

Analisi dei dati



ROMA 6/10/20



DAVIDE LA ROCCA



Analisi dei dati di sequenziamento

Chiariamo alcuni concetti introduttivi

Cos'è un file fastq?
Cos'è il phred score?

```
@M03865:12:000000000-D7FLL:1:1101:18034:2006 1:N:0:AAGAGGCA+CTCTCTAT
CCACGGCGTGCATGCTTCACGGTGCAAGCAGCCGTCCAGCAACTGCTCGTAAGTCCTCTGGT
+
AAAA?@?1D10>FFCGGD1100F00AA1000AA/EAA/1A0ABGAAGD/EEF0FEBGFGBG
@M03865:12:000000000-D7FLL:1:1101:17653:2012 1:N:0:AAGAGGCA+CTCTCTAT
TTTTATTTCGTTTTTCGCTATCGAACTGTGAAATGGAAATGGATGGAGAAGAGTTAATGAATGATATGGTCCTTTTGT
CATTCTCA
+
AAAAAFAFFDC1AEEFDA1E0EA000BGBGB221F11111F111F1/0010A10FGD2GBD1G21F2FD1GGCGHHBHH
FFHHFHFD
```



Analisi dei dati di sequenziamento

Chiariamo alcuni concetti introduttivi

Keep calm

```
@M03865:13:000000000-J3DTD:1:1104:22979:3134
GTATGTGAGTACACATGTTGACAAACCGTTCCTGTGCCCCATGAAGACGATTGCTCACAGTGGCTTGAACGGCGCTT
TAACTTTTGATTGCAAAAAGCCGTCTGTTCCGGTTGCAGTGCCGCTTTGGCCCTTGGCATACCCACTTACCATTGTGTT
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<:
```



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

GTATGTGAGTACACATGTTGACAAACCGTTCCTGTGCCCCATGAAGACGATTGCTCACAGTGGCTTGGAAACGGCGCTT
TAACTTTTGATTGCAAAAAGCCGCTCTGTTCCGGTTGCAGTGCCGCTTTGGCCCTTGGCATACCCACTTACCATTGT
ACCGTAA

+

C-B@CCG9CFGFGGFGGEGGFFFGF<D:CFG@CEE@FFEFGFFA<9E@FCCCE9F,@F<FCFG?EE7:DCECFGG+8:
DfE9E9E<EFD<FDGFGF<=<CFFGG7FF,C=EGGC@FDA=FGFGGGGF,EDF<84?A5EE@E8FF8?F<,,?AFFDC@
D?BCF<:



Analisi dei dati di sequenziamento

Cos'è il formato file fastq?

Header

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control
number>:<sample number>

```
@M03865:12:000000000-D7FLL:1:1101:18034:2006 1:N:0:AAGAGGCA+CTCTCTAT
```



Analisi dei dati di sequenziamento

Phred/Q score

Come viene generato il Phred ? **Durante una corsa di sequenziamento, un punteggio di qualità viene assegnato a ciascuna chiamata di base per ogni cluster, su ogni tile, per ogni ciclo di sequenziamento.** Ci sono 3 fasi:

- misurazione di vari aspetti correlati alla qualità dell'identificazione delle basi come il rapporto singolo-rumore e l'intensità della luce. Sulla base di questi parametri viene calcolato un valore predittivo di qualità
- Il valore predittivo di qualità deve essere tradotto in un punteggio di qualità con l'aiuto di una tabella di Qualità. La tabella Q basata su una curva di calibrazione statistica è derivata da dati empirici inclusi vari campioni umani e non umani ben caratterizzati, principalmente utilizzando una versione del cosiddetto algoritmo Phred
- il punteggio di qualità (per ciclo) viene registrato in comune con l'identificazione delle basi in un file di identificazione delle basi (.bcl)



Analisi dei dati di sequenziamento

Phred/Q score

La memorizzazione dei punteggi Q, come singoli caratteri (o byte) ha fornito una codifica semplice ma ragionevolmente efficiente in termini di spazio. Affinché il file sia leggibile dall'uomo e facilmente modificato, ciò ha limitato le scelte ai caratteri stampabili della tabella ASCII 32–126 (decimali) e poiché ASCII 32 è il carattere spazio, i file Sanger FASTQ utilizzano ASCII 33 (tabella codifica ascii da 32 non considerando lo spazio vuoto)–126 (tutti i caratteri possibili) per codificare le qualità PHRED da 0 a 93 (cioè punteggi PHRED con un offset ASCII di 33).

.. Il PHRED Next-Gen è stato adattato su quello sanger ma non arriva a 93 ma a 42



Analisi dei dati di sequenziamento

Phred/Q score

- E' un valore intero che rappresenta la probabilità stimata di un errore di sequenziamento, ovvero di una base non corretta
- $P = 10^{-Q/10} \iff Q = -10 \log_{10}(P)$
- Da questa formula otteniamo
 $Q = \text{ASCII_CODE} - \text{ASCII_BASE}$ (formula per convertire i caratteri in interi)

C-B@CCG9CFGFGGGGEGGFFFGF<D:CFG@CEE@FFEFGGFA<9E@FCCCE9F,@F<FCFG?EEC7:DCECFGG+8:
DFE9E9E<EFD<FDGFGF<=<CFFGG7FF,C=EGGC@FDA=FGFGGGGF,EDF<84?A5EE@E8FF8?F<,,??AFFDC@
D?BCF<:

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			



Analisi dei dati di sequenziamento

Chiariamo alcuni concetti introduttivi

La strategia di sequenziamento può essere:

- single-end (frammenti corti da una direzione)
- **paired-end** (frammenti corti da entrambe le direzioni)
- mate-pairs (frammenti lunghi da entrambe le direzioni)



Analisi dei dati di sequenziamento

Chiariamo alcuni concetti introduttivi

Il file .fastq R1 sta ad indicare che la direzione di sequenziamento è 5'-3'

```
3558-5-BCL_S1_L001_R1_001.fastq  
3558-5-BCL_S1_L001_R2_001.fastq
```

Il file .fastq R2 sta ad indicare che la direzione di sequenziamento è 3'-5'



Analisi dei dati di sequenziamento

****Piccola parentesi****

Cos'è un algoritmo?

Una serie di operazioni o “passi” da eseguire in un determinato ordine sui nostri dati, al fine di ottenere dei risultati



Ok ora possiamo cominciare



Esperimento pilota

Step 1. Selezione dei target

- ❖ 5 MDR
- ❖ 3 Campioni di campo
- ❖ 8 Libraries
- ❖ 55 Ampliconi

LIBRARIES	Amplicone/target	Dimensione degli Ampliconi (bp)	Numero ampliconi per library
Cotone MON15985	Acp1	143	5
	NptII	180	
	Cry1Ab/Cry1ac	141	
	23579	261	
	trnI	148-175	
Mais 89034	HMG	146	5
	Tnos	171	
	FMV	145	
	23579	261	
	trnI	148-175	
Soia A5547-127	LEC	141	5
	PAT	135	
	T35s	140	
	23579	261	
	trnI	148-175	
Babrbabietola H7-1	GS	185	7
	CTP-CP4EPSPS	129	
	CTP2-CP4EPSPS	243	
	Te9	154	
	FMV	145	
Patata EH92-527-1	23579	261	5
	trnI	148-175	
	UGPasi	156	
	NptII	180	
	JUNCTION P NOS/NPTII	237	
20008280 (mangime complementare)	23579	261	9
	trnI	148-175	
	Acp1	143	
	GS	185	
	Tnos	171	
20008271 (mangime semplice)	CP4EPSPS	212	8
	CTPCP4EPSPS	129	
	CTP2 CP4EPSPS	243	
	PAT	135	
	23579	261	
20005309 (mangime complementare)	trnI	148-175	11
	HMG	146	
	LEC	141	
	Acp1	143	
	Tnos	171	
	CP4EPSPS	212	
	CTPCP4EPSPS	129	
	CTP2 CP4EPSPS	243	
	PAT	135	
	BAR	127	
	23579	261	
	trnI	148-175	



Istituto Zooprofilattico Sperimentale
del Lazio e della Toscana M. Aleandri



Centro di Riferenza Nazionale
per la Ricerca di OGM

Applicazioni di metodiche di NGS all'analisi degli OGM (Progetti in corso al CROGM)

- ❑ Saggio multiscreening in NGS su piattaforma Illumina, CA5 (progetto in collaborazione con FEM)
- ❖ Sequenziamento in parallelo di un elevato numero di target (37 target di cui 11 geni endogeni, 14 elementi di screening, 8 costrutti, 2 donatori, 2 saggi identificatori di specie)
- ❖ Strategia NGS utilizzata; amplicon sequencing
- ❖ Rilevazione eventi GM di specie vegetali non considerate nella fase di analisi taxon-specifica;
- ❖ Rilevazione eventi GM non autorizzati;
- ❖ Rilevazione presenza nella matrice alimentare di organismi donatori naturali appartenenti a numerose specie (vegetali e non)
- ❖ Identificazione di varianti di sequenza (SNP) correlabili ad uno specifico evento GM



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Demultiplexing



- P5 e P7 sono adattatori della flow cell
- **i5 e i7 sono gli index**
- SP1 e SP2 siti di legame dei primer di sequenziamento



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Demultiplexing : è il processo tramite cui le reads vengono divise e raggruppate in diversi file .fastq separati, per ogni index associato a quel determinato campione. La sequenza dell'index verrà rimossa dalla read(riga 2) e annotata nell'header

Questa operazione è diversa per i diversi sequenziatori, ad esempio nel mondo

Illumina:

- Mi-seq: viene eseguita automaticamente utilizzando il PC di bordo
- NextSeq500 e HiSeq: viene eseguito su BaseSpace (la risorsa basata su cloud di Illumina)



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

I programmi che fanno parte del sequenziatore quando generano il file .fastq, possono effettuare una rimozione degli adattatori (**P5 e P7**) dalle reads contenute nel , **perciò in teoria non dovremmo ritrovarli nelle reads.** Tuttavia è sempre bene controllare sia attraverso programmi che valutano la qualità (FastQC) ma anche ad esempio attraverso la ricerca testuale della sequenza con comandi unix/linux (ad esempio 'grep') all'interno del file



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Demultiplexing e Rimozione anche di adattatori

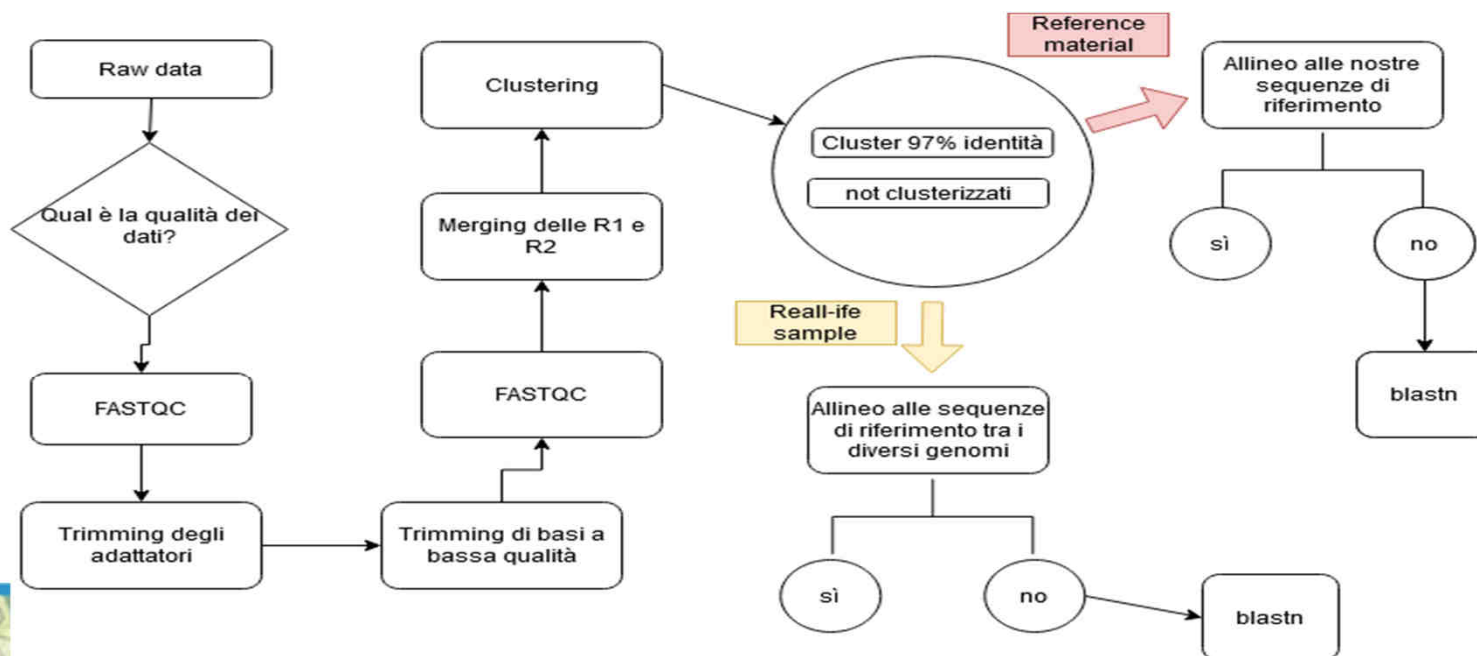


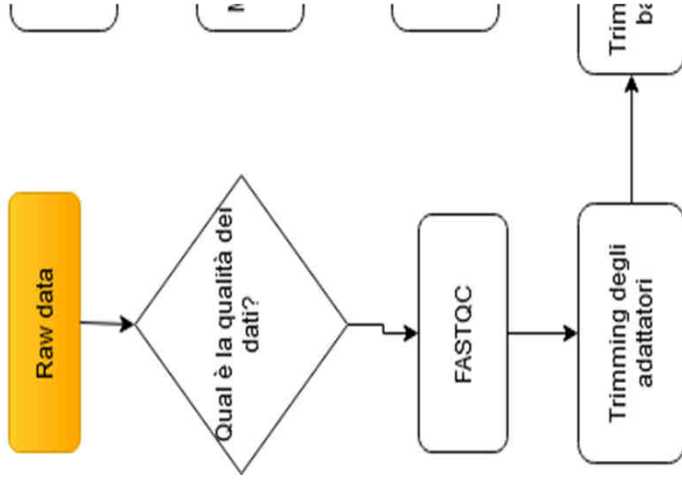
- **P5 e P7** sono adattatori della flow cell
- **i5 e i7** sono gli index
- SP1 e SP2 siti di legame dei primer di sequenziamento



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi





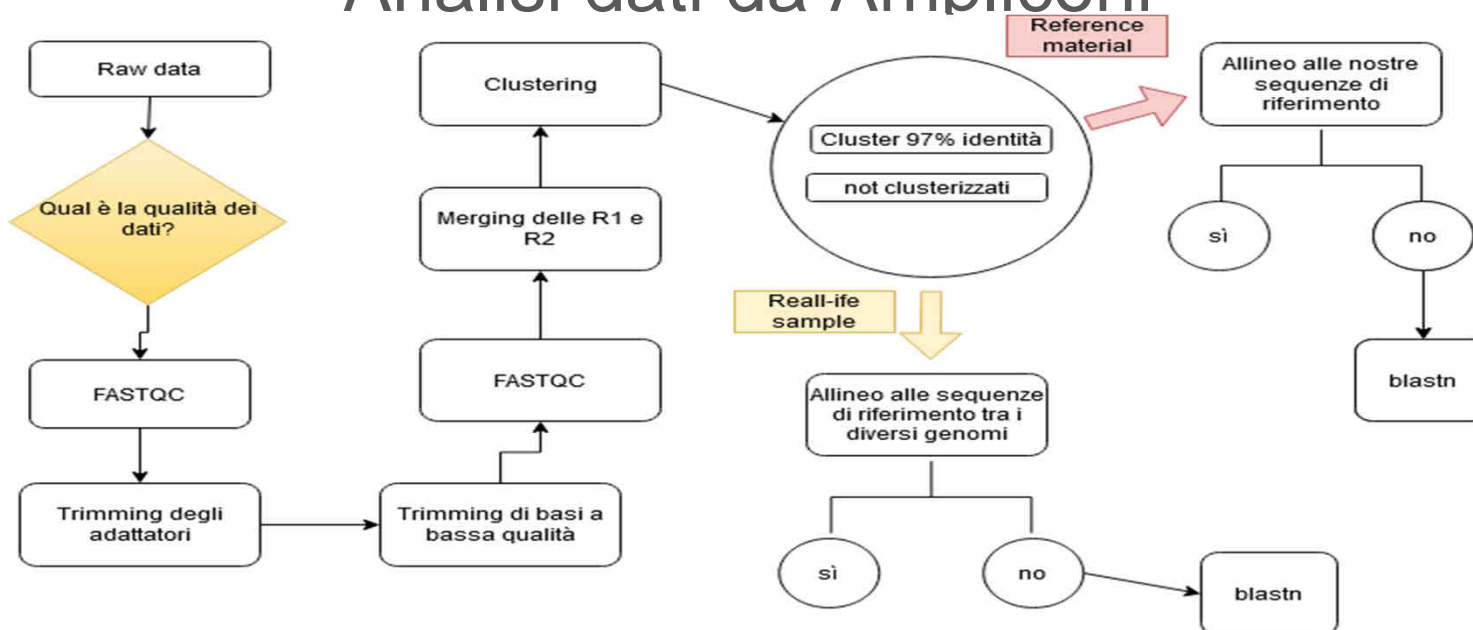
```

@M03865:12:000000000-07FLL:1:1101:18034:2006 1:N:0:AAGAGGCA+CTCTCTAT
CCACGGGTGCATGCTTCACGGTGCAGCAGCCGTCAGCAACTGCTCGTAAGTCCTCTGGT
+
AAAA?@?1D10>FFCGD1100F00AA1000AA/EAA/1A0ABGAAGD/EEF0FEBGFCBCG
@M03865:12:000000000-07FLL:1:1101:17653:2012 1:N:0:AAGAGGCA+CTCTCTAT
TTTTTATTCGGTTTCGCTATCGAACTGTGAAATCGAAATCGATGGAGAGAGTTAATGAATGATATGTCCTTTTGT
CATTCTCA
+
AAAAAFFFFDC1AEFFDA1E0EA000BGB8221F11111F1111F1/0010A10FGD2GBD1G21F2FD1GGCGHHBHH
FFHHFFHD
  
```



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

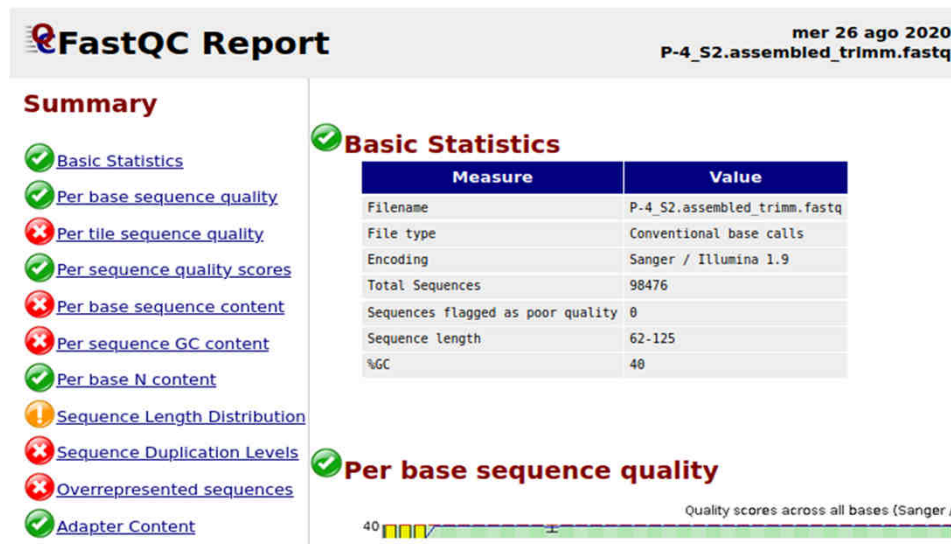


Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati?

