Centro de Investigación en Dinámica Celular

BCVI
Biología Computacional y Virómica Integrativa

Indentificación de Secuencias Virales (Metagenómica Viral y Profagos)

Dr. Gamaliel López Leal
Centro de Investigación en Dinámica Celular-UAEM

Links a utilizar a lo largo del curso

Acceso remoto a la sesión

(Por favor revise previamente y actualice su versión de zoom)

ID de reunión: **899 6000 8592**
Código de acceso: **661568**

https://us06web.zoom.us/j/89960008592?pwd=3WXhx6ENvGsszXeakfatCJRLDgZgRt.1

Acceso cuenta Github BCVI

(Repositorio disponible 26 febrero 2024)

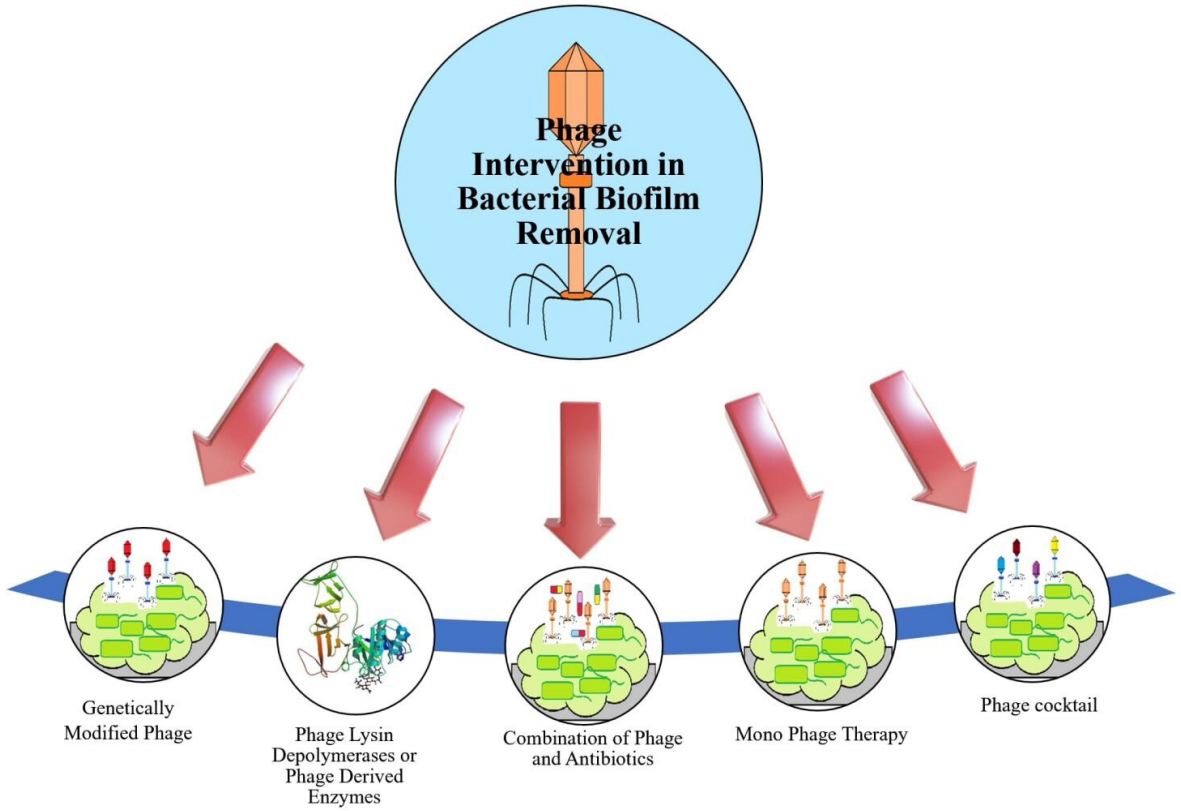
https://github.com/BCVI/Taller-Virtual-Bioinformatica-Genomica-Viral

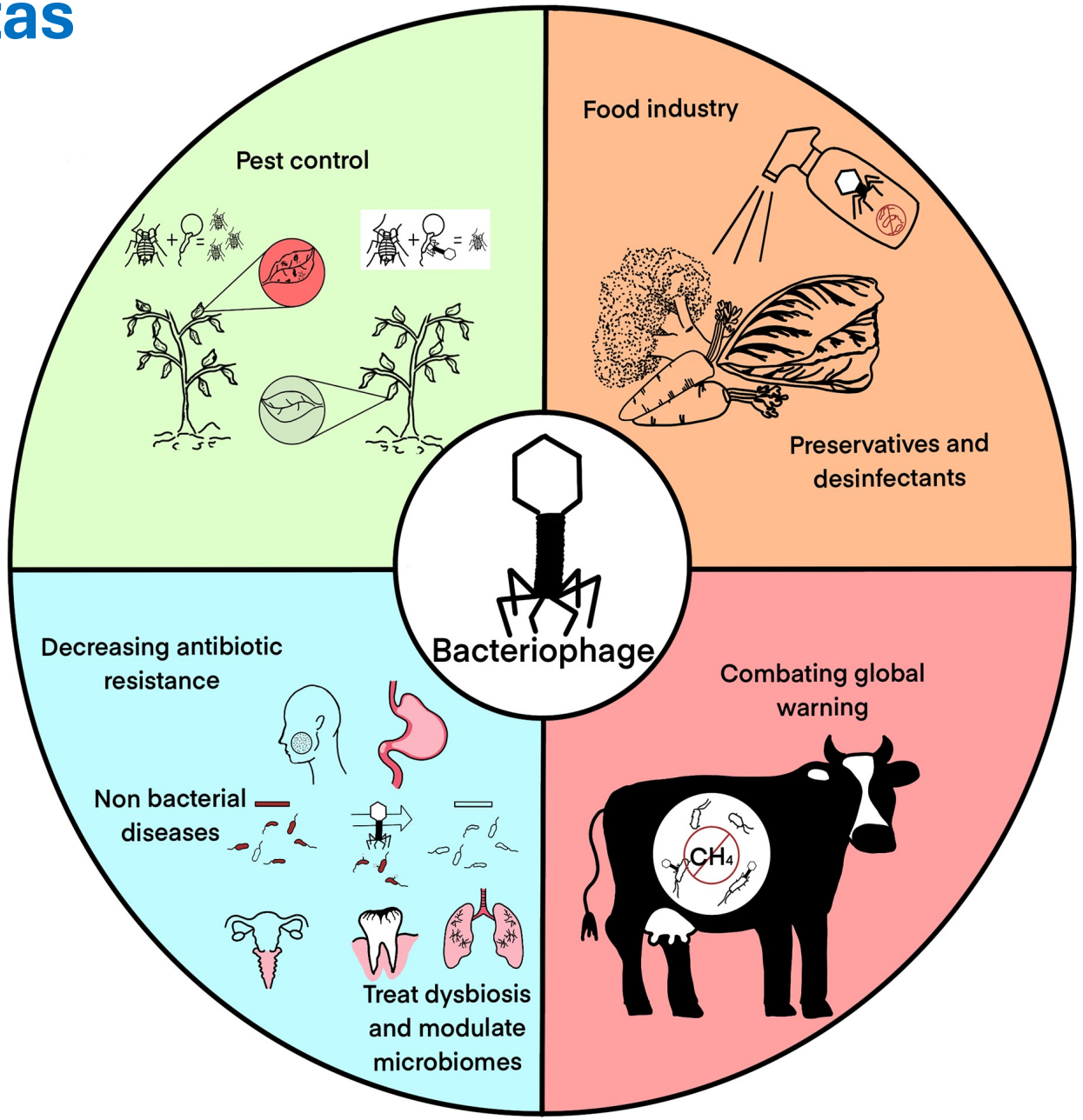
Acceso Slack

https://join.slack.com/t/slack-smo2540/shared_invite/zt-2d62iq1fr-lketSaPu~IeUyZ5LIghRqA

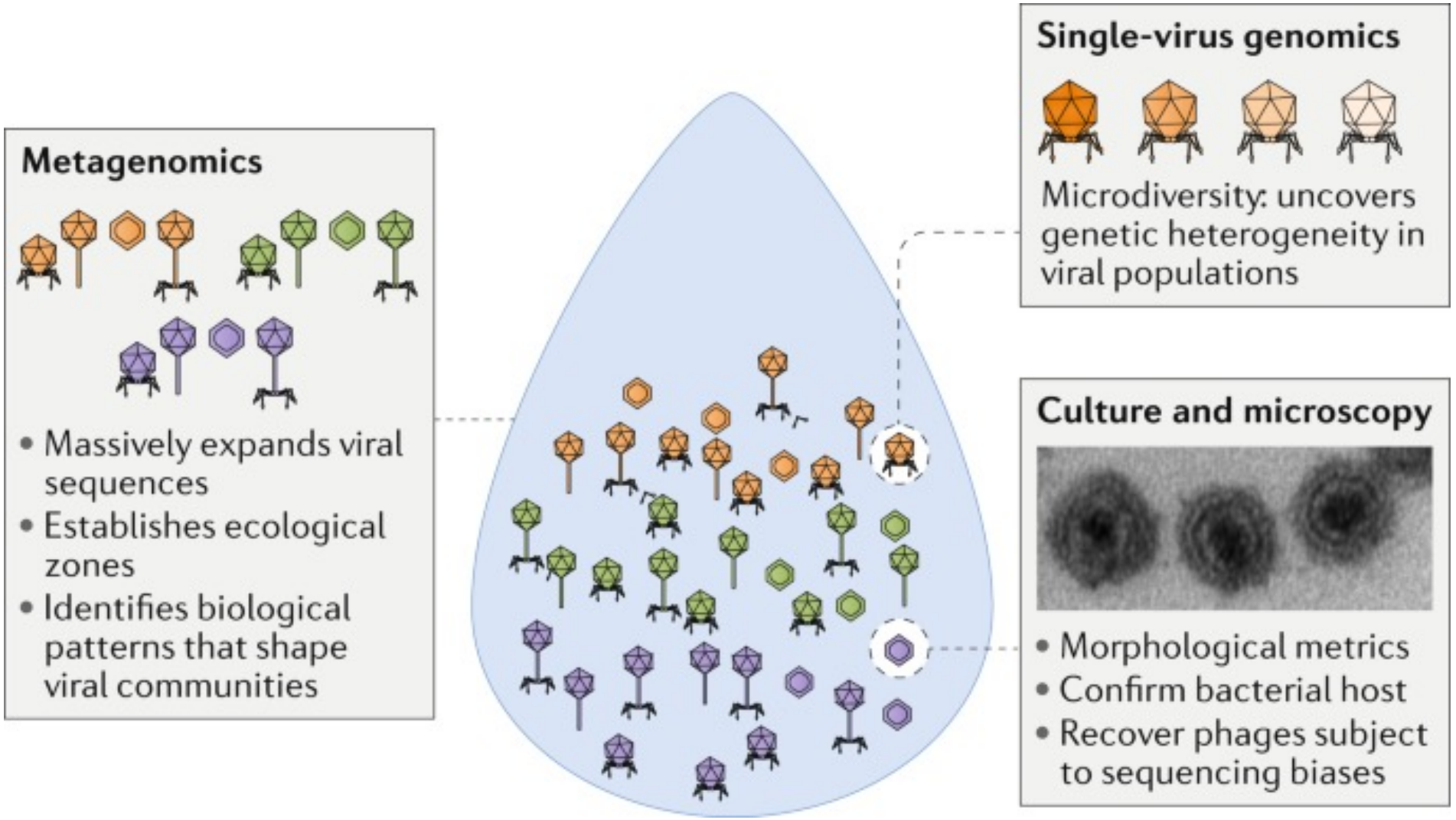
Los bacteriófagos son cosmopolitas y son considerados como las entidades biológicas más abundantes de la tierra.

