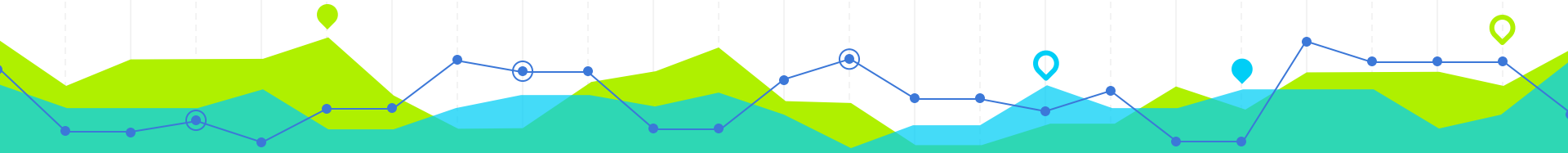
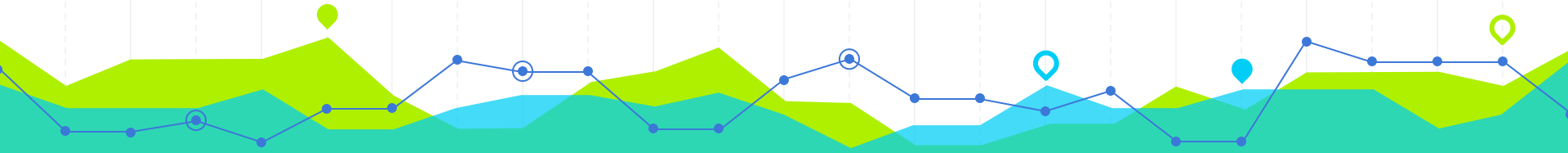
BowBin



Pipeline per l’estrazione di virus da dati metagenomici correlati ad IBS

Alessandro Aquino & Nicolapio Gagliarde



Introduzione

Pillole teoriche

1

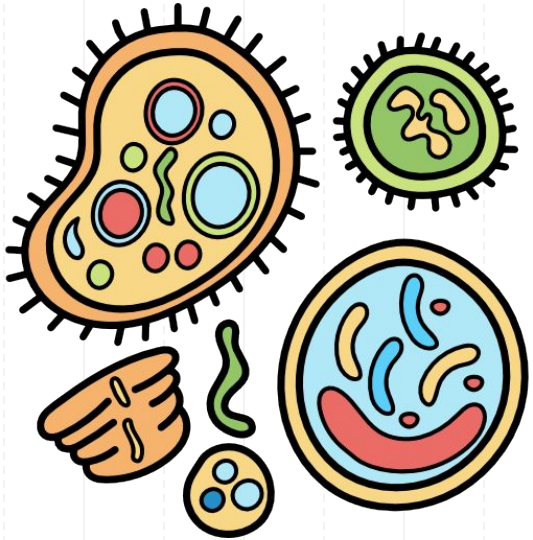
Introduzione

La sindrome dell'intestino irritabile (IBS) è un disturbo gastrointestinale prolungato e invalidante con un tasso di incidenza del 11% nel mondo.

Tra le possibili cause: stress, dieta, alterazioni del **microbioma intestinale**

Introduzione

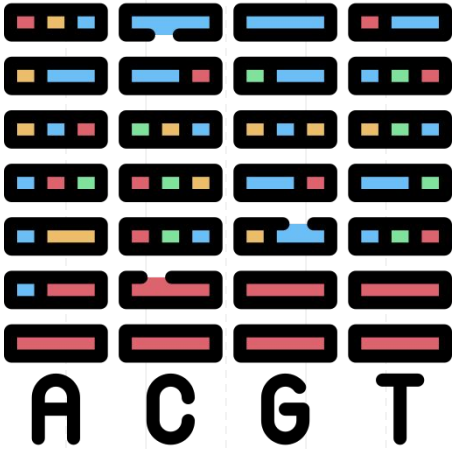
La metagenomica è lo studio del materiale genetico recuperato direttamente da campioni ambientali o clinici mediante un metodo chiamato sequenziamento



Introduzione

L’**allineamento di sequenze** è una procedura bioinformatica con cui vengono messe a confronto ed allineate due o più sequenze di aminoacidi, DNA o RNA.

Utilizzato da BowBin per allineare reads metagenomiche con scaffolds virali.



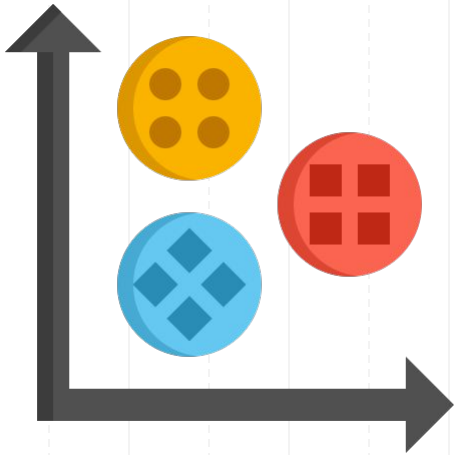
Introduzione

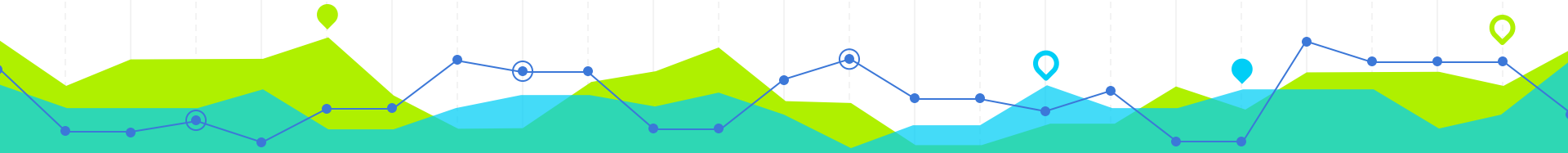
La **coverage table** è caratterizzata, oltre ai dati degli scaffolds, come : nome, lunghezza … anche dalla media e dalla deviazione standard per ogni file.

scaffoldName	scaffoldLen	totalAvgDepth	file1.sort.bam	file1.sort.bam.var	file2.sort.bam	file2.sort.bam.var
nome1	|||||||	M1 + M2	M1	V1	M2	V2
nome2	|||||||	M'1 + M'2	M'1	V'1	M'2	V'2
nome3	|||||||	M"1 + M"2	M"1	V"1	M"2	V"2
nome4	|||||||	M""1 + M""2	M""1	V""1	M""2	V""2

Introduzione

Il **binning** è il processo di raggruppamento di reads, contigs o scaffolds e di assegnazione a singoli genomi .

