Secuenciación del genoma de bacterias: ensamblado y anotación

Unidad de Bioinformática
Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

20-24 Mayo 2024, 11ª Edición
Programa Formación Continua, ISCIII

Ensamblado

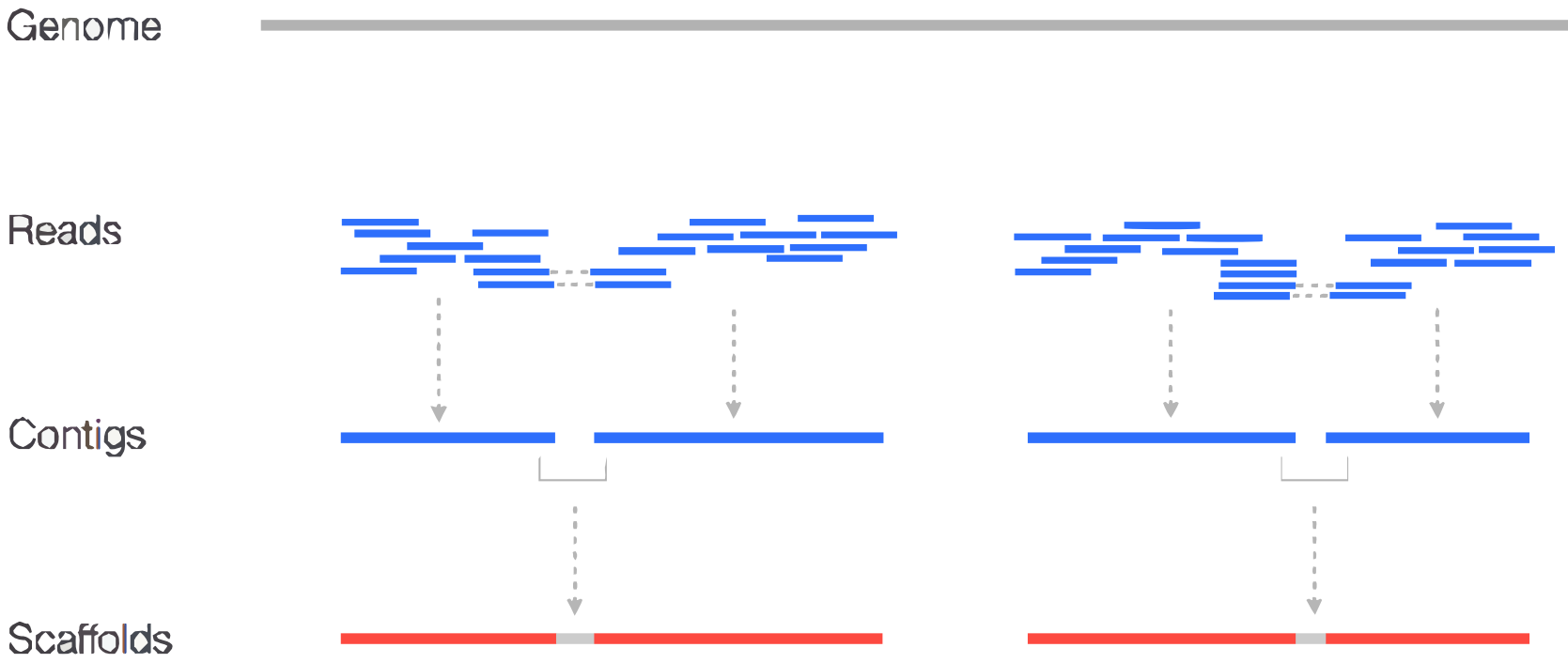
Reconstruir la **secuencia de DNA original** a partir de **lecturas** o secuencias de mucho menos tamaño.

- ***De novo:*** sin ningún tipo de conocimiento previo a cerca del genoma a ensamblar. Busca lecturas cuyo final coincida con el principio de otra para formar fragmentos del mayor tamaño posible.

- ***Usando Referencia:*** se usa un genoma como guía que suponemos es similar al que se quiere ensamblar.

Ensamblado: contig y scaffold

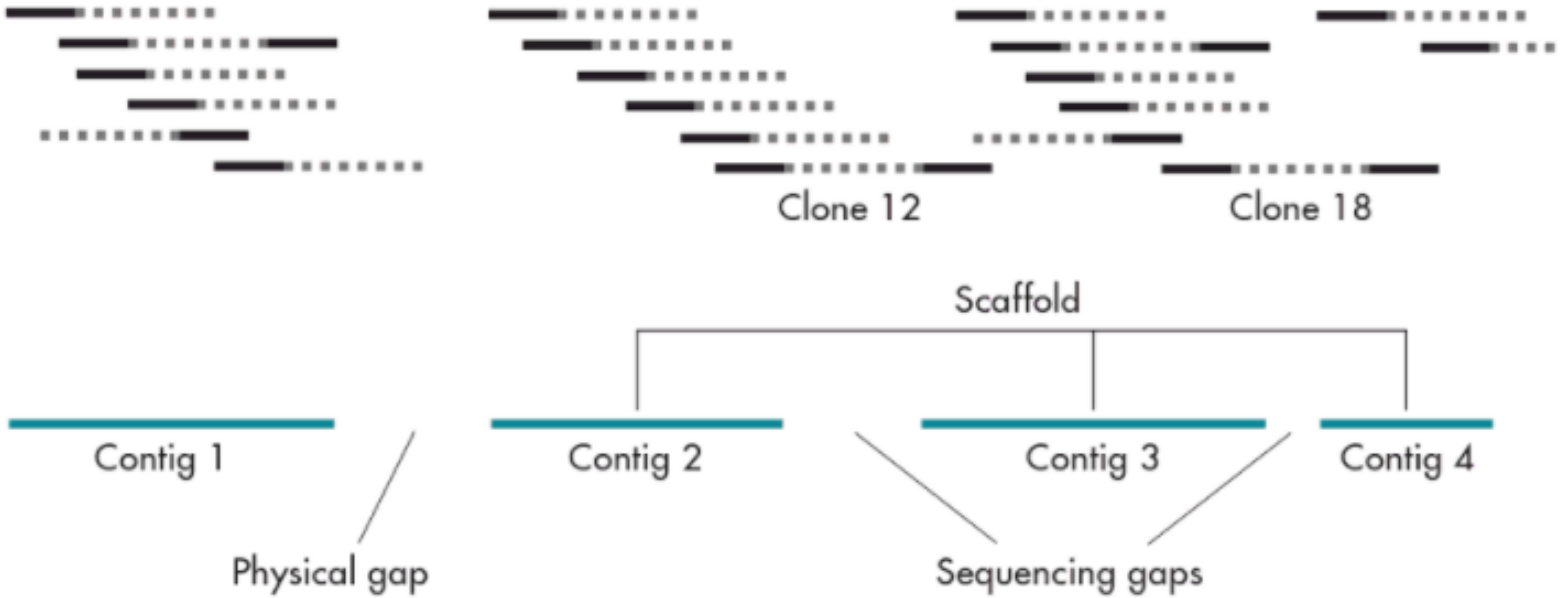
- **Contig**: secuencia continua del genoma formada por lecturas solapantes
- **Scaffold**: dos o más contigs unido por información de longitudes conocidas (pair-end, mate pair, referencia)