- Virus que solo atacan a los procariotas.





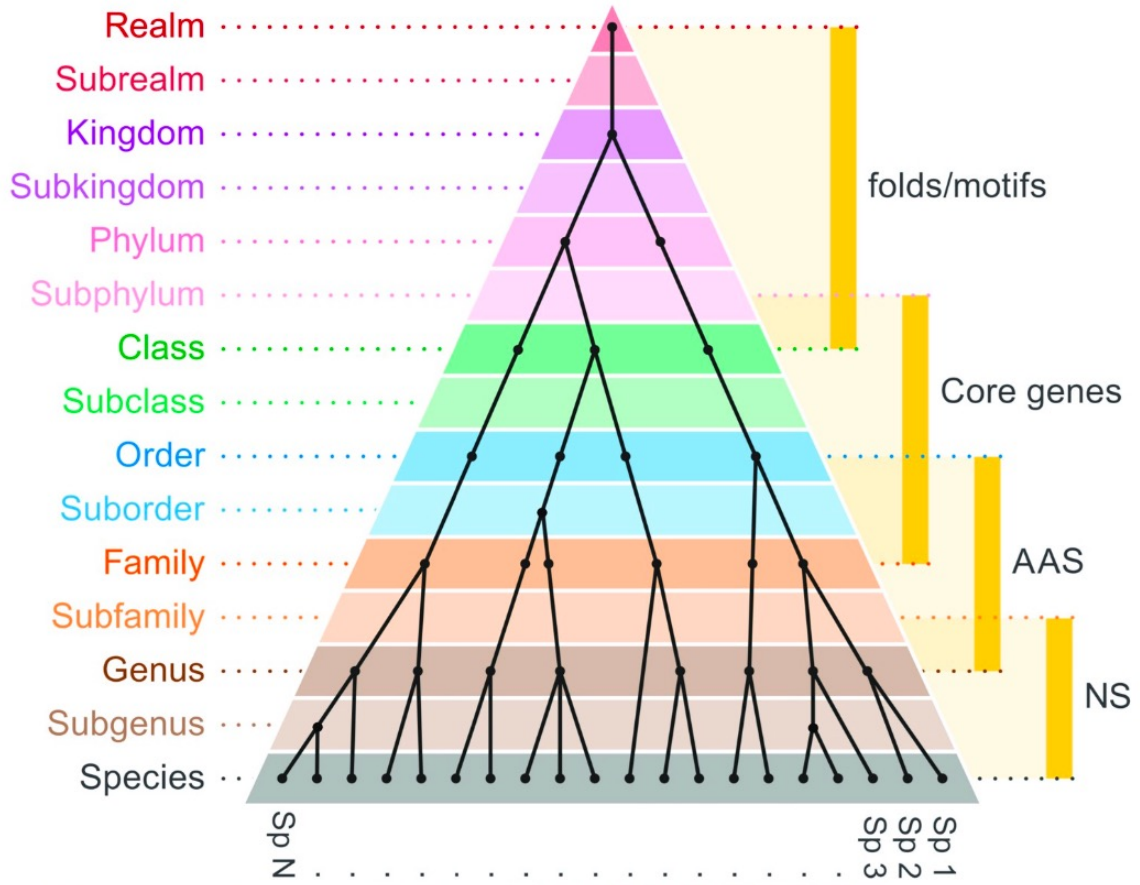
Estrategias para identificarlos



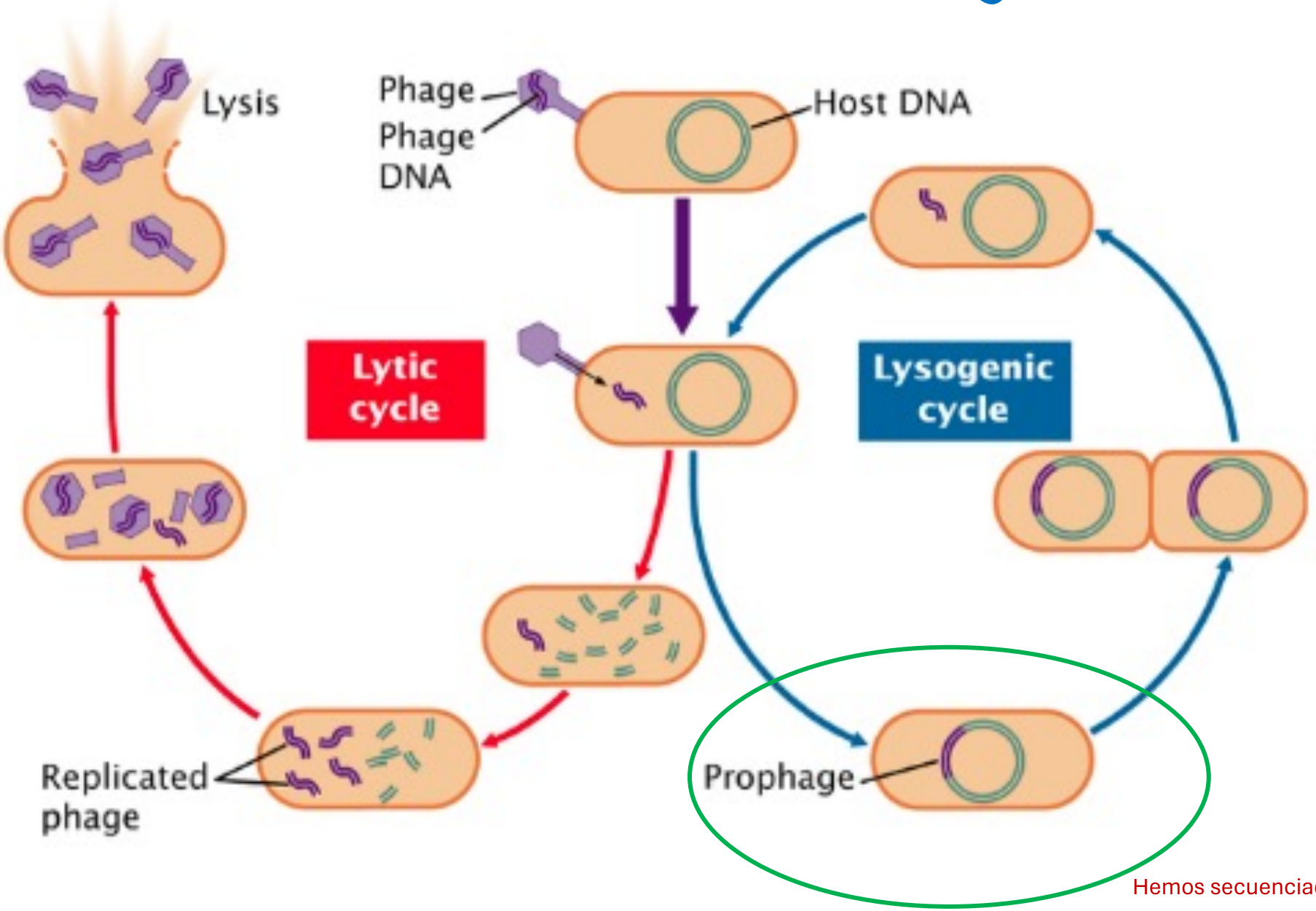
El sesgo de las bases de datos

Kingdom ☐ Bacteria (552,527)