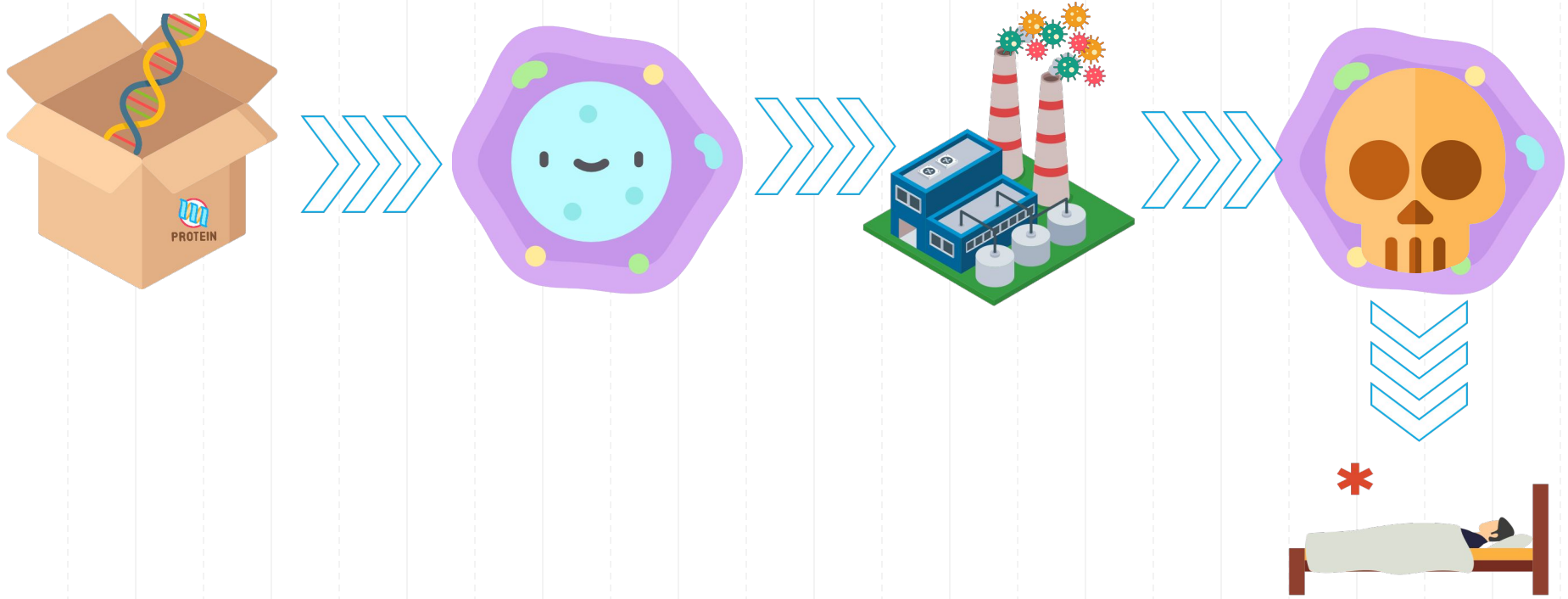




Problema 2

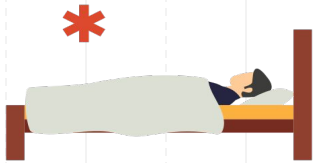
Scendiamo nel dettaglio

PROBLEMA



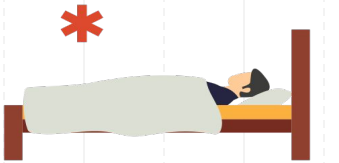
PROBLEMA

virus è un "pacchetto" di geni racchiuso in una scatola di proteine



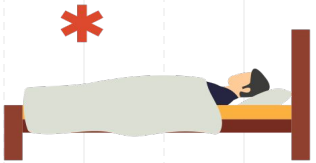
PROBLEMA

Cellula ospitante del virus



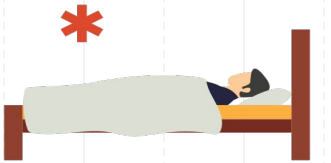
PROBLEMA

Prende il controllo e la cellula diventa una fabbrica di virus

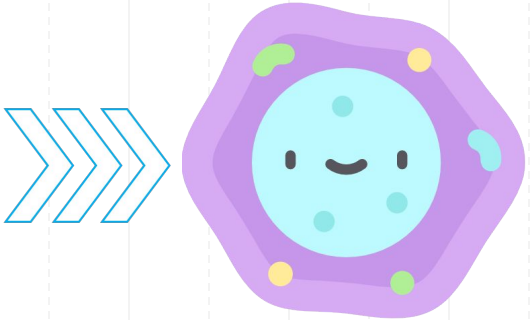


PROBLEMA

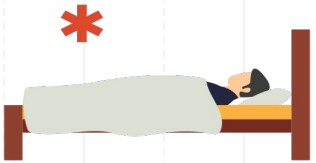
La cellula muore



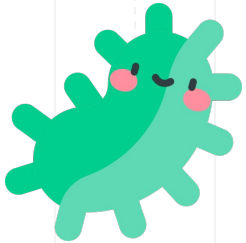
PROBLEMA

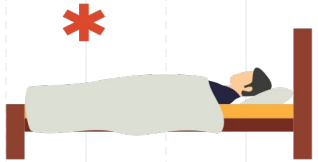


Si genera una malattia



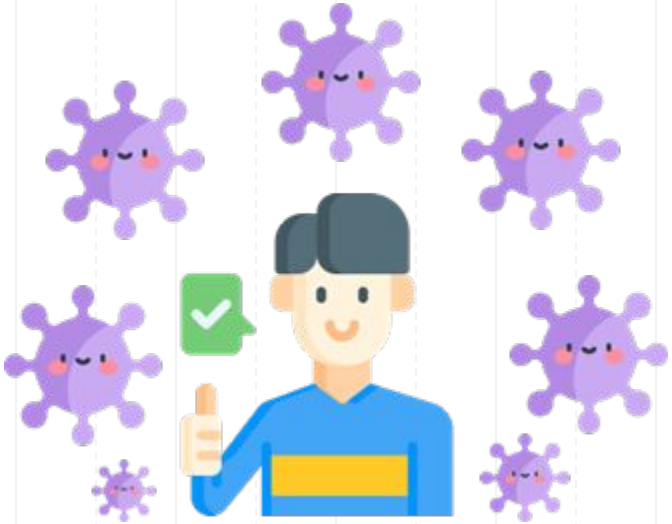
PROBLEMA



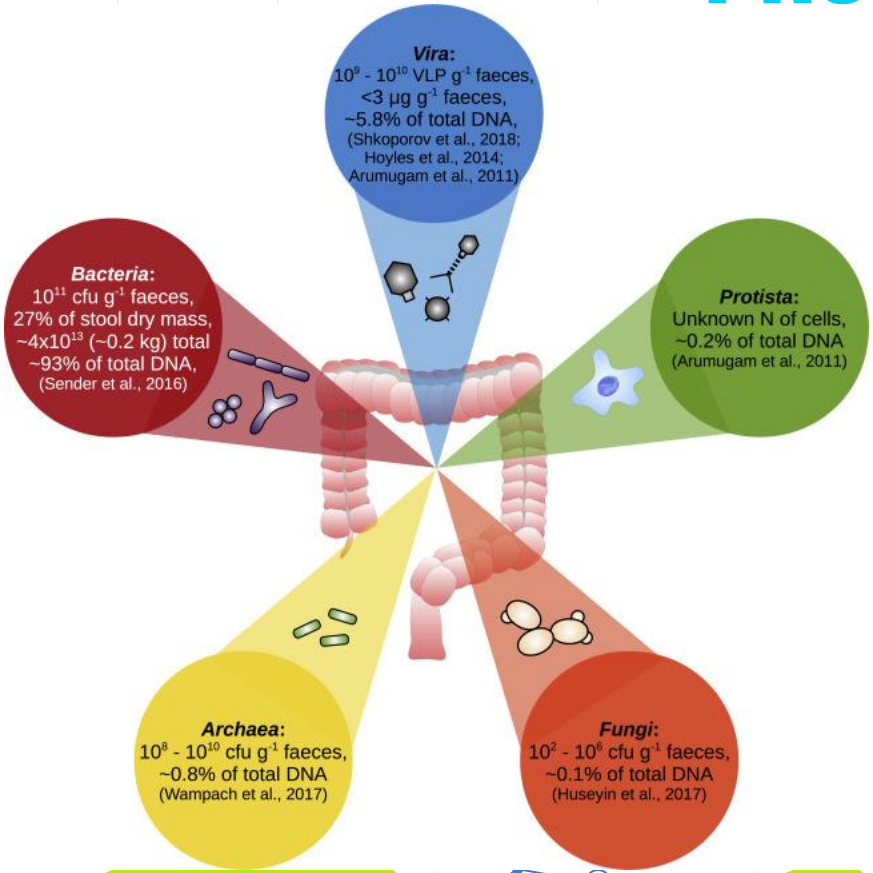


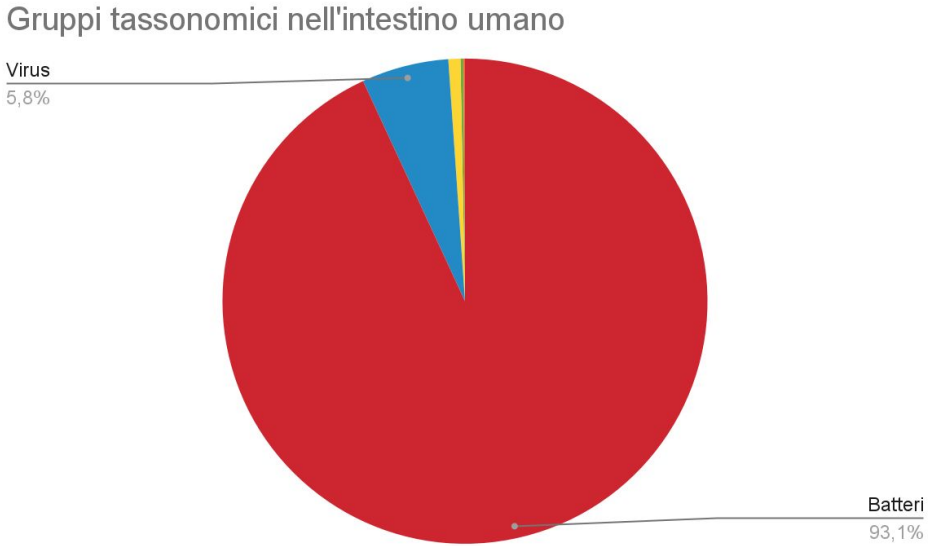
Batteriofagi

OPPURE



PROBLEMA





cfu g^{-1} indica milioni di unità che formano colonie per grammo
VLP g^{-1} indica milioni di particelle virus-like per grammo

OBIETTIVO

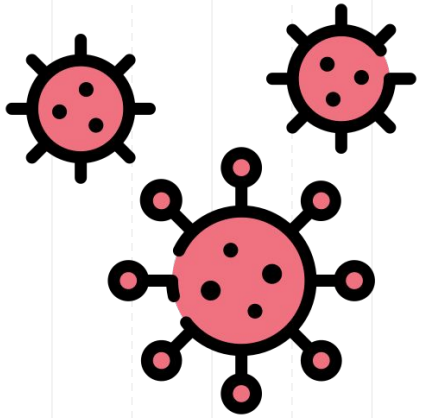
Pertanto abbiamo deciso, partendo da un metagenoma, di estrarne i virus e classificarli mediante dei tools che formeranno la nostra pipeline: BowBin.

Soluzione 3

Tools

SOLUZIONE

Nel corso degli anni il machine learning ha contribuito allo sviluppo della classificazione dei virus, mediante l'utilizzo di tools specifici al problema.

