



#### Session 2.3 - Ensamblado

#### **Isabel Cuesta**

BU-ISCIII

<u>Unidades Comunes Científico Técnicas - SGSAFI-ISCIII</u>

14 al 18 Noviembre 2022 2ª Edición Programa Formación Continua, ISCIII



### Assembly

Reconstruct the sequence of the original DNA from shorter DNA sequences or small fragments known as reads

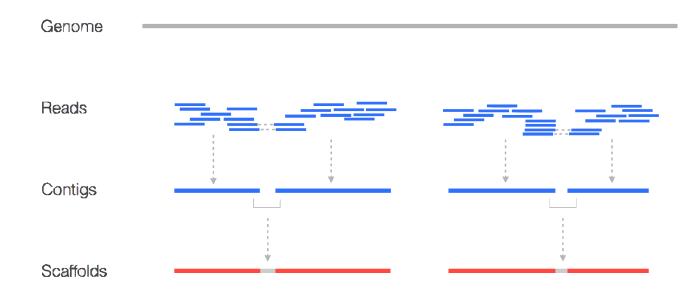
- **De novo:** with no previous knowledge of the genome to be assembled. It overlap the end of the end of each read in order to create a longer sequence.
- Assembly with reference: A similar but not identical genome guides the assembly process. Map reads over supplied genome.





### Assembly: contig y scaffold

- Contig: continuous sequence made up of overlaping shorter sequences
- Scaffold: two or more contigs located and rearranged according to spatial information (pair-end, mate pair, reference)



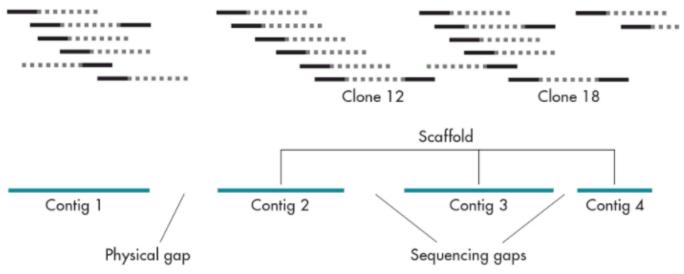
https://www.biostars.org/p/253222/





### Assembly: gaps

- Sequencing gaps: Position and orientation known by spatial information
- Physical gaps: No information about adjacent contigs



Gene Cloning, Lodge et al.



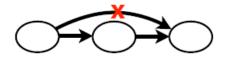


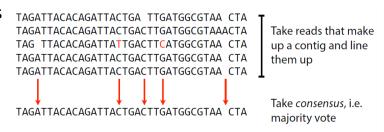
### Assembly: Algorithms

- Overlap, Layout, Consensus (OLC overlap graph):
  - O first overlaps among all the reads are found
  - L then it carries out a layout of all the reads and overlaps information on a graph
    - Removes redundant and low quality overlaps
  - C and finally the consensus sequence is inferred

Ex. Newbler, Mira, Celera Assembler, CAP3, PCAP, Phrap, Phusion.









### Assembly: Algorithms

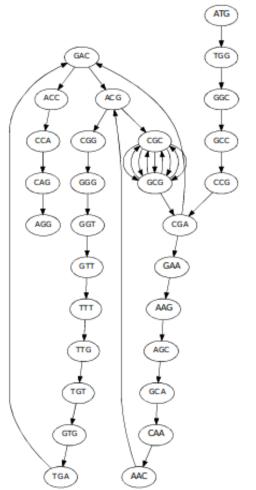
• De Brujin Graph (DBG: k-mer graph)

Chopping reads into much shorter k-mers (fixed length fragments) and then using all the k-mers to form a DBG and infer the contigs.