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

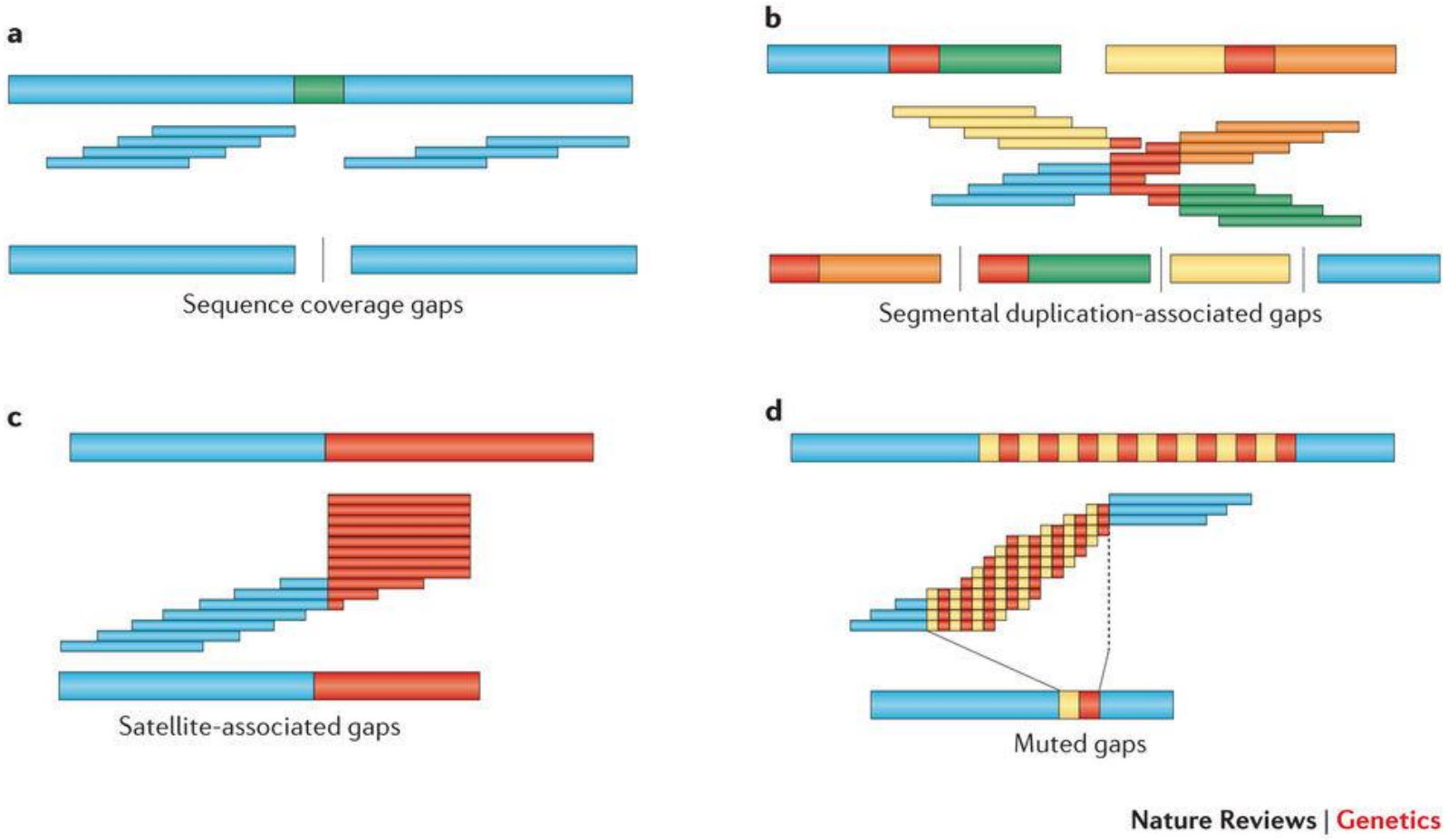
Ensamblado: gaps

- **Sequencing gaps**: sabemos el orden y orientación de los contigs por tener al menos un par que cubre ambos contigs
- **Physical gaps**: no tenemos información entre contigs adyacentes



Gene Cloning, Lodge *et al*.

Ensamblado: Errores



- **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

Ensamblado: Algoritmos

- **Overlap, Layout, Consensus (OLC - overlap graph):**

Overlap: Busca todos los pares de secuencia que solapan; Layout: Quita solapamientos redundantes y de baja calidad; Consensus: Alinea las secuencias que solapan solo entre ellas.

Ej. Newbler, Mira….

- **De Brujin (k-mer graph)**

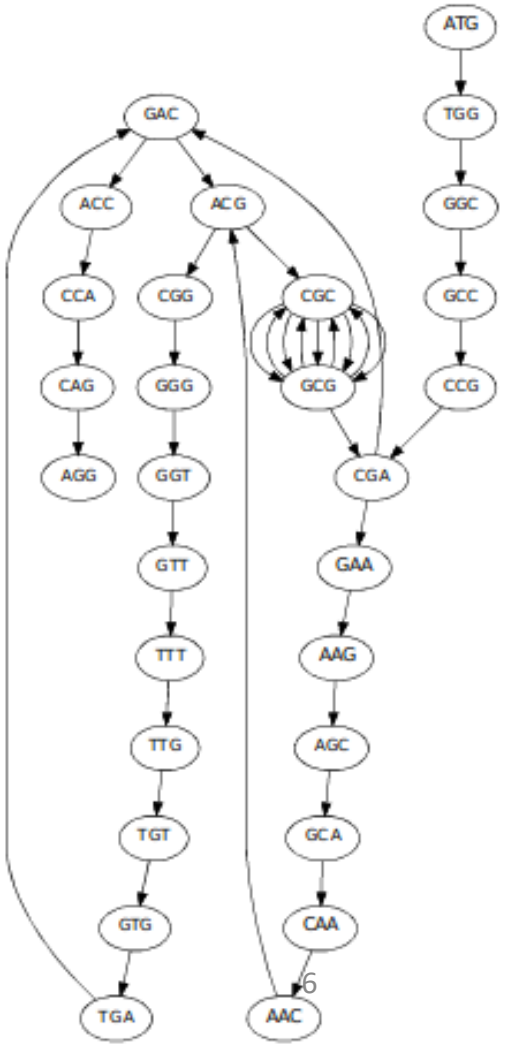
Grafos de Brujin: Elaboración de un grafo de k-mers (fragmentos de secuencia de longitud fija) donde se representan todos los solapamientos entre k-mers. Se unen nodos, burbujas y selección del mejor camino hasta un grafo irreducible del que se obtienen los contigs.

Ej. SPAdes, ABySS, Velvet, AllPaths, Soap….

- **Burrows Wheeler transform (FM-index):**

OLC usando el algoritmo “Ferragina-Manzine index” para encontrar todos los pares de secuencias que solapan de manera eficiente (rápida).

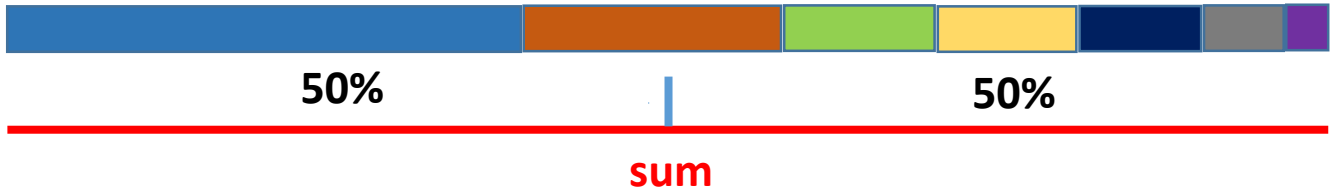
Ej. Assembler SGA, String Graph…

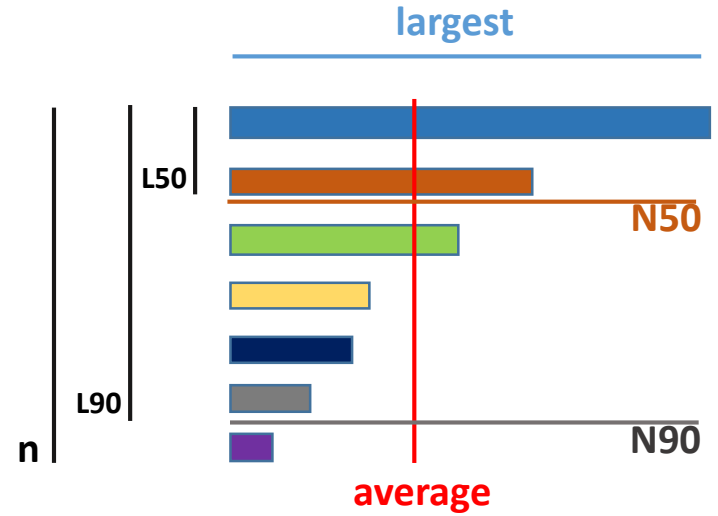


https://medium.com/@han_chen

Ensamblado: Métricas

- **sum** = numero total de bases
- **n** = numero total de contigs
- **average** = promedio de longitud de los fragmentos
- **largest** = bases en el fragmento mas largo
- **N50** = el tamaño mas corto de los contigs en donde el 50% de sum (el total de bases) esta contenido.
- **L50** = numero de contigs en donde tengo el 50% del genoma
- **N90** = el tamaño mas corto de los contigs en donde el 90% de sum (el total de bases) esta contenido. Un buen ensamblado el N90 a veces es casi igual al tamaño promedio de contig.
- **L90** = numero de contigs en donde tengo el 90% del genoma





