dei
conting

Rifinire gli assemblaggi: Pilon identifica le
incongruenze tra il genoma assemblato e le
informazione delle reads. Riempimento di
gap e correzione errori assemblaggio





Valutazione qualità assemblaggi del genoma: QUAST

Numero di contig

Lunghezza totale genoma

N50: lunghezza (in bp) del contig più piccolo del gruppo di contig che coprono metà della lunghezza del genoma assemblato

L50: numero di contig la cui lunghezza è maggiore o uguale al valore di lunghezza dell'N50

GC% importante perché specie specifico

All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length").

Worst Median Best ☒ Show heatmap

	18006523_2LSM_2019.TE.2942.1....	18032346_2_D_2019.TE.22810.1....
Statistics without reference		
# contigs	58	1514
# contigs (≥ 0 bp)	58	1514
# contigs (≥ 1000 bp)	46	616
# contigs (≥ 5000 bp)	24	133
# contigs (≥ 10000 bp)	18	68
# contigs (≥ 25000 bp)	15	46
# contigs (≥ 50000 bp)	11	22
Largest contig	476 350	204 461
Total length	3 238 956	4 952 434
Total length (≥ 0 bp)	3 238 956	4 952 434
Total length (≥ 1000 bp)	3 230 676	4 326 505
Total length (≥ 5000 bp)	3 190 471	3 371 557
Total length (≥ 10000 bp)	3 145 833	2 920 628
Total length (≥ 25000 bp)	3 100 471	2 582 583
Total length (≥ 50000 bp)	2 969 007	1 692 441
N50	344 454	27 994
N75	226 682	2782
L50	4	43
L75	7	225
GC (%)	37.61	40.06
Mismatches		
# N's	96	2037
# N's per 100 kbp	2.96	41.13

Criteri minimi di qualità (POS DIG 039 INT rev 0):

Lunghezza de contig > 500 bp

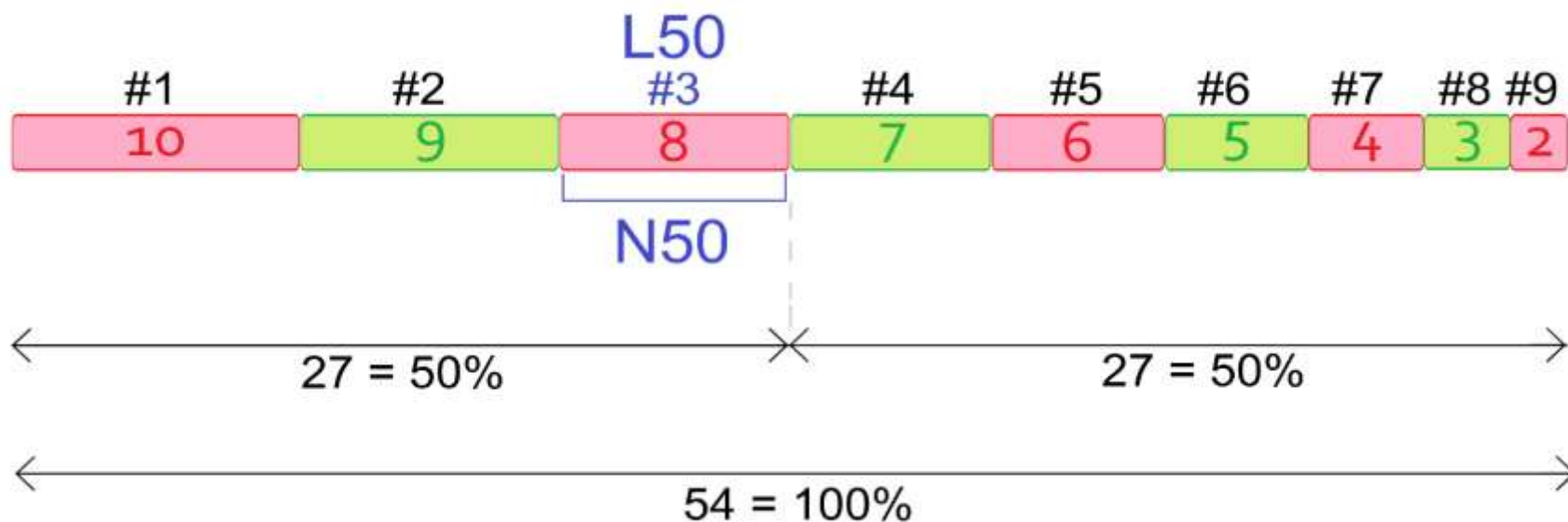
Numero di contig < 300

N50 > 15000 bp



N50: lunghezza (in bp) del contig più piccolo del gruppo di contig che coprono metà della lunghezza del genoma assemblato