- Nodes in the graph are k-mers
- Edges represent consecutive k-mers (which overlap by k-n symbols)

Ex. SPAdes, ABySS, Velvet, AllPaths, Soap....

https://medium.com/@han\_chen



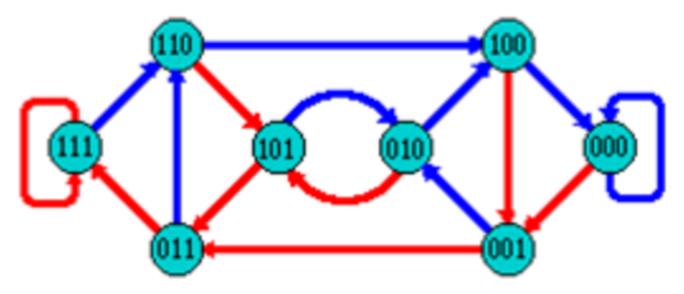




### de Bruijn Graphs

- A directed graph of sequences of symbols
- Nodes in the graph are k-mers
- Edges represent consecutive k-mers (which overlap by k-1 symbols)

Consider the 2 symbol alphabet (0 & 1) de Bruijn Graph for k =3

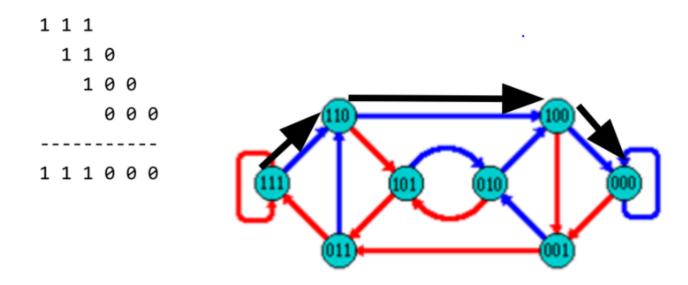






## Producing sequences

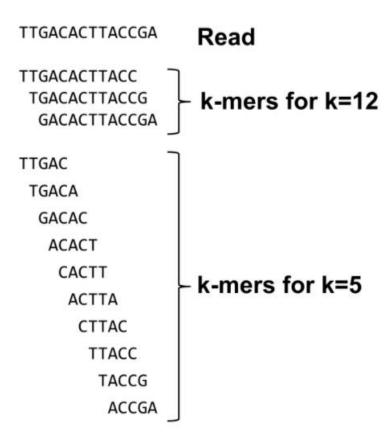
Sequences of symbols are produced by moving through the graph







### What are K-mers?









Example #1:

HAPPI PINE INESS APPIN







Example #1:

HAPPI PINE INESS APPIN

All 4-mers:

HAPP PINE INES APPI

APPI NESS PPIN

Unique 4-mers:

HAPP APPI PINE PPIN INES NESS







Example #1:

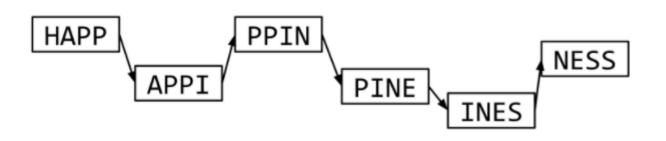
HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

PINE PPIN

**INES NESS** 









Example #1:

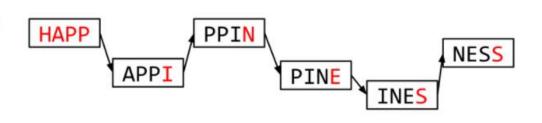
HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

PINE PPIN

**INES NESS** 



#### **HAPPINESS**

### Easy!







Example #2: MISSIS SSISSI SSIPPI







```
Example #2:
```

MISSIS SSISSI SSIPPI

All 4-mers (9):

MISS SSIS SSIP

ISSI SISS SIPP

SSIS ISSI IPPI

Unique 4-mers (7):

MISS SSIS SSIP ISSI SISS SIPP IPPI



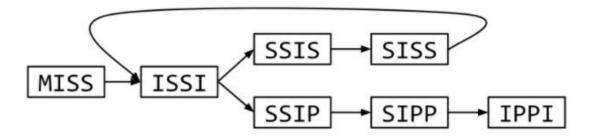


Example #2:

MISSIS SSISSI SSIPPI

All 4-mers:

MISS ISSI SSIS SISS SSIP SIPP IPPI







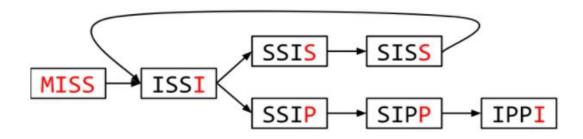


Example #2:

MISSIS SSISSI SSIPPI

All 4-mers:

MISS ISSI SSIS SISS SSIP SIPP IPPI



MISSISSIPPI or MISSISSISSIPPI or ...