Ensamblado: Scaffolding – Genoma completo

- **A partir del draft:**

  Ordenar contigs (Nucmer, si hay **referencia** la usamos para alinear y orientar contigs)

  Completar los GAPs (GapFiller, rellena los gaps de los contigs – sequencing gap)

  Resolver ambigüedades por repeticiones (Expander)

  Volver a secuenciar con una librería de mayor fragmento y/o distinta plataforma

- **Herramientas que mejoran los ensamblados**

  SSPACE (hace Scaffolding) REAPR (Evalúa el scaffolding, rompiendo los scaffolds incorrectos)

- **Visualizar un ensamblado**

  Artemis, ACT (comparación de dos o más secuencias)

Ensamblado: Evaluación

- Software que evalúa diferentes algoritmos y parámetros

  iMetAMOS, *Koren et al., BMCBioinformatics 2014, 15:126*

  GAGE-B, *Magoc et al.*, Bioinformatics 2013,29(14):1718-25

- Evaluación del ensamblado: **Quast**, *Gurevich et al., Bioinformatics 2013, 29:8*

- **Criterios elección mejor ensamblado:**

  N50 mas grande

  Num. total de bases más cercano a lo esperado

  Menos contigs totales

  Menos contigs tanto en L50 como L90

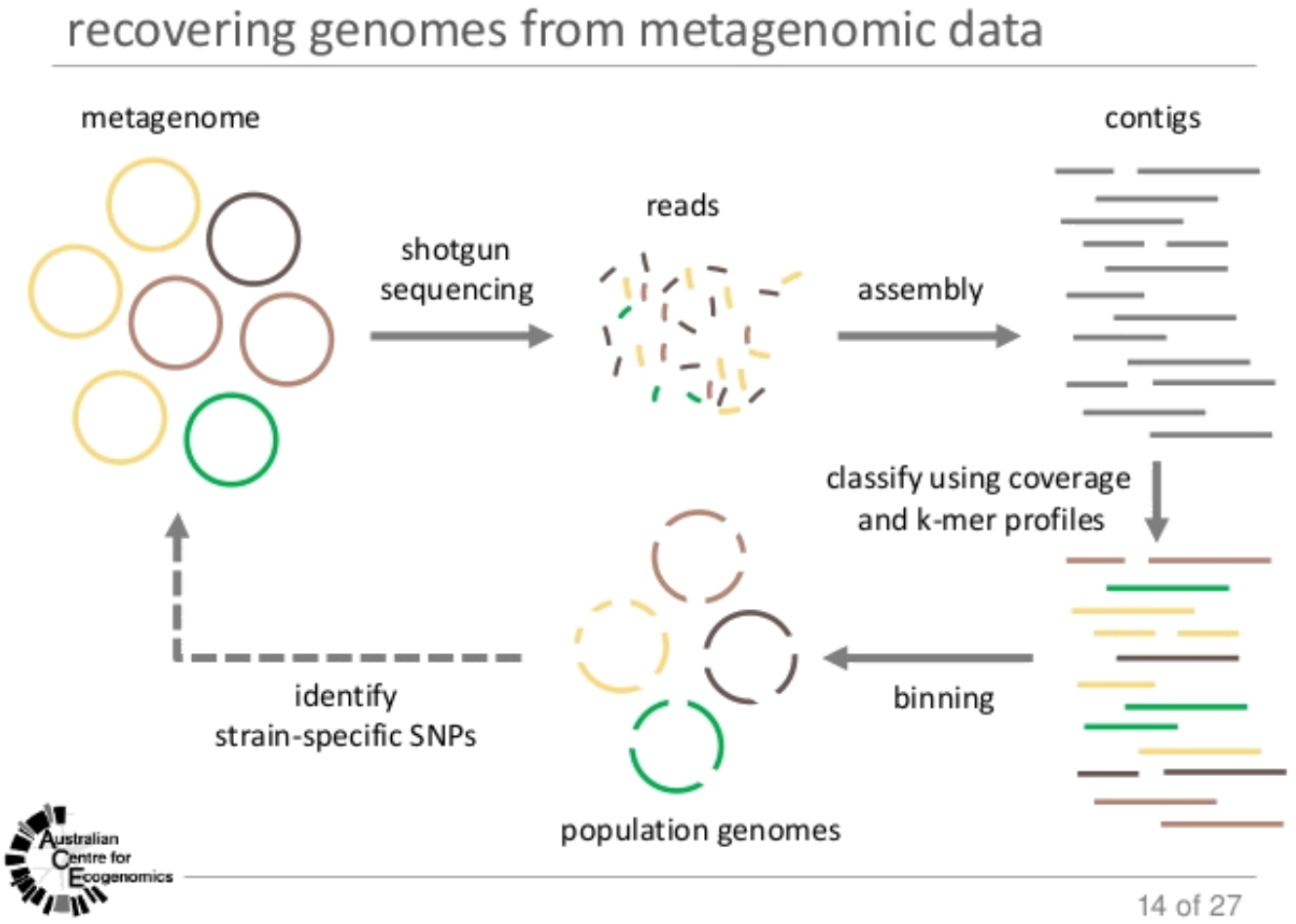
Ensamblado: Ensambladores

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

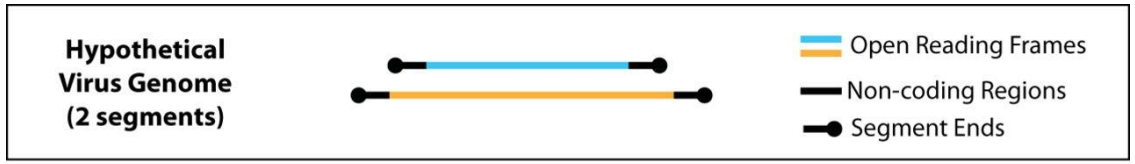
***Licences:** OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics

Ensamblado: Ensamblados especiales

- Genomas diploides
- Metagenomas
- Plásmidos
- Transcriptoma



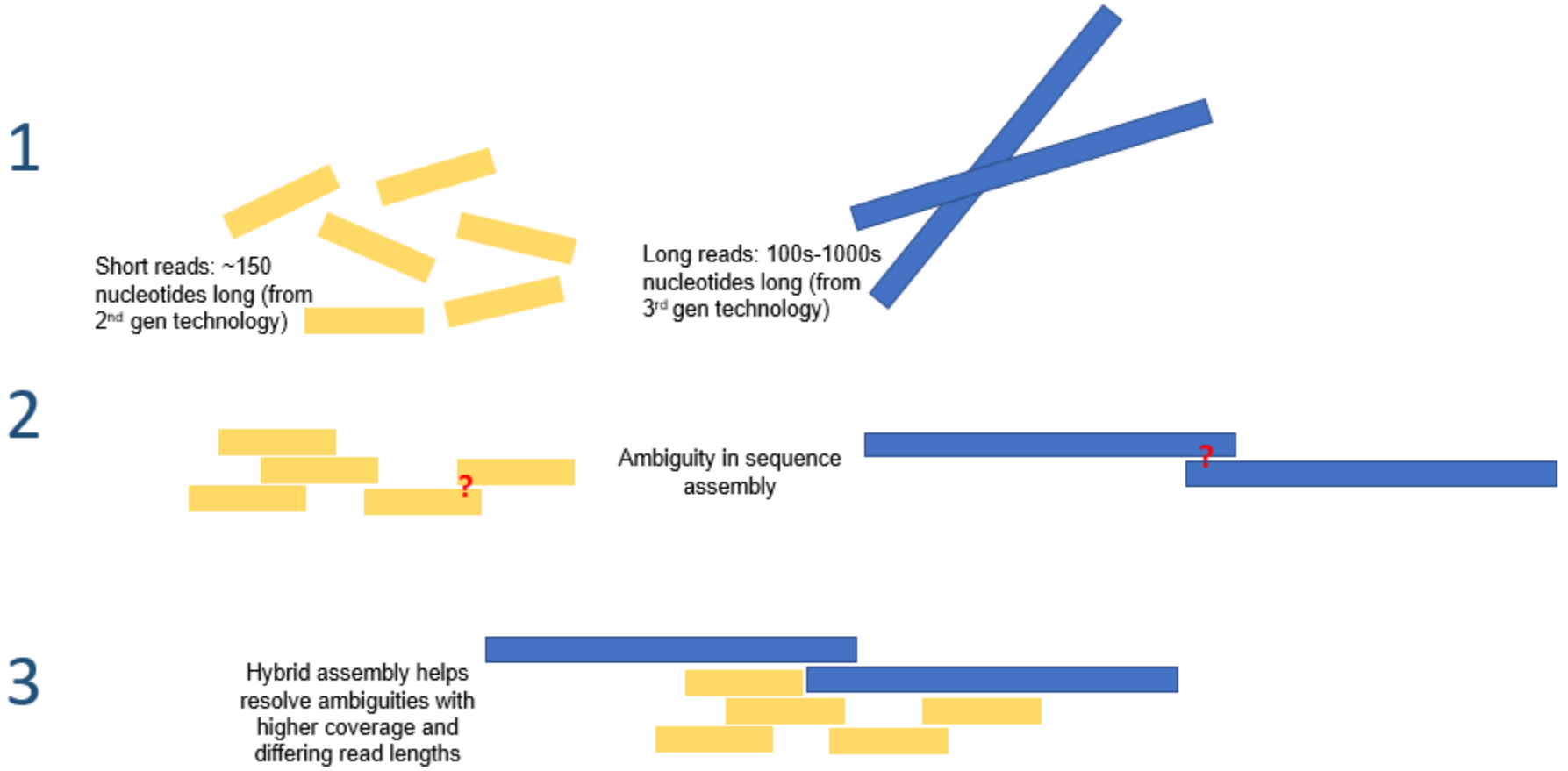
Ensamblado: Categorías



Category		*Potential Uses*
Standard Draft (SD) - Fragmented segments		- Taxonomic identification - Design of inclusivity tests
High Quality (HQ) - Single contig per segment - Incomplete ORFs		- Comparative genomics
Coding Complete (CC) - Complete ORFs - Missing ends		- Development of immunological assays
Complete - Full genome		- Design of exclusivity tests - Reverse genetics - Microbial forensics
Finished - Characterization of population-level variability		- Countermeasure development - Animal model development

Standards for Sequencing Viral Genomes in the Era of HighThroughput Sequencing. Ladner *et al*.

Hybrid genome assembly – short and long reads



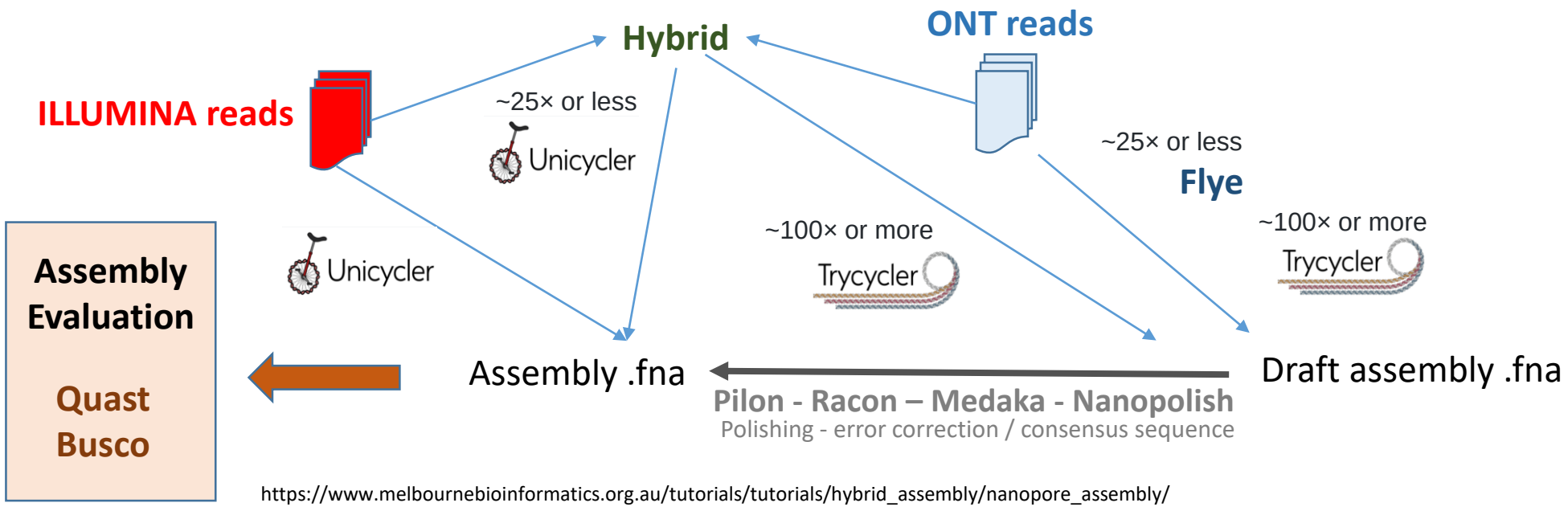
https://en.wikipedia.org/wiki/Hybrid_genome_assembly

Hybrid genome assembly - nanopore and illumina

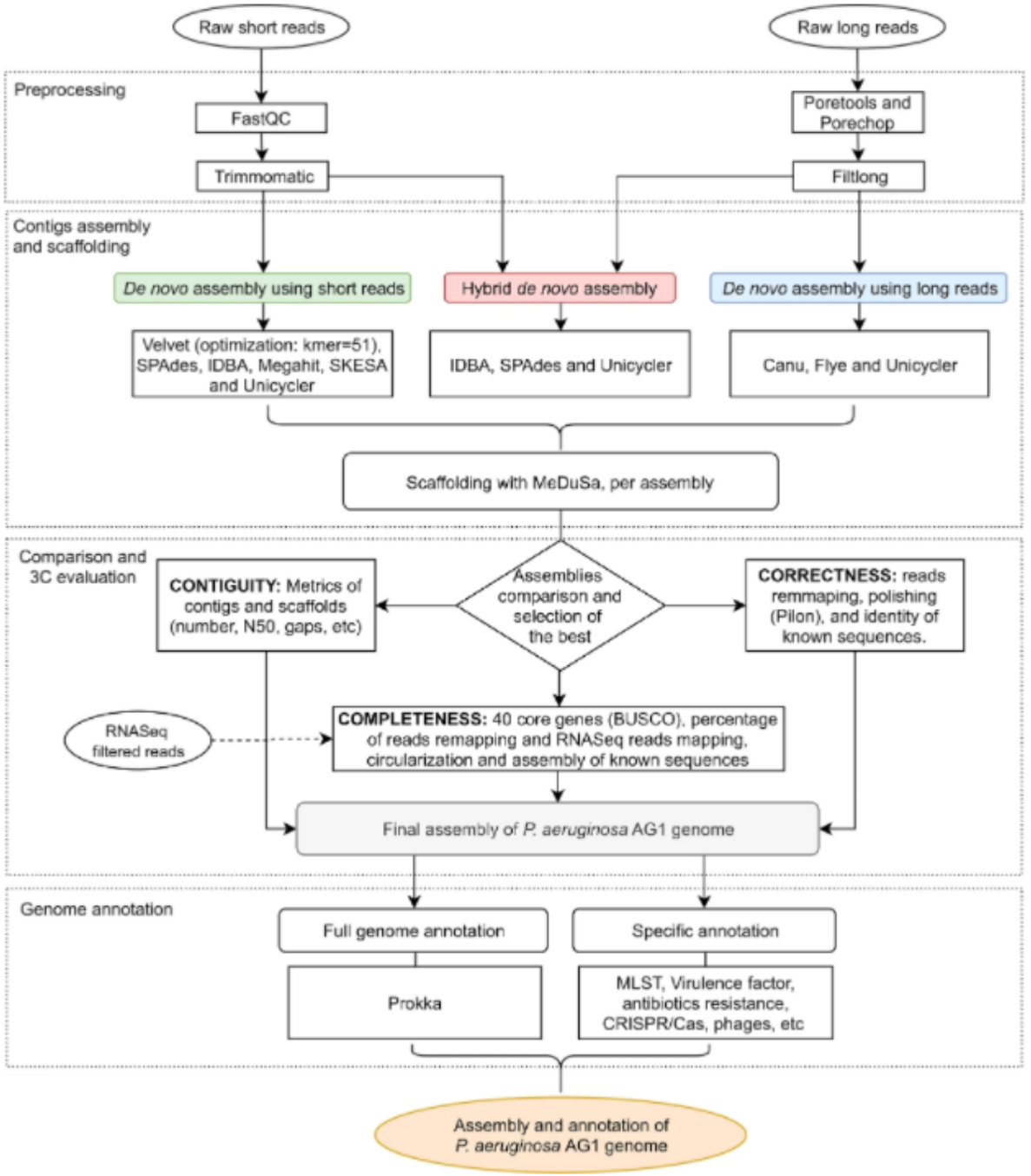
Short reads (ILLUMINA) + Long reads (ONT) → deNovo assembly (De novo assembly is the process of assembling a genome from scratch using only the sequenced reads as input - no reference genome is used.) → **high-quality assembly**