L50: numero di contig la cui lunghezza è maggiore o uguale al valore di lunghezza dell'N50





Identificazione geni di virulenza:

- Virulence finder

Identificazione di geni per la resistenza agli antibiotici

- ResFinder

Annotazione di tutti geni presenti nel genoma (funzionalità cellula)

- Prokka

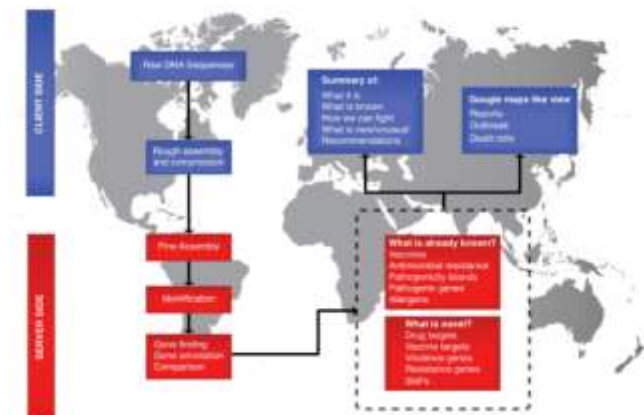
Servono dei **database di riferimento** verificati e curati (in continuo aggiornamento)

Si basano su BLAST, per il riconoscimento delle sequenze tra genoma e database

Caratterizzazione



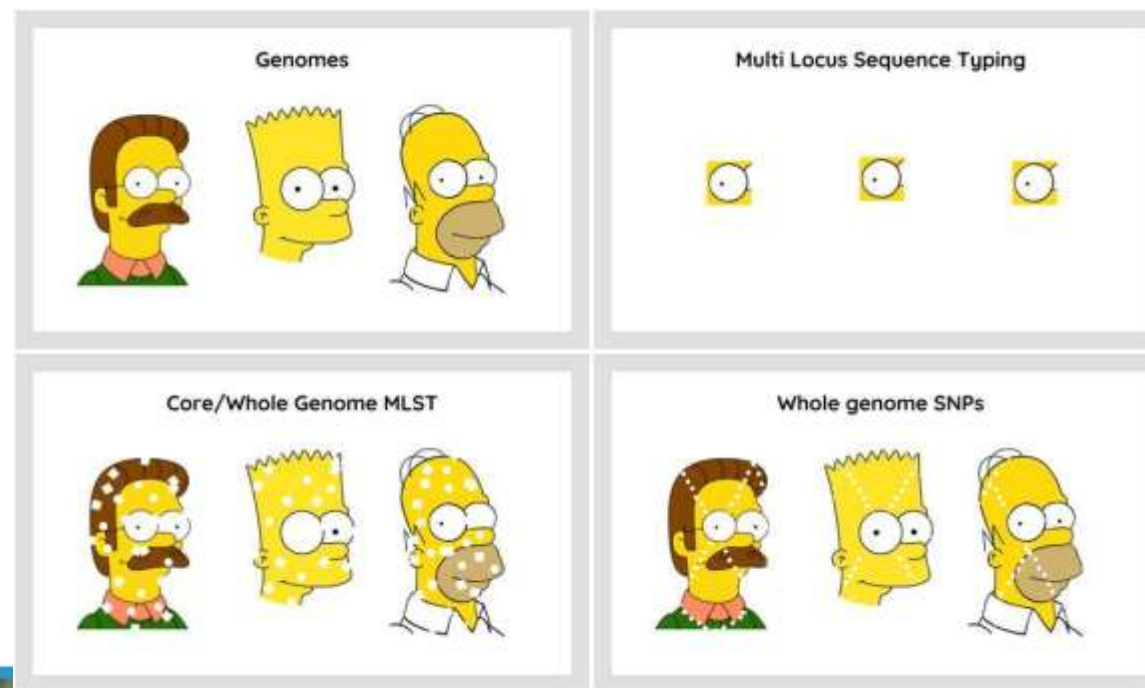
age will experience down time during next week (Week 42)





Metodologie molecolari per identificazione e confronto ceppi

- Identificazione sierogruppi e sierotipi molecolari: ricerca di geni correlati a formule somatiche (es. *Salmonella*, *Listeria*)
- Multi Locus Sequence Type (MLST): mlst-finder, analizza un set di geni la cui combinazione allelica assegna MLST
- Core Genome MLST: chewbbaca, analizza set di geni «core» (essenziali e funzionali) sempre presenti nella specie considerata (es. 3002 geni in *Salmonella*, 1701 *Listeria*) la loro comparazione ci indica la distanza «allelica» dei ceppi. Soglia specifica per considerare focolaio (es. 7 alleli per *L. monocytogenes*)
- Single Nucleotide Polymorphism (SNPs): CFSAN pipeline, CSIphylogeny, indentificano i polimorfismi a singolo nucleotide, differenze puntiformi tra genomi. Indica la quantità di distanza genetica tra ceppi e la relazione filogenetica