Example #2a:

MISSIS SSISSI SSIPPI





Example #2a:

MISSIS SSISSI SSIPPI

All 5-mers (6):

MISSI SSISS SSIPP

ISSIS SISSI SIPPI

Unique 5-mers (6, no duplicates):
MISSI ISSIS SSISS SISSI SSIPP SIPPI



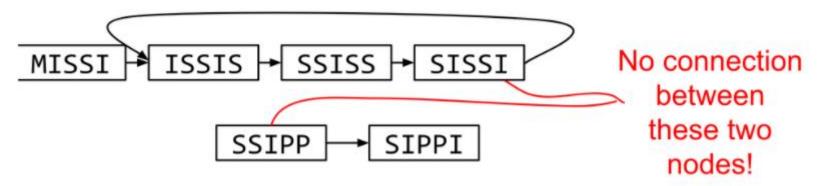


Example #2a:

MISSIS SSISSI SSIPPI

This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI





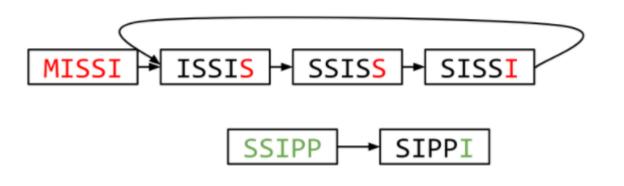


Example #2a:

MISSIS SSISSI SSIPPI

This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI



MISSISSIS

SSIPPI





# Choose k wisely

- Lower k
  - More connections
  - Less chance of resolving small repeats
  - Higher k-mer coverage
- Higher k
  - Less connections
  - More chance of resolving small repeats
  - Lower k-mer coverage

#### Optimum value for k will balance these effects.



### Read errors

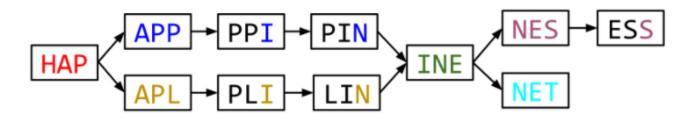


Example #3:

HAPPI INESS APLIN PINET

k = 3 k-mers:

HAP APP PPI INE NES ESS APL PLI LIN PIN NET



6 contigs: HAP APPIN APLIN INE NESS NET





# More coverage



- Errors won't be duplicated in every read
- Most reads will be error free
- We can count the frequency of each k-mer
- Annotate the graph with the frequencies
- Use the frequency data to clean the de Bruijn graph

#### More coverage depth will help overcome errors!



### **SPAdes**

 de Bruijn graph assembler by Pavel Pevzner's group out of St. Petersburg



- Uses multiple k-mers to build the graph
  - Graph has connectivity and specificity
  - Usually use a low, medium and high k-mer size together.
- Performs error correction on the reads first
- Maps reads back to the contigs and scaffolds as a check
- Under active development
- Much slower than Velvet
- Should be used in preference to Velvet now.



### Assembly: Scaffolding

From draft:

Order contigs (Nucmer, if there is reference it can be used to align and guide)

Fill the GAPs (GapFiller, fill sequencing gap (not physical gap)

Solve repeated sequence ambiguities (Expander)

Resequence with different library:

- Longer fragments and/or distance
- Tools for assembly improvement

SSPACE (Scaffolding) REAPR (evaluate scaffolding, breaking incorrect scaffolds)

Assembly visualyzing

Artemis, ACT (compare two or more sequences), Icarus (Quast)





### A move back to OLC

- New long read technologies
  - PacBio and MinIon
- Assemblers: HGap, CANU
  - Use overlap, layout consensus approach
- CANU can perform hybrid assemblies with long and short reads