ONT: >40.000b, higher error rate – **genome structure**
ILLUMINA: 300b, lower error rate – **high base-level accuracy**

Higher COST



https://www.melbournebioinformatics.org.au/tutorials/tutorials/hybrid_assembly/nanopore_assembly/
https://github.com/rrwick/Unicycler
https://denbi-nanopore-training-course.readthedocs.io/en/latest/index.html



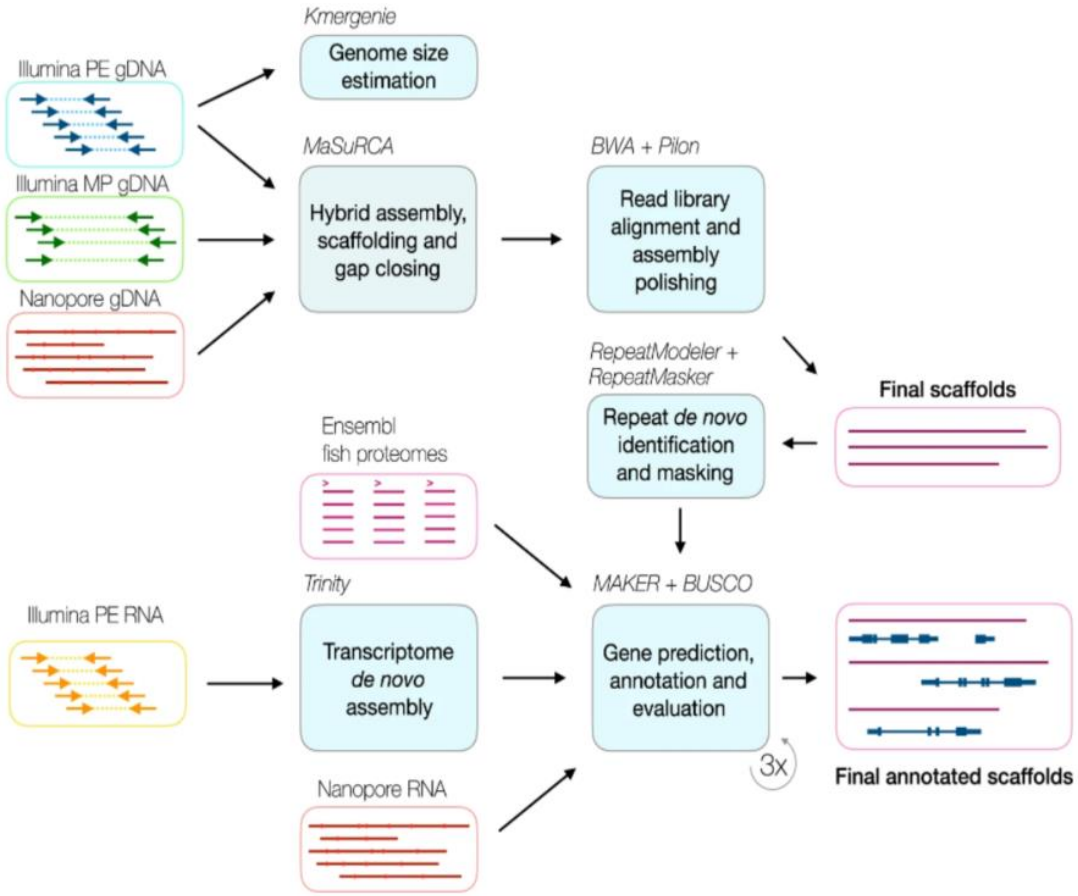
General bioinformatic pipeline to assemble, compare and annotate the *Pseudomonas aeruginosa* AG1 genome using short and long reads as well as hybrid approaches.

Molina-Mora et al., Scientific Reports 2020

Hybrid genome assembly and annotation of *Danionella translucida*

Fig. 2

From: Hybrid genome assembly and annotation of *Danionella translucida*

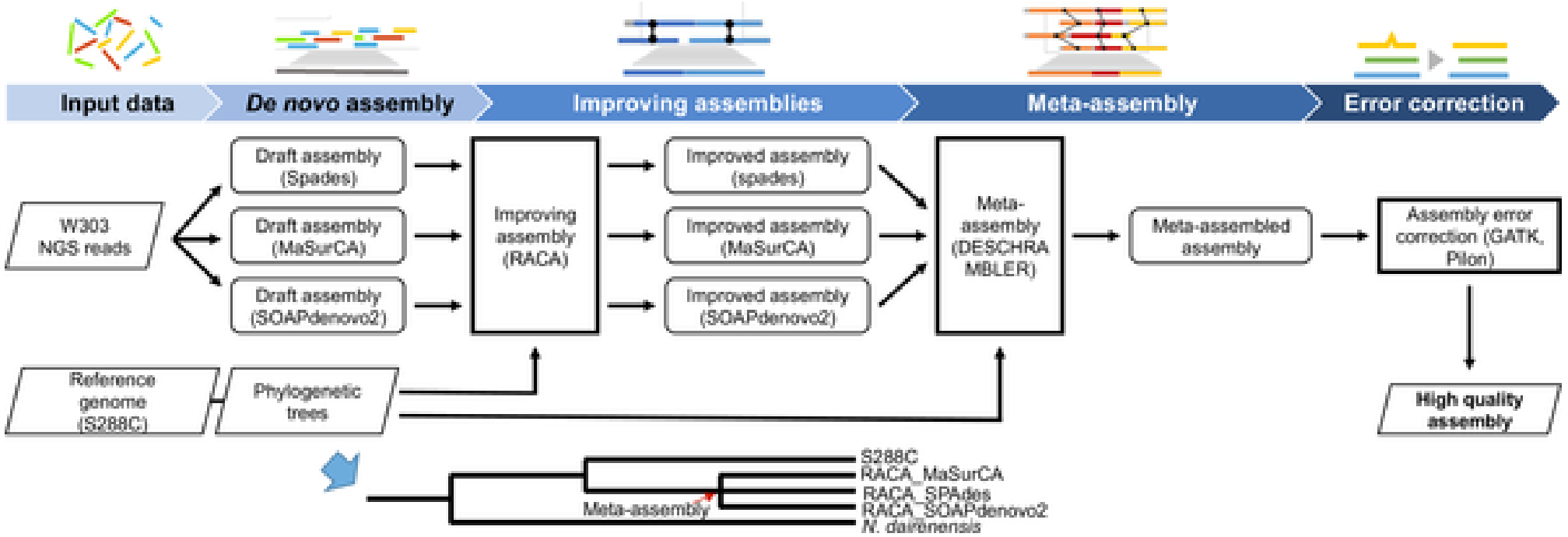


DT genome assembly and annotation pipeline. **PE**, paired-end; **MP**, mate-pair.

Genome assembly statistics	
Total scaffolds	27,639
Total contigs	36,005
Total scaffold sequence	735.303 Mb
Total contig sequence	725.703 Mb
Gap sequences	1.306%
Scaffold N50	340.819 kb
Contig N50	133.131 kb
Longest scaffold	3.085 Mb
Longest contig	995.155 kb
Fraction of genome in >50 kb scaffolds	88.3%
BUSCO genome completeness score	
Complete	91.5%
Single	87.0%
Duplicated	4.5%
Fragmented	3.6%
Missing	4.9%
Total number of Actinopterygii orthologs	4,584

Kadobianskyi et al., Scientific Data 2019

Fig 1. Data flow chart of the integrative meta-assembly pipeline (IMAP).



