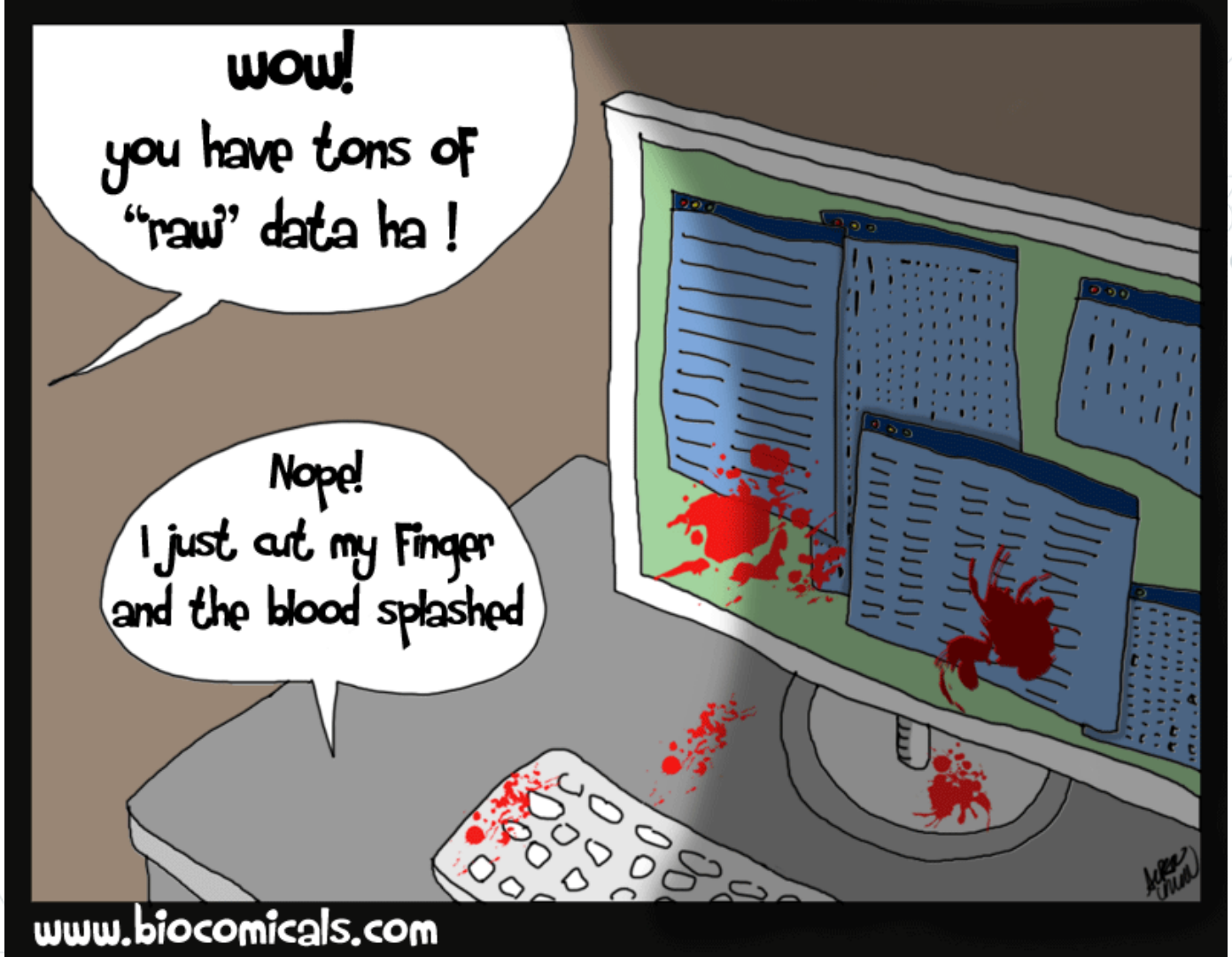
ENSAMBLAJE Y ANOTACION DE GENOMAS BACTERIANOS

Rafael Puche Q.
Sección de Bioinformática del Servicio de Secuenciación de ADN
Unidad de Estudios Genéticos y Forenses (UEGF)
Centro de Microbiologia y Biologia Celular
Instituto Venezolano de Investigaciones Científicas (IVIC)

UEGF Unidad de Estudios Genéticos y Forenses

IVIC

Mincyt Ministerio del Poder Popular para Ciencia y Tecnología


wow!
you have tons of "raw" data ha !
Nope!
I just cut my Finger and the blood splashed
www.biocomicals.com

Ensamblaje y Anotación de Genomas

Metodologías - Aplicaciones

Manos al Ensamblaje

Actividad lúdica de reconstrucción de un genoma

- Deben ensamblar 2 genomas usando solo tijeras
  - Un solo grupo hará un ensamblaje con referencia
  - El segundo grupo lo hará *de novo*

Seemann, (2010). *Genome Assembly Strategies.*

Flujo de Trabajo

Detras del boton "rojo" existe un flujo de trabajo

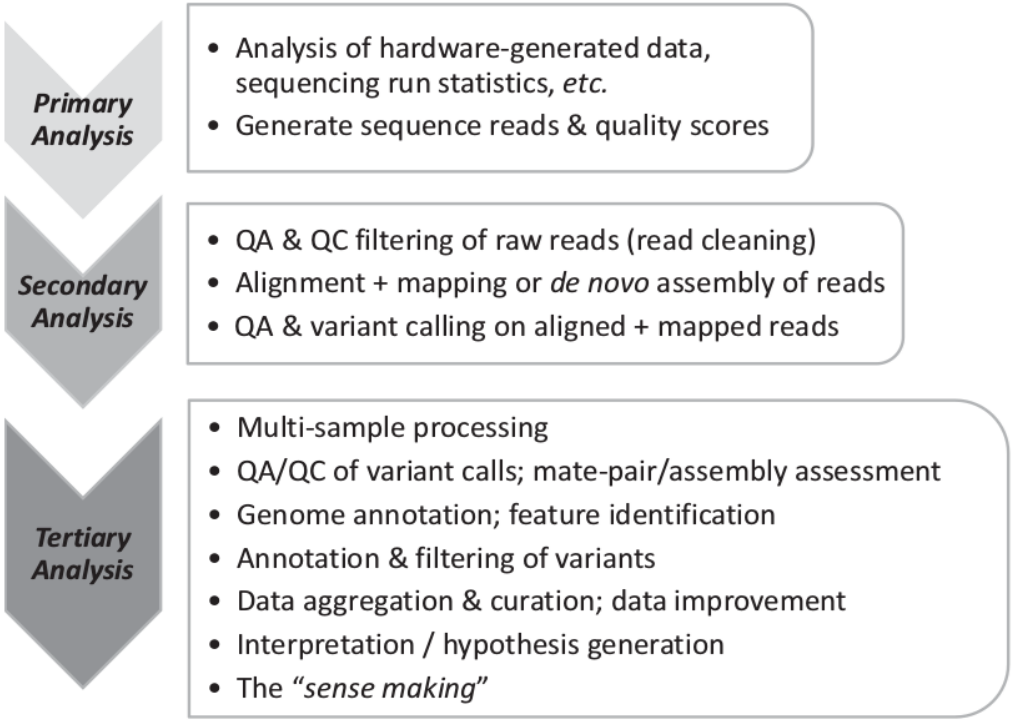