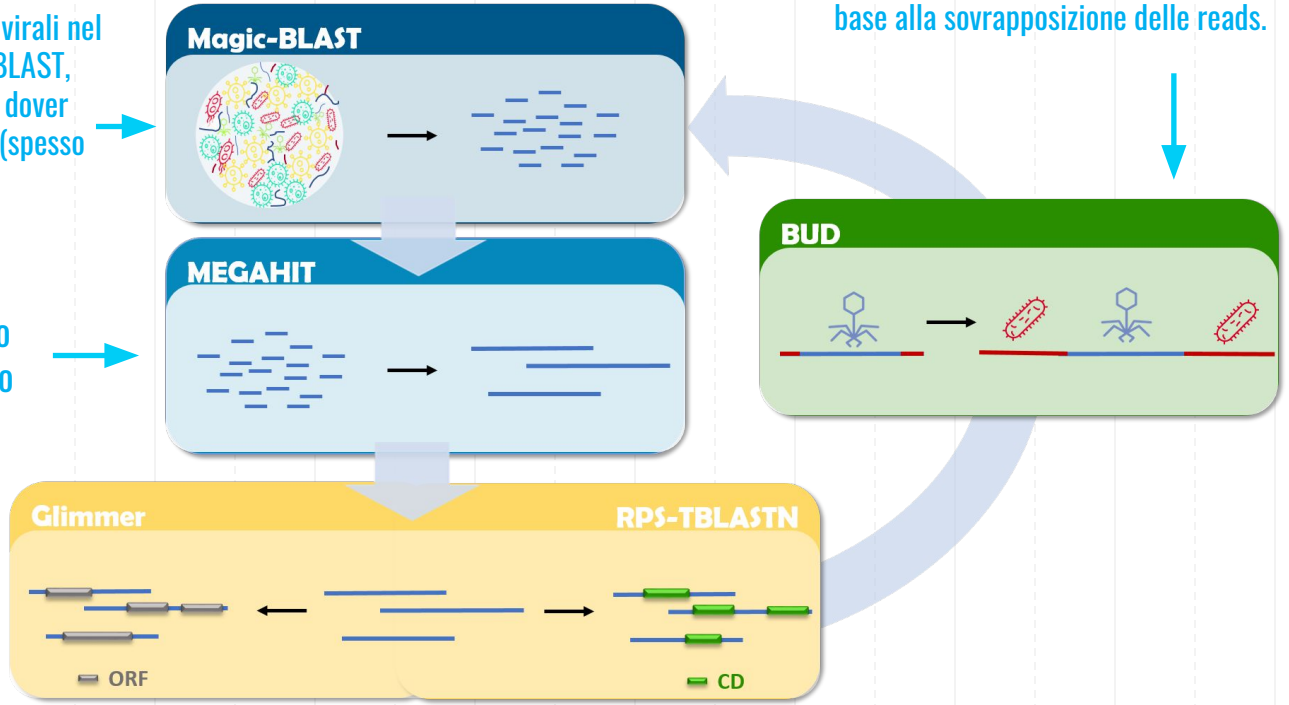


ViruSpy

Il primo passaggio identifica le reads virali nel campione metagenomico con Magic-BLAST, che consente questo passaggio senza dover scaricare il set di dati metagenomici (spesso piuttosto grandi).

Le reads grezze estratte vengono assemblate in contigs utilizzando MEGAHIT.

Una volta assemblate sono annotate per geni da Glimmer.

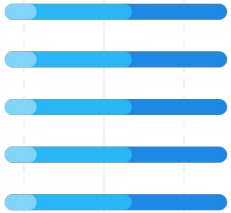


Dopo l'annotazione, l'algoritmo Building Up Domains (BUD) consente di estendere i contigs in base alla sovrapposizione delle reads.

Bowtie2

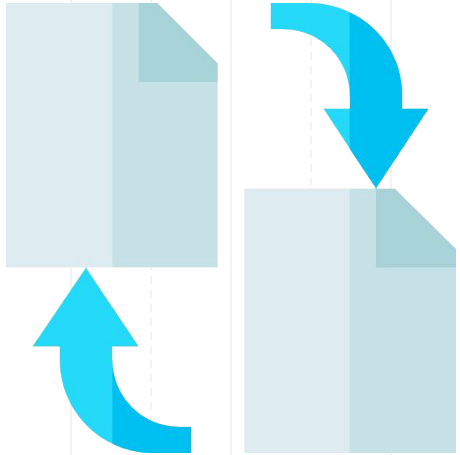
Bowtie 2 è un tool ultraveloce ed efficiente in termini di memoria per allineare le reads di sequenziamento a lunghe sequenze di riferimento, nel nostro caso a scaffolds.

Prende in input un indice Bowtie2 e una serie di reads del sequenziamento e restituisce una serie di allineamenti in formato SAM.



SAMtools

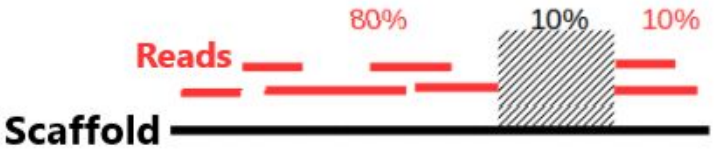
SAMtools è un insieme di utilità per l'interazione e la post-elaborazione di allineamenti di reads di brevi sequenze di DNA nei formati SAM (Sequence Alignment/Map) e BAM (Binary Alignment/Map).



Metabat

Metabat è pensato per il raggruppamento di grandi frammenti genomici assemblati da sequenze metagenomiche consentendo lo studio dei singoli organismi e delle loro interazioni.

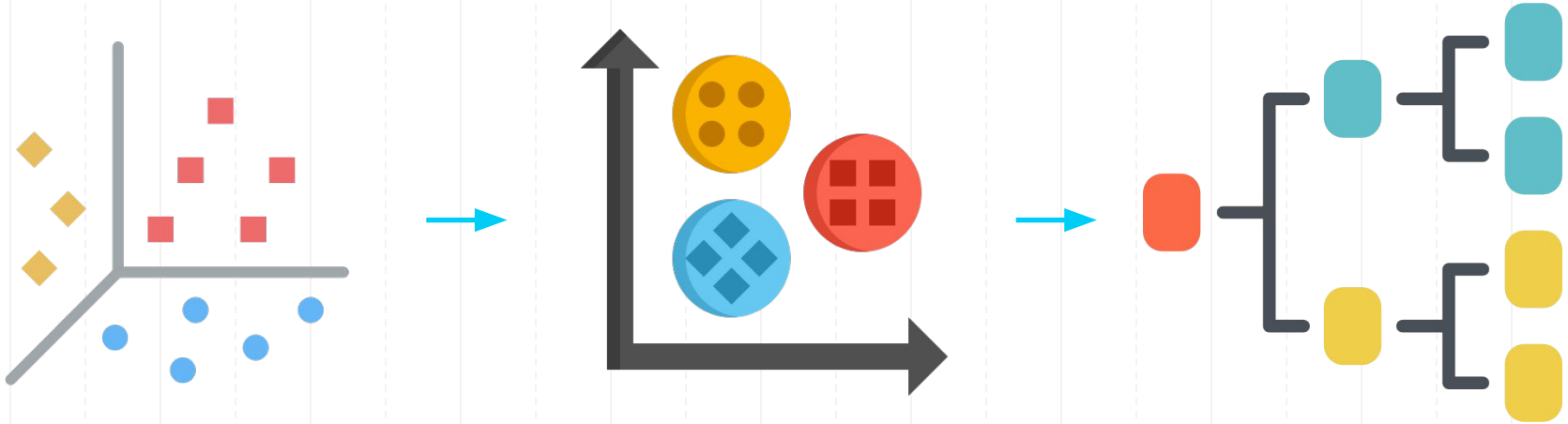
Nel nostro caso Viene utilizzato "jgi_summarize_bam_contig_depths " di Metabat2 che, dato in input il file BAM ordinato, genera la coverage table e la salva nel file depth.txt

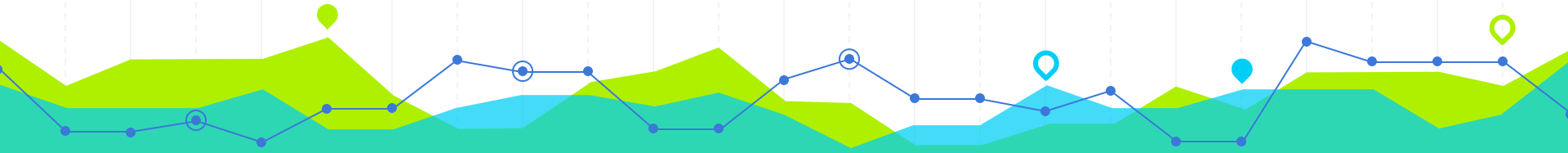


$$C = \frac{\text{\# Area coperta dalle reads}}{\text{\#Scaffold area}}$$

vRhyme

vRhyme è un tool per il binning di genomi dei virus partendo da metagenomi.