Viruses **- 11677 complete genomes**
filtered by host 'bacteria': 4195 genomes

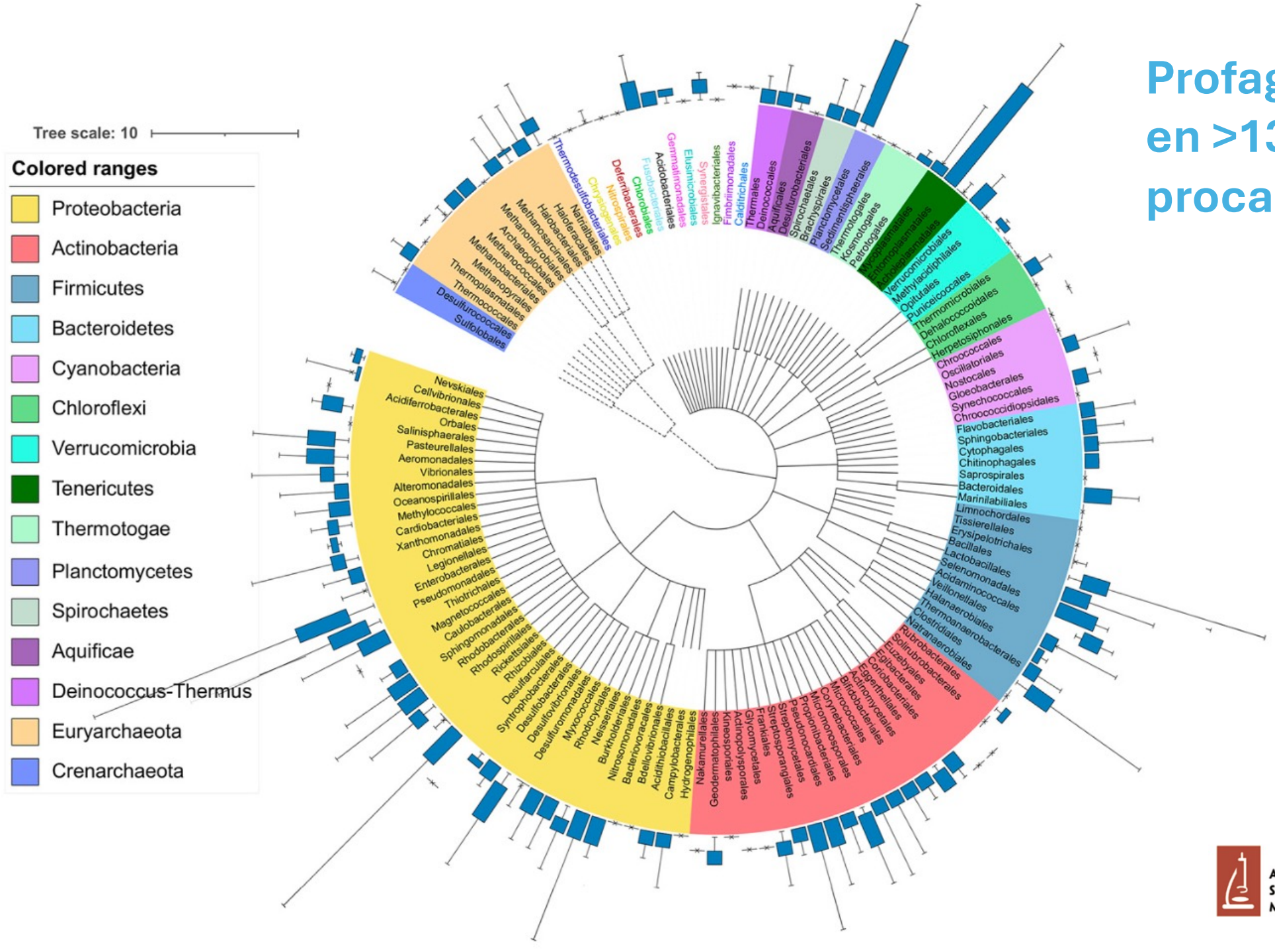


Estilo de vida de los bateriofágos



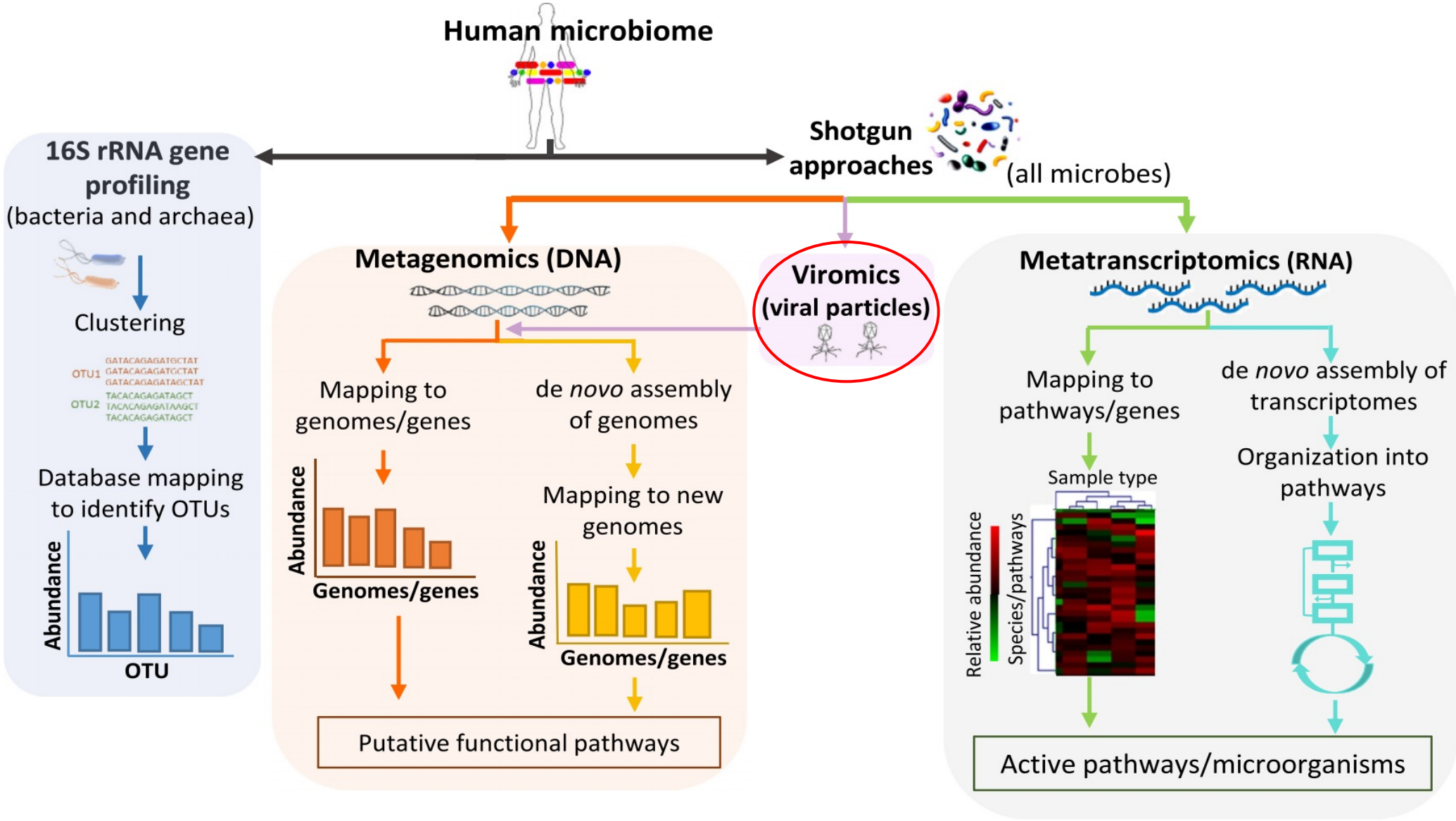
Hemos secuenciado miles de profagos

Profagos identificados en >13,000 genomas procariontes



AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

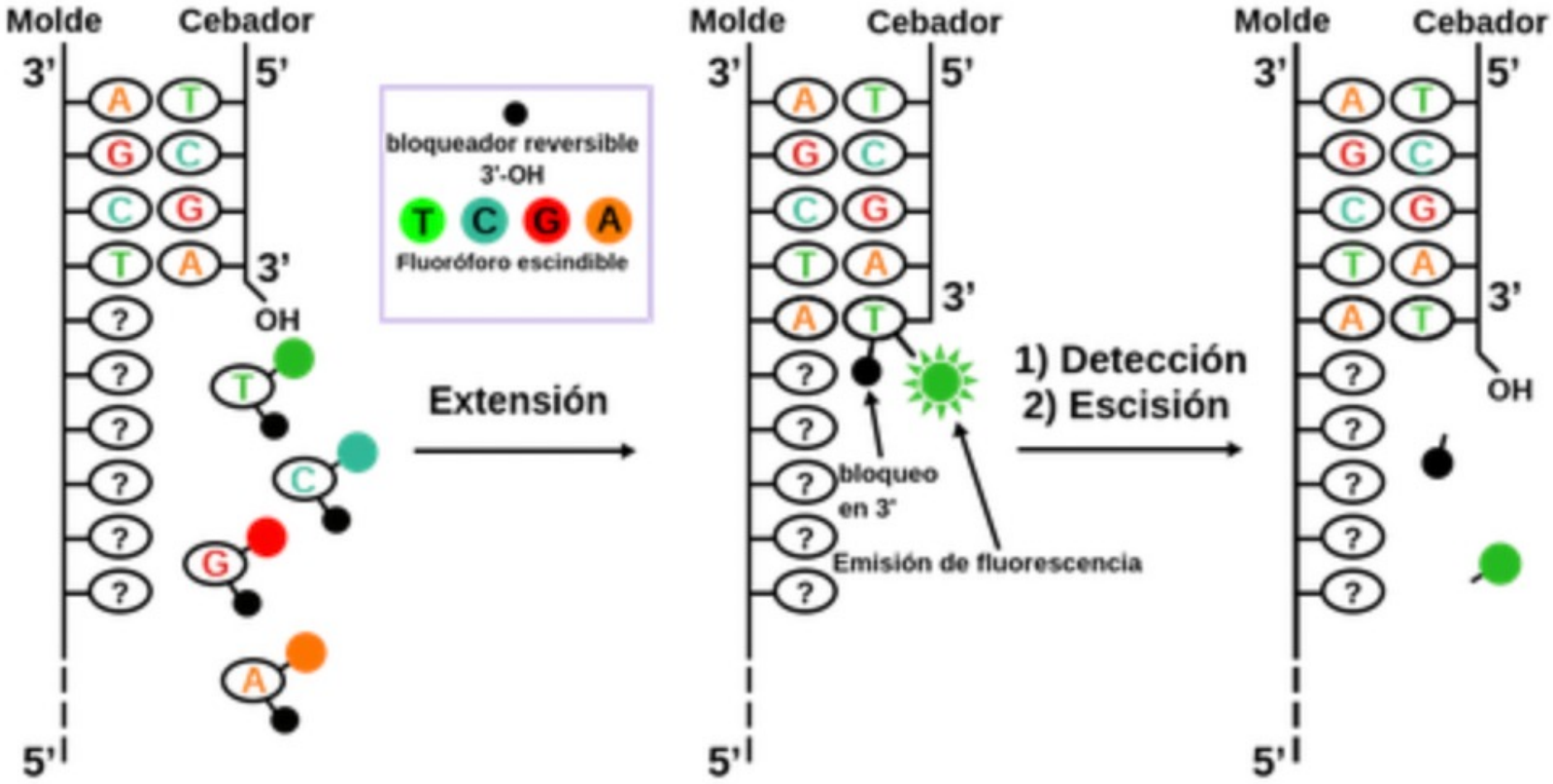
La Metagenómica y sus diversas aproximaciones