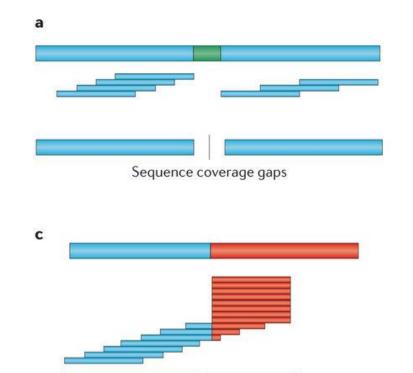




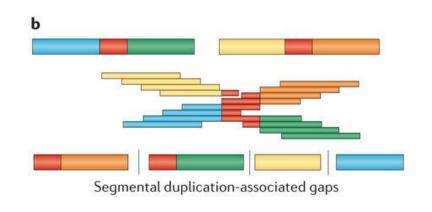


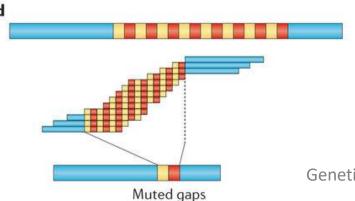


#### Ensamblado: Errores



Satellite-associated gaps





- A. Gaps región del genoma sin secuenciar
- B. Duplicaciones de gran tamaño
  - Quimeras
- Regiones repetidas colapsadas
  - C. Terminales
  - D. Intersticiales

Genetic variation and the de novo assembly of human genomes

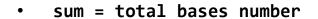
Chaisson et al. 2015

Nature Reviews | Genetics



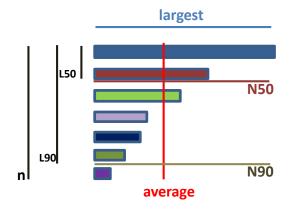


### Assembly: Metrics



- n = contigs number
- average = average contig length
- largest = largest contig
- N50 = length of the shortest contig where 50% of sum is held
- L50 = number of contigs which have 50% of the genome
- N90 = length of the shortest contig where 90% of sum is held.
- L90 = number of contigs which have 90% of the genome







#### Assembly: Evaluation

- Software that evaluate differets algorithms & parameters iMetAMOS, Koren et al., BMCBioinformatics 2014, 15:126 GAGE-B, Magoc et al., Bioinformatics 2013,29(14):1718-25
- **Graph evaluation**: Bandage, Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015)
- Assembly evaluation: Quast, Gurevich et al., Bioinformatics 2013, 29:8
- Metrics for a good assembly:

Large N50 Sum closest to expected Low n Low L50





### Assembly: Evaluation - Quast

• Assembly evaluation: Quast, Gurevich et al., Bioinformatics 2013, 29:8

Genome statistics	■ RA_L2073_paired_assembly	■ RA_L2391_paired_assembly	RA_L2677_paired_assembly	■ RA_L2978_paired_assembly	RA_L2281_paired_assembly	RA_L2450_paired_assembly	RA_L2701_paired_assembly
Genome fraction (%)	81.079	88.828	84.92	90.172	85.733	88.172	92.463
Duplication ratio	1	1	1.001	1.001	1.001	1	1
# genomic features	1736 + 824 part	2113 + 600 part	1881 + 768 part	2157 + 611 part	1992 + 637 part	2073 + 643 part	2368 + 412 part
Largest alignment	16612	33 033	21 336	25 068	29 638	30 305	40 471
Total aligned length	2 405 510	2 635 297	2 519 300	2 675 166	2 543 440	2 615 874	2743222
NGA50	3176	6162	4234	5948	5104	5358	9519
LGA50	267	151	219	153	166	166	96
Misassemblies							
# misassemblies	23	1	14	2	17	12	4
Misassembled contigs length	84193	9611	45 868	6390	111 490	72 879	37 962
Mismatches							
# mismatches per 100 kbp	17	18.78	15	16.71	341.39	15.75	13.49
# indels per 100 kbp	1.21	1.25	1.87	1.94	7.27	1.45	0.87
# N's per 100 kbp	0	0	0	0	0	0	0
Statistics without reference	e						
# contigs	748	546	684	569	569	584	392
Largest contig	16612	33 033	21 336	25 068	30 915	30 305	40 471
Total length	2 440 656	2 676 227	2 562 578	2714287	2 629 607	2 618 624	2787129
Total length (>= 1000 bp)	2 439 127	2 676 227	2 559 569	2714287	2 628 029	2 615 105	2 785 415
Total length (>= 10000 bp)	257 236	739 181	320 638	811 392	700516	658 31 9	1 419 641
Total length (>= 50000 bp)	0	0	0	0	0	0	0