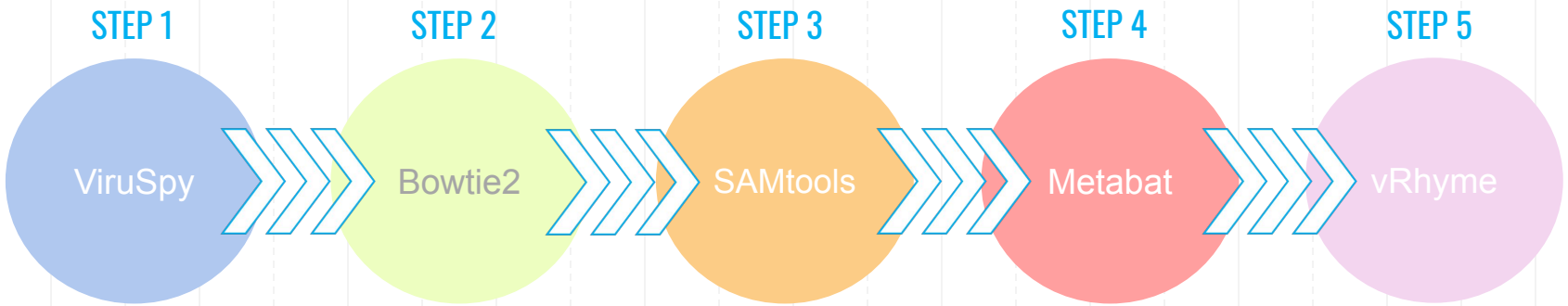


Pipeline

BowBin

4

Pipeline


STEP 1
ViruSpy
STEP 2
Bowtie2
STEP 3
SAMtools
STEP 4
Metabat
STEP 5
vRhyme

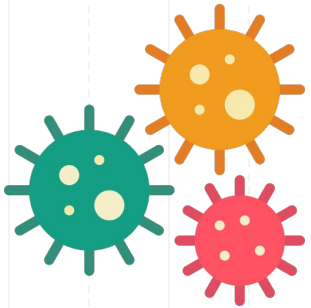
Pipeline

Metagenoma

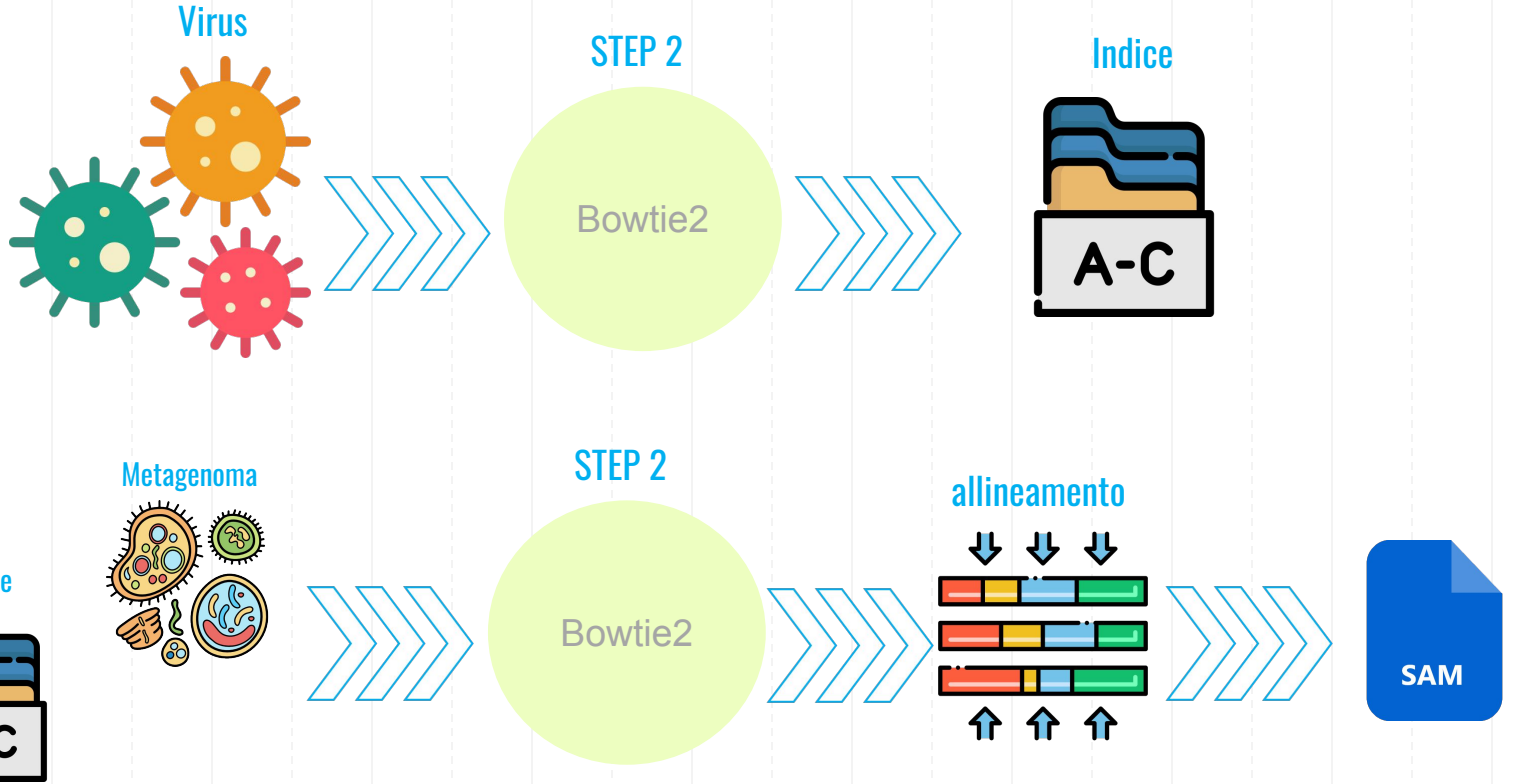
STEP 1

ViruSpy

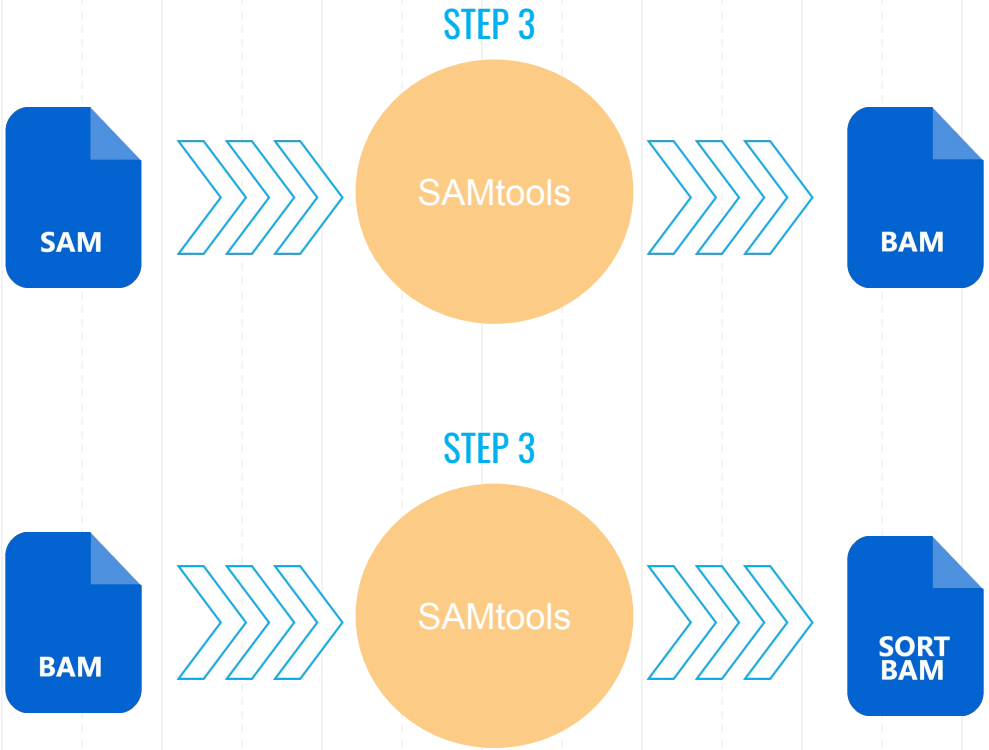
Virus



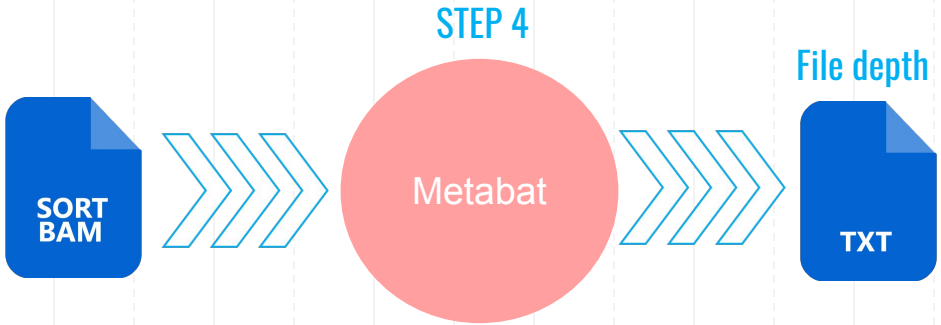
Pipeline


Virus
STEP 2
Bowtie2
Indice
A-C
Indice
A-C
Metagenoma
STEP 2
Bowtie2
allineamento
SAM