Song G, Lee J, Kim J, Kang S, Lee H, et al. (2019) Integrative Meta-Assembly Pipeline (IMAP): Chromosome-level genome assembler combining multiple de novo assemblies. PLOS ONE 14(8): e0221858. https://doi.org/10.1371/journal.pone.0221858
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221858

Table 1. Assembly evaluation metrics and results.

Dataset (W303)		MIN (bp)	MAX (bp)	N50 (bp)	Total length (bp)	Mapped reads (%)	Proper pairs (%)
De novo assembly	SPAdes	80	515,973	187,035	13,901,101	99.40	96.24
	MaSurCa	300	784,921	273,283	11,838,299	82.86	97.88
	SOAPdenovo2	200	61,911	13,286	11,749,637	50.60	88.48
RACA assembly	RACA-SPAdes	80	1,058,428	716,084	13,905,771	99.40	96.24
	RACA-MaSurCa	300	1,436,612	706,991	11,842,202	82.86	97.88
	RACA-SOAPdenovo2	200	1,076,849	69,631	11,772,637	50.60	88.49
Meta assembly	Meta	80	1,448,740	702,641	13,773,679	98.56	96.19
Final assembly	Corrected assembly	80	1,450,556	705,629	13,847,490	98.57	97.10
PacBio	PacBio	3,688	1,575,129	929,095	12,433,409	99.15	98.73

https://doi.org/10.1371/journal.pone.0221858.t001

Song G, Lee J, Kim J, Kang S, Lee H, et al. (2019) Integrative Meta-Assembly Pipeline (IMAP): Chromosome-level genome assembler combining multiple de novo assemblies. PLOS ONE 14(8): e0221858. https://doi.org/10.1371/journal.pone.0221858
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221858

Human Reference Consortium

General	
Assembly name	GRCh38.p13
Release date	2019-03-01
Assembly type	haploid-with-alt-loci
Release type	patch
Assembly units	38
Total bases	3,272,116,950
Total non-N bases	3,110,748,599
Primary assembly N50	67,794,873
Regions	
Total regions	358
Regions with alternate loci	178
Regions with FIX patches	113
Regions with NOVEL patches	72
Regions as PAR	4
Alternate loci and patches	
Alternate loci	261
Alternate loci aligned to primary assembly	261
FIX patches	113
FIX patches aligned to primary assembly	112
NOVEL patches	72
NOVEL patches aligned to primary assembly	72

https://www.ncbi.nlm.nih.gov/grc/human

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.28_GRCh38.p13/README.txt

Index of /genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38

Name	Last modified	Size
Parent Directory		-
GCA_000001405.15_GRCh38_assembly_structure/	2016-09-20 09:37	-
GO_TO_CURRENT_VERSION/	2020-04-15 12:35	-
seqs_for_alignment_pipelines.ucsc_ids/	2021-03-02 22:14	-
GCA_000001405.15_GRCh38_assembly_regions.txt	2016-10-13 08:52	25K
GCA_000001405.15_GRCh38_assembly_report+ucsc_names.txt	2014-11-21 17:18	49K
GCA_000001405.15_GRCh38_assembly_report.txt	2016-10-13 08:52	53K
GCA_000001405.15_GRCh38_assembly_stats.txt	2016-10-13 08:52	64K
GCA_000001405.15_GRCh38_genomic.fna.gz	2015-02-10 13:55	902M
GCA_000001405.15_GRCh38_genomic.gaps.gz	2014-04-10 21:33	11K
GCA_000001405.15_GRCh38_genomic.gbff.gz	2015-02-10 13:55	120K
GCA_000001405.15_GRCh38_genomic.gff.gz	2015-02-10 13:55	18K
GCA_000001405.15_GRCh38_protein.faa.gz	2014-08-04 06:40	2.5K
GCA_000001405.15_GRCh38_protein.gpff.gz	2014-08-04 06:40	6.4K
GCA_000001405.15_GRCh38_rm.out.gz	2015-02-10 13:55	170M
GCA_000001405.15_GRCh38_rm.run	2014-08-04 04:18	873
README.txt	2020-09-02 16:26	43K
assembly_status.txt	2021-05-24 08:42	16
md5checksums.txt	2019-03-12 16:48	124K

Human Genome Resources at NCBI

https://www.ncbi.nlm.nih.gov/genome/guide/human/

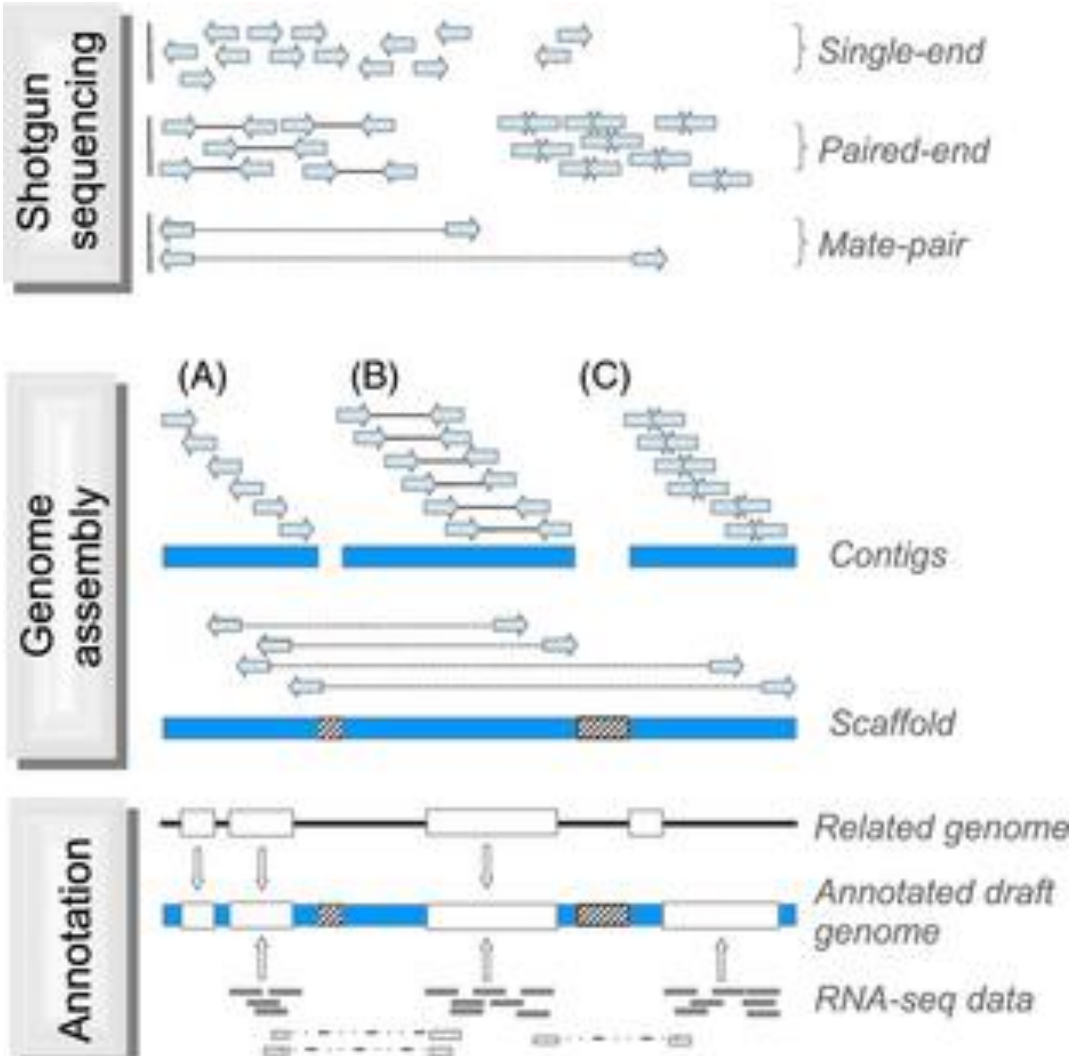
Download	GRCh38	GRCh37
Reference Genome Sequence	Fasta	Fasta
RefSeq Reference Genome Annotation	gff3	gff3
RefSeq Transcripts	Fasta	Fasta
RefSeq Proteins	Fasta	Fasta
ClinVar	vcf	vcf
dbSNP	vcf	vcf
dbVar	vcf	vcf