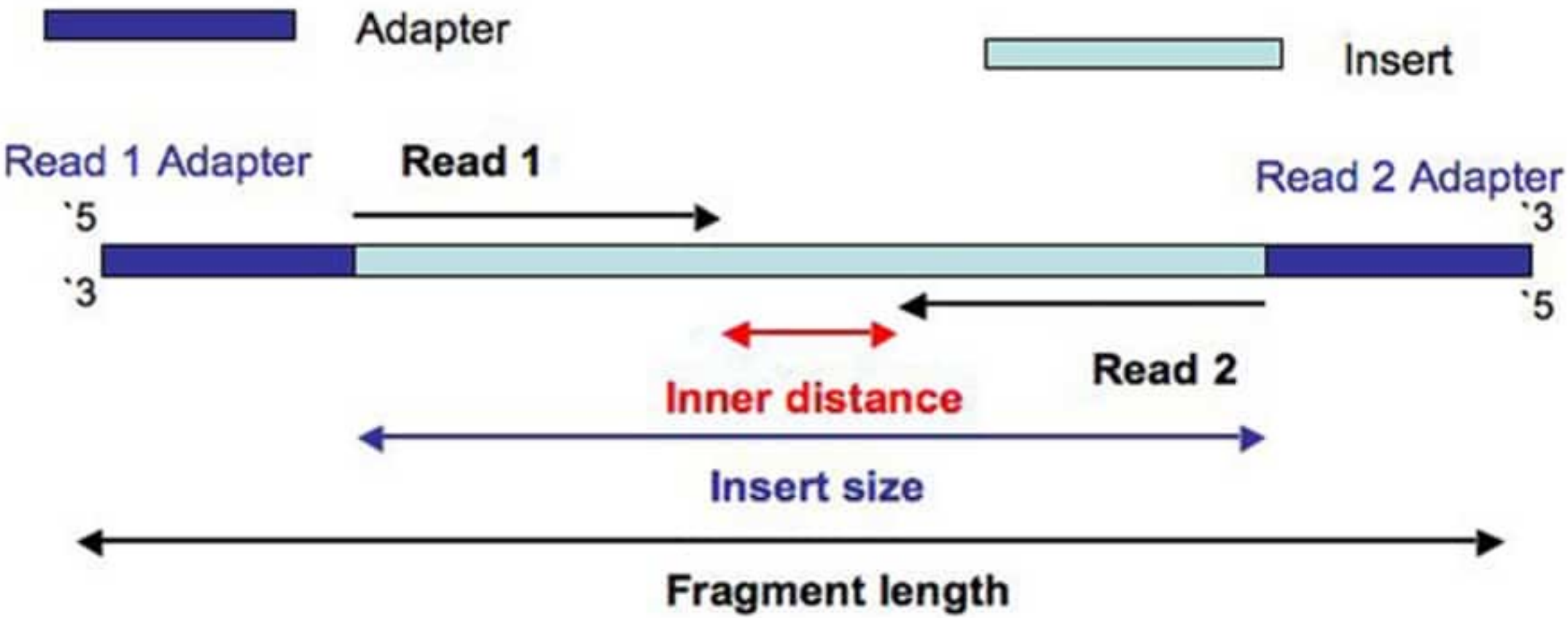
Bikel et al 2015 Nat. Rev. Microbiology

Análisis de calidad de lecturas

Background: Secuenciación por Illumina



Paired-End y Single-End



Estructura de una lectura

Etiqueta

Secuencia

```
@HWI-ST999:102:D1N6AACXX:1:1101:1235:1936 1:N:0:
ATGTCTCCTGGACCCCTCTGTGCCCAAGCTCCTCATGCATCCTCCTCAGCAACTTGTCCTGTAGCTGAGGCTCACTGACTACCAGCTGCAG
+
1:DAADDDF<B<AGF=FGIEHCCD9DG=1E9?D>CF@HHG??B<GEBGHCG;;CDB8==C@@>>GII@@5?A?@B>CEDCFCC:;?CCCAC
```

Calidad de la lectura

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN...........................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
0.2......................26...31........41
0..................20........30........40........50
0..................20........30........40........50..........................................93
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
N - Nanopore      Phred+33,  Duplex reads typically (0, 50)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

FastQC Report

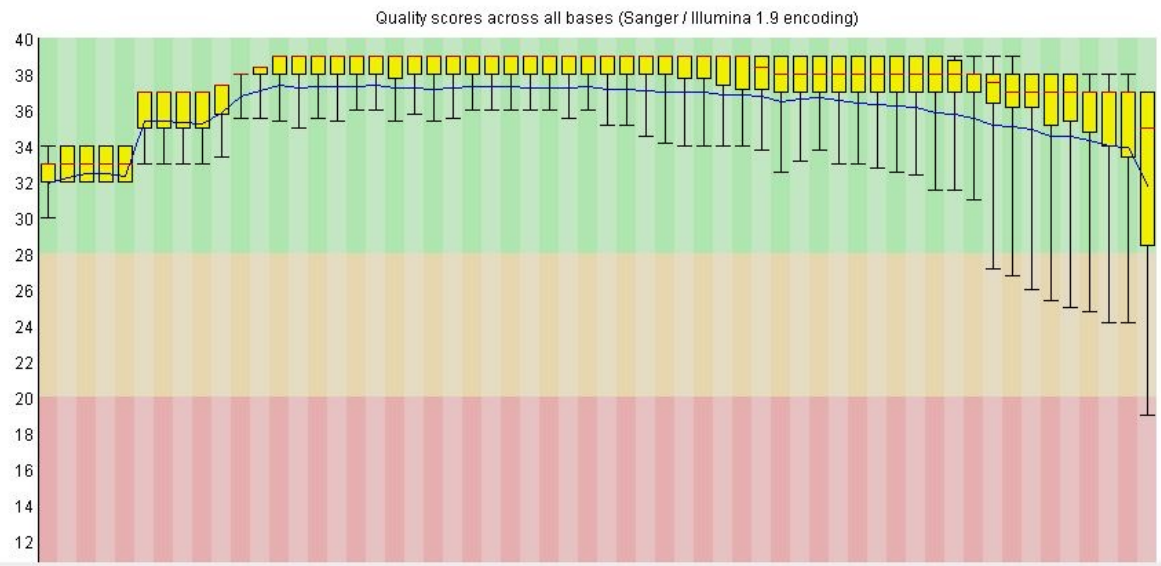
Mi 19 Sep 2018
L14-10_1.fastq.gz

Summary

Basic Statistics

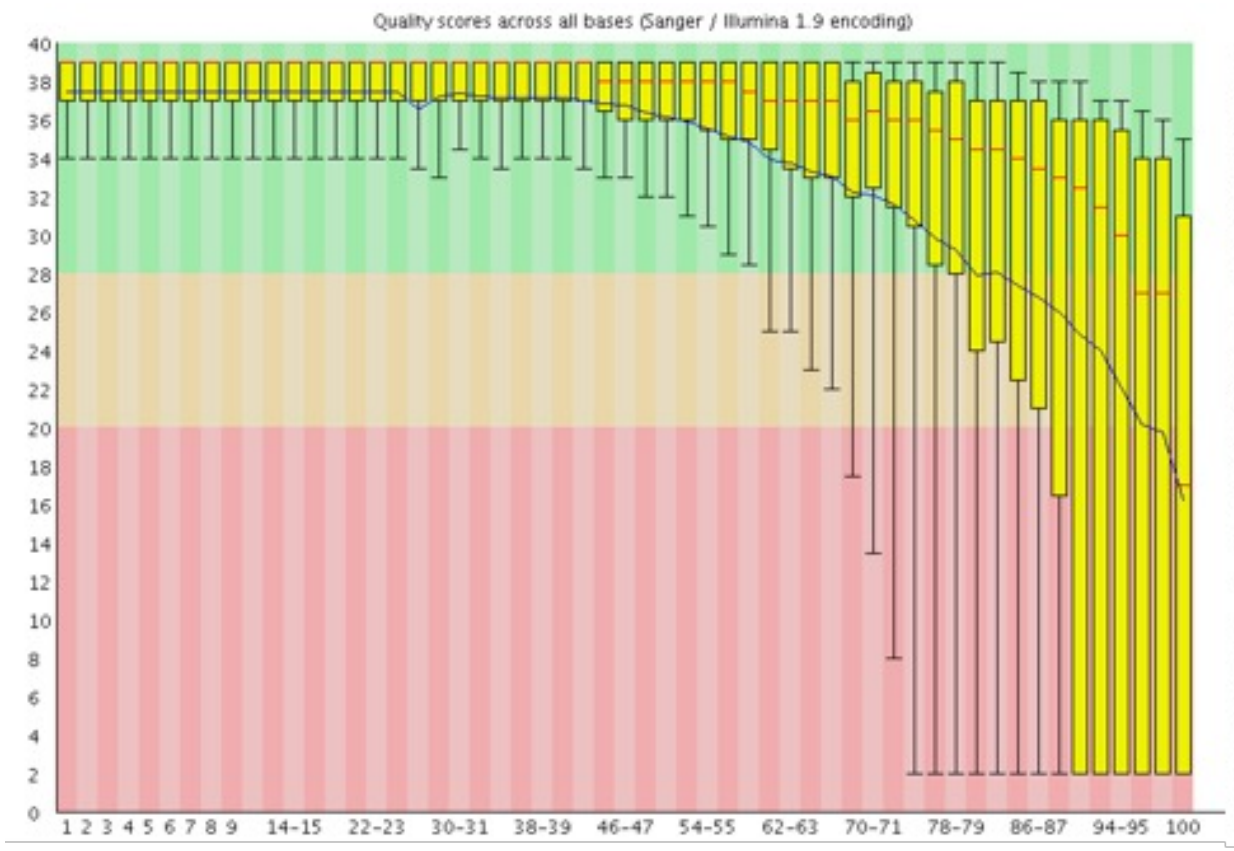
Measure	Value
Filename	L14-10_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4950
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	39