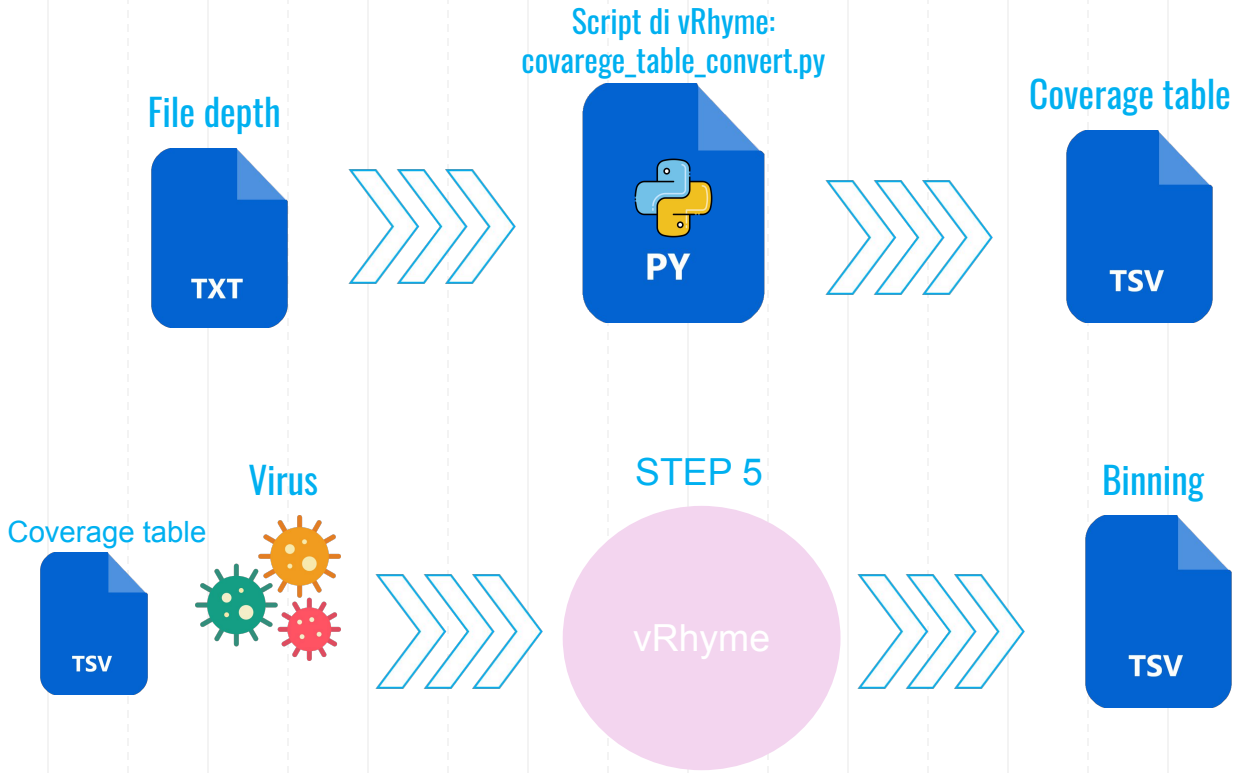
Pipeline



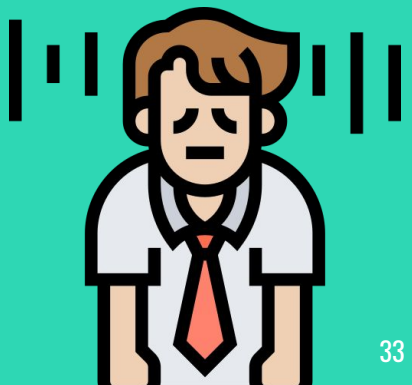
Pipeline



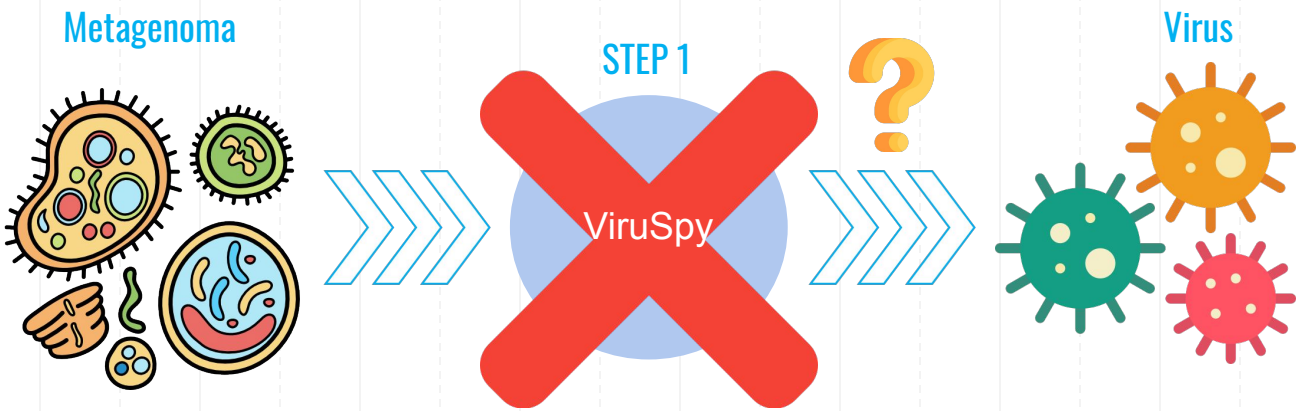
Pipeline


File depth
TXT
Script di vRhyme:
covarege_table_convert.py
PY
Coverage table
TSV
Coverage table
TSV
Virus
STEP 5
vRhyme
Binning
TSV

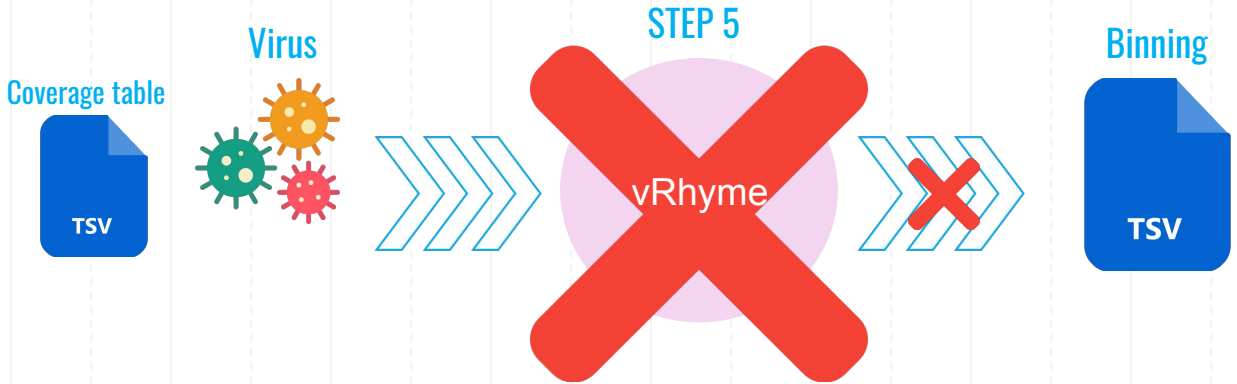
INVECE NON E' ANDATA PROPRIO COSI



Problemi della Pipeline



Problemi della Pipeline



Risoluzione dei problemi

Oltre a provare ulteriori tools con problematiche analoghe a quelle di ViruSpy o per eccessiva capacità di archiviazione richiesta, abbiamo deciso di prendere direttamente i virus completi

Risoluzione dei problemi

BinSanity

BinSanity contiene una suite di script progettati per raggruppare i contig generati dall'assemblaggio metagenomico in presunti genomi.