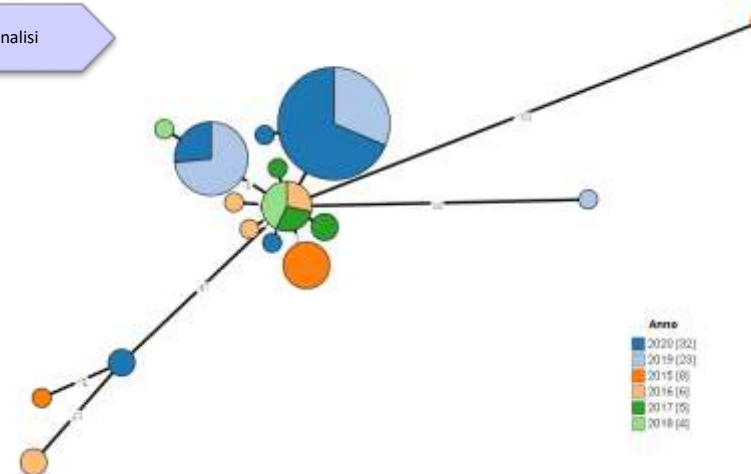


Source: Torsten Seeman, University of Melbourne Australia.

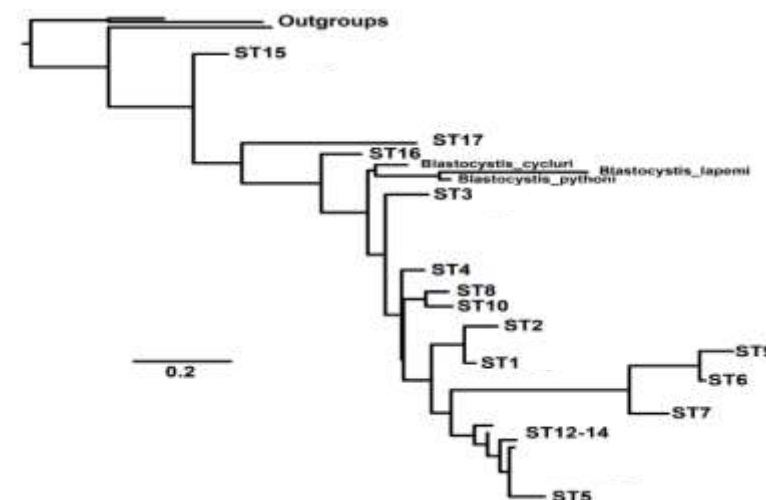


Visualizzare le differenze

Minimum Spanning Tree: albero di copertura di costo minimo, rete che collega i punti che da il valore più piccolo possibile (cgMLST)