Per base sequence quality

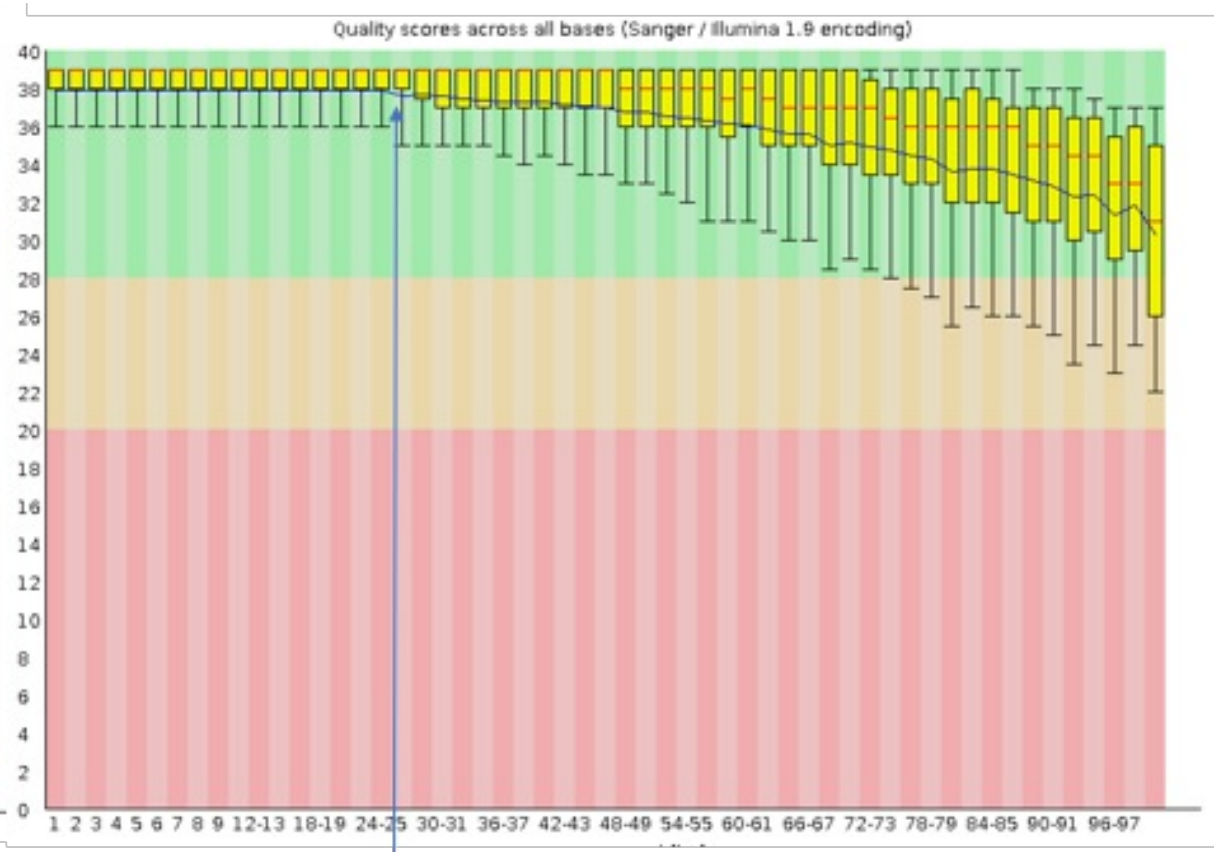


Análisis de Calidad

Antes del trimming



Después del trimming



Media de la calidad

TRIM GALORE

Tool to eliminate the sequencing adapters and use the quality parameters (2pb)

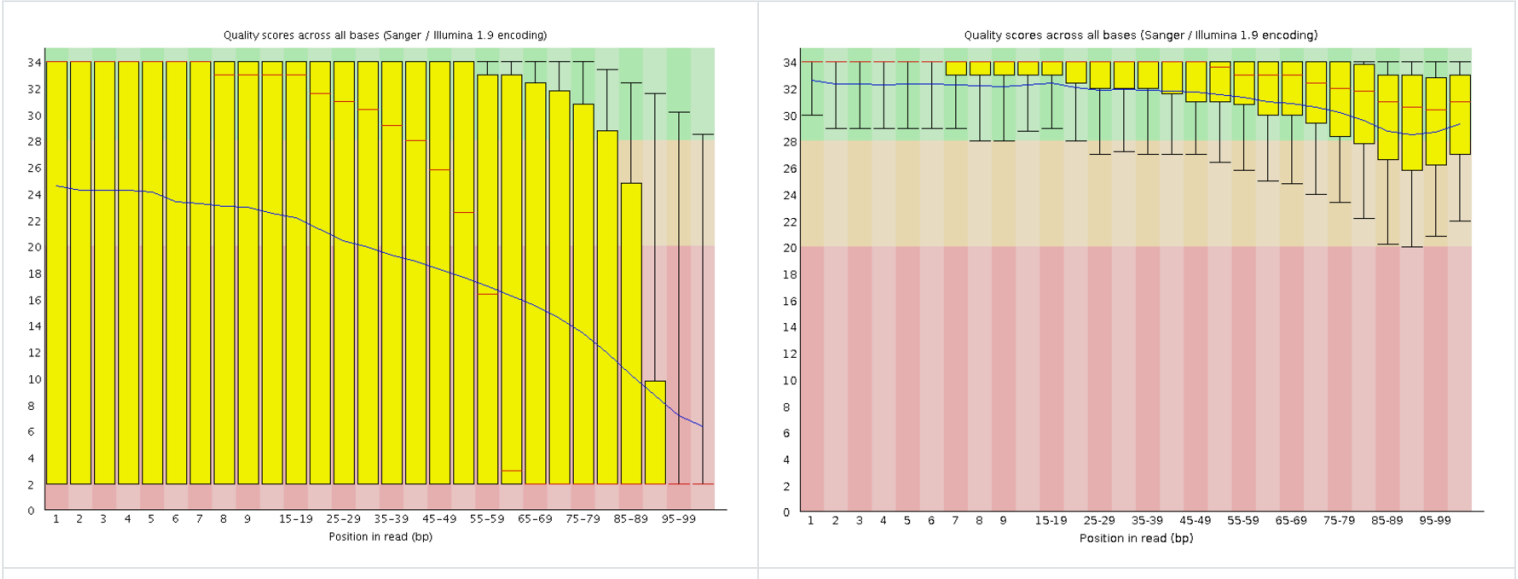
This tool use Cutadapt y FastQC packages

Remove de adapters (12 – 13 pb del extremo 3' TruSeq y Sanger iTag)

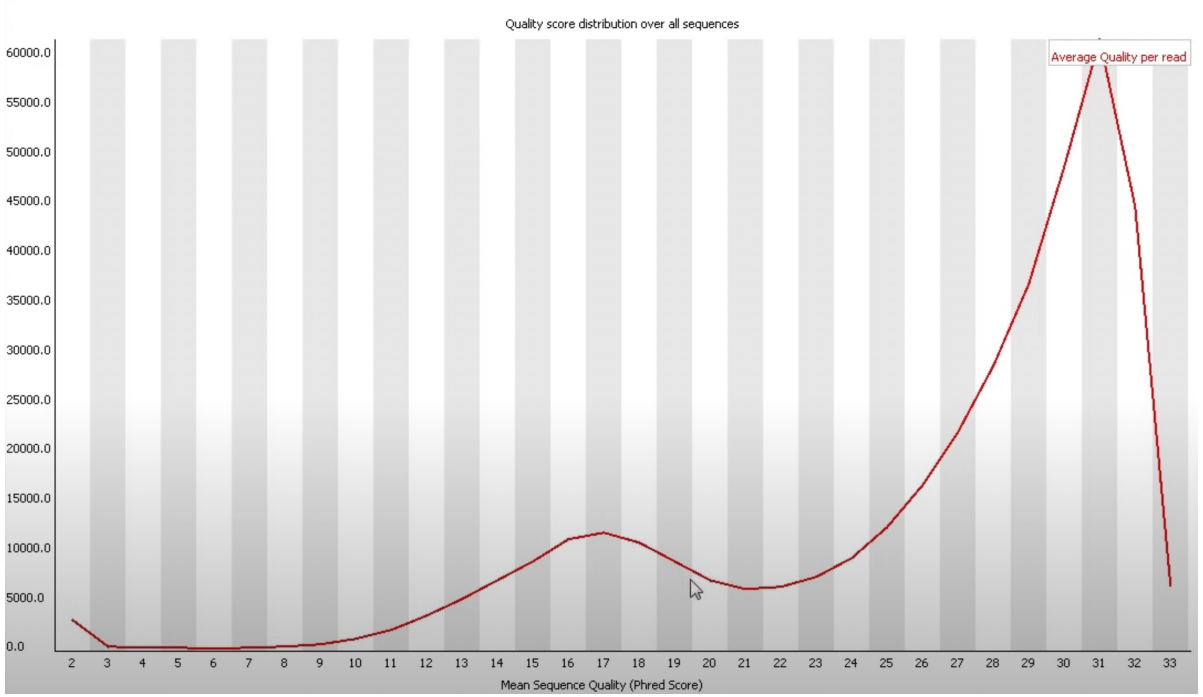
```
Illumina:    AGATCGGAAGAGC
Small RNA:   TGGAATTCTCGG
Nextera:     CTGTCTCTTATA
```