Risoluzione dei problemi

BinSanity

File depth

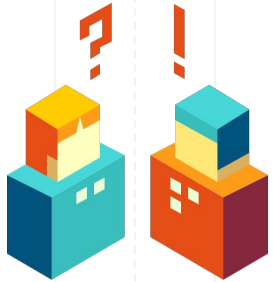
TXT

Script di vRhyme:
coverage_table_convert + multiplyAvg + multiplyStdev

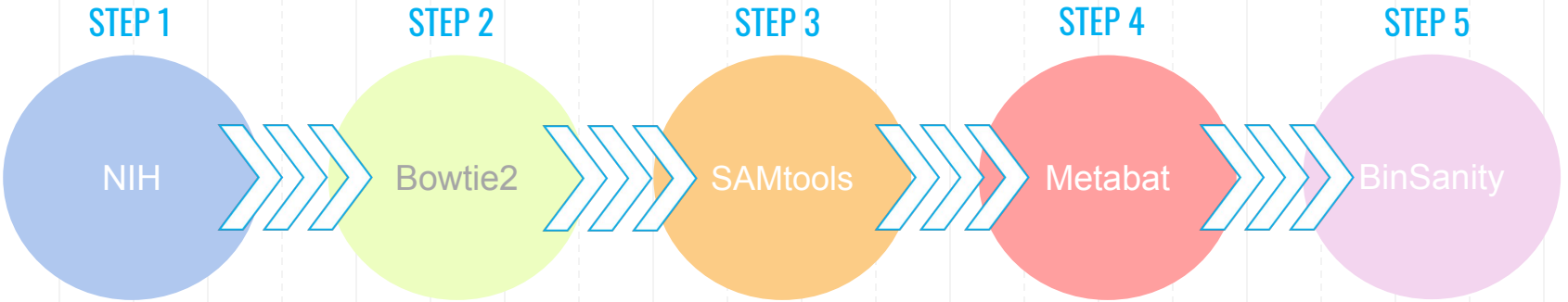
PY

Coverage table

TSV

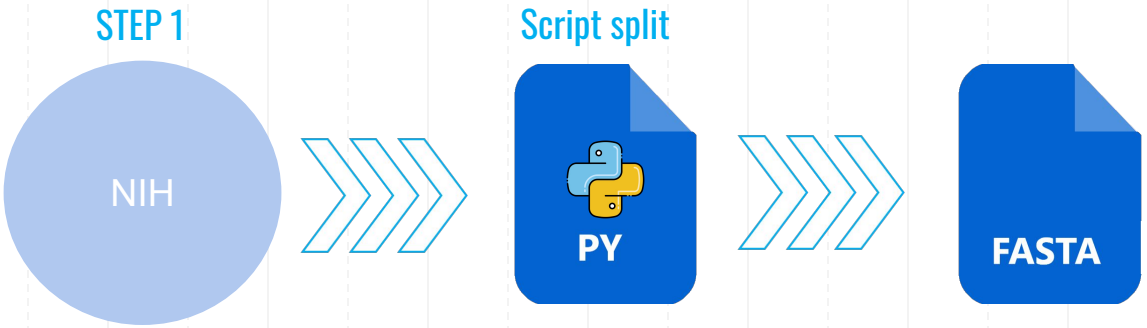


New Pipeline

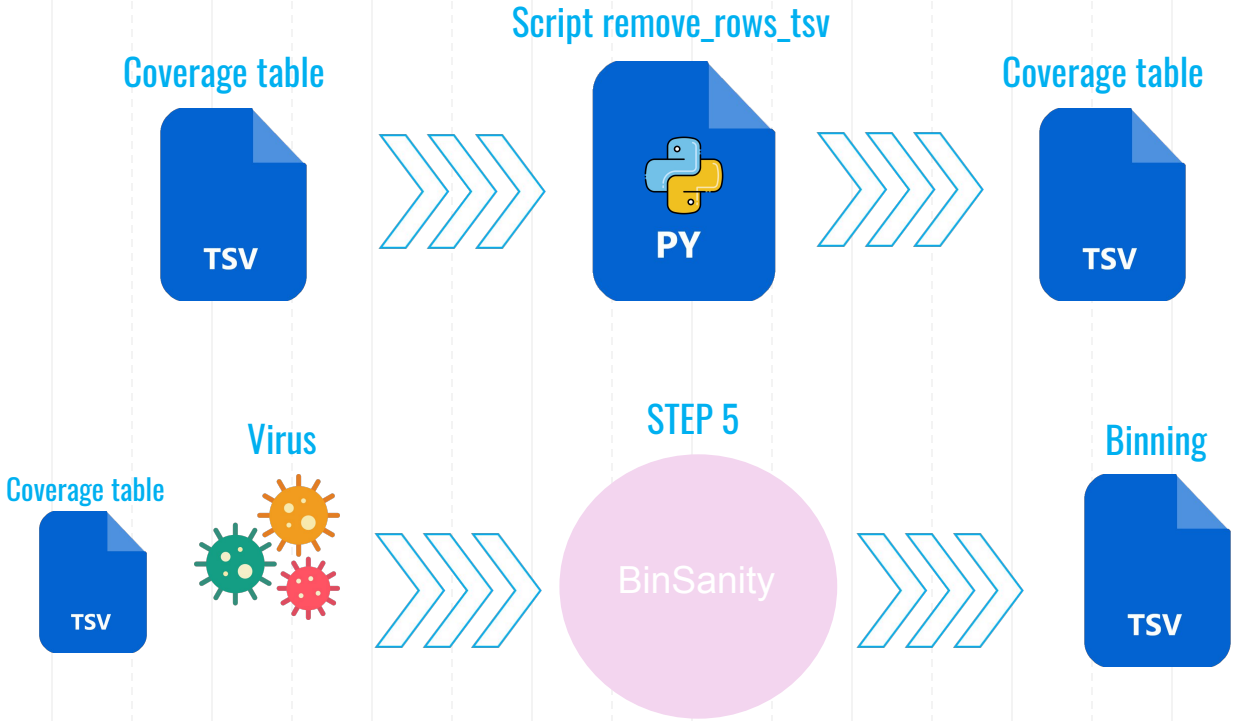


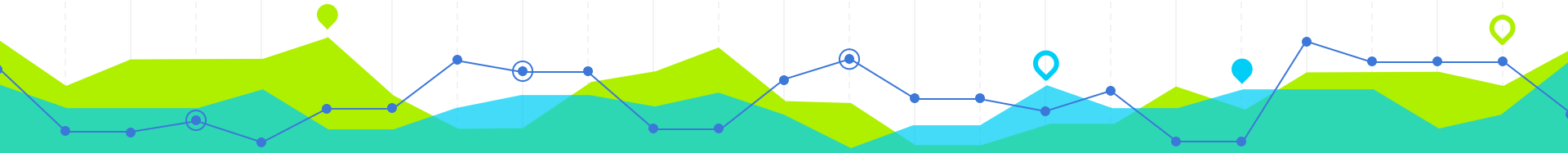
Scendiamo nel dettaglio...