Extended report





### Assembly: Evaluation - Quast

 Assembly evaluation: Quast, Gurevich et al., Bioinformatics 2013, 29:8







### Assembly: Assemblers

Name	Туре	Technologies	Author	Presented /Last updated	Licence*	Homepage
<u>DNASTAR</u> Lasergene Genomics Suite	(large) genomes, exomes, transcriptomes, metagenomes, ESTs	Illumina, ABI SOLiD, Roche 454, Ion Torrent, Solexa, Sanger	DNASTAR	2007 / 2016	С	link
Newbler	genomes, ESTs	454, Sanger	454/Roche	2004/2012	С	link
<u>Canu</u>	Small and large, haploid/diploid genomes	PacBio/Oxford Nanopore reads	Koren et al. <sup>[8]</sup>	2001 / 2018	os	link
<u>SPAdes</u>	(small) genomes, single- cell	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	Bankevich, A et al.	2012 / 2017	os	link
<u>Velvet</u>	(small) genomes	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007 / 2011	OS	link

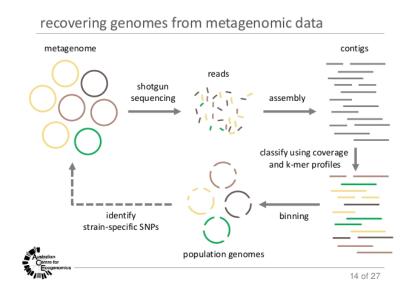
<sup>\*</sup>Licences: OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics





### Assembly: Specials assemblers

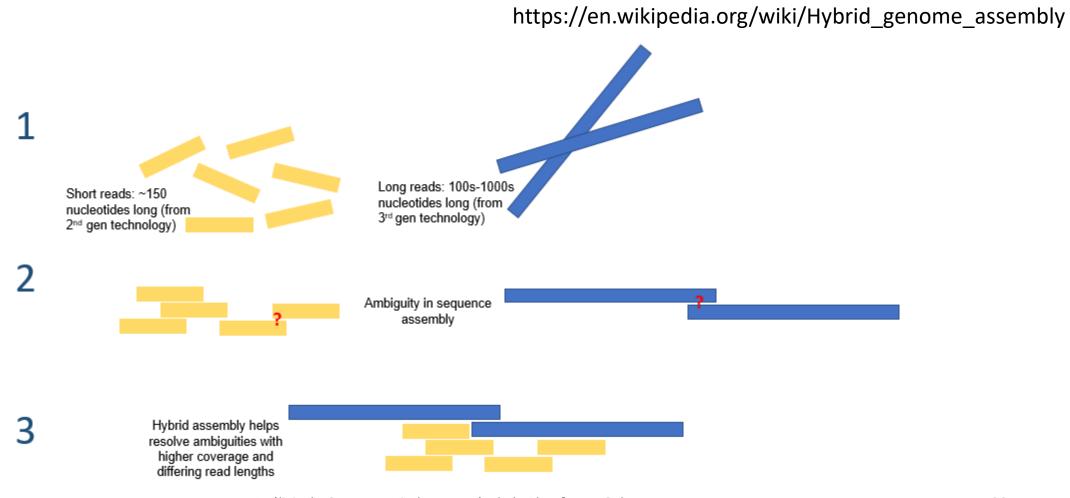
- Diploid genomes
- Metagenomics
- Plasmids
- Transcriptome
- Virus
  - VICUNA: population consensus genome assembly
  - IVA: assembler for RNA viruses







### Hybrid genome assembly - short and long reads







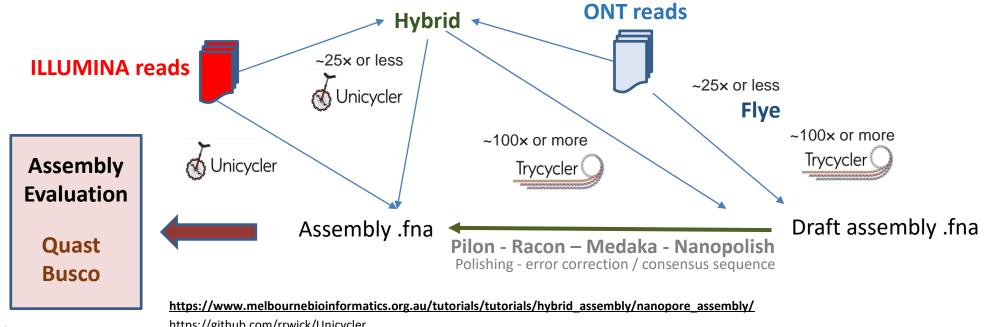
### Hybrid genome assembly - nanopore and illumina