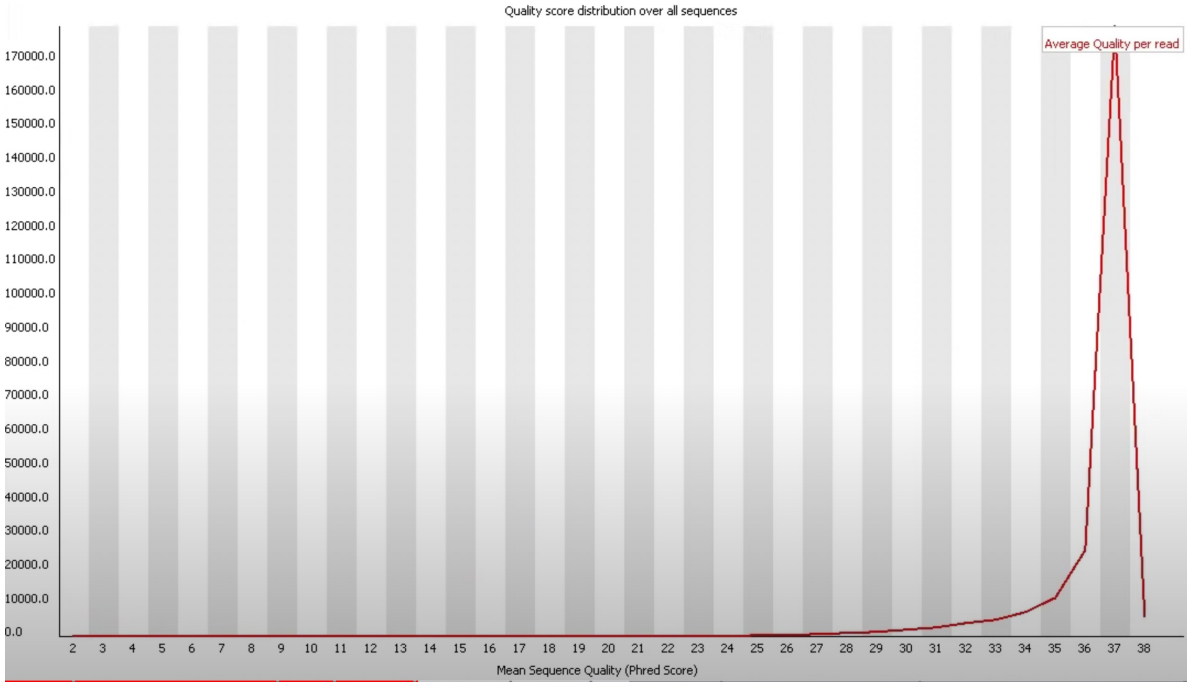
```
-a/--adapter <STRING>
```

```
-a2/--adapter2 <STRING>
```