FastQC fornisce un 'Quality-Check' report che può evidenziare eventuali problemi originati nel sequenziatore o nel materiale di library di partenza



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

- File type: Dice se il file ha o meno dati di spazi colorati da convertire in base calls
- Encoding: Mostra il tipo di codifica ASCII
- Total Sequences: Conteggio del numero di sequenze processate
- Filtered sequences: Mostra le sequenze rimosse se il programma è eseguito in Casava mode
- Sequence Length: Fornisce il valore della lunghezza della sequenza più corta e di quella più lunga
- %GC: La percentuale GC media su tutte le basi di tutte le sequenze

✓ Basic Statistics

Measure	Value
Filename	P-4_S2.assembled_trimm.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	98476
Sequences flagged as poor quality	0
Sequence length	62-125
%GC	40



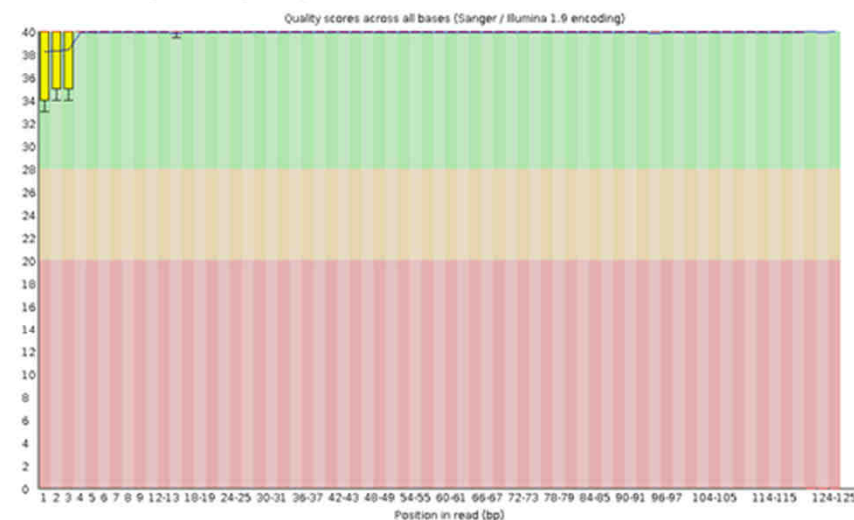
Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

- La linea centrale rossa è il valore mediano
- Il box giallo rappresenta il range inter-quartile (25-75%)
- Il baffo più alto e quello più basso rappresentano il 90% e il 10%
- La linea blu rappresenta la qualità media

L'asse y nel grafico mostra i quality scores. Più alto è il punteggio migliore è la base call

✓ Per base sequence quality



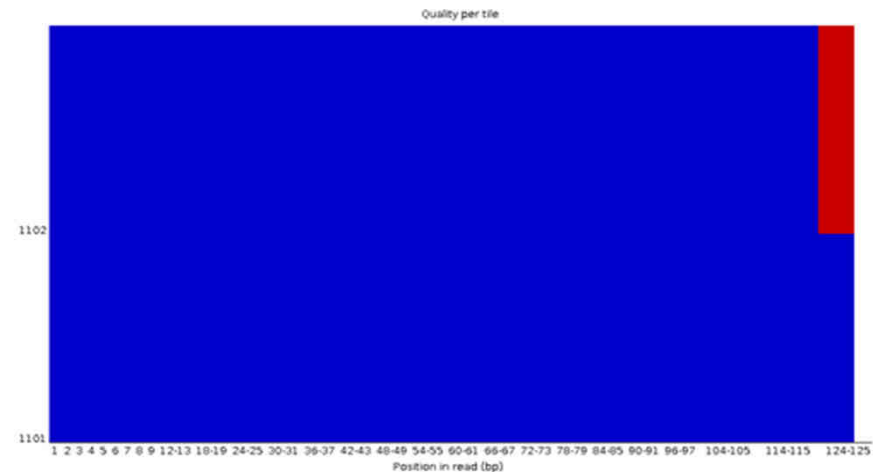
Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati?

Il grafico permette di valutare i quality scores di tutte le basi da ogni tile, per vedere se è presente una perdita di qualità localizzata in una parte della flowcell.

❌ Per tile sequence quality

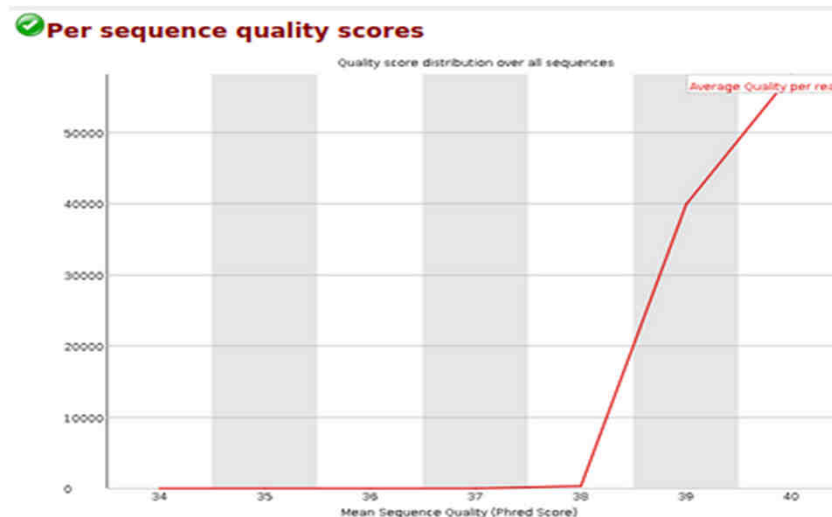


Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati?

Mostra se un subset delle nostre sequenze ha universalmente valori di bassa qualità.

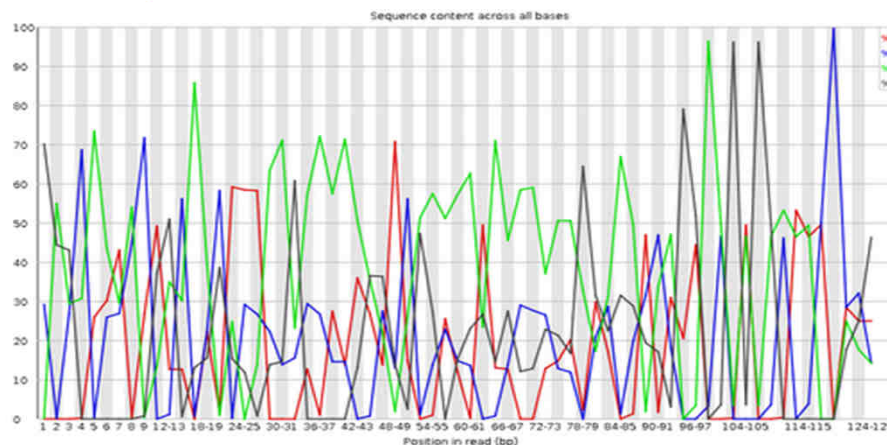


Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati ? ➡ **Per base sequence content**

Grafico che mostra il relativo ammontare di ogni base(ogni base ha un colore), riflettendo l'ammontare medio di queste basi.



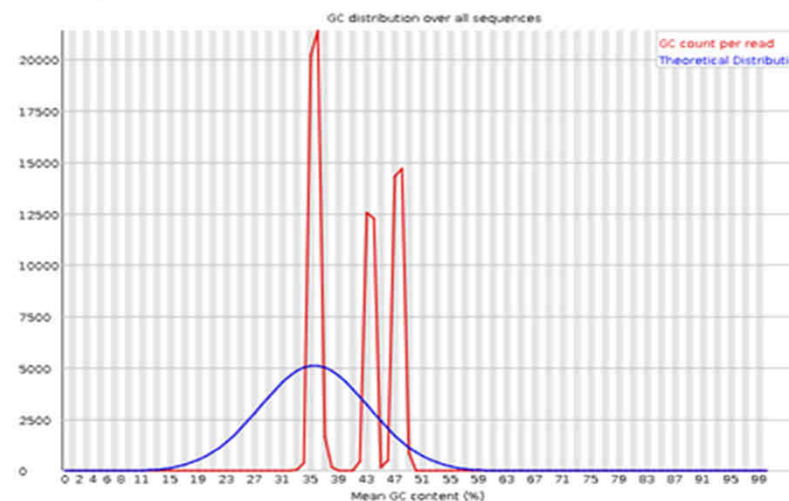
Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati ?

Questo modulo misura il contenuto di GC per tutta la lunghezza di ogni sequenza in un file e lo compara ad una distribuzione normale modellata sul contenuto di GC

✖ Per sequence GC content



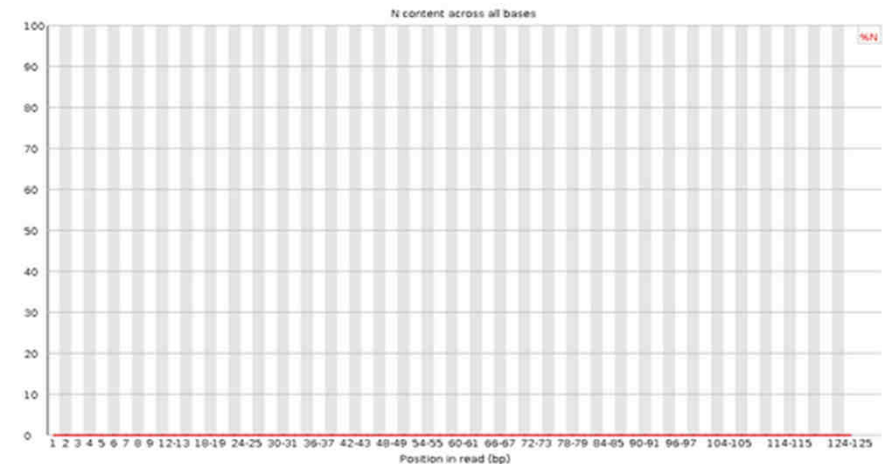
Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati ?

Questo modulo traccia la percentuale di identificazione delle basi in ciascuna posizione per la quale è stato chiamato un N.

✓ Per base N content



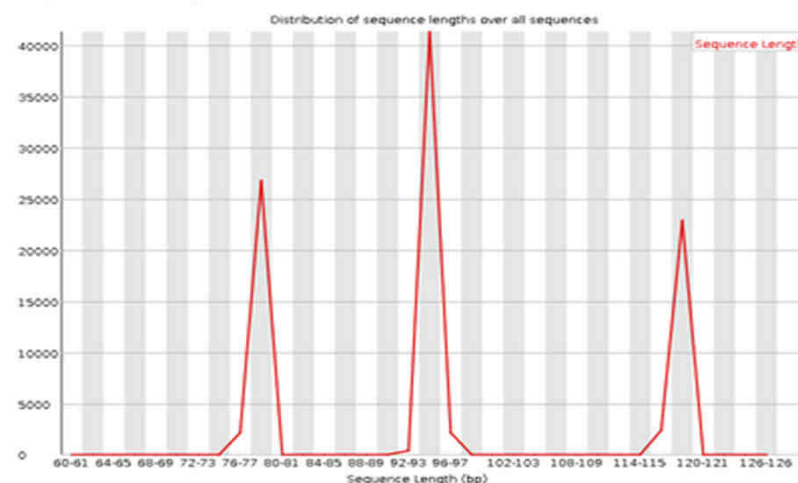
Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati?

Questo modulo genera un grafico che mostra la distribuzione delle dimensioni dei frammenti presenti nel file.

Sequence Length Distribution

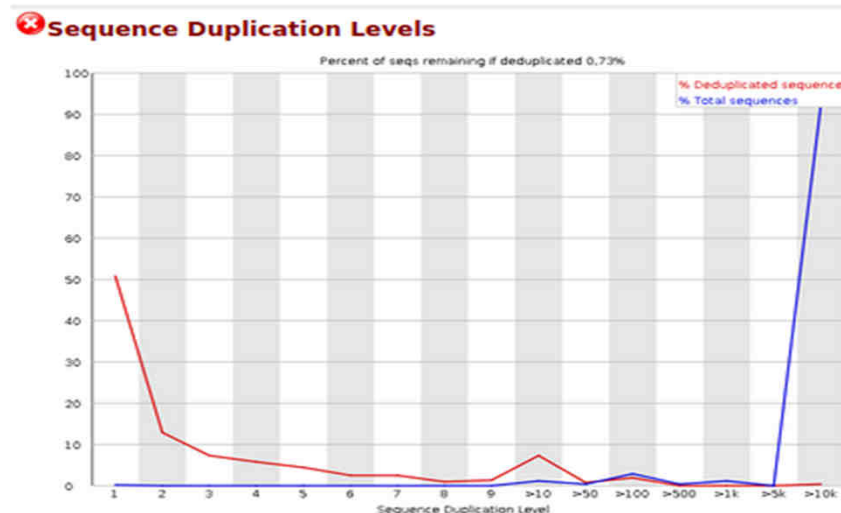


Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati?

Questo modulo conta i gradi di duplicazione per ogni sequenza in una library e crea un grafico che mostra il relativo numero di sequenze con diversi gradi di duplicazione.



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Qual è la qualità dei nostri dati?

Questo modulo lista tutte le sequenze che appaiono più dello 0.1% del totale

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGGCAATCCTGAGCCAACCTCTTTTTCARAGAAAAAATAAGGATTC	40725	41.355254072058166	No Hit
CAAAATAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTA	26746	27.15991713717048	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	23378	23.73979446768756	No Hit
GGCAATCCTGAGCCAACCTCTTTTTCARAGAAAAAATAAGGATTC	1157	1.1749855607457655	No Hit
CAAAATAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTAT	508	0.5158617328079939	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	393	0.39908200982980624	No Hit
GGGCAATCCTGAGCCAACCTCTTTTTCARAGAAAAAATAAGGATTC	372	0.37775701693813724	No Hit
GGGCAATCCTGAGCCAACCTCTTTTTCARAGAAAAAATAAGGATTC	314	0.3188594175230513	No Hit
CAATTAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTAT	242	0.24574515618018605	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	194	0.19708231528494255	No Hit
GGGCAATCCTGAGCCAACCTCTTTTTCARAGAAAAAATAAGGATTC	182	0.18481660506113165	No Hit
CAAAATAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTAT	177	0.17973922580121043	No Hit
CAAAATAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTAT	165	0.16755351557739956	No Hit
ACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	153	0.15536780535358868	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	152	0.15435232950160446	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	129	0.13099638490596693	No Hit
CAAAATAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTAT	119	0.12084162638612454	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	111	0.11271781957025062	No Hit
GACCTCCATATTACTGAAGGAAGCCAAAAGGGATCAATTAAGTGTCTAC	110	0.11170234371826637	No Hit
AAATAACGTGGAAAAGAGCTGCTGACAGCCCACTCAATGCGTAT	103	0.1045940127543767	No Hit

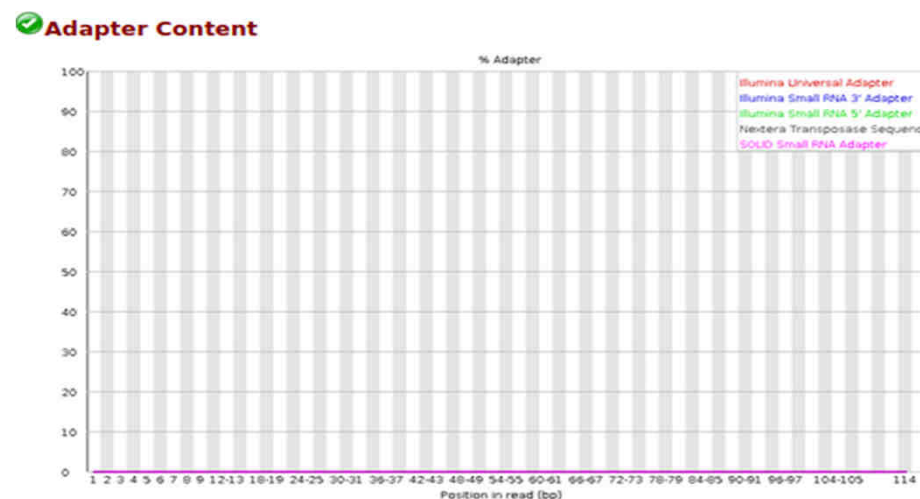


Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

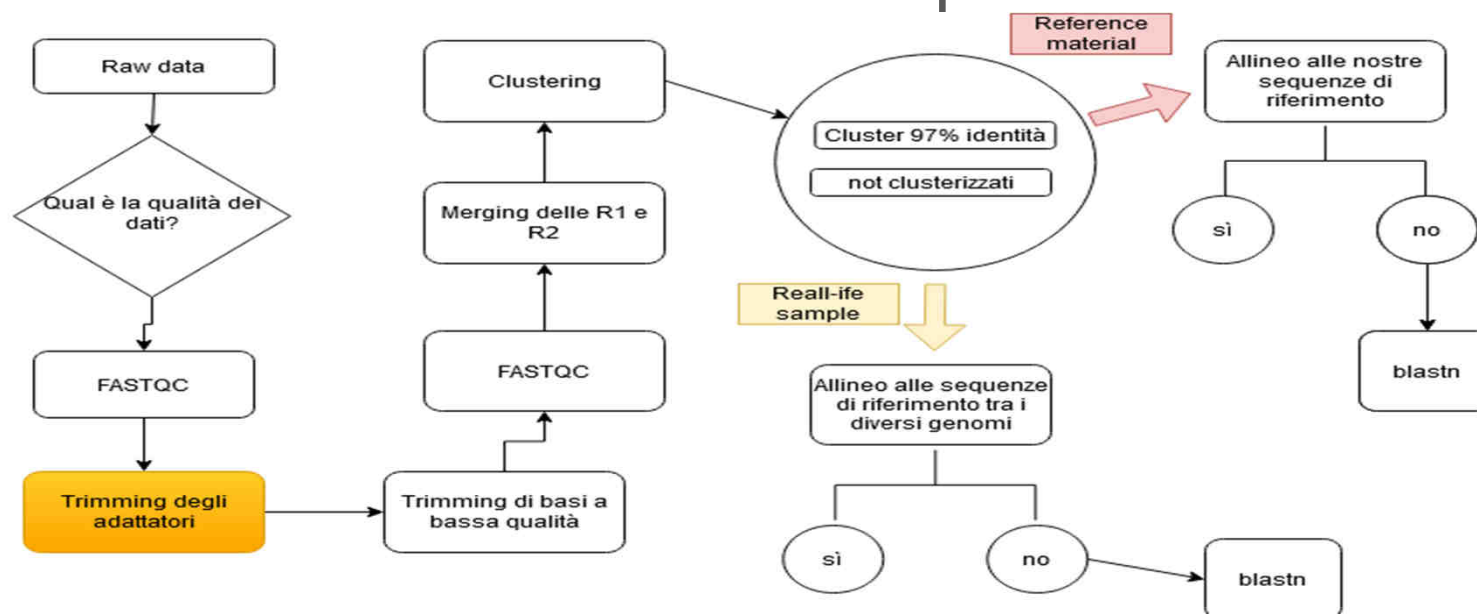
Qual è la qualità dei nostri dati?