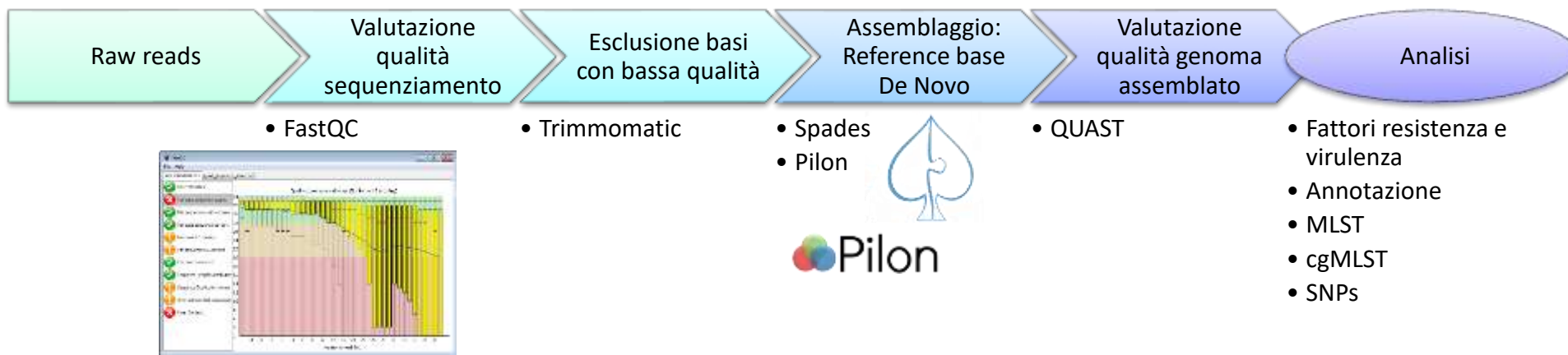
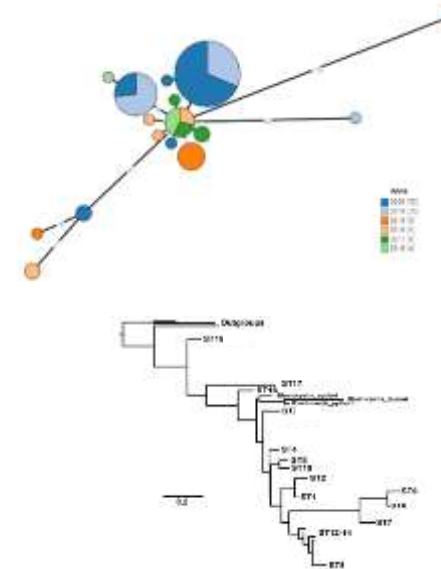
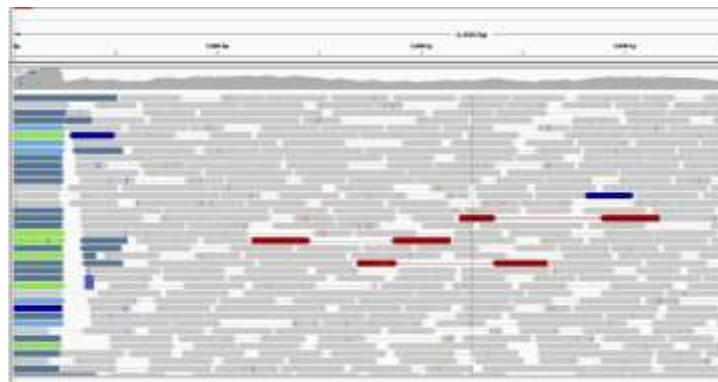
Alberi filogenetici: rappresentazione grafica delle relazioni tra gli individui o ceppi, le distanze nell'albero riflettono le differenze genetiche calcolate tra i campioni (SNPs)

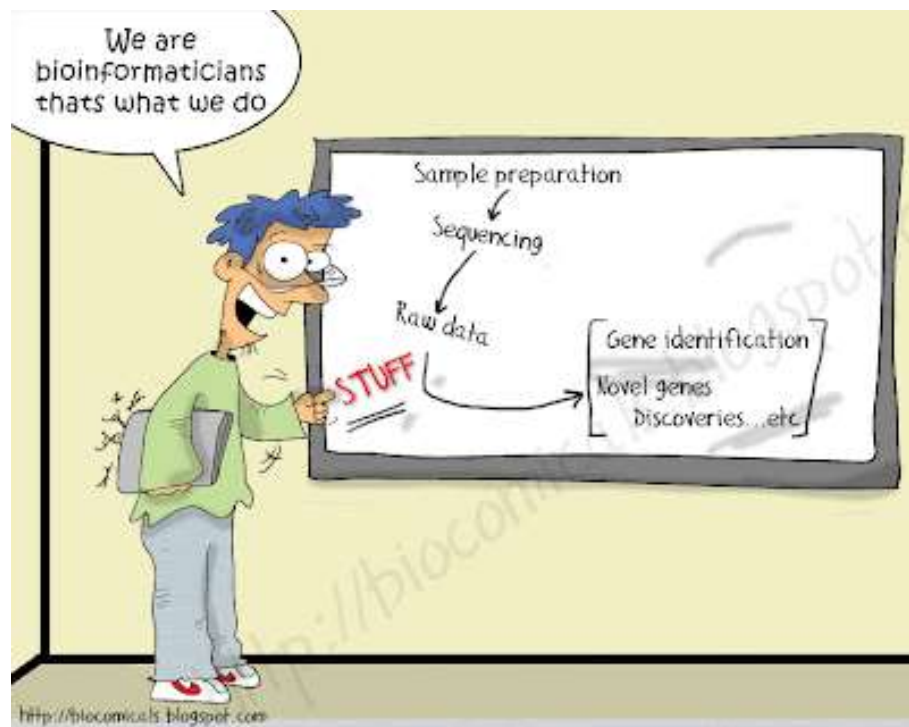


Riassumendo...



ACTTGG GTCATA ACCTAC GTATAG CATGCC ATGCAG
CCCGAT GCATGC CGACAT TGCAGT CCAGCA TGCTAA
ATCGAT ATGCAC TGAATC





Grazie per l'attenzione!