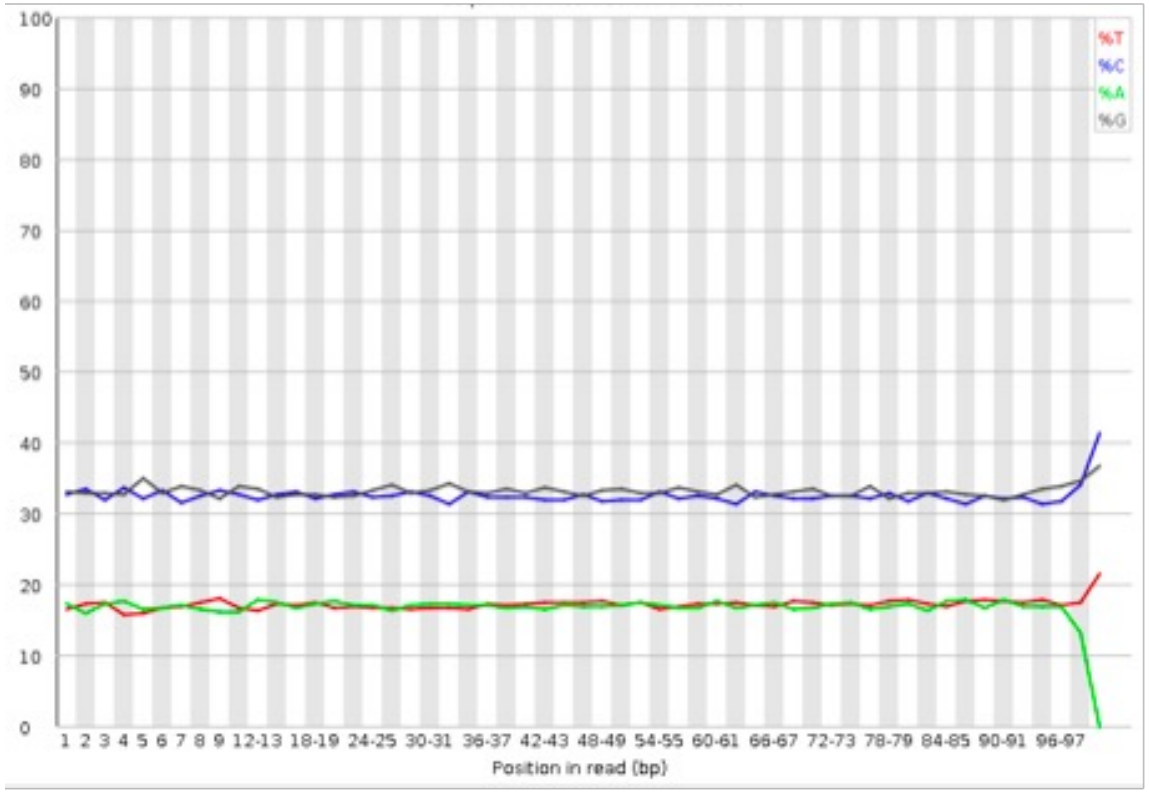
Distribución de los índices de calidad



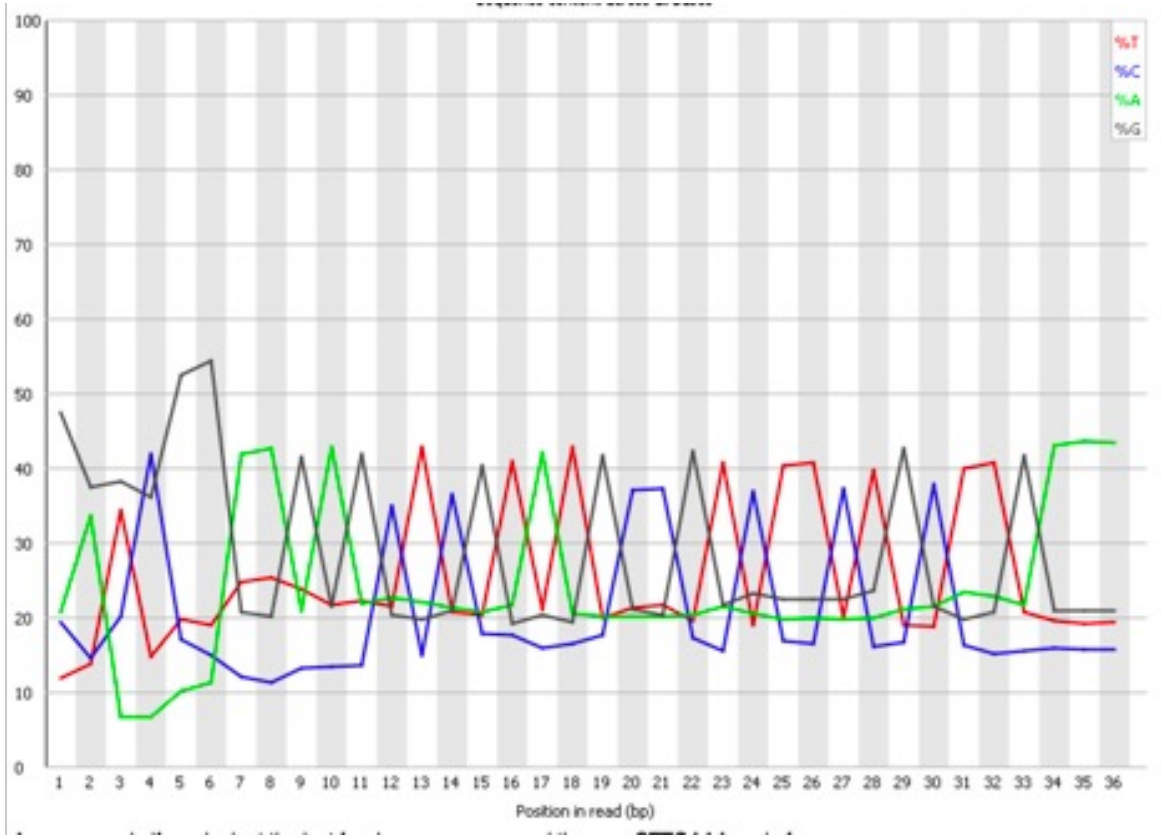


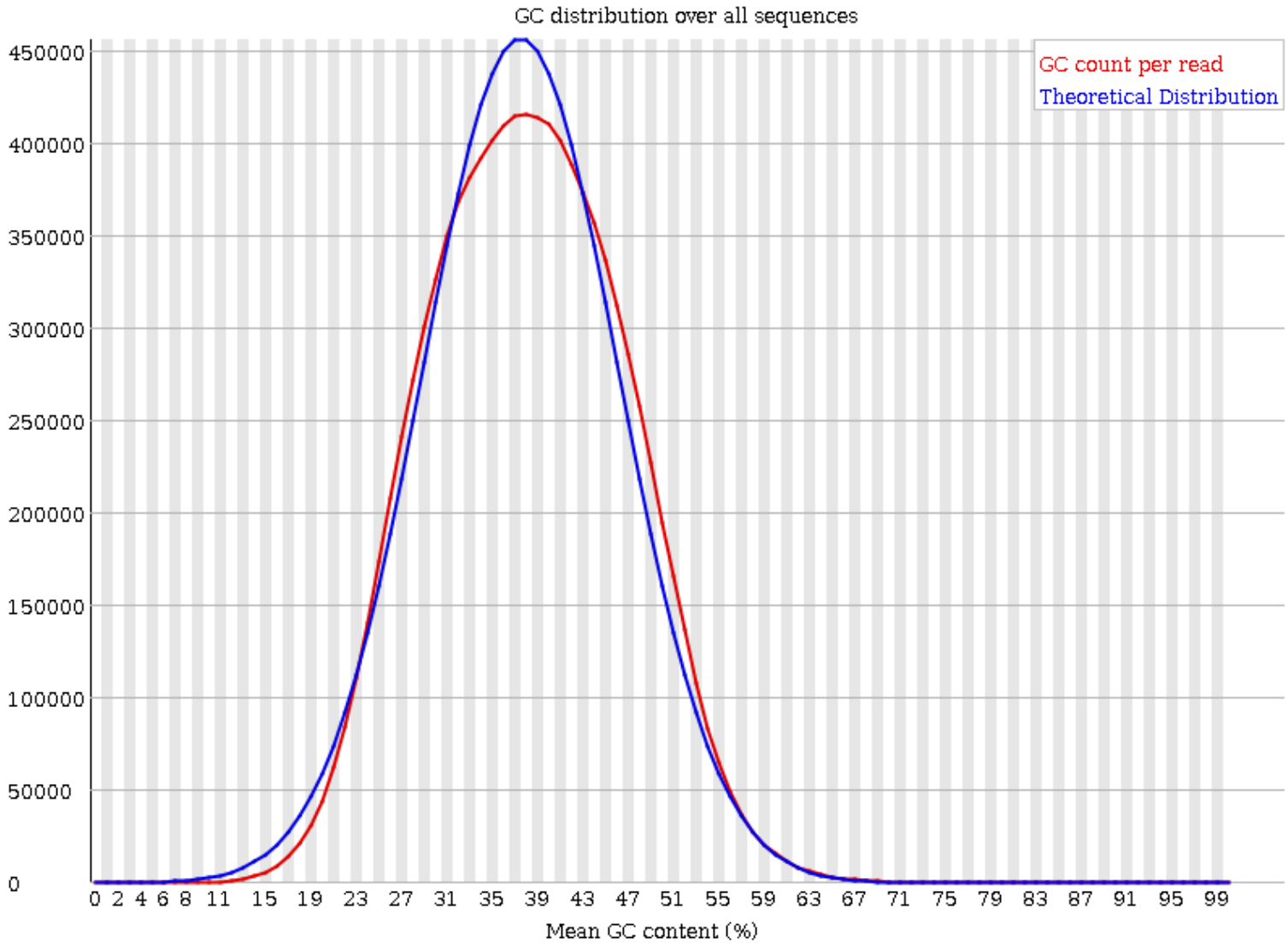
Base content

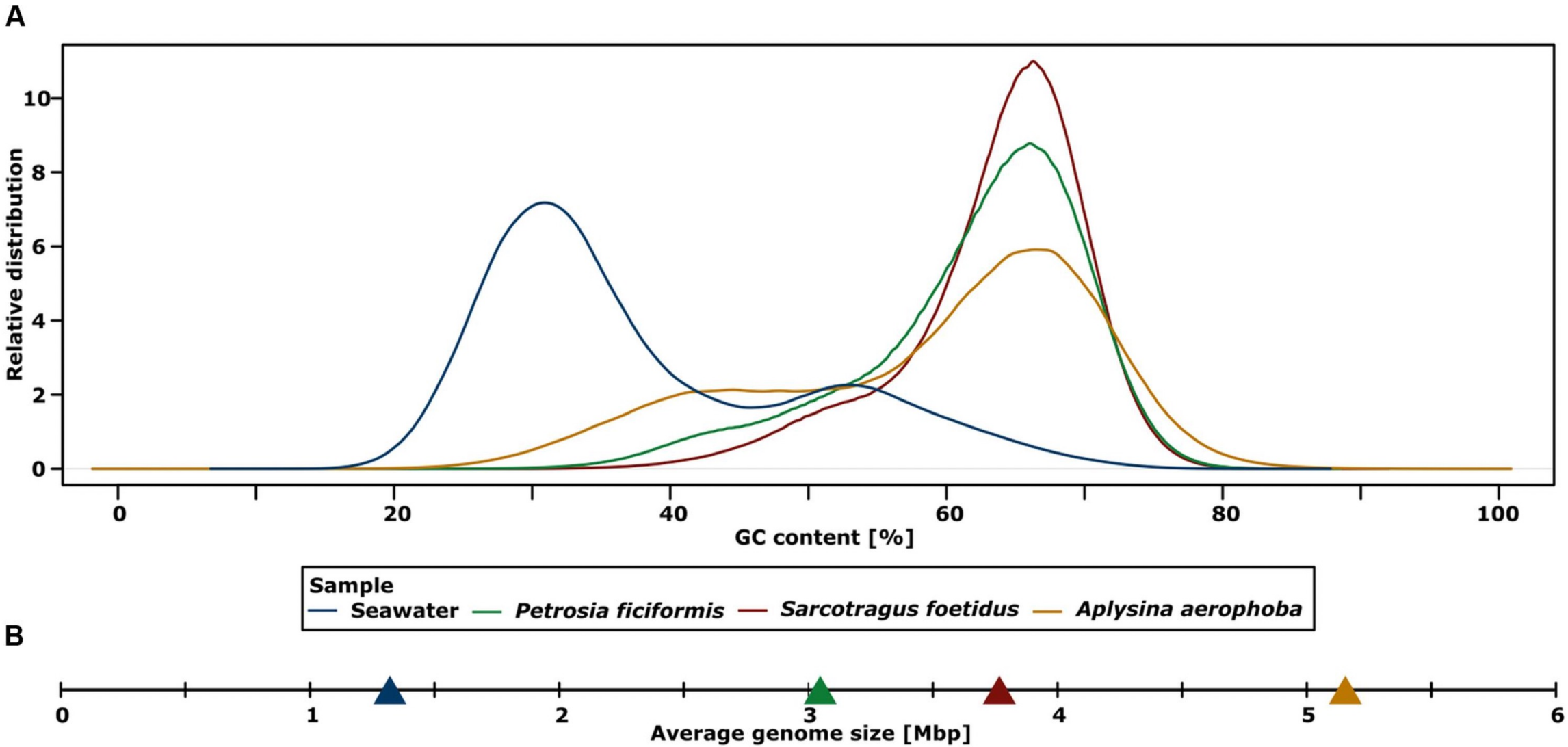
Good base content



Bad base content




GC distribution over all sequences
GC count per read
Theoretical Distribution
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)


A
Relative distribution
0
2
4
6
8
10
0
20
40
60
80
100
GC content [%]
Sample
Seawater
Petrosia ficiformis
Sarcotragus foetidus
Aplysina aerophoba
B
0
1
2
3
4
5
6
Average genome size [Mbp]

Herramientas Bioinformáticas para Caracterizar Bacteriófagos

Ensambles *de Novo* y/o identificación de contigs virales (profagos)

Validación y verificación de contigs virales

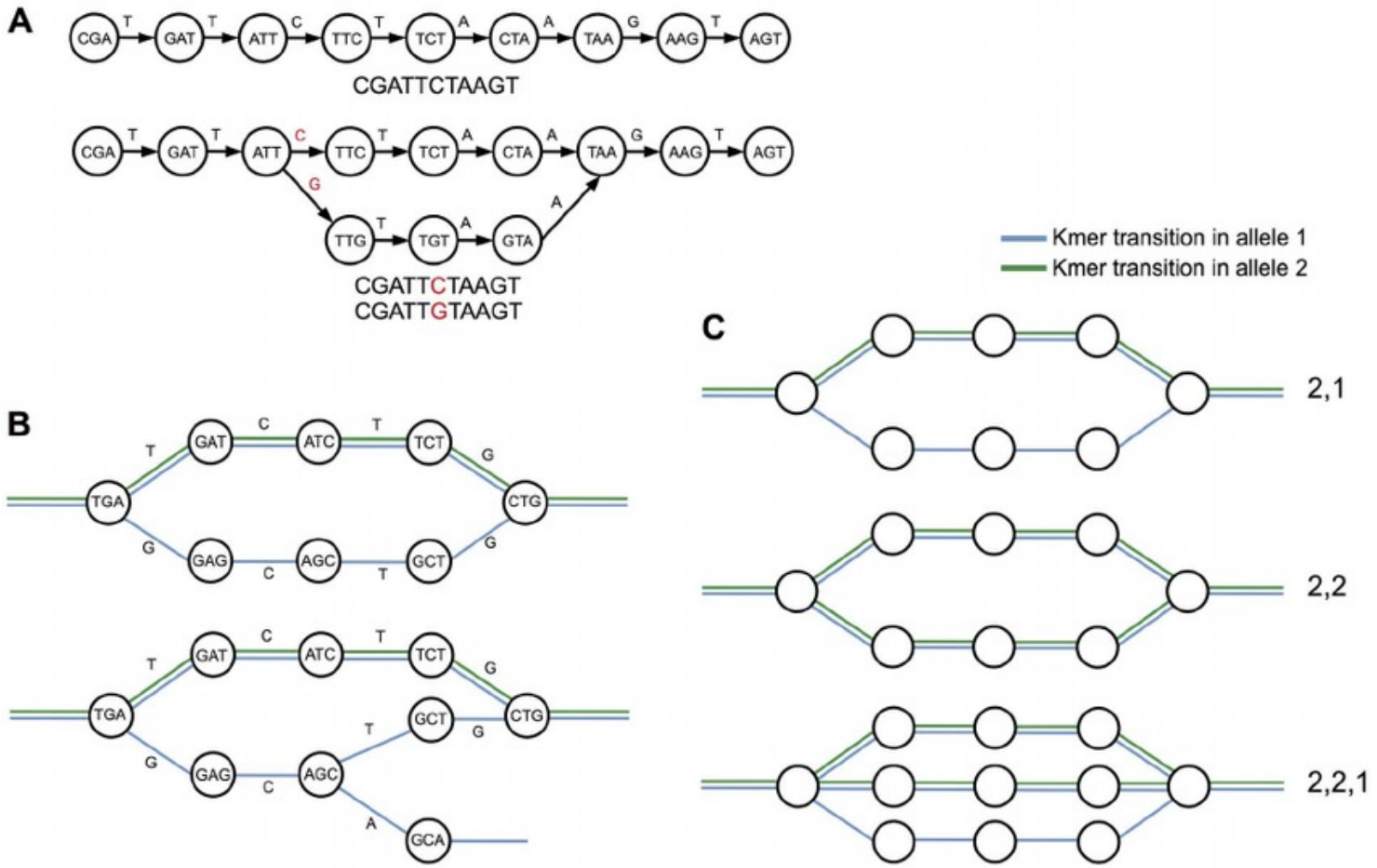
Ensambles *de Novo* y/o identificación de contigs virales (profagos)