Short reads (ILLUMINA) + Long reads (ONT)  $\rightarrow$  deNovo assembly (De novo assembly is the process of assembling a genome from scratch using only the sequenced reads as input - no reference genome is used.)  $\rightarrow$  high-quality assembly

**ONT**: >40.000b, higher error rate – **genome structure** 

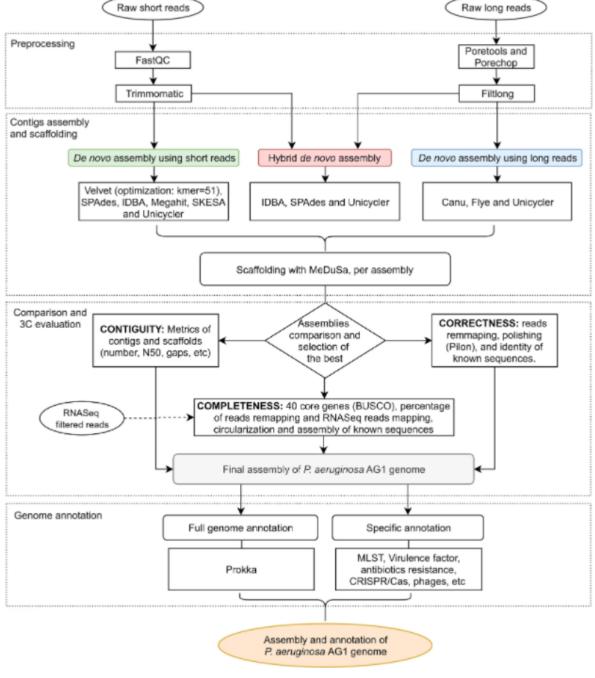
**ILLUMINA**: 300b, lower error rate – **high base-level accuracy** 

**Higher COST** 







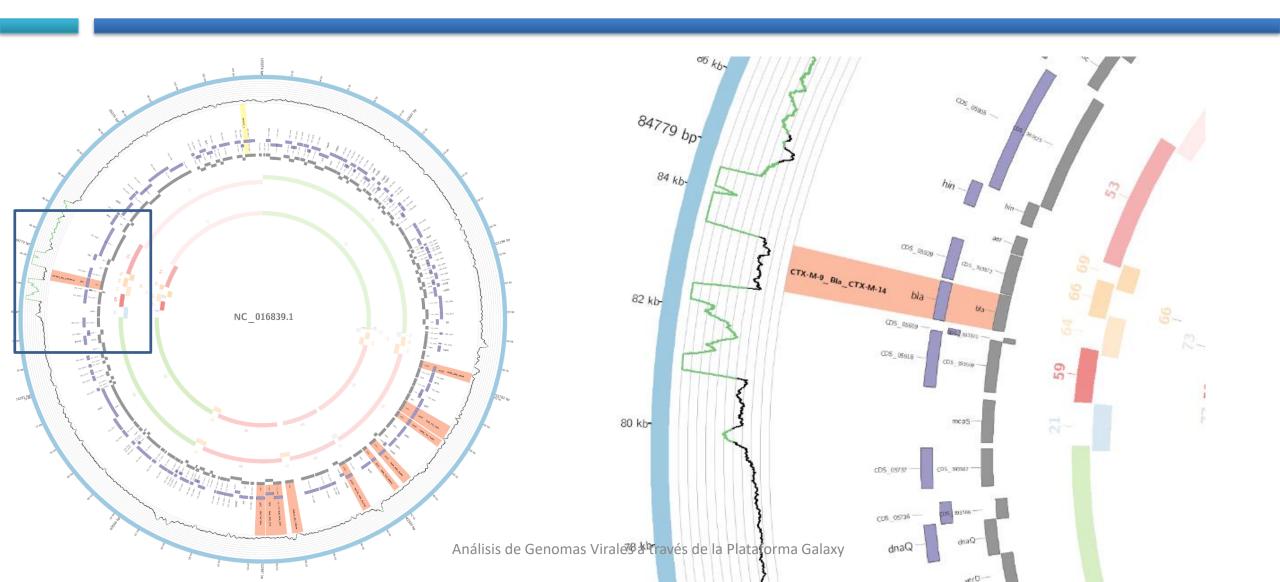


Molina-Mora et al., Scientific Reports 2020





#### PlasmidID







### Thanks for your attention!

### Questions?