Adattamento dei dati



Adattamento dei dati

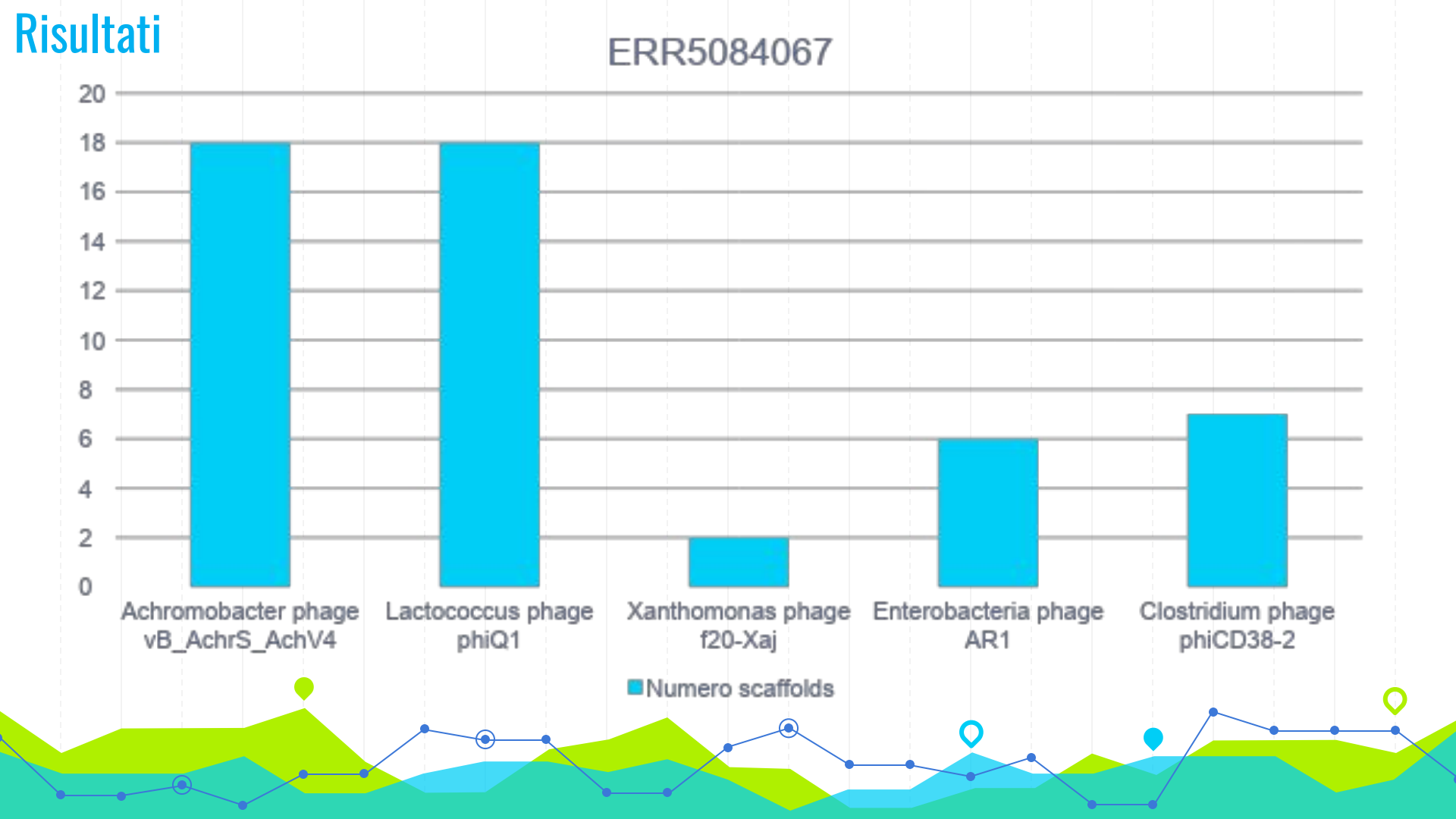




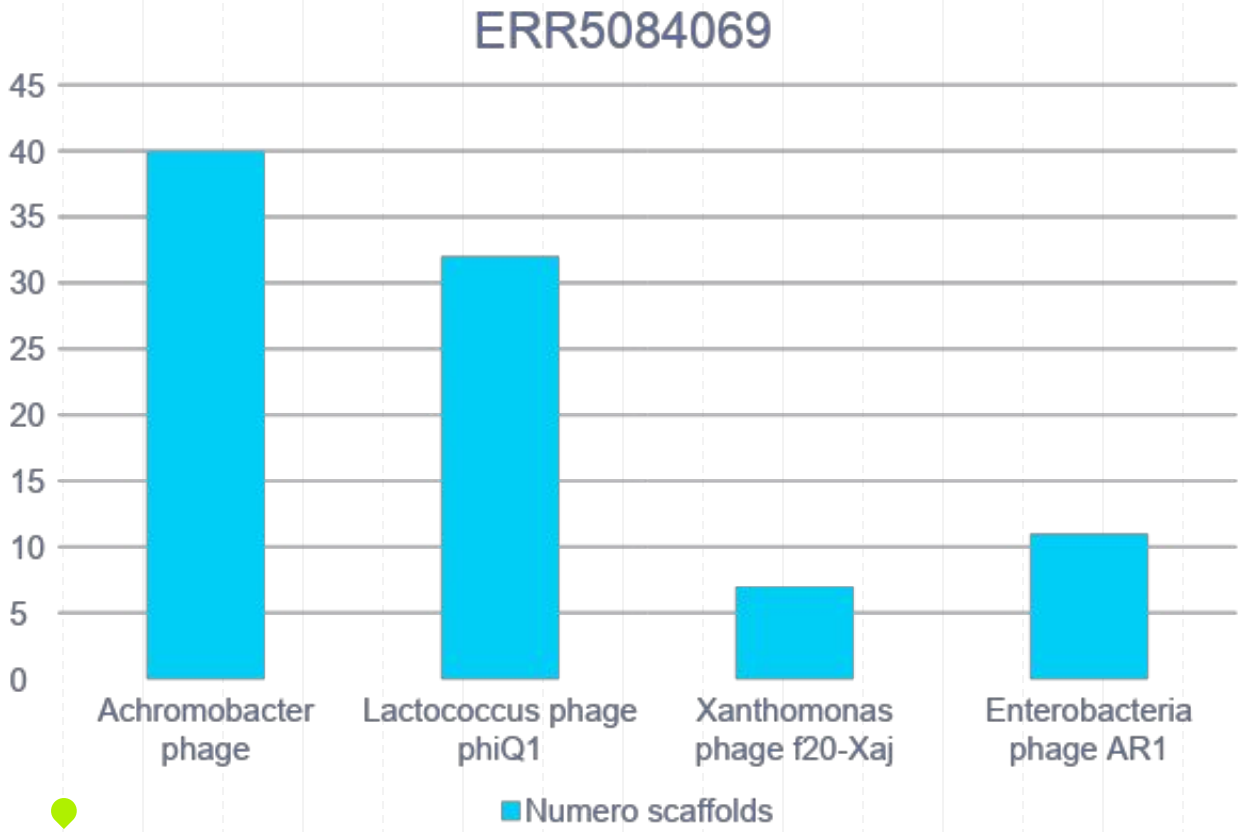
Analisi dei risultati

5

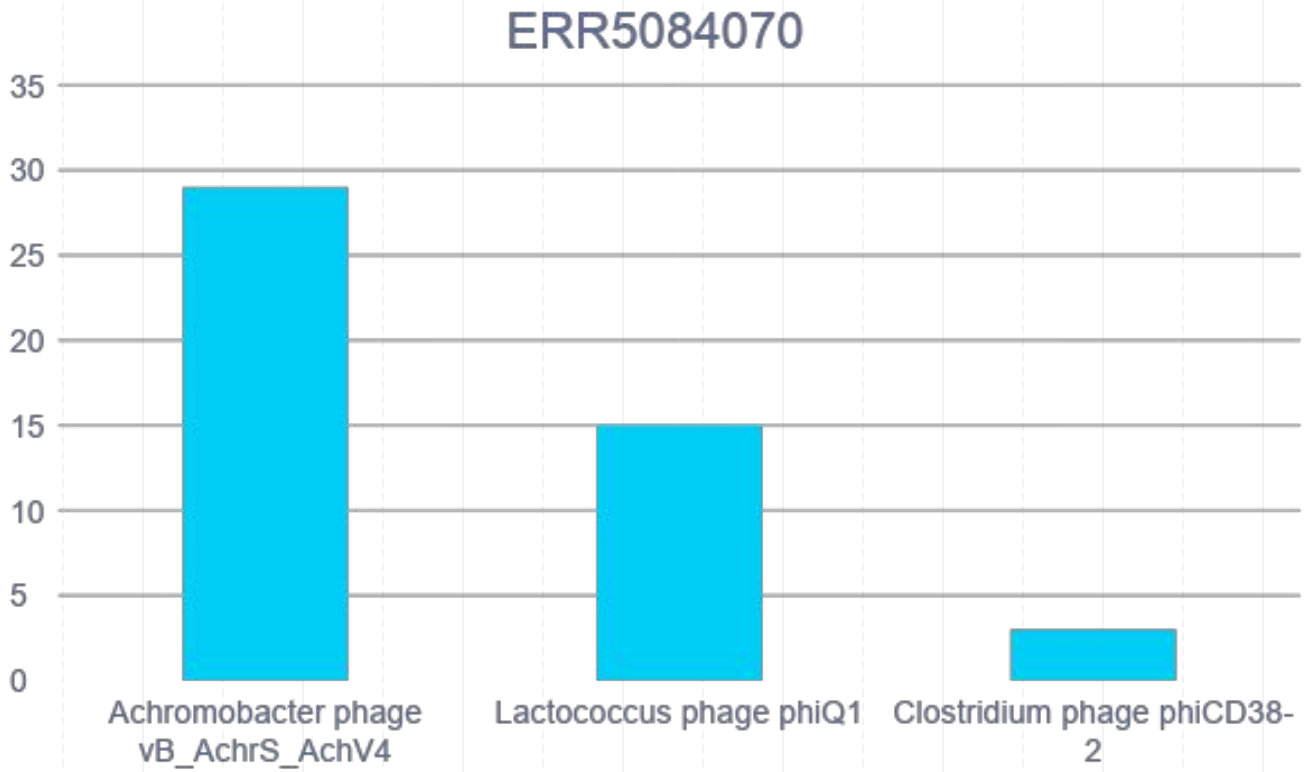
Risultati


ERR5084067
20
18
16
14
12
10
8
6
4
2
0
Achromobacter phage vB_AchrS_AchV4
Lactococcus phage phiQ1
Xanthomonas phage f20-Xaj
Enterobacteria phage AR1
Clostridium phage phiCD38-2
Numero scaffolds

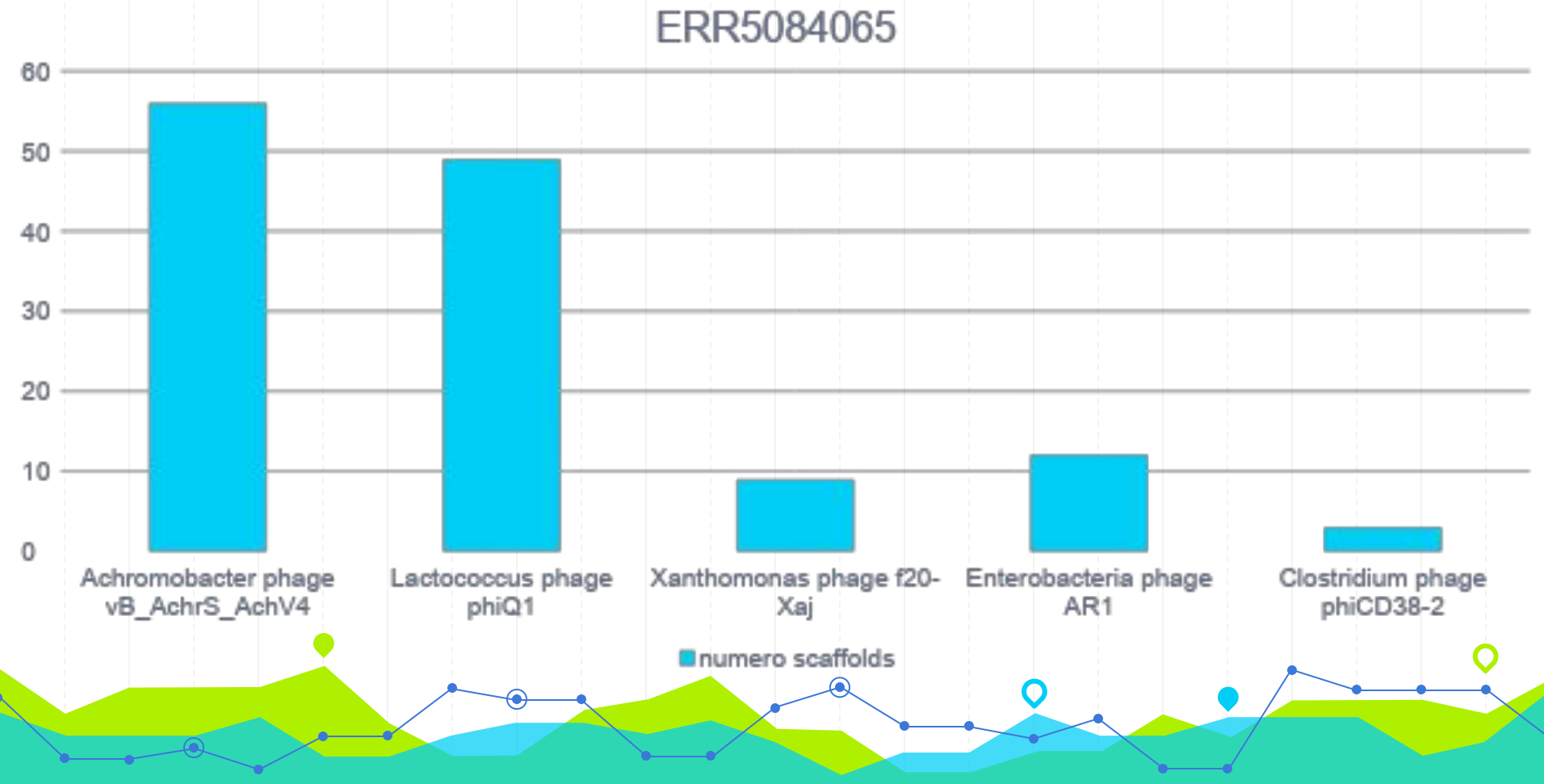
Risultati


ERR5084069
45
40
35
30
25
20
15
10
5
0
Achromobacter phage
Lactococcus phage phiQ1
Xanthomonas phage f20-Xaj
Enterobacteria phage AR1
Numero scaffolds

Risultati


ERR5084070
35
30
25
20
15
10
5
0
Achromobacter phage vB_AchrS_AchV4
Lactococcus phage phiQ1
Clostridium phage phiCD38-2
Numero scaffolds

Risultati


ERR5084065
60
50
40
30
20
10
0
Achromobacter phage vB_AchrS_AchV4
Lactococcus phage phiQ1
Xanthomonas phage f20-Xaj
Enterobacteria phage AR1
Clostridium phage phiCD38-2
numero scaffolds

Risultati

ERR5084067

Nome virus	Numero scaffolds	Lineage
Achromobacter Phage vB_AchrS_AchV4	12	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
CrAssphage cr7_1	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Crassvirales; Suoliviridae; Oafivirinae; Burzaovirus; Burzaovirus coli
Lactococcus phage phiQ1	6	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Teubervirus; Teubervirus Q1
Lactococcus phage 16802	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus sv16802
Xanthomonas phage f20-Xaj	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Autographiviridae; Pradovirus; Pradovirus f20
Shigella phage SfII	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Myoviridae; unclassified Myoviridae
Enterobacteria phage AR1	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Straboviridae; Tevenvirinae; Tequatrovirus; Tequatrovirus ar1
Clostridium phage phiCD38-2	3	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Leicestervirus; Leicestervirus CD382

Nome virus	Numero scaffolds	Lineage
Achromobacter phage vB_AchrS_AchV4	3	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
Lactococcus phage phiQ1	3	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Teubervirus; Teubervirus Q1
Enterobacteria phage AR1	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Straboviridae; Tevenvirinae; Tequatrovirus; Tequatrovirus ar1
Clostridium phage phiCD38-2	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Leicestervirus; Leicestervirus CD382

Risultati

ERR5084069

Nome virus	Numero scaffolds	Lineage
Achromobacter Phage vB_AchrS_AchV4	8	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
Lactococcus phage fd13	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus fd13
Lactococcus phage phiQ1	11	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Teubervirus; Teubervirus Q1
Lactococcus phage 56003	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus sv56003
Lactococcus Phage ASCC281	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus ASCC281
Xanthomonas phage f20-Xaj	5	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Autographiviridae; Pradovirus; Pradovirus f20
Enterobacteria phage AR1	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Straboviridae; Tevenvirinae; Tequatrovirus; Tequatrovirus ar1

Nome virus	Numero scaffolds	Lineage
Achromobacter Phage vB_AchrS_AchV4	3	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
CrAssphage cr56_1	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Crassvirales; Suoliviridae; Oafivirinae; Burzaovirus; Burzaovirus faecalis
Lactococcus phage phiQ1	5	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Teubervirus; Teubervirus Q1
Lactococcus phage 13w11L	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus sv13w11L
Lactococcus phage 16802	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus sv16802
Lactococcus phage CHPC965	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus CHPC965
Xanthomonas phage f20-Xaj	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Autographiviridae; Pradovirus; Pradovirus f20

Risultati

ERR5084070

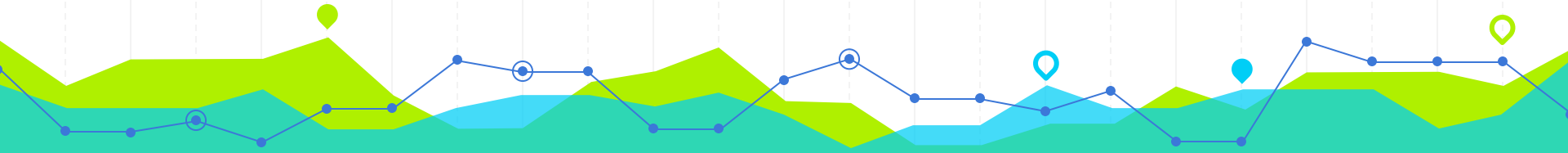
Nome virus	Numero scaffolds	Lineage
Achromobacter Phage vB_AchrS_AchV4	5	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
Enterobacteria phage AR1	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Straboviridae; Tevenvirinae; Tequatrovirus; Tequatrovirus ar1
Clostridium phage phiCD38-2	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Leicestervirus; Leicestervirus CD382

Risultati

ERR5084065

Nome virus	Numero scaffolds	Lineage
Achromobacter Phage vB_AchrS_AchV4	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
Lactococcus phage phiQ1	6	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Teubervirus; Teubervirus Q1
Lactococcus phage 936 group phage PhiL.6	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus L6
Xanthomonas phage f20-Xaj	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Autographiviridae; Pradovirus; Pradovirus f20
Enterobacteria phage AR1	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Straboviridae; Tevenvirinae; Tequatrovirus; Tequatrovirus ar1

Nome virus	Numero scaffolds	Lineage
Achromobacter Phage vB_AchrS_AchV4	6	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Casjensviridae; Gediminasvirus; Gediminasvirus AchV4
Lactococcus phage fd13	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus fd13
Lactococcus phage phiQ1	24	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Teubervirus; Teubervirus Q1
Lactococcus phage 16802	1	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Skunavirus; Skunavirus sv16802
Xanthomonas phage f20-Xaj	4	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Autographiviridae; Pradovirus; Pradovirus f20
Enterobacteria phage AR1	4	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Straboviridae; Tevenvirinae; Tequatrovirus; Tequatrovirus ar1
Clostridium phage phiCD38-2	2	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Leicestervirus; Leicestervirus CD382



Comparativa

Genome Detective

6

Genome Detective

La pipeline di Genome Detective di è progettata per identificare e classificare in modo rapido e accurato i virus presenti in dati NGS.

Comparativa

ERR5084065 & ERR5084067 & ERR5084069 & ERR5084070
vs
Genome Detective

- Lactococcus phage 936 group phage PhiB1127
- Lactococcus phage 936 group phage Phi91127
- Lactococcus lactis phage p272
- Lactococcus phage LP9207
- Skunavirus sv3R16S
- Skunavirus sv30804
- Skunavirus sv16802
- Lactococcus phage CB19
- Lactococcus phage R31
- Lactococcucs phage CB20
- Leuconostoc phage phiLN6B
- Leuconostoc phage Lmd1
- Salmonella phage vB_SenS_Sasha
- Streptococcus phage YMC-2011

Virus uguali

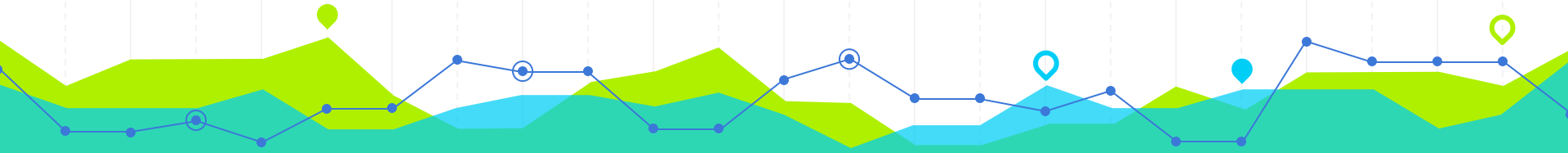
Comparativa

ERR5084065 & ERR5084067 & ERR5084069 & ERR5084070

vs

Genome Detective

Genome Detective	BinSanity
CrAss-like virus sp.	CrAssphage 50_1 - 10_1 - 4_1- 6_1 - 114_1 - 125_1
Burzaovirus Faecalis	CrAssphage cr7_1
Brigitvirus Brigit	Faecalibacterium phage FP_Brigit
Taranisvirus Taranis	Faecalibacterium phage FP_Taranis
Skunavirus fd13	Lactococcus phage fd13
Toutatisvirus Toutatis	Faecalibacterium phage FP_Toutatis



Conclusioni

7

Conclusioni

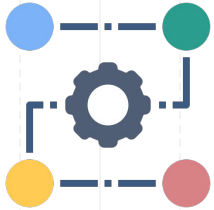
Scalabilità: oltre all'estrazione di virus dai dati metagenomici permette l'estrazione di batteri, funghi e di qualsiasi altro regno.

Velocità : grazie a bowtie2 l'esecuzione dell'allineamento è rapida

Limitatezza : pochi virus a disposizione

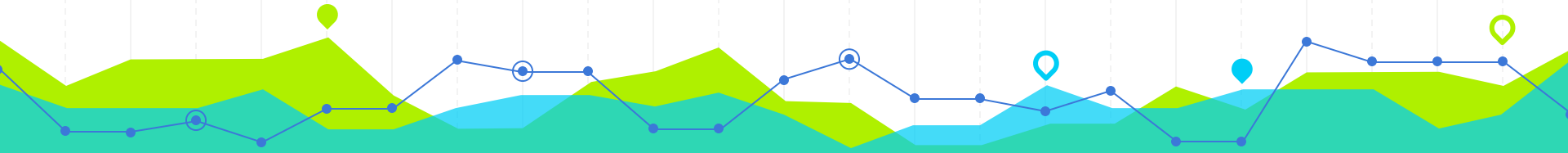
Implementazioni future

Migliorare la fase di pre-processing, ad esempio inserire la fase di trimming



Uno script per l'invocazione automatica dei vari tools.

Utilizzare hardware più potente per avere una maggiore potenza computazionale e quindi poter disporre di più virus



GRAZIE A TUTTI PER L'ATTENZIONE