Bioinformatics, 36(14), 2020, 4126–4129
doi: 10.1093/bioinformatics/btaa490
Advance Access Publication Date: 15 May 2020
Original Paper

Genome analysis

METAVIRALSPADES: assembly of viruses from metagenomic data

Dmitry Antipov[1,*], Mikhail Raiko[1], Alla Lapidus[1] and Pavel A. Pevzner[2]



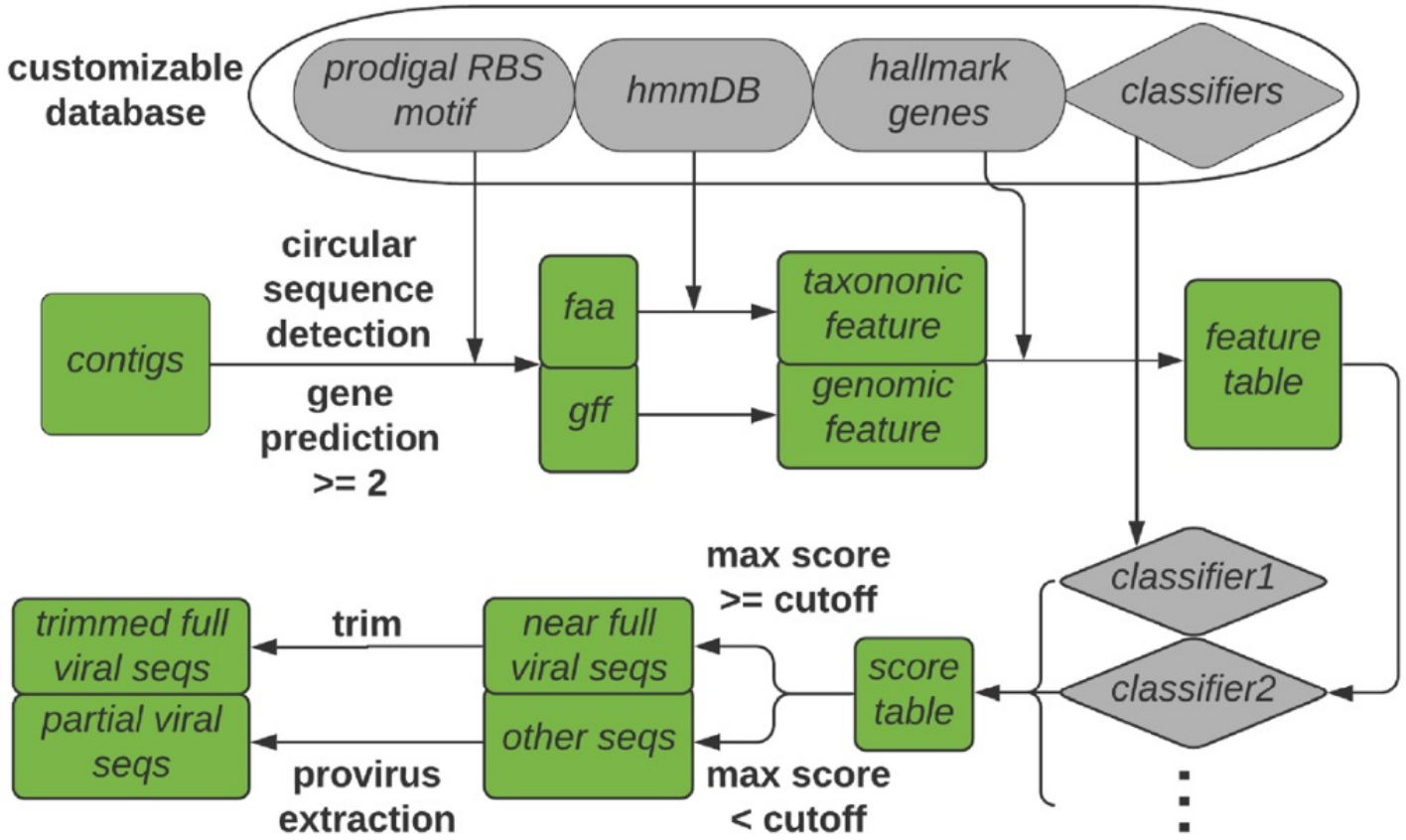
Ensambles *de Novo* y/o identificación de contigs virales (profagos)

SOFTWARE ARTICLE Open Access

VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses

Jiarong Guo[1], Ben Bolduc[1], Ahmed A. Zayed[1], Arvind Varsani[2,3], Guillermo Dominguez-Huerta[1], Tom O. Delmont[4], Akbar Adjie Pratama[1], M. Consuelo Gazitúa[5], Dean Vik[1], Matthew B. Sullivan[1,6,7*] and Simon Roux[8*]

- VirSorter2 (vs2)
- VirSorter (vs1)
- VirFinder (vf)
- DeepVirFinder (dvf)
- MARVEL (mv)
- VIBRANT (vb)



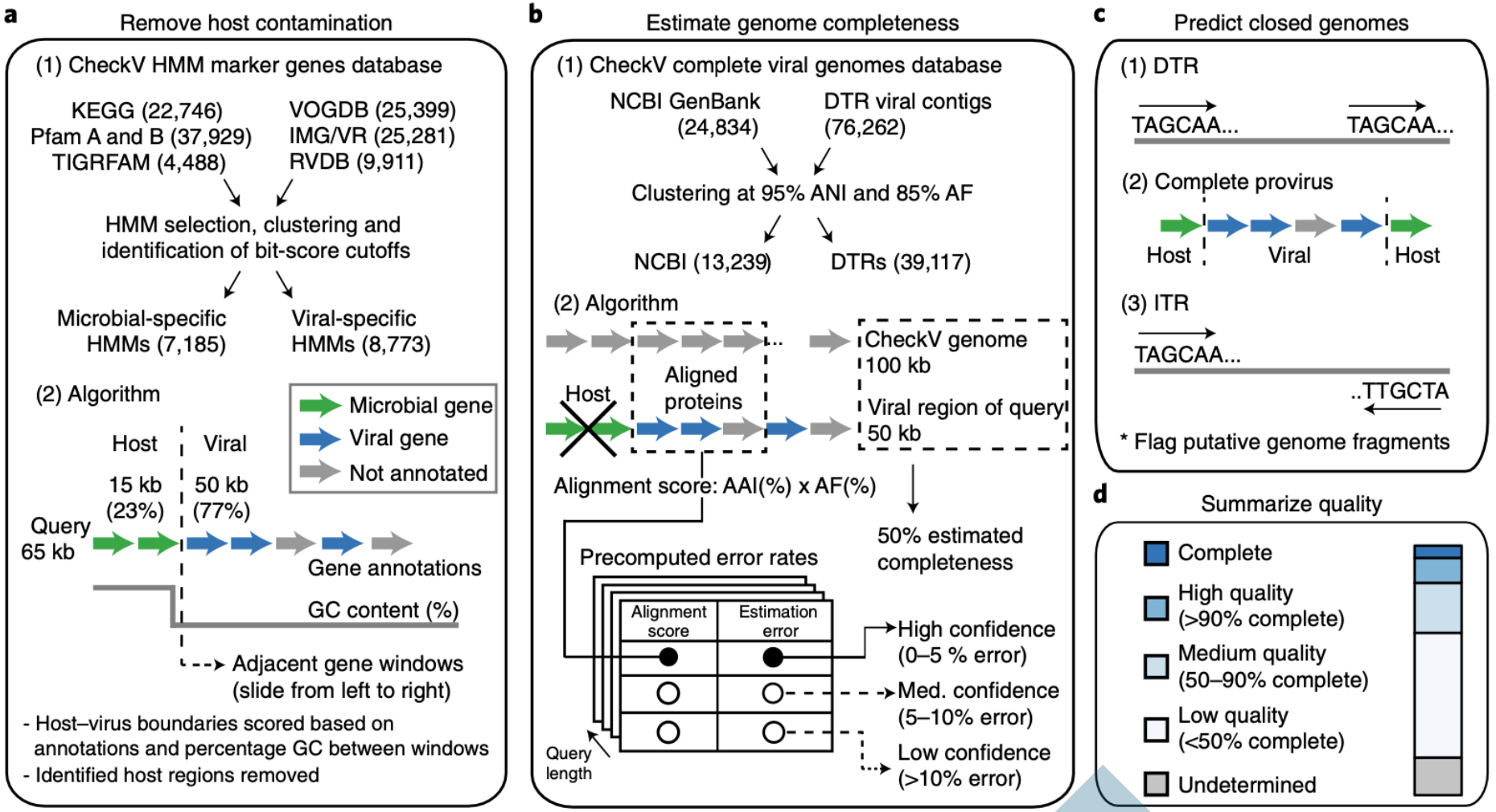
Validación y verificación de contigs virales

ARTICLES
https://doi.org/10.1038/s41587-020-00774-7
nature biotechnology
Check for updates

OPEN

CheckV assesses the quality and completeness of metagenome-assembled viral genomes

Stephen Nayfach[1✉], Antonio Pedro Camargo[2], Frederik Schulz[1], Emiley Eloe-Fadrosh[1], Simon Roux[1] and Nikos C. Kyrpides[1✉]



Sesión práctica