The human reference genome continues to change

- Ongoing efforts to fill "gaps" and properly/thoroughly represent complex structures and loci in the genome (e.g., Major Histocompatibility Complex)
- Each improvement leads to a new genome "build". Currently on build 38.
- Experimental and computational methods provide new genome annotations
  - New gene models, transcription factor binding sites, and loci where human individuals differ (i.e., polymorphisms)
- Therefore, the human reference genome is by no means "complete"!
- How does the same genome yield such phenotypic diversity across tissue types?

https://github.com/quinlan-lab/applied-computational-genomics

Anotación



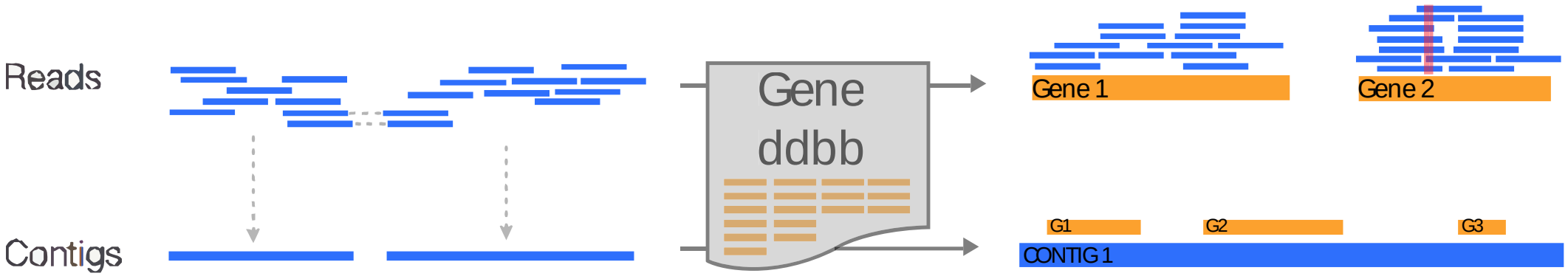
A field guide to whole-genome sequencing, assembly and annotation. Ekblom R, Wolf J

Anotación

- Identificación y/o localización de regiones codificantes en un genoma, determinando la función de cada uno.
  - Identificar elementos genómicos codificantes
  - Asignar función biológica a esos elementos
- Anotación estructural
  - ORFs y su localización
  - Regiones codificantes (cds)
  - Promotores y elementos reguladores
- Anotación funcional
  - Asignar función biológica a esos elementos

Anotación funcional

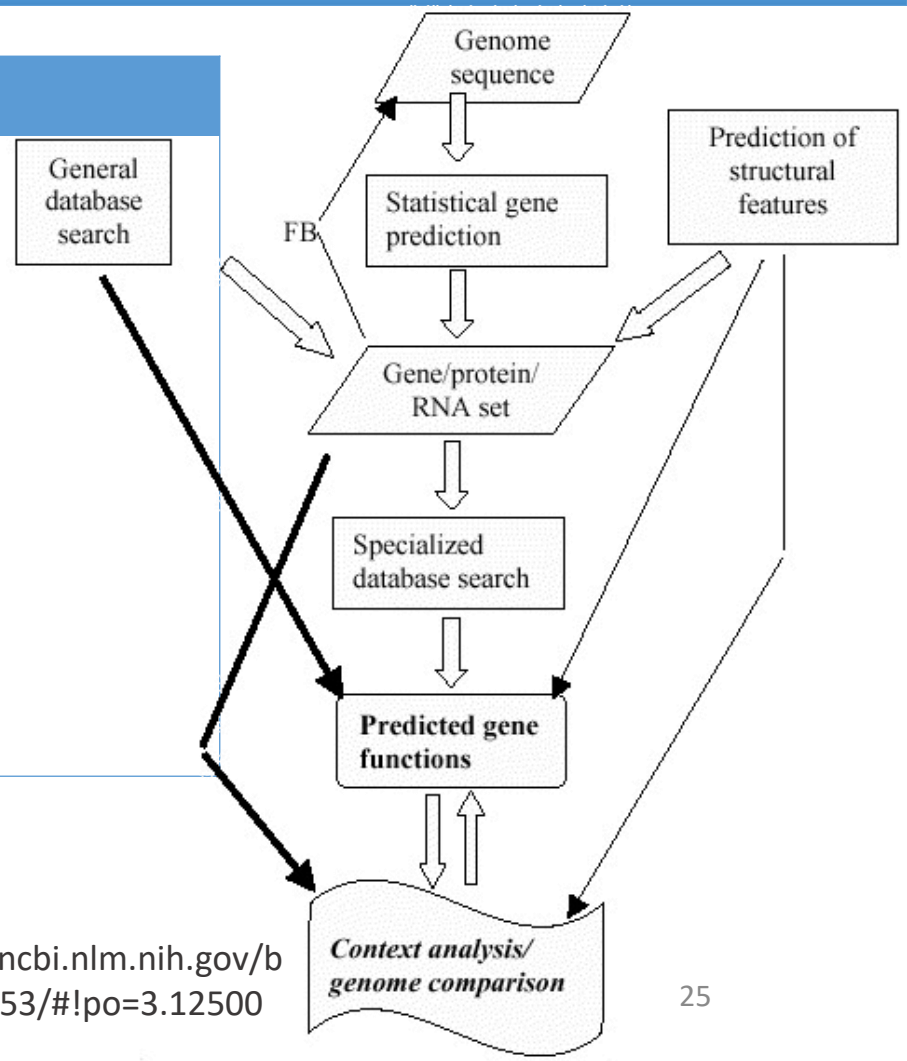
- Requiere una base de datos con la que comparer
  - Encyclopedia of DNA elements (ENCODE)
  - Entrez Gene
  - Ensembl
  - GENCODE
  - Gene Ontology Consortium
  - GeneRIF
  - RefSeq
  - Uniprot
  - Vertebrate and Genome Annotation Project (Vega)
  - Pfam
- Mapado (srst2) o Alineamiento Local –BLAST- (Prokka)



Anotación: Prokka

Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen et al. , 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen et al. , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA
BLAST+ (Camacho *et al.* , 2009)	Specific function or name
	Personal database

- Anotación automática - Anotación manual (Curado)



https://www.ncbi.nlm.nih.gov/books/NBK20253/#!po=3.12500

Formato ficher fna, faa

- fna:

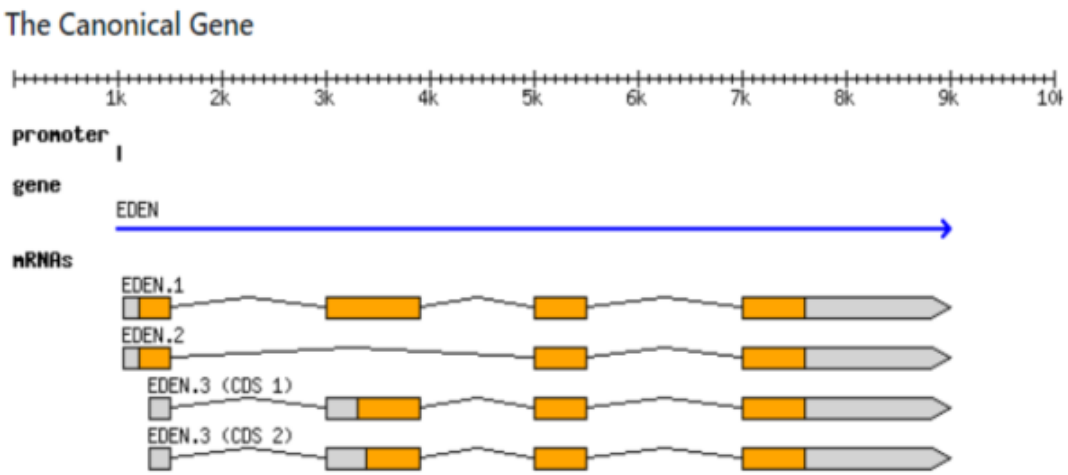
```
>seqid | atributes
ATCGATCGATCG
```

- faa:

```
>seqid | atributes
MAGCTYWHDEGGGML
```

Formato del fichero GFF (General feature format)