Questo modulo può trovare un numero di differenti fonti di bias nella library che può includere la presenza di adattatori che si accumulano alla fine delle nostre sequenze.



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Trimming : termine utilizzato per descrivere la procedura di rimozione di adattatori, primer di sequenziamento ancorati alle estremità delle reads ma anche rimozione di basi a bassa qualità di sequenziamento.



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Trimming adattatori(controllo)/primer

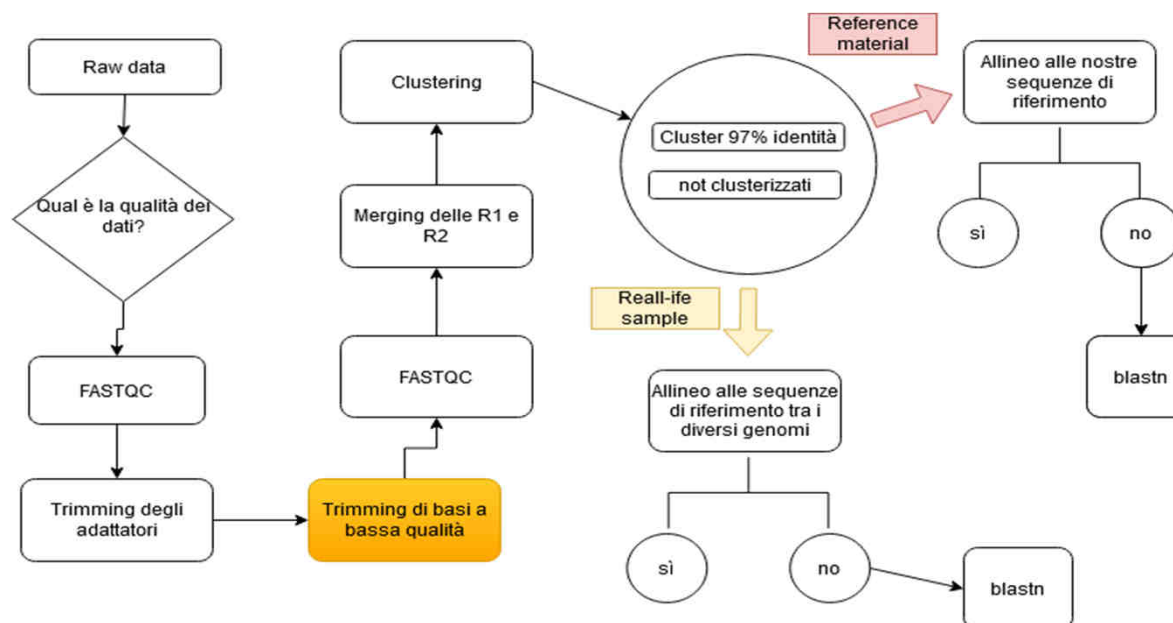


- P5 e P7 sono adattatori della flow cell
- i5 e i7 sono i barcode
- SP1 e SP2 siti di legame dei primer di sequenziamento



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

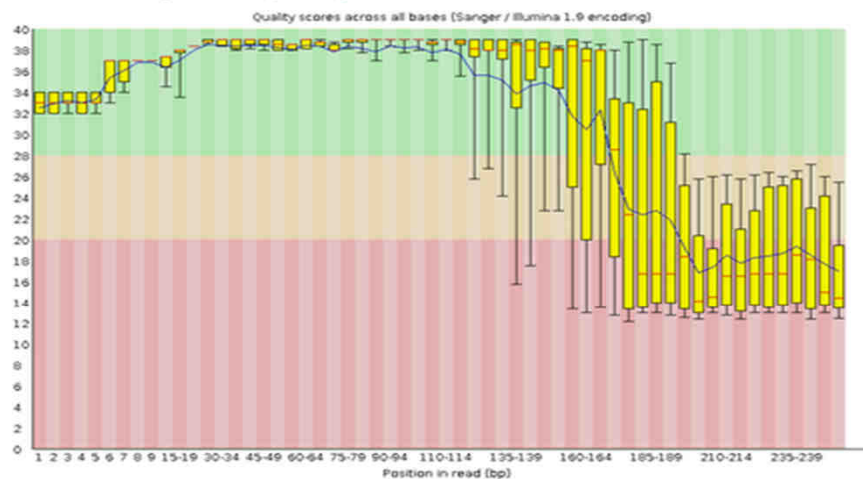


Analisi dei dati di sequenziamento

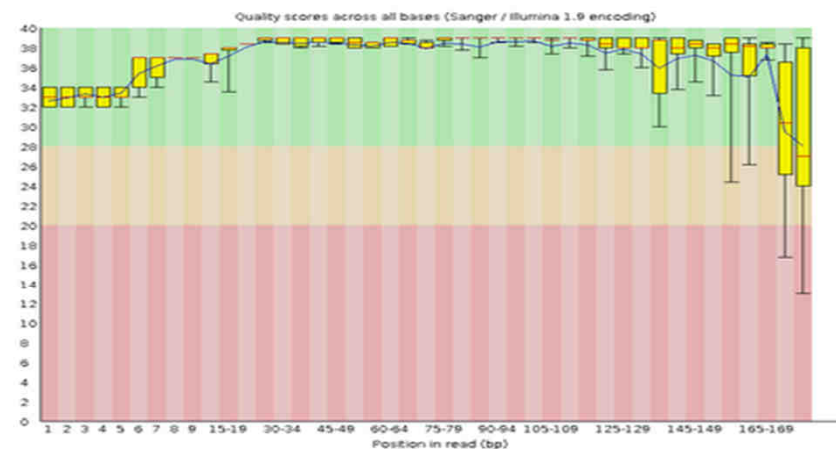
Analisi dati da Ampliconi

Trimming adattatori/ basi a bassa qualità in sequenza

✖ Per base sequence quality

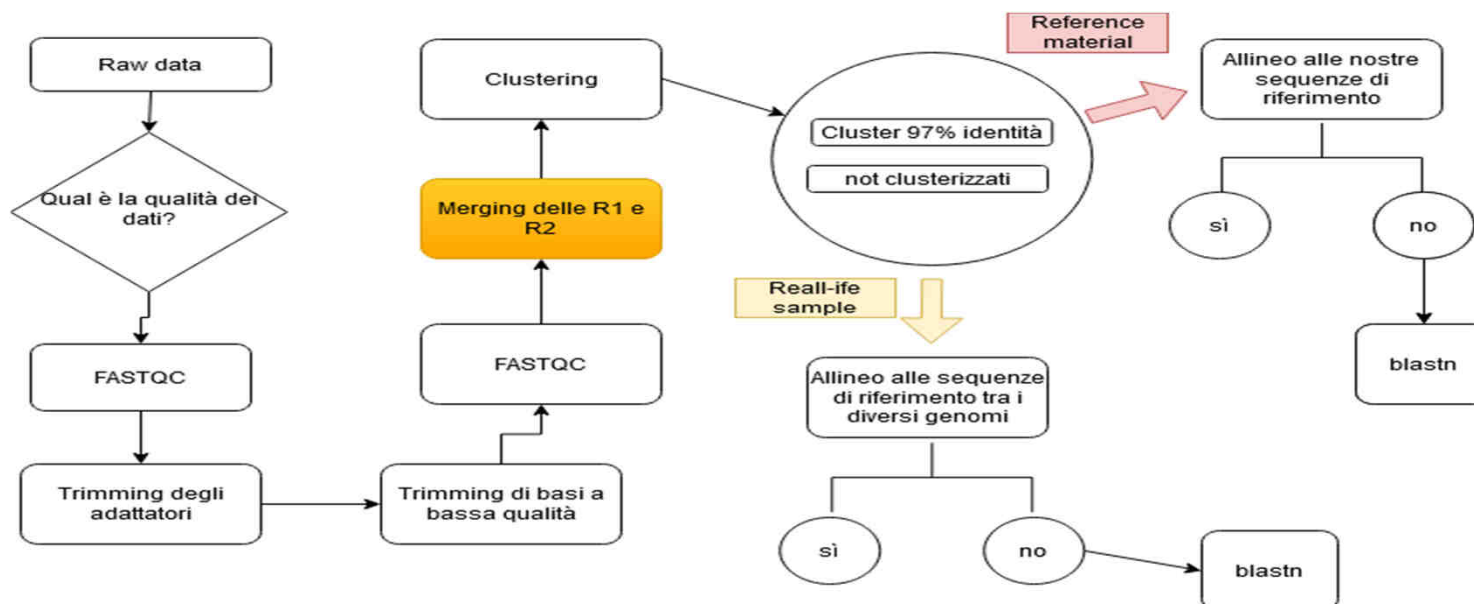


✔ Per base sequence quality



Analisi dei dati di sequenziamento

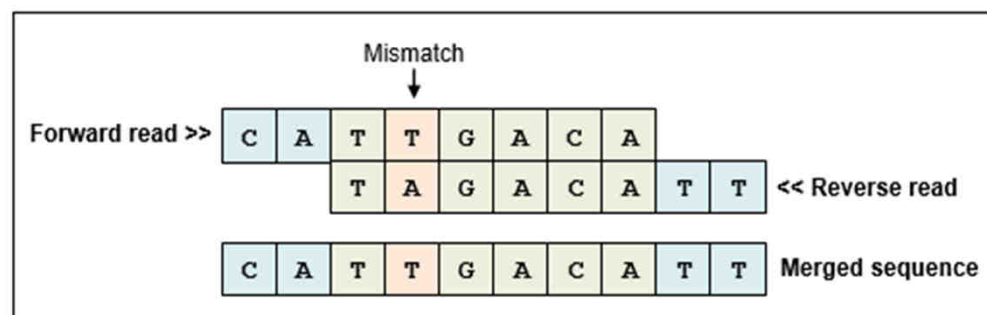
Analisi dati da Ampliconi



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Merging delle reads R1 e R2: termine usato per descrivere l'operazione di unione della read R1 e la sua corrispondente R2 sfruttando la porzione di overlap (quella comune a entrambe, che si trova all'estremità, quindi al 3' della R1 e al 5' della R2) generando una singola read

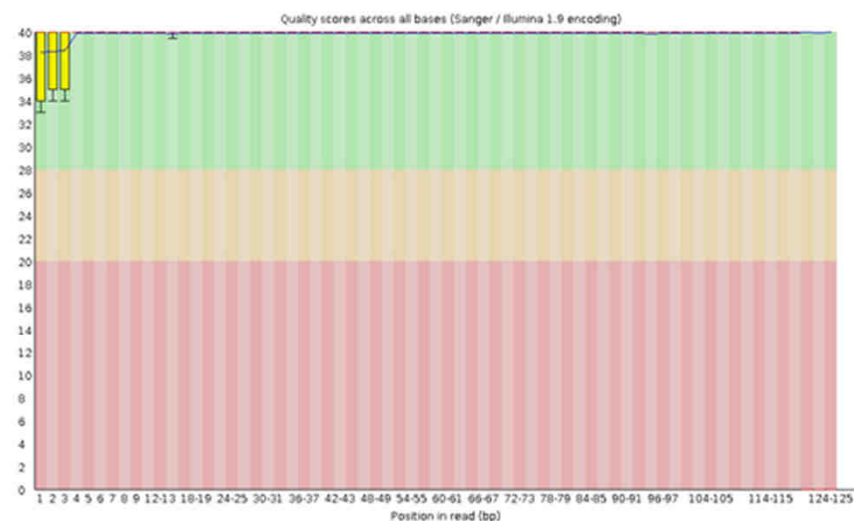


Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

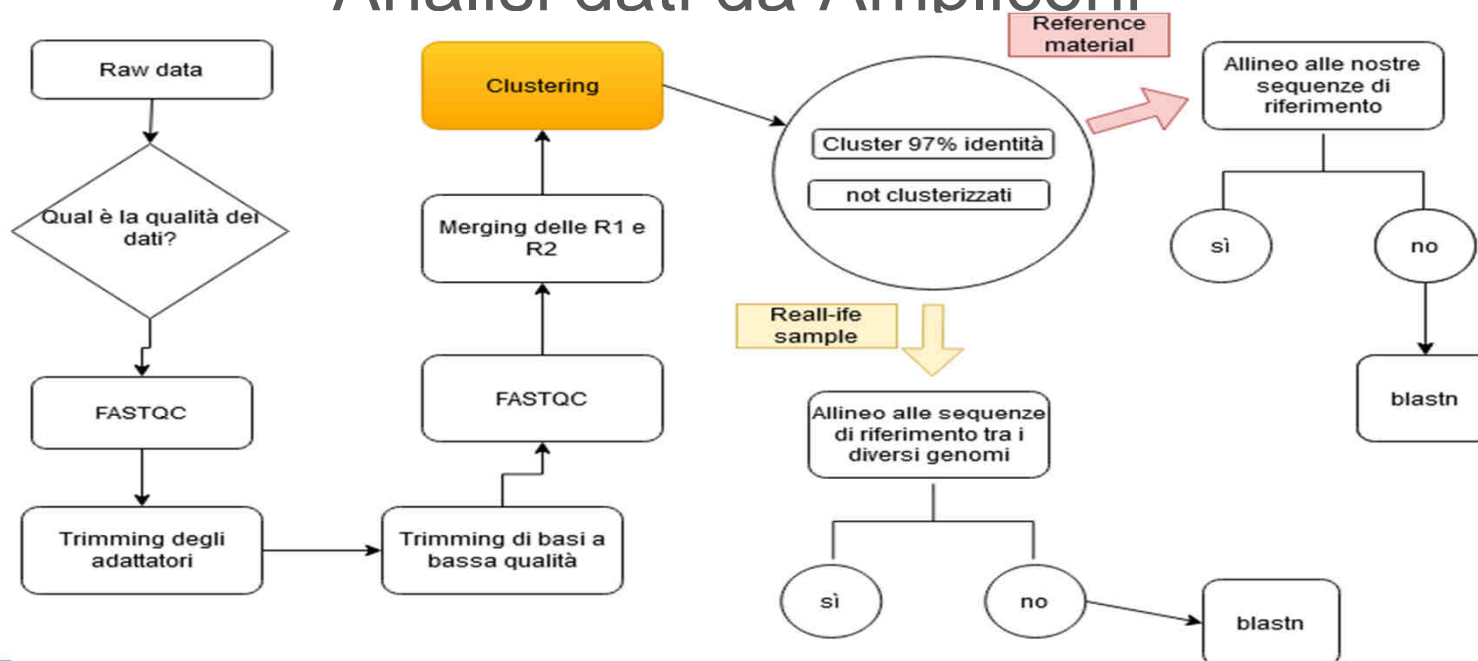
Merging delle reads R1 e R2

✓ **Per base sequence quality**



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi

Clustering

Operazione fondamentale per ridurre i tempi di calcolo.

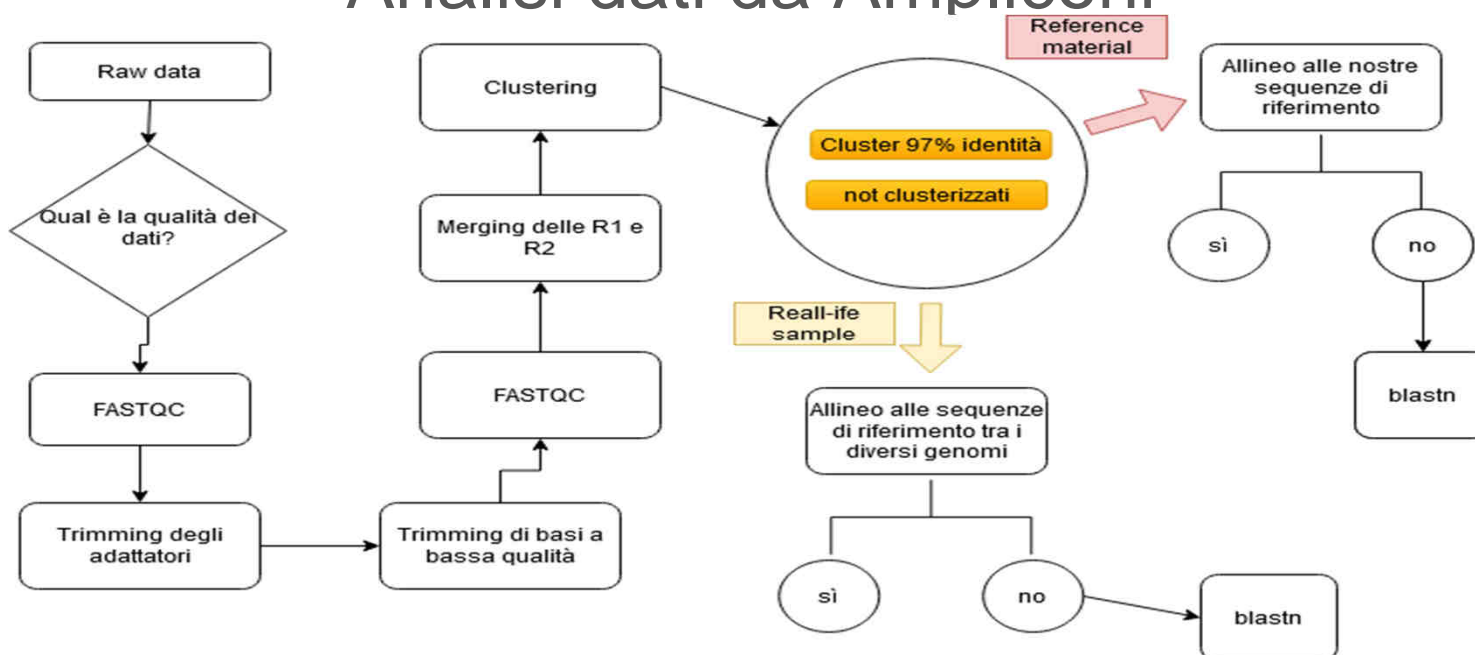
Premesso che esistono diversi algoritmi che utilizzano diversi approcci statistici per clusterizzare le sequenze: Per quanto riguarda l'approccio 'centroid-based'

- Ogni cluster avrà una sequenza rappresentativa che è definita centroide, che sarà quella che meglio rappresenta il cluster
- tutti i membri del cluster devono avere una determinata percentuale di identità



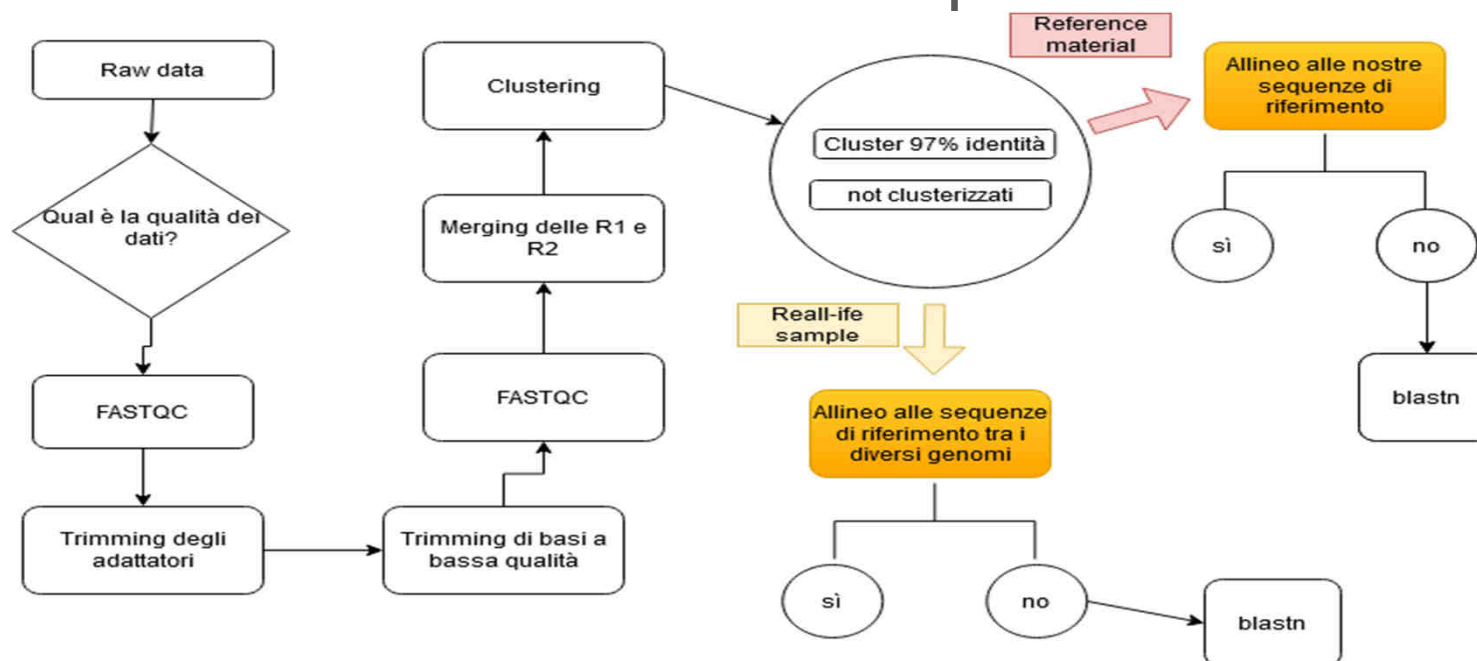
Analisi dei dati di sequenziamento

Analisi dati da Ampliconi



Analisi dei dati di sequenziamento

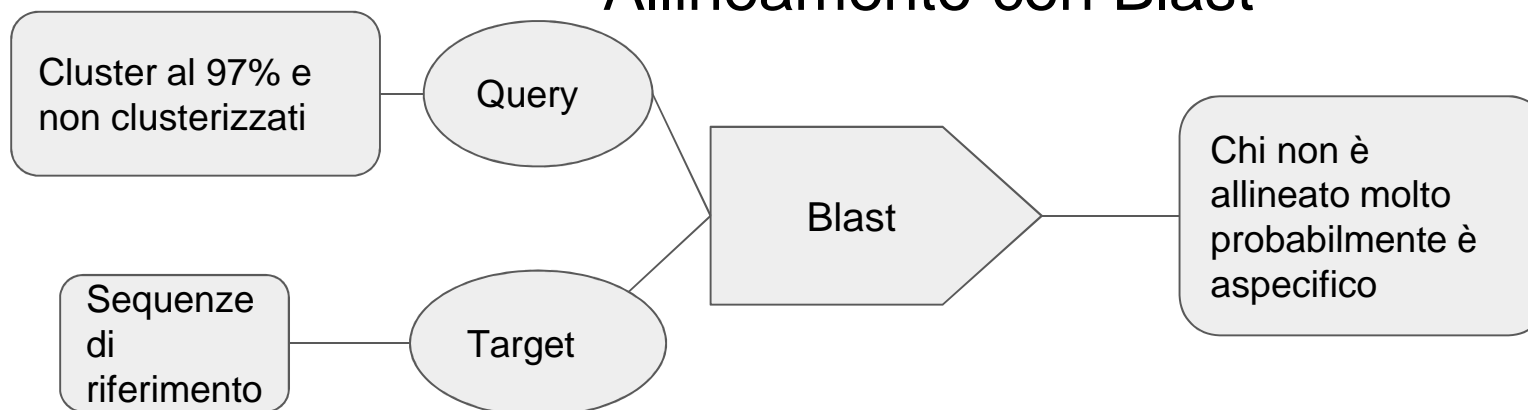
Analisi dati da Ampliconi



Analisi dei dati di sequenziamento

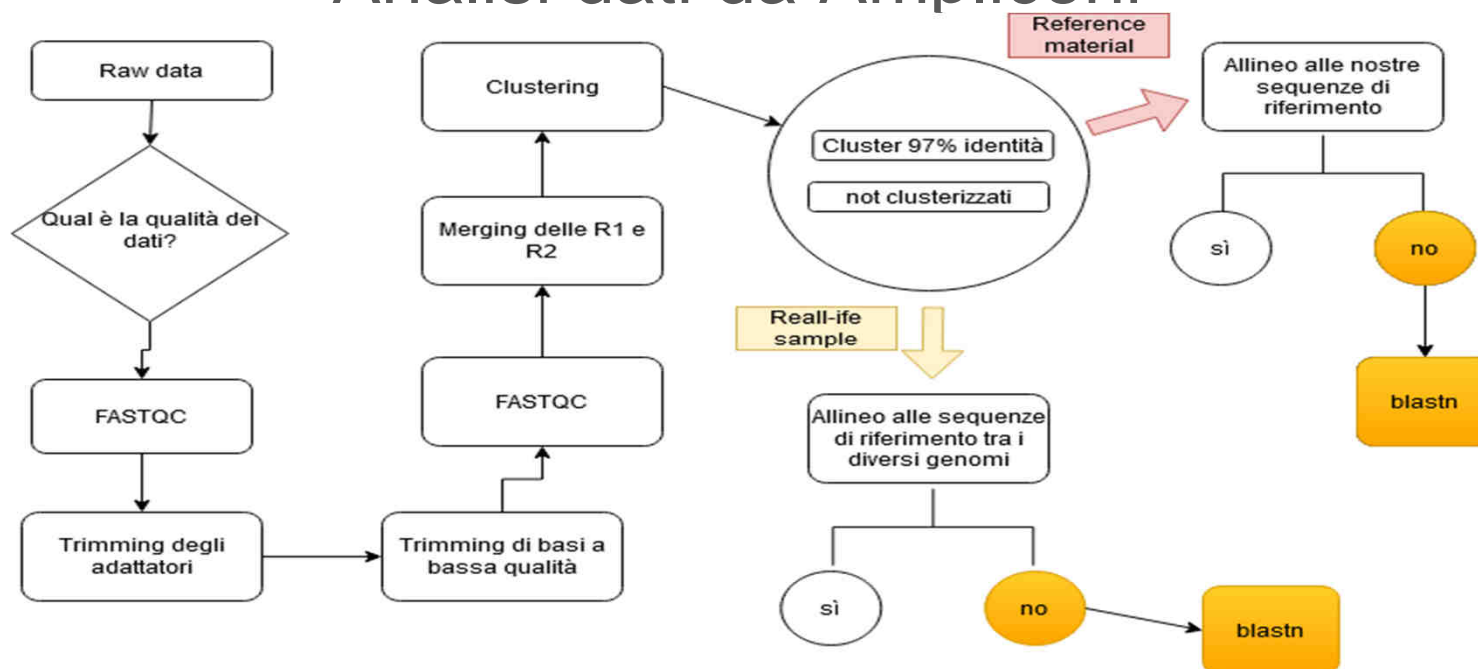
Analisi dati da Ampliconi

Allineamento con Blast



Analisi dei dati di sequenziamento

Analisi dati da Ampliconi



Microrganismi Geneticamente Modificati (MOGM)

☐ Definizione e campo di applicazione

Direttiva 2009/41CE

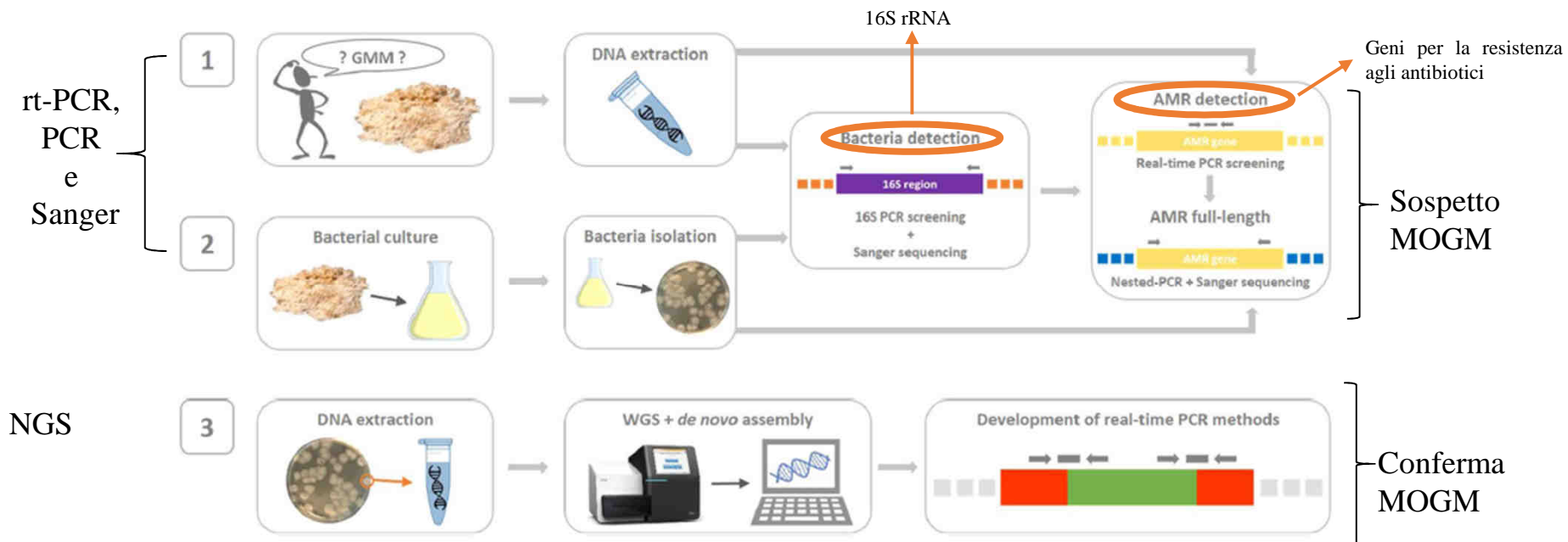
MOGM: un microrganismo il cui materiale genetico è stato modificato in un modo non naturale mediante moltiplicazione e/o ricombinazione naturale;

IMPIEGO CONFINATO: ogni attività nella quale i microrganismi sono modificati geneticamente o nella quale tali MGM sono messi in coltura, conservati, trasportati, distrutti, smaltiti o altrimenti utilizzati, e per la quale vengono usate misure specifiche di contenimento al fine di limitare il contatto degli stessi con la popolazione e con l'ambiente e per garantire a questi ultimi un livello elevato di sicurezza

Utilizzati nei processi industriali (panificazione, produzione della birra ecc) per produrre enzimi, additivi.



Strategia



Fraiture et al,2020, «Identification of an unauthorized genetically modified bacteria in food enzyme through whole genome sequencing». Sci Rep, 10:7094.



Analisi dei dati di sequenziamento

Analisi dati da WGS



Molecular characterization of an unauthorized genetically modified *Bacillus subtilis* production strain identified in a vitamin B₂ feed additive



Valentina Paracchini^{a,1}, Mauro Petrillo^{a,1}, Ralf Reiting^b, Alexandre Angers-Loustau^a, Daniela Wahler^c, Andrea Stolz^c, Birgit Schöning^c, Anastasia Matthies^c, Joachim Bendiek^c, Dominik M. Meinel^d, Sven Pecoraro^d, Ulrich Busch^d, Alex Patak^a, Joachim Kreysa^a, Lutz Grohmann^{c,a}

^aEuropean Commission, Joint Research Centre, Ispra, Italy

^bHessian State Laboratory Kassel (LHL), Kassel, Germany

^cFederal Office of Consumer Protection and Food Safety (BVL), Genetic Engineering Department, Berlin, Germany

^dBavarian Health and Food Safety Authority (LGL), Oberschleissheim, Germany

RESEARCH ARTICLE

Open Access



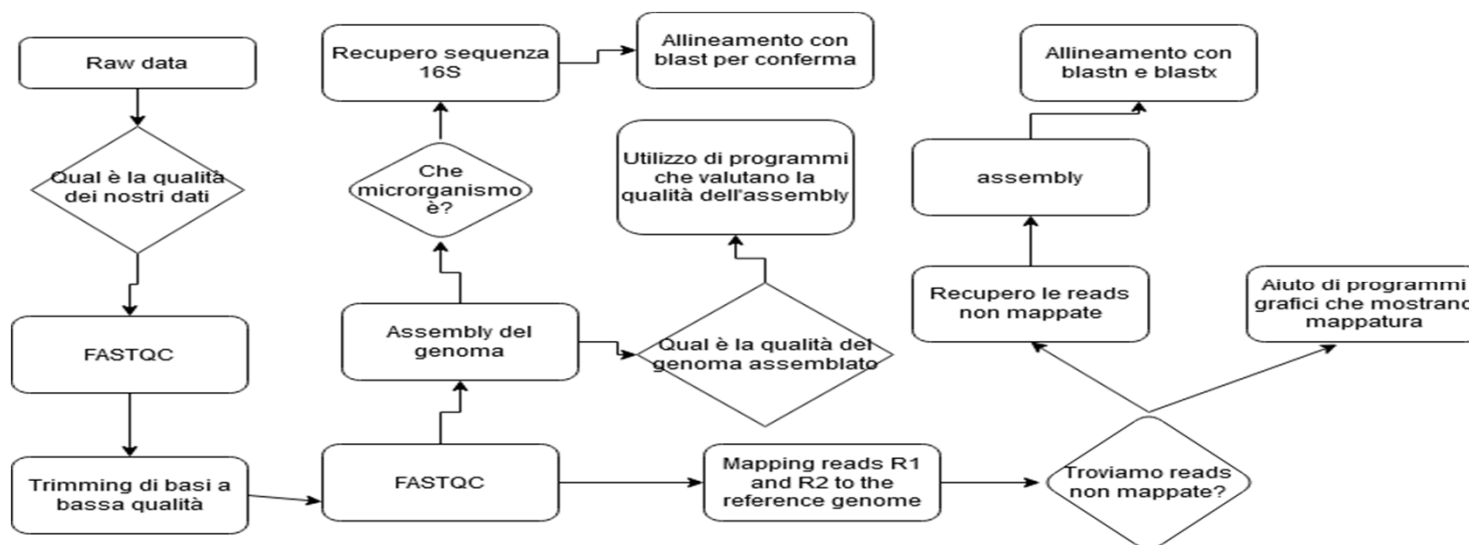
Use of next generation sequencing data to develop a qPCR method for specific detection of EU-unauthorized genetically modified *Bacillus subtilis* overproducing riboflavin

Elodie Barbau-piednoir¹, Sigrid C. J. De Keersmaecker¹, Maud Delvoye¹, Céline Gau², Patrick Philipp² and Nancy H. Roosens^{1*}



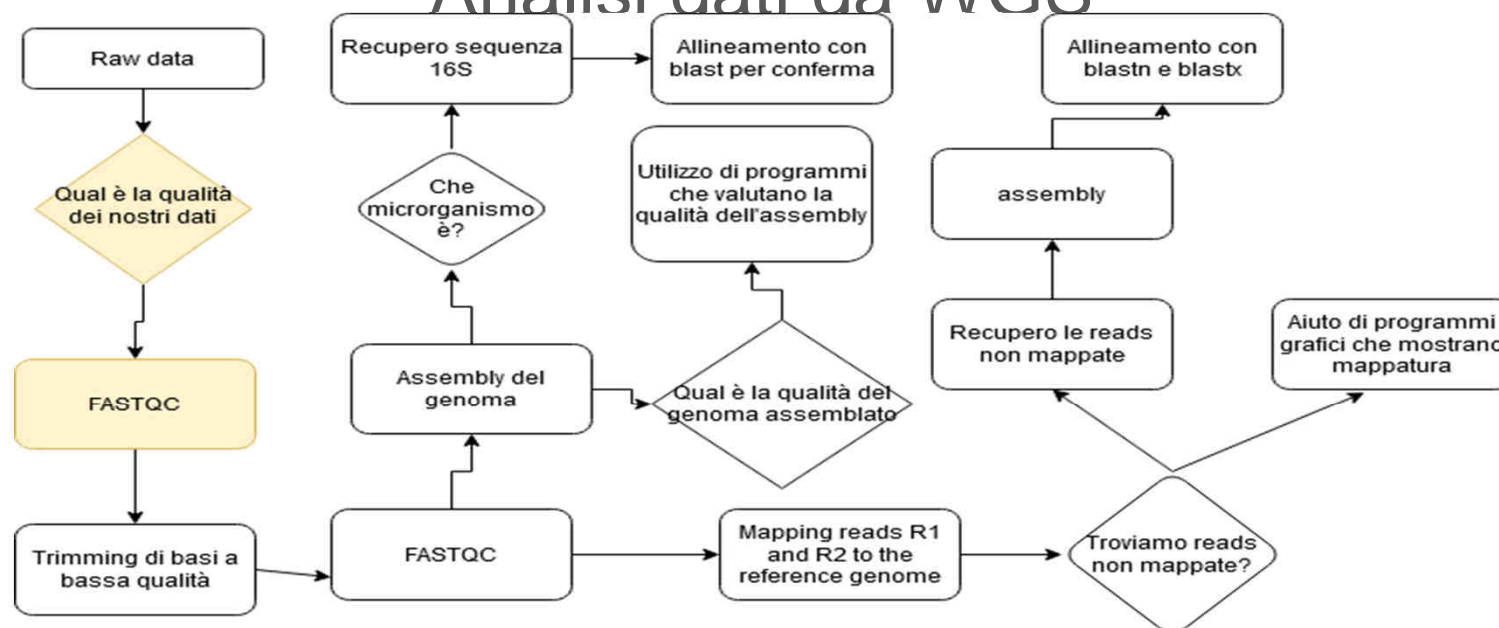
Analisi dei dati di sequenziamento

Analisi dati da WGS



Analisi dei dati di sequenziamento

Analisi dati da WGS



Analisi dei dati di sequenziamento

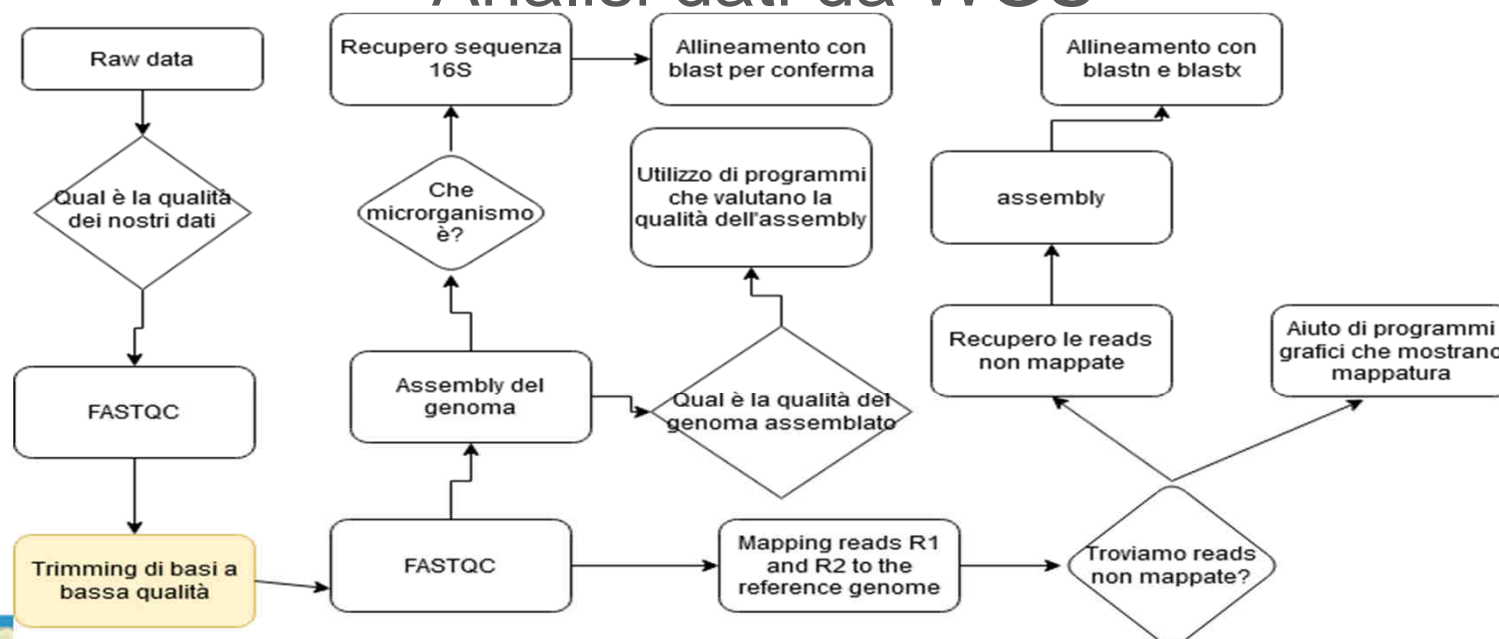
Analisi dati da WGS

Qual è la qualità delle mie reads?



Analisi dei dati di sequenziamento

Analisi dati da WGS

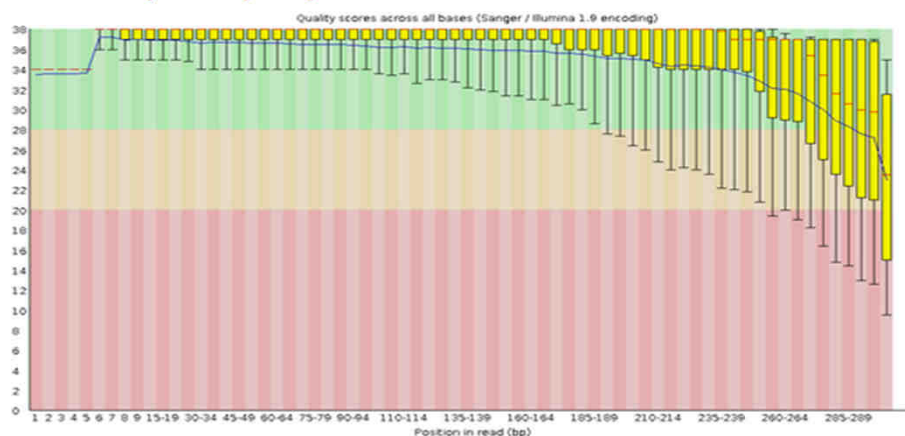


Analisi dei dati di sequenziamento

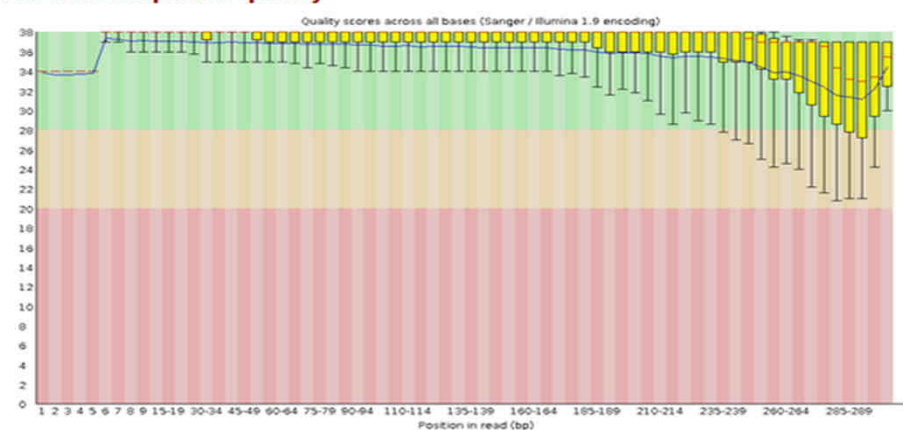
Analisi dati da WGS

Trimming di basi a bassa qualità

❌ Per base sequence quality

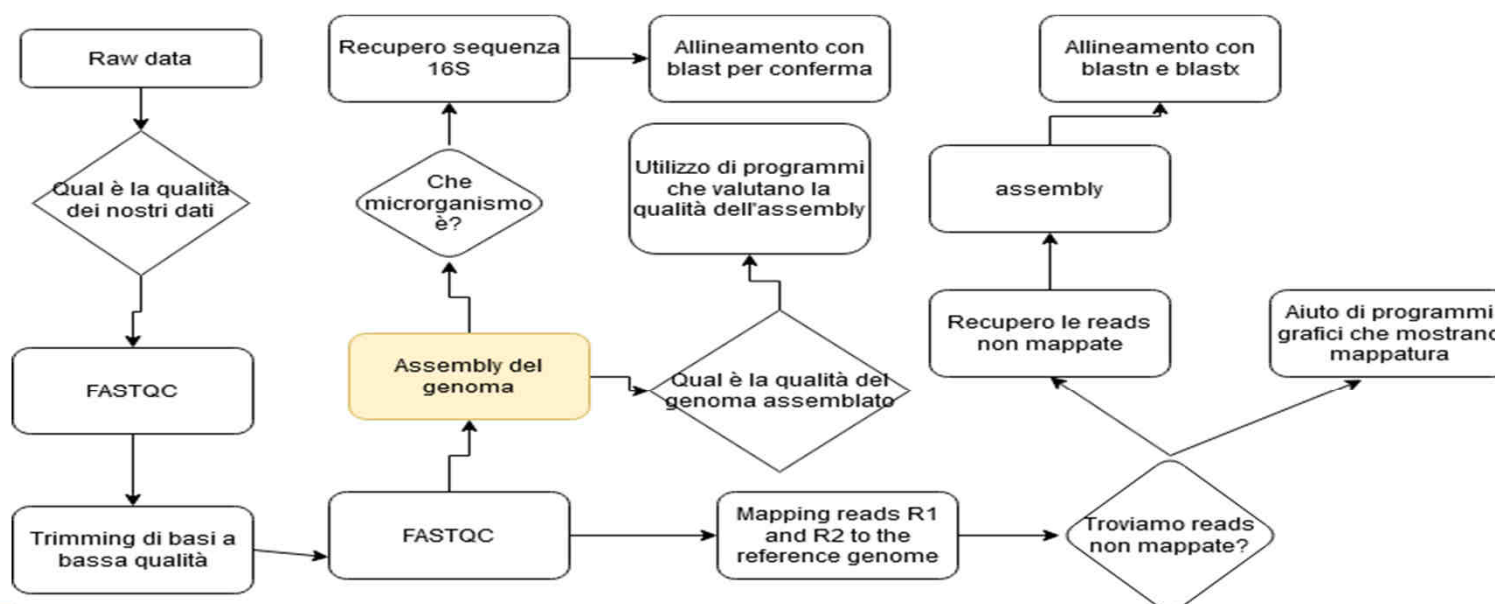


✅ Per base sequence quality



Analisi dei dati di sequenziamento

Analisi dati da WGS

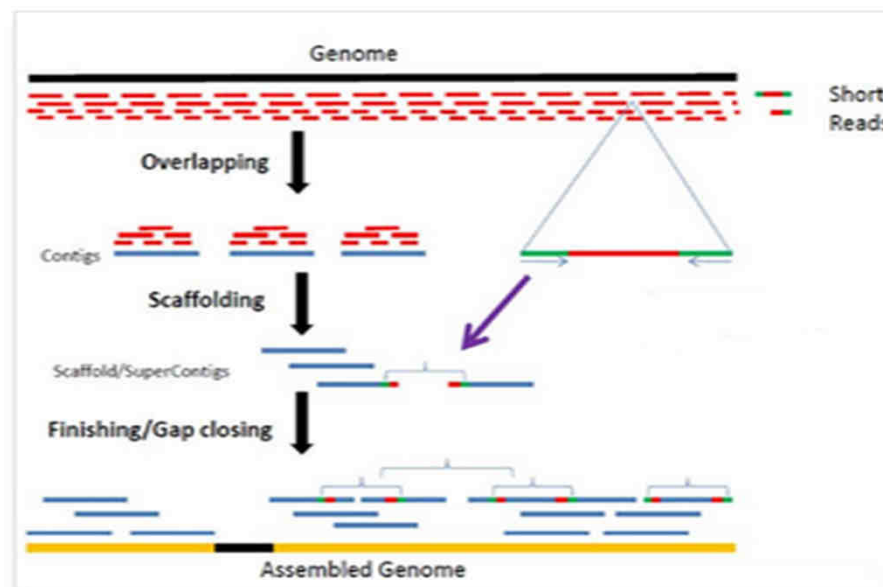


Analisi dei dati di sequenziamento

Analisi dati da WGS

Assemblaggio delle reads

I programmi 'assembler de novo' assemblano sequenze nucleotidiche corte in sequenze più lunghe senza l'uso di un genoma di riferimento. In questo caso è stato utilizzato Spades che è consigliato per genomi di piccola taglia tra cui batteri

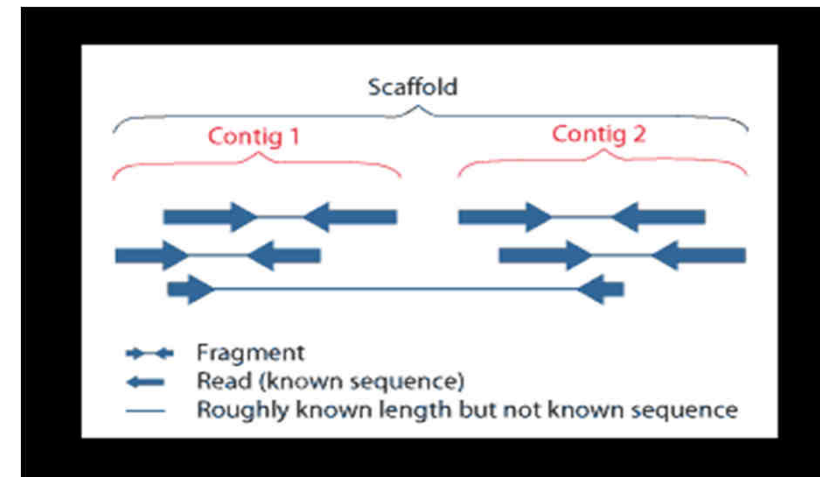


Analisi dei dati di sequenziamento

Analisi dati da WGS

Assemblaggio delle reads

1. Contig: set di segmenti di DNA che sovrappongono e rappresentano quindi una regione *consensus* di DNA
2. Scaffold: raggruppamento di un insieme di contig, cercando di colmare al massimo i gap esistenti.

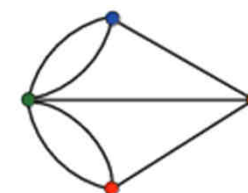
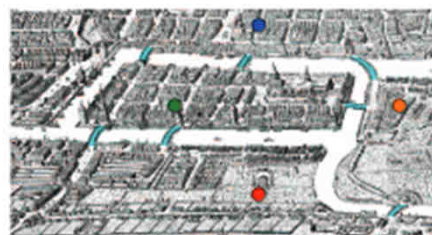


Analisi dei dati di sequenziamento

Analisi dati da WGS

Assemblaggio delle reads/De Bruijn Graph

Città prussiana di Königsberg (l'attuale Kaliningrad, Russia). Il problema dei ponti di Königsberg: i cittadini volevano, attraversare la città passando per tutti i ponti solo una volta, ritornando al punto di partenza. Ma Euler trovò la soluzione!!!



ogni pezzo di terra è un nodo(un punto), ogni ponte è una linea.. Euler ha creato un grafo(una rete di punti collegato da linee)



Analisi dei dati di sequenziamento

Analisi dati da WGS

Assemblaggio delle reads/De Bruijn Graph

Un k-mer è una sequenza di caratteri di lunghezza k, contenuta nelle sequenze biologiche.

Ad esempio: le possibili combinazioni di un 3-mer nucleotidico sono 4^3 possibili combinazioni. Tuttavia in una sequenza di un genoma tali 3-mer possono ripetersi tante volte.



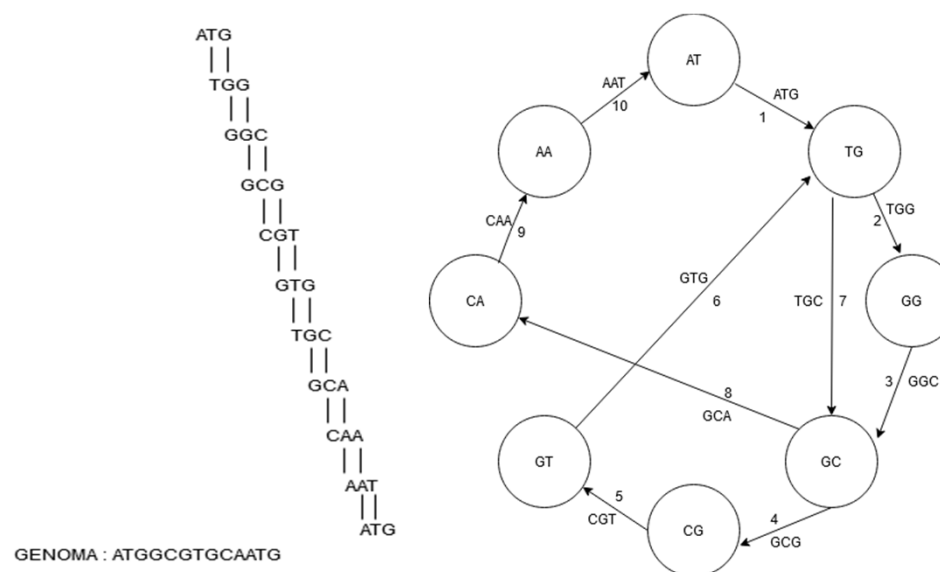
E' compito, di chi utilizza questi approcci, trovare un giusto compromesso tra la lunghezza impostata e la possibilità che si formino tanti pattern che disturbano il risultato finale



Analisi dei dati di sequenziamento

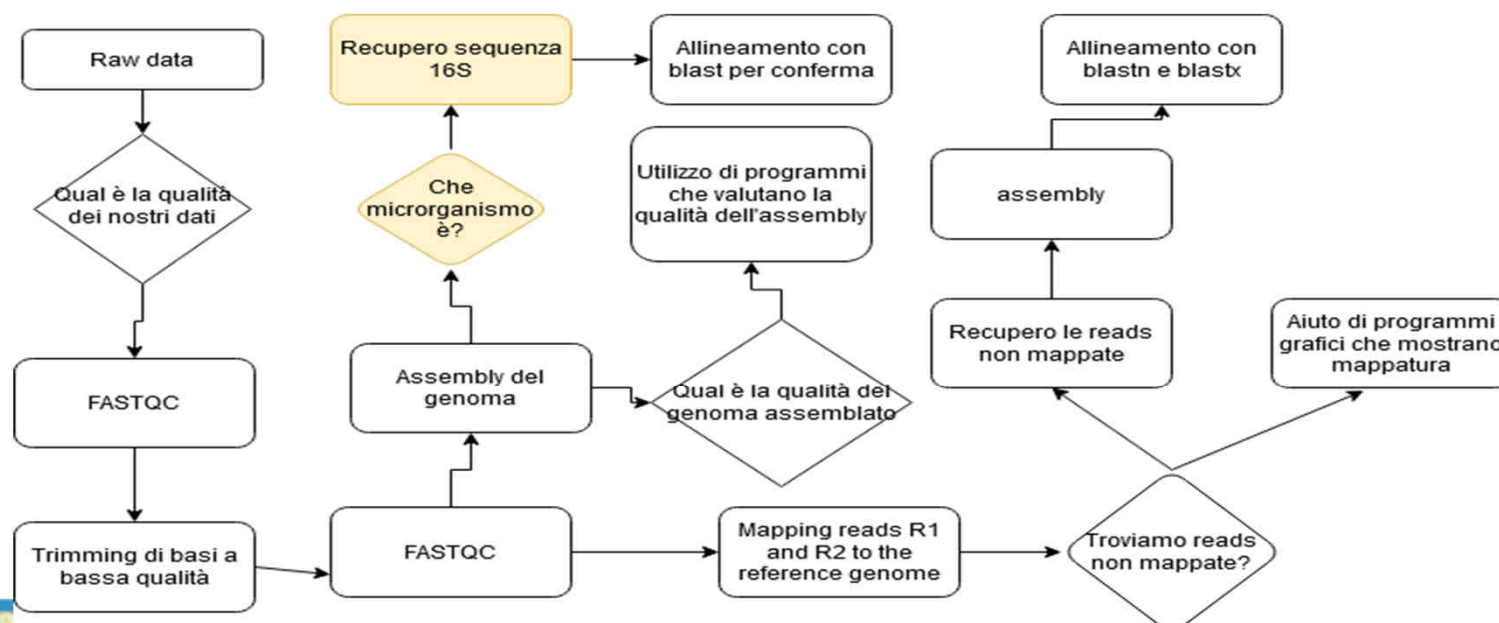
Analisi dati da WGS

Viene assegnato a ciascuno dei k-mer (con $k = 3$ ad esempio) un arco(edge), in primo luogo, e si forma un nodo per ogni prefisso o suffisso distinto di un k-mer, di lunghezza $k - 1$. In questo caso, dai possibili 3-mer della sequenza, i prefissi/suffissi possibili sono AT, TG, GG, GC, CG, GT, CA, AA i quali possono apparire solo una volta come nodo del grafico. Quindi, si collega il nodo x al nodo y con una linea diretta se tra i k-mer ce ne sia qualcuno (ad esempio, ATG) che ha il prefisso x (ad esempio, AT) e il suffisso y (ad esempio, TG), e si etichetta il bordo con questo k-mer.



Analisi dei dati di sequenziamento

Analisi dati da WGS

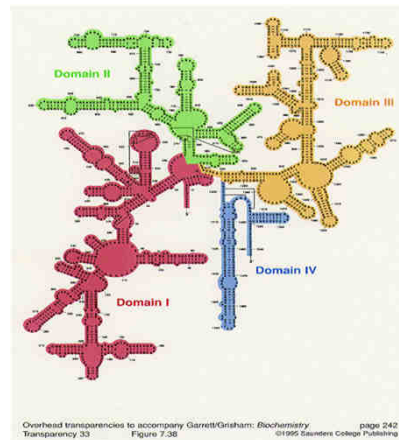


Analisi dei dati di sequenziamento

Analisi dati da WGS

Recupero sequenza rRNA16S

rRNA 16S è una subunità del rRNA 30S procariote. Può essere considerato un identificatore di specie e siccome ha un basso tasso evolutivo è usato per ricostruire filogenie

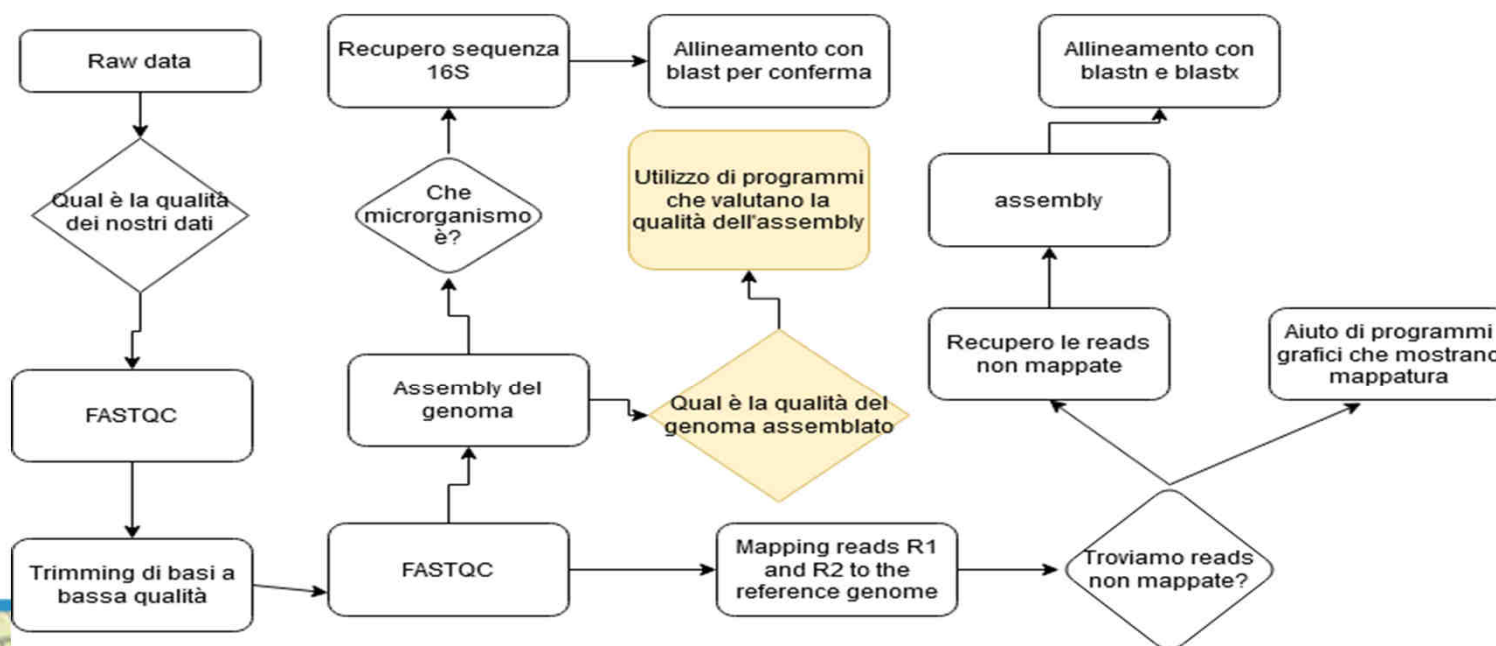


Tramite programmi di predizione più o meno complessi (ad esempio RNAmmer, barnap), si possono recuperare le sequenze ribosomiali tra cui il 16S.



Analisi dei dati di sequenziamento

Analisi dati da WGS



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come valuto la qualità dell'assembly?

Il programma rilascia diversi file di report: In questo file le informazioni da tenere d'occhio sono:

- **Total Length**
- **Largest contig**
- **N50**
- **L50**
- **contig (≥ 0 bp)**

Statistics without reference	scaffolds
# contigs	60
# contigs (≥ 0 bp)	99
# contigs (≥ 1000 bp)	53
# contigs (≥ 5000 bp)	39
# contigs (≥ 10000 bp)	34
# contigs (≥ 25000 bp)	26
# contigs (≥ 50000 bp)	18
Largest contig	423 985
Total length	4 179 972
Total length (≥ 0 bp)	4 191 346
Total length (≥ 1000 bp)	4 175 641
Total length (≥ 5000 bp)	4 146 796
Total length (≥ 10000 bp)	4 114 982
Total length (≥ 25000 bp)	3 983 766
Total length (≥ 50000 bp)	3 704 005
N50	315 503
N75	126 244
L50	6
L75	12
GC (%)	43.44
Mismatches	
# N's	198
# N's per 100 kbp	4.74
Predicted genes	
# predicted genes (unique)	2741
# predicted genes (≥ 0 bp)	2739 + 2 part
# predicted genes (≥ 300 bp)	2645 + 2 part
# predicted genes (≥ 1500 bp)	593 + 2 part
# predicted genes (≥ 3000 bp)	99 + 1 part



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come valuto la qualità dell'assembly?

statistica N50: la qualità dell'assembly in termini di contiguità dato un set di contig, *N50* è definito **la lunghezza di sequenza del contig più piccolo** considerando il **50% della lunghezza totale del genoma**

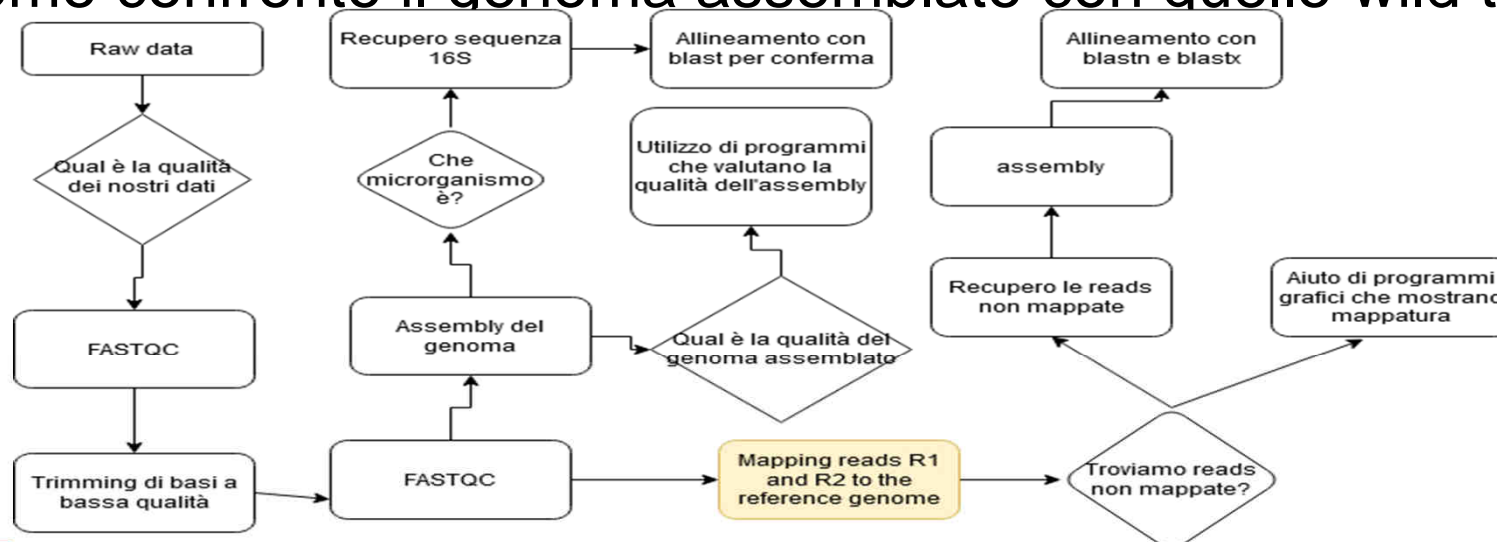
statistica L50: dato un set di contigs, ognuno con la sua propria lunghezza, il conteggio *L50* è definito come **il più piccolo numero di contig, le cui lunghezze sommate raggiungono la metà della dimensione del genoma.**



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come confronto il genoma assemblato con quello wild type?



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come confronto il genoma assemblato con quello wild type?

Utilizzo i programmi di mapping; in questo caso ho utilizzato BWA ("*Burrow Wheeler Aligner*") che ha restituito file .SAM

```

350          SN:NC_000964.3      LN:4215606
APC       ID:bwa PN:bwa      VN:0.7.17-r1188 CL:bwa mem /home/crognm/Scrivania/CA5/wg
S/N_C000964_Bacillus.fa -t 4 /home/crognm/Scrivania/CA5/wgs/trimm_wgs/3558-S-BCL
_S1_L001001_trimm_1P.fastq /home/crognm/Scrivania/CA5/wgs/trimm_wgs/3558-S-BCL_S
1_L001001_trimm_2P.fastq
M03865:13:000000000-J3DTD:1:1101:11562:2280    99      NC_000964.3      1886144
        60      120M      =      1886212 119      CTTGAAAAGAGTAACTGGGTAGACTTTTCCC
ATGAAGATATTGATCTCTCTATTCCCTTTAATTTTTTAGAGAAAAAAGTGGAAGAAACATAAAAAAGATGATGGAATCGCA
TGATACACAA      CCCCCFGGGGFCFGGGGGGGGGGGGGGGGGGGFFGGGGFE@FEGGGGGGGGGGGGFCGGGGGG
GGGGGFGGGGGGGGGG:CFFGGFGCEGGGCGE@CFFGGGGGGGGGGGGGGG8FFGGGG      NM:i:0 MD:Z:12
9      MC:Z:51M      AS:i:120      XS:i:0
M03865:13:000000000-J3DTD:1:1101:11562:2280    147     NC_000964.3      1886212
        60      51M      =      1886144 -119      AGAGGAAAAGTGGGAAAAAAAAAAAAAAGTG
ATGGAAGTCAAAGGATACAA      6<6<@F@C,,<<<@F:@8++@@ECACC69@C,C6C<,;;C,C<C,8-B@ N
4:i:4 MD:Z:19C0T19C1T8      MC:Z:120M      AS:i:31 XS:i:0
M03865:13:000000000-J3DTD:1:1101:18847:2298      99      NC_000964.3      4109740
        60      139M      =      4109822 137      TAGTACAATCATATAAAAAAGAAAGTAAGCGG
ATTTCCTCAGAACCATCTAGAACCAAGCACCAAGAAAGGAGGAAGCTGTTCTGATTGGAGAGCAGCACTGATGTGAAAG
TCATACAGAATTAATTTTGAAGATGTTGTA      CC@CCFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
CFGGDFGGGFGGGGGGGGGGGF@C=-FGFFGFCFFGGGGG@FGFGGGGGGGGGGGGGGGGGBF6GFCEGGGFDDFFGGG9
F9EFF-FCGGEFAE      NM:i:0 MD:Z:139      MC:Z:55M      AS:i:139      XS:i:0
M03865:13:000000000-J3DTD:1:1101:18847:2298      147     NC_000964.3      4109822
        60      55M      =      4109740 -137      GATTGGAGAGAAGAAGCTTGATGTGAAGTGCA
TAGCAATTAATTTTGAAGATGTTGA      C,9@<EC@,<@,<,9<@E9F9<FC;;,;C,E,AFE,<C6,;;,C,C
F8,,,AA      NM:i:2 MD:Z:10C2C41      MC:Z:139M      AS:i:45 XS:i:0
M03865:13:000000000-J3DTD:1:1101:23612:2316      83      NC_000964.3      2820816
        78M      =      282082 -77      GCCCAGACCAAGCACCATACACACATAATCTGCATC

```



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma assemblato con quello wild type?

[illegible]

SAM

Il file .SAM ha un
equivalente in formato
binario che è il file .BAM

[illegible]

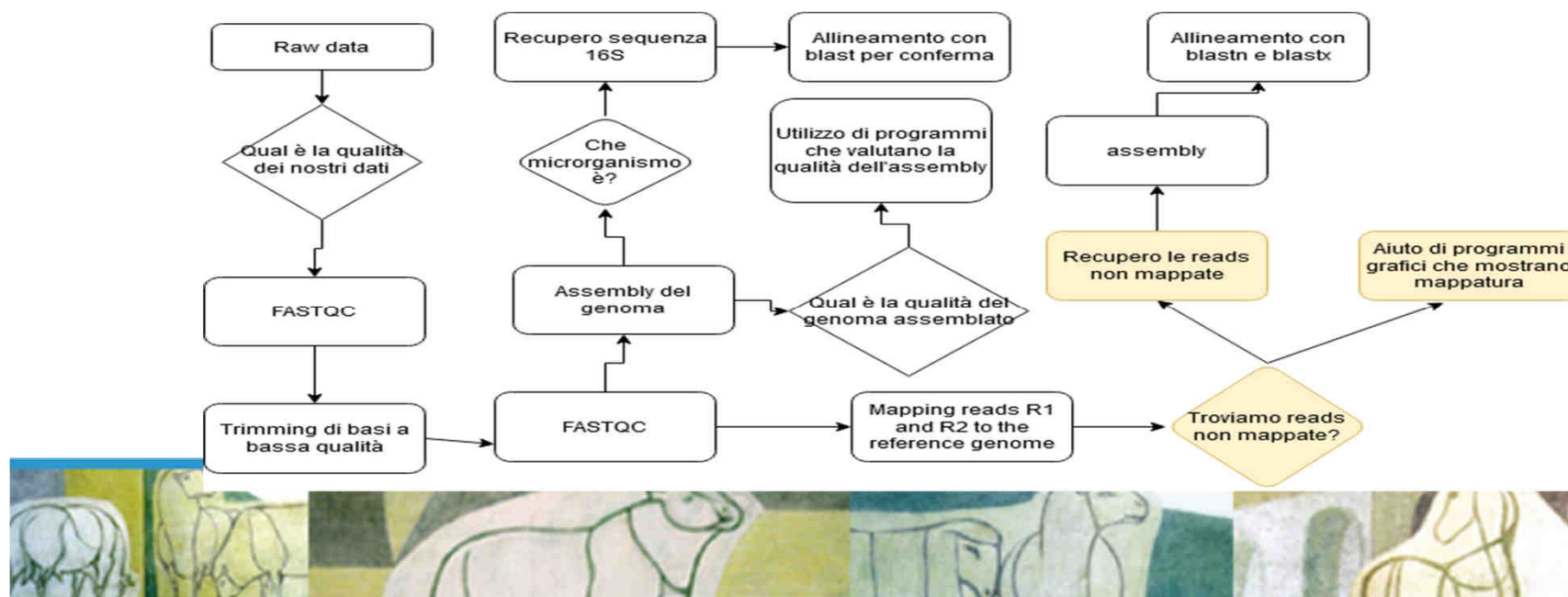
BAM



Analisi dei dati di sequenziamento

Analisi dati da WGS

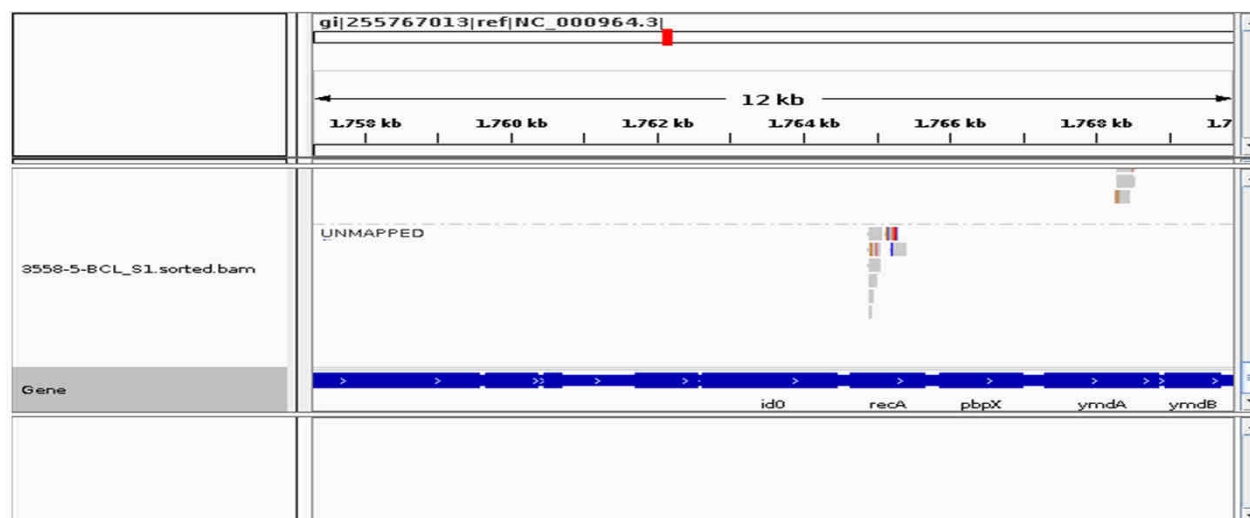
Come trovo la parte che non è del genoma wild type?



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come trovo la parte che non è del genoma wild type?



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

Il file .SAM, oltre all'intestazione, ha almeno 11
campi standard che saranno utili per avere
informazioni per descrivere la mappatura
(allineamento) rispetto alla sequenza di
riferimento

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!~?A-Z]{1,254}	Query template NAME
2	FLAG	Int	[0, 2 ¹⁶ - 1]	bitwise FLAG
3	RNAME	String	* [:rname:^**] [:rname:]	Reference sequence NAME ¹¹
4	POS	Int	[0, 2 ³¹ - 1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0, 2 ⁸ - 1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* [:rname:^**] [:rname:]	Reference name of the mate/next read
8	PNEXT	Int	[0, 2 ³¹ - 1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ + 1, 2 ³¹ - 1]	observed Template LENGTH
10	SEQ	String	* [A-Za-z=]+	segment SEQUENCE
11	QUAL	String	[!~]+	ASCII of Phred-scaled base QUALity+33



Analisi dati da WGS

Come interpreto differenze tra il genoma assemblato con quello wild type?

- FLAG
- POS
- MAPQ
- stringa CIGAR

[illegible]

Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma assemblato con quello wild type?

Il FLAG: Il campo FLAG viene visualizzato come un singolo intero, ma in un file ritroviamo la somma dei valori interi flag bit per bit per denotare più attributi di un allineamento di lettura. Ogni attributo denota un bit nella rappresentazione binaria dell'intero.

Il valore equivale a 2 elevato alla posizione in cui si trova 1 (da destra verso sinistra).

Bitwise Flags		
Integer	Binary	Description (Paired Read Interpretation)
1	000000000001	template having multiple templates in sequencing (read is paired)
2	000000000010	each segment properly aligned according to the aligner (read mapped in proper pair)
4	000000000100	segment unmapped (read1 unmapped)
8	000000001000	next segment in the template unmapped (read2 unmapped)
16	000000010000	SEQ being reverse complemented (read1 reverse complemented)
32	000000100000	SEQ of the next segment in the template being reverse complemented (read2 reverse complemented)
64	000001000000	the first segment in the template (is read1)
128	000010000000	the last segment in the template (is read2)
256	000100000000	not primary alignment
512	001000000000	alignment fails quality checks
1024	010000000000	PCR or optical duplicate
2048	100000000000	supplementary alignment (e.g. aligner specific, could be a portion of a split read or a tied region)



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

Quindi ad esempio: la riga SAM risultante
da un record FASTQ paired Illumina con il
valore FLAG 2145 indicherebbe queste
caratteristiche:

Flag Value	Meaning	Flag Sum
1	read is paired	1
32	read2 was reverse complemented	33
64	read1	97
2048	Supplementary alignment	2145



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

POS: rappresenta la posizione del nucleotide della sequenza di riferimento
(spostata di uno a sinistra) da dove comincia l'allineamento con la read,
descritto dalla stringa CIGAR



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma assemblato con quello wild type?

La stringa **CIGAR** è formata da <intero><op>. Descrive l'allineamento tra read e sequenza di riferimento

Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

Esempio CIGAR:

```
RefPos:    1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19  
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C  
Read:  ACTAGAATGGCT
```



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

Esempio CIGAR: Allineamento della read con la sequenza di riferimento

RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Read:					A	C	T	A	G	A	A		T	G	G	C	T		



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

Esempio CIGAR:

- Il POS indica che la lettura si allinea a partire dalla posizione 5 sul riferimento.
- le prime 3 basi nella sequenza di lettura si allineano con il riferimento.
- La base successiva nella lettura non esiste nel riferimento.
- Altre 3 basi si allineano con il riferimento.
- La base di riferimento successiva non esiste nella sequenza di lettura,
- Altre 5 basi si allineano con il riferimento.

POS: 5
CIGAR: 3M1I3M1D5M



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma
assemblato con quello wild type?

RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Read:					A	C	T	A	G	A	A		T	G	G	C	T		

Alla posizione 14, la base nella lettura è diversa dal riferimento, ma conta ancora come una M poiché si allinea a quella posizione



Analisi dei dati di sequenziamento

Analisi dati da WGS

Come interpreto differenze tra il genoma assemblato con quello wild type?

MAPQ : è un intero che indica la probabilità che la posizione di mapping sia errata.. È uguale a $-10 \log_{10} Pr$, arrotondato all'intero più vicino. Un valore 255 indica che la qualità della mappatura non è disponibile.

Tuttavia i diversi programmi di allineamento utilizzano una propria scala di valori:

- per Bowtie2 il migliore valore di MAPQ è 42
- per Bwa Mem il migliore valore di MAPQ è 60

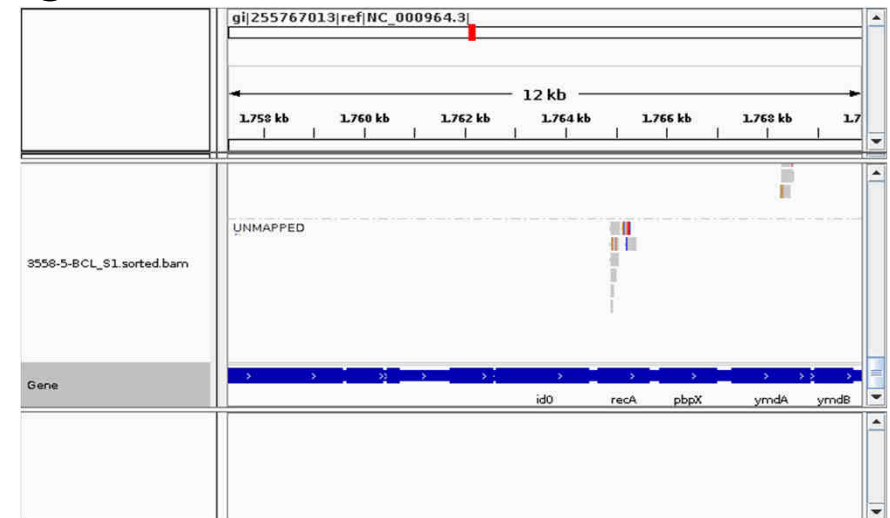


Analisi dei dati di sequenziamento

Analisi dati da WGS

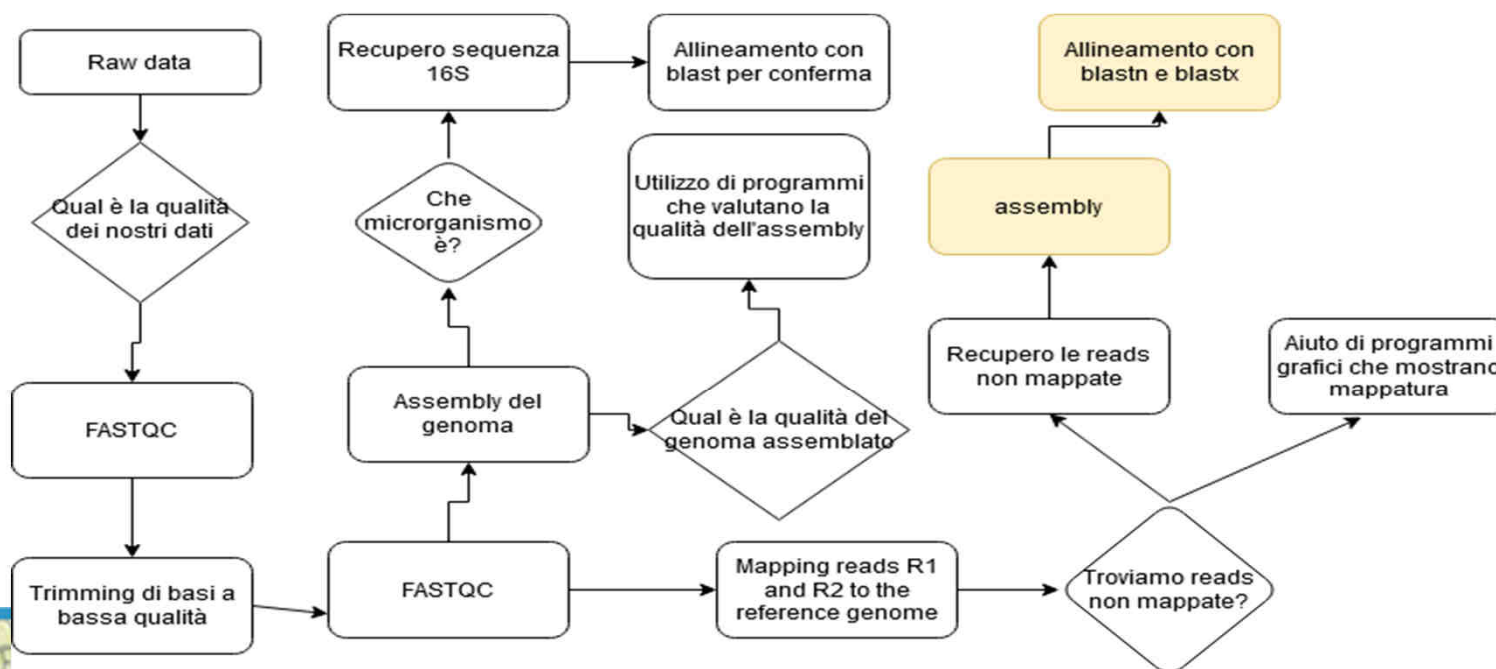
Come trovo la parte che non è del genoma
wild type?

Posso estrarre le reads che non mappano con la sequenza di riferimento tramite il programma Samtools (opera con file tra cui il formato BAM e SAM) e al tempo stesso usare un programma di visualizzazione grafica (IGV in questo caso) in modo tale da poter valutare quelle di mio interesse



Analisi dei dati di sequenziamento

Analisi dati da WGS



Analisi dei dati di sequenziamento

Analisi dati da WGS

Cos'è la parte che non è del genoma wild type?



Analisi dei dati di sequenziamento



Come utilizzo tutti questi programmi ???



Unix

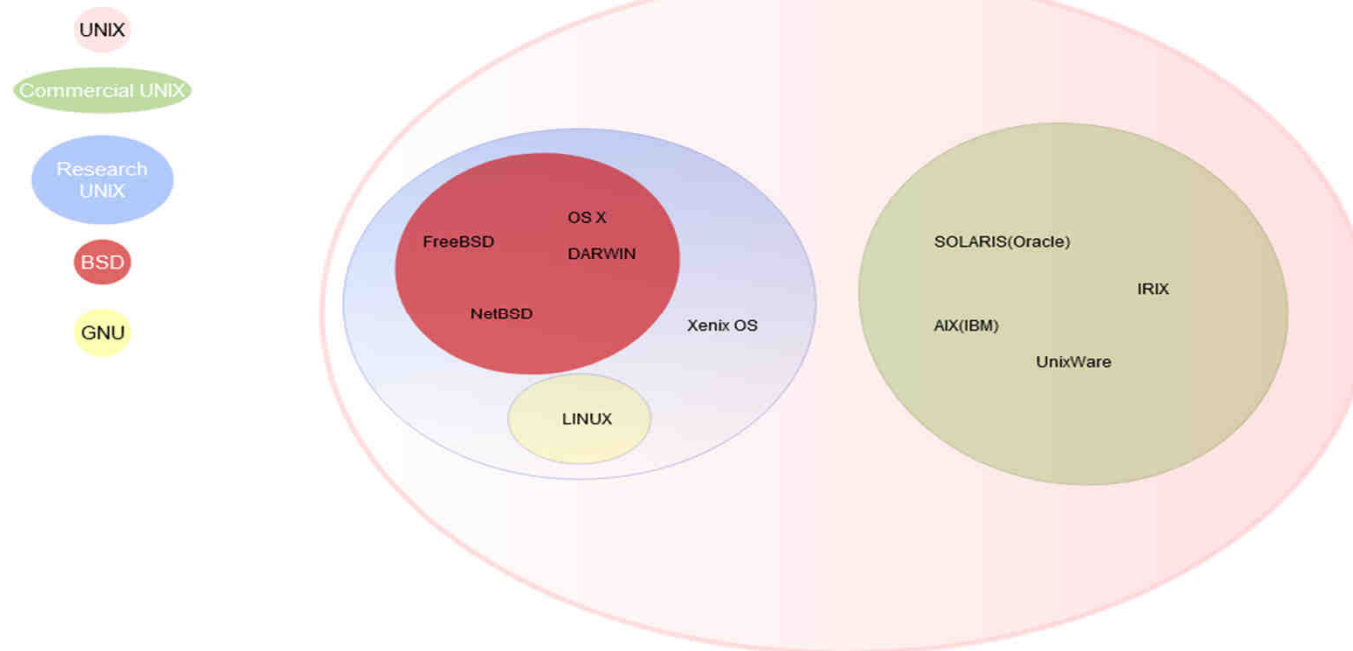
Introduzione e concetti più importanti

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»

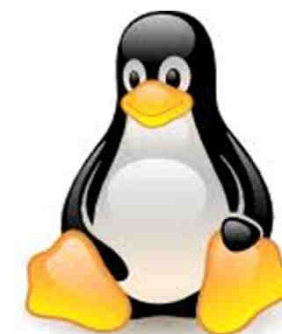
(Richard M. Stallman)



GNU Linux

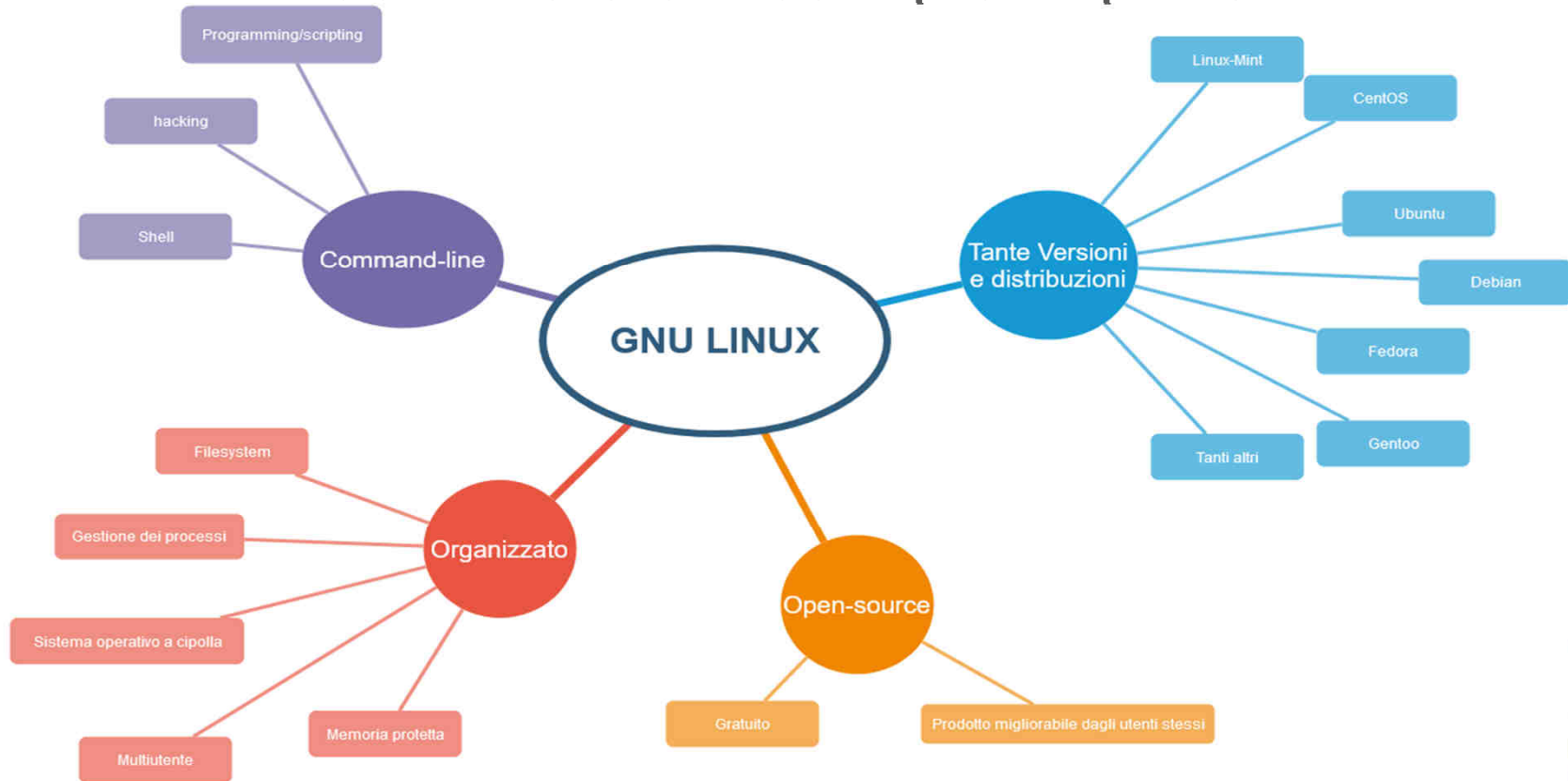
Introduzione e concetti più importanti

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



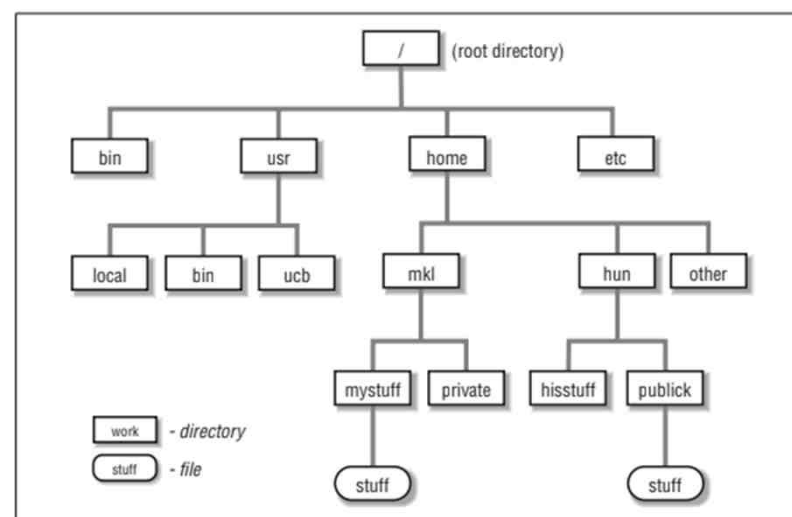
GNU Linux

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

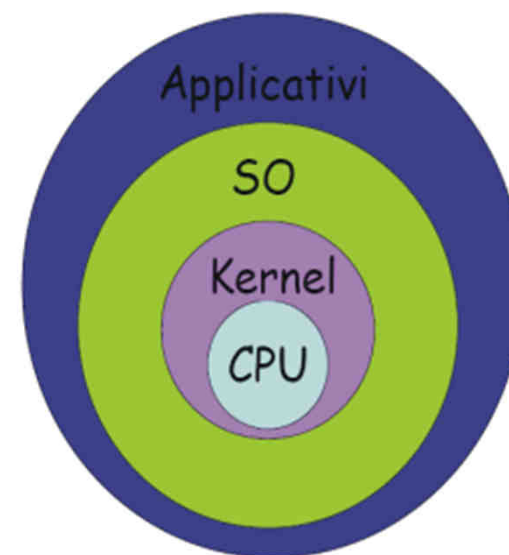
```
crogm@crogm-VirtualBox: ~  
File Modifica Visualizza Cerca Terminale Aiuto  
crogm@crogm-VirtualBox:~$ for i in {1..10};do echo "CIAO AL PARTECIPANTE NUM$i"  
; done; echo "SONO CONTENTO DI VEDERVI, QUI NON CI SONO VESPE";  
CIAO AL PARTECIPANTE NUM1  
CIAO AL PARTECIPANTE NUM2  
CIAO AL PARTECIPANTE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



GNU Linux

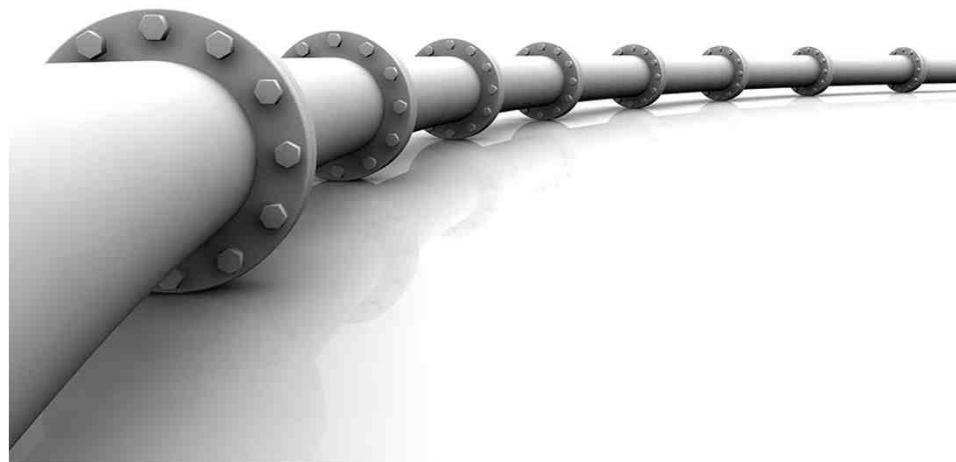
Introduzione e concetti più importanti

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).



Analisi dei dati di sequenziamento

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



Unix/GNU Linux

Introduzione e concetti più importanti

Che cos'è una pipeline?

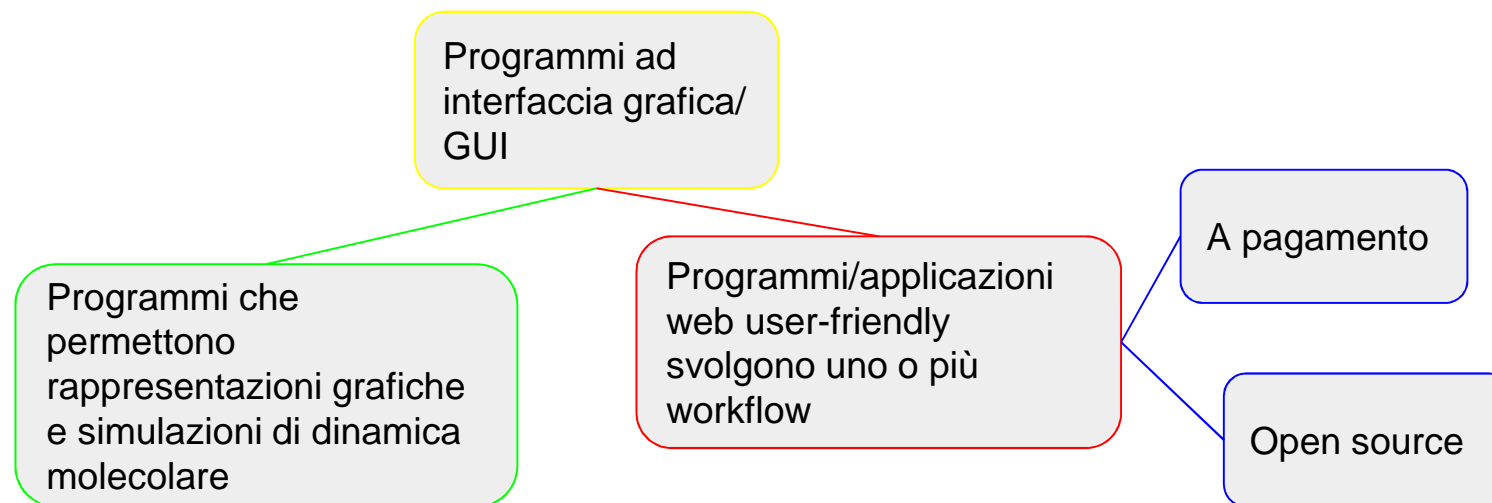
- Questo termine (in inglese *tubatura* — composta da più elementi collegati — o *condotto*) viene utilizzato per indicare un insieme di componenti software collegati tra loro in cascata, in modo che il risultato prodotto da uno degli elementi (output) sia l'ingresso di quello immediatamente successivo (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
```



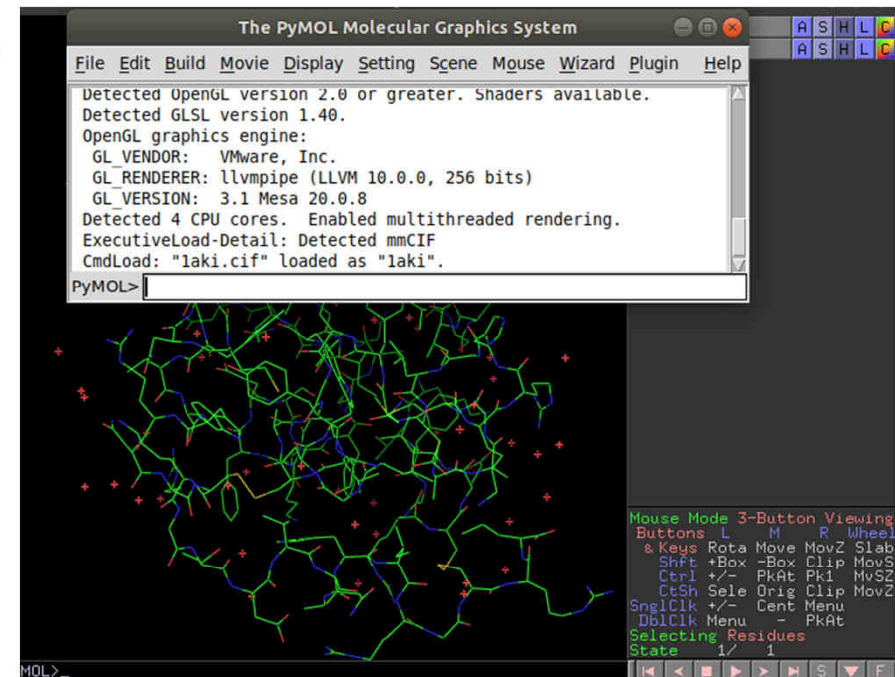
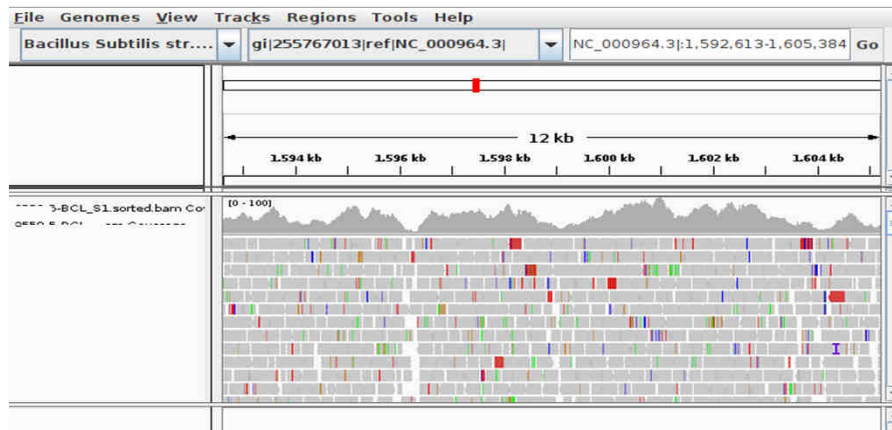
Programmi ad interfaccia grafica

Introduzione e concetti più importanti



Programmi ad interfaccia grafica

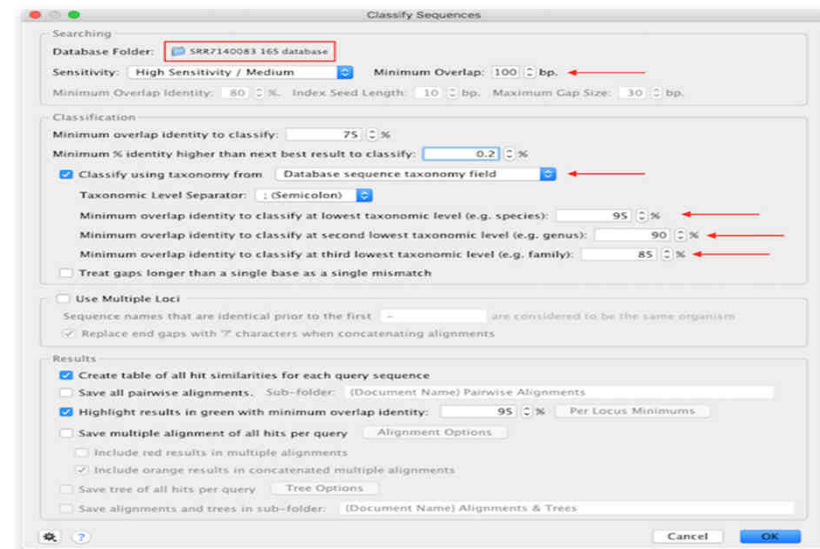
I programmi che permettono rappresentazione grafiche o che eseguono simulazioni hanno un pannello di comando dove si caricano i file di interesse e si impostano i valori e successivamente si vedono i risultati sulla stessa finestra o su altre finestre



Programmi ad interfaccia grafica

Introduzione e concetti più importanti

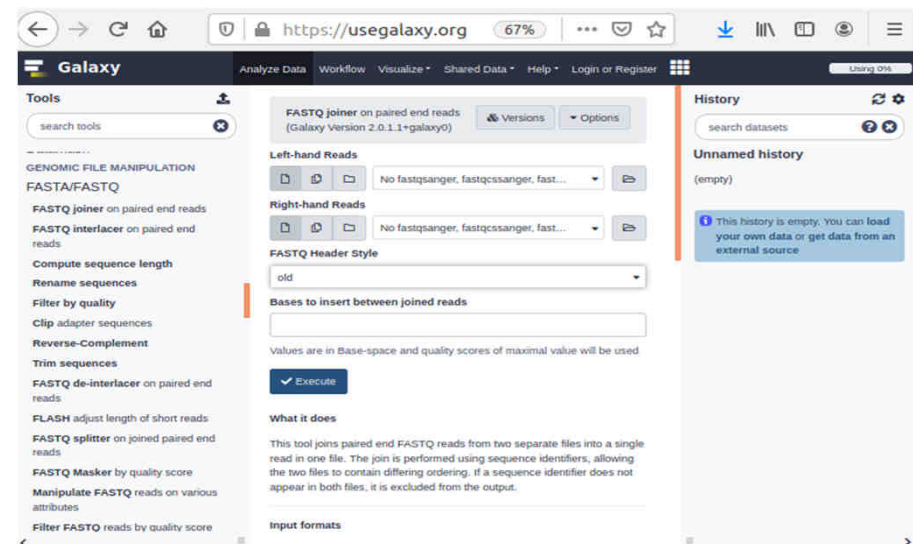
I programmi user-friendly come ad esempio 'Geneius', fanno le stesse cose di una pipeline/ o dell'esecuzione di più programmi da linea di comando. L'utente vede soltanto un pannello su cui imposterà i parametri per eseguire quella determinata operazione.



Programmi ad interfaccia grafica

Introduzione e concetti più importanti

Esistono anche applicazioni web che sono gratuite e user-friendly. In questo caso il vantaggio è poter utilizzare tool progettati da altri ma al tempo stesso si possono creare tool aggiuntivi che possono essere utilizzati da chiunque



Linea di comando o Interfaccia grafica?

Vantaggi della **linea di comando**:

- hai un maggiore controllo sull'esecuzione dei programmi
- strumenti di sviluppo e di analisi del codice; **debugging** di applicazioni; manipolazione di dati tramite la **redirezione** e il piping.

Vantaggi dei **programmi ad interfaccia grafica/GUI**:

- Sono 'user-friendly'
- Vedi solo il risultato



Unix o Interfaccia grafica?

Piccolo esempio pratico: Ricerca del promotore 35S della cassetta transgenica in *Oryza sativa*

```
File Modifica Visualizza Cerca Terminale Aiuto
crogm@crogm-VirtualBox:~$ esearch -query "P35S AND Oryza sativa transgenic cassette" -db "Nuccore" | efetch -format fasta
>JX139719.1 Oryza sativa transgenic GM cassette, partial sequence
ATTTTGGTTTTAGGAATTAGAAATTTATTGATAGAAGTATTTTACAAATACAAATACATACTAAGGGTT
TCTTATATGCTCAACACATGAGCGAAACCCATAAGAACCCTAATTCCTTATCTGGGAACACTACAC
ATTATTATAGAGAGAGATAGATTTGTAGAGAGAGACTGGTGATTTTCAGCGGGCATGCCTGCAGGTCGACT
CAGATCTCGGTGACGGGCAGGACGGACGGGCGGTACCGGCAGGCTGAAGTCCAGCTGCCAGAAACCCA
CGTCATGCCAGTTCCTCGTCTTGAAGCCGGCGCCCGCAGCATGCCCGGGGGGCATATCCGAGCGCCTC
GTGCATGCCAGCCTCGGTCGTTGGGCGAGCCGATGACAGCGACACGCTCTTGAAGCCCTGTGCCTCC
AGGACTTCAGCAGGTGGGTGTAGAGCGTGGAGCCAGTCCCGTCCGCTGGTGGCGGGGGGAGACGTACA
CGGTGCACTCGGCGCTCCAGTCTGAGGCGTTCGCTGCCTTCCAGGGGCCCGCTAGGCGATGCCGGCGAC
CTCGCGTCCACCTCGGCGACGAGCCAGGGATAGCGCTCCCGCAGACGGACGAGGTCGTCCGTCCTCC
TGCGGTTCCTGCGGCTCGGTACGGAAGTTGACCGTCTTGTCTCGGTGTAGTGGTTGACGATGGTGCGA
CCGCCGGCATGTCCGCTCGGTGGCAGGGCGGATGTCCGGCGGGCGTCTTCTGGGCTCATTGCTGGATC
CGGTACCCGTCTCTCCAAATGAAATGAATCTCTTATATAGAGGAAGGCTCTTGGCAAGGATAGTGGG
ATTGTGGTCACTCCCTTACGTCACTGGAGATATCATCAATCCACTTGTCTTGAAGACGTGGTTGGAAC
GTCTTCTTTTCCACGATGCTCCTCGTGGGTGGGGTCCATCTTTGGGACCACTGTGGGCAGAGGCATCT
TCAAC
crogm@crogm-VirtualBox:~$
```

GenBank: JX139719.1

Oryza sativa transgenic GM cassette, partial sequence

FASTA Graphics

Go to: []

LOCUS	JX139719	985 bp	DNA	linear	SYN 09-APR-2015
DEFINITION	Oryza sativa transgenic GM cassette, partial sequence.				
ACCESSION	JX139719				
VERSION	JX139719.1				

gene 209..761
/gene="bar"
/note="phosphinothricin-N-acetyltransferase"

regulatory 779..985
/regulatory_class="promoter"
/note="P35S-CaMV"

ORIGIN

```
1 attttggttt taggaattag aaattttatt gatagaagta ttttacaat acaaatatcat
61 actaagggtt tcttatatgc tcaacacatg agcgaacccc tataagaacc ctaattccct
121 tatctgggaa ctactcacac attattatag agagagatag attgttagag agagactggt
181 gatttcagcg ggcattgcctg caggctcgact cagatctcgg tgacgggacg gaccggacgg
241 ggcgggtacc gacggctgaa gtccagctgc cagaaaccca cgtcatgcca gtccccgtgc
301 ttgaagcgg ccgcccgag catgcccggg ggggcatatc cgagcgcttc gtgcatgcgc
361 acgctcgggt cgttgggag cccgatgaca gcgaccacgc tcttgaagcc cgtgccttc
421 agggacttca gcaggtgggt gtagagcggt gagccagtc ccgtcgctgt gtggcggggg
481 gagacgtaca cggtcgactc ggccgtccag tcgtaggcgt tgcgtgcctt ccaggggccc
541 gcgtaggcga tgccggcgac ctgccgtccc acctcggcga cgagccaggg atagcgctcc
601 cgcagacgga cgaggtcgtc cgtccactcc tgcggttccc gcggtcgtg acggaagtgt
661 accgtgcttg tctcggtgta gtggttgacg atggtgcaga ccgcggcatg gtccgcttcg
721 gtggcacggc ggtatgtcgc cgggcgctgt tctgggtcca ttgctggatc cgtatccctg
781 tctctcccaa atgaaatgaa ctctcttata tagagggaag gtcttgcgaa ggtatgtggg
841 attgtgcgtc atcccttacc tcagttgaga tatcatatca atccacttgc tttgaagacg
901 tggttggaac gtctctcttc tcacgatgac tctctgtggg tgggggttca ctttggggac
961 cactgtcgcc agaggcatct tcaac
```



Database GMO



Database GMO

JRC GMO-Amplicon



Database 2015, 1-11
doi: 10.1093/database/bav101
Original Article

Original Article

JRC GMO-Amplicons: a collection of nucleic acid sequences related to genetically modified organisms

Mauro Petrillo^{*†}, Alexandre Angers-Loustau[†], Peter Henriksson,
Laura Bonfini, Alex Patak and Joachim Kreysa

Molecular Biology and Genomics Unit, Joint Research Centre, European Commission, Ispra, Italy

^{*}Corresponding author; Telephone: +39 0332 786232; Fax: +39 0332785483; Email: mauro.petrillo@ec.europa.eu

[†]These authors contributed equally to this work.

Citation details: Petrillo, M., Angers-Loustau, A., Henriksson, P. et al. JRC GMO-Amplicons: a collection of nucleic acid sequences related to genetically modified organisms. *Database* (2015) Vol. 2015: article ID bav101; doi:10.1093/database/bav101

Può essere interrogato tramite 3 interfacce per l'utente:

- Amplicon finder (interrogare il database riguardo le caratteristiche dell'amplicone)
- Search by target ID (controllare sequenze target dalle banche dati pubbliche)
- Blast amplicons (effettuare ricerche per similarità contro dataset di JRC GMO-Amplicons)



Database GMO

Euginius

La sua banca dati contribuisce all'applicazione della legislazione dell'UE sugli organismi geneticamente modificati (OGM). L'attenzione si concentra sugli OGM non autorizzati a livello mondiale che non sono ancora sicuri per l'uomo, gli animali e l'ambiente



Database GMO

BCH

Il Biosafety Clearing-House (BCH) è un meccanismo istituito dal Protocollo di Cartagena sulla biosicurezza per facilitare lo scambio di informazioni sugli organismi viventi modificati (LMO) e aiutare le Parti a rispettare meglio i loro obblighi ai sensi del Protocollo. L'accesso globale a una varietà di informazioni scientifiche, tecniche, ambientali, legali e di rafforzamento delle capacità è fornito nelle sei lingue ufficiali delle Nazioni Unite.



Grazie per l'attenzione!