- Formato standard para describir genes o transcritos…. 9 col, tab-delimited, plain text files



seqid	source	type	start	end	score	strand	phase	atributes
0	##gff-version 3.1.26							
1	##sequence-region ctg123 1 1497228							
2	ctg123	gene	1000	9000	.	+	.	ID=gene00001;Name=EDEN
3	ctg123	TF_binding_site	1000	1012	.	+	.	ID=tfbs00001;Parent=gene00001
4	ctg123	mRNA	1050	9000	.	+	.	ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5	ctg123	mRNA	1050	9000	.	+	.	ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6	ctg123	mRNA	1300	9000	.	+	.	ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7	ctg123	exon	1300	1500	.	+	.	ID=exon00001;Parent=mRNA00003
8	ctg123	exon	1050	1500	.	+	.	ID=exon00002;Parent=mRNA00001,mRNA00002
9	ctg123	exon	3000	3902	.	+	.	ID=exon00003;Parent=mRNA00001,mRNA00003
10	ctg123	exon	5000	5500	.	+	.	ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11	ctg123	exon	7000	9000	.	+	.	ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12	ctg123	CDS	1201	1500	.	+	0	ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13	ctg123	CDS	3000	3902	.	+	0	ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14	ctg123	CDS	5000	5500	.	+	0	ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15	ctg123	CDS	7000	7600	.	+	0	ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16	ctg123	CDS	1201	1500	.	+	0	ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17	ctg123	CDS	5000	5500	.	+	0	ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18	ctg123	CDS	7000	7600	.	+	0	ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19	ctg123	CDS	3301	3902	.	+	0	ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20	ctg123	CDS	5000	5500	.	+	1	ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21	ctg123	CDS	7000	7600	.	+	1	ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22	ctg123	CDS	3391	3902	.	+	0	ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23	ctg123	CDS	5000	5500	.	+	1	ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24	ctg123	CDS	7000	7600	.	+	1	ID=cds00004;Parent=mRNA00003;Name=edenprotein.4

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

Formato fichero GTF (General Transfer Format)

- Es exacto al fichero GFF versión 2

https://www.ensembl.org/info/website/upload/gff.html

Contenido de ficheros fna, gff, gtf

- *_genomic.fna.gz file

FASTA format of the genomic sequence(s) in the assembly. Repetitive sequences in eukaryotes are masked to lower-case (see below). The FASTA title is formatted as sequence accession.version plus description. The genomic.fna.gz file includes all top-level sequences in the assembly (chromosomes, plasmids, organelles, unlocalized scaffolds, unplaced scaffolds, and any alternate loci or patch scaffolds). Scaffolds that are part of the chromosomes are not included because they are redundant with the chromosome sequences; sequences for these placed scaffolds are provided under the assembly_structure directory.

- *_genomic.gbff.gz file

GenBank flat file format of the genomic sequence(s) in the assembly. This file includes both the genomic sequence and the CONTIG description (for CON records), hence, it replaces both the .gbk & .gbs format files that were provided in the old genomes FTP directories.

- *_genomic.gff.gz file

Annotation of the genomic sequence(s) in Generic Feature Format Version 3 (GFF3). Sequence identifiers are provided as accession.version. Additional information about NCBI's GFF files is available at ftp://ftp.ncbi.nlm.nih.gov/genomes/README_GFF3.txt.

- *_genomic.gtf.gz file

Annotation of the genomic sequence(s) in Gene Transfer Format Version 2.2 (GTF2.2). Sequence identifiers are provided as accession.version.

- *_genomic_gaps.txt.gz

Tab-delimited text file reporting the coordinates of all gaps in the top-level genomic sequences. The gaps reported include gaps specified in the AGP files, gaps annotated on the component sequences, and any other run of 10 or more Ns in the sequences. See the "Description of files" section below for details of the file format.

Contenido de ficheros fna, faa, gff, gtf

- *_protein.faa.gz file

FASTA format sequences of the accessioned protein products annotated on the genome assembly. The FASTA title is formatted as sequence accession.version plus description.

- *_protein.gpff.gz file

GenPept format of the accessioned protein products annotated on the genome assembly

- *_rm.out.gz file

RepeatMasker output; Provided for Eukaryotes

- *_rm.run file

Documentation of the RepeatMasker version, parameters, and library; Provided for Eukaryotes

- *_rna.fna.gz file

FASTA format of accessioned RNA products annotated on the genome assembly; Provided for RefSeq assemblies as relevant (Note, RNA and mRNA products are not instantiated as a separate accessioned record in GenBank but are provided for some RefSeq genomes, most notably the eukaryotes.) The FASTA title is provided as sequence accession.version plus description.

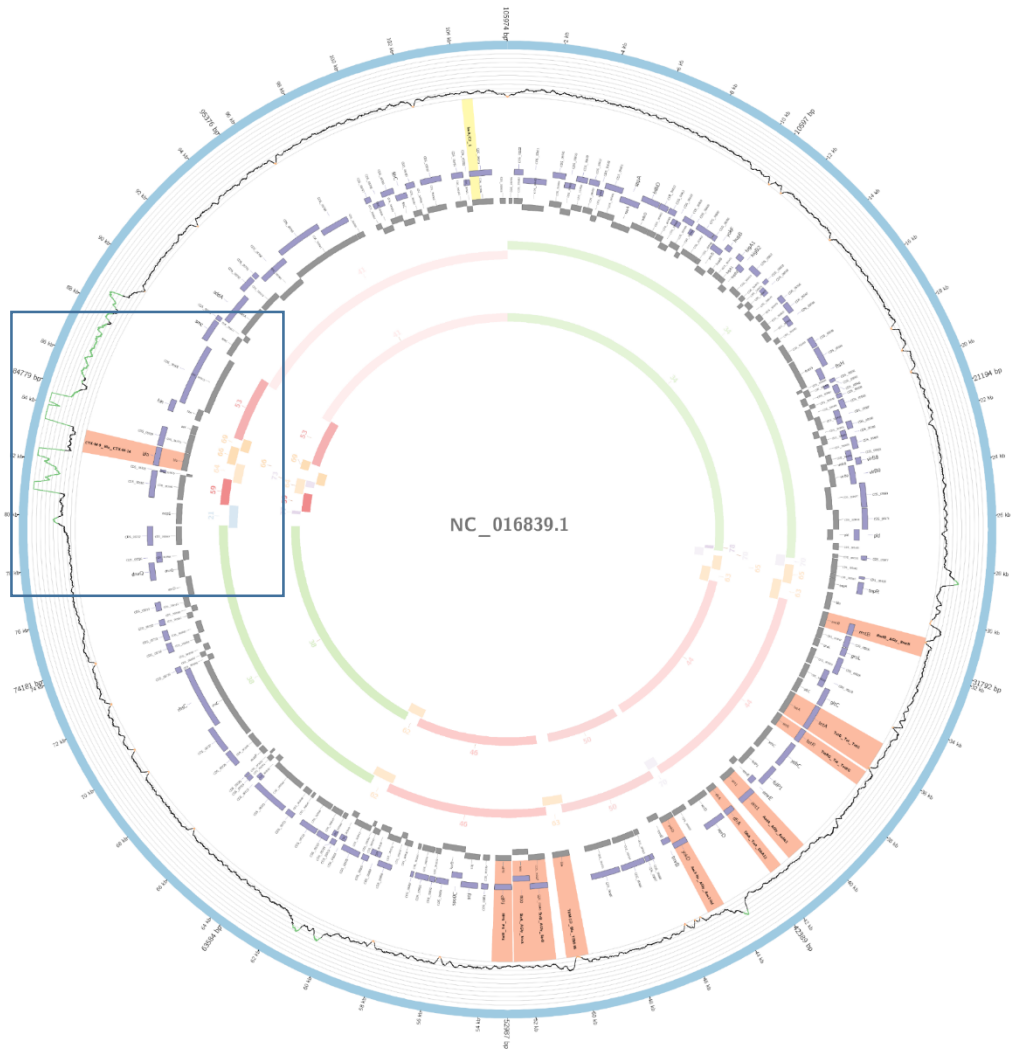
- *_rna.gbff.gz file

GenBank flat file format of RNA products annotated on the genome assembly; Provided for RefSeq assemblies as relevant

- *_rna_from_genomic.fna.g

FASTA format of the nucleotide sequences corresponding to all RNA features annotated on the assembly, based on the genome sequence. See the "Description of files" section below for details of the file format.

PlasmidID


NC_016839.1


86 kb
84779 bp
84 kb
82 kb
80 kb
78 kb
CTX-M-9_Bla_CTX-M-14
bla
hin
dnaQ
mcpS
CDS_05929
CDS_05918
CDS_05737
CDS_05736
53
59
66
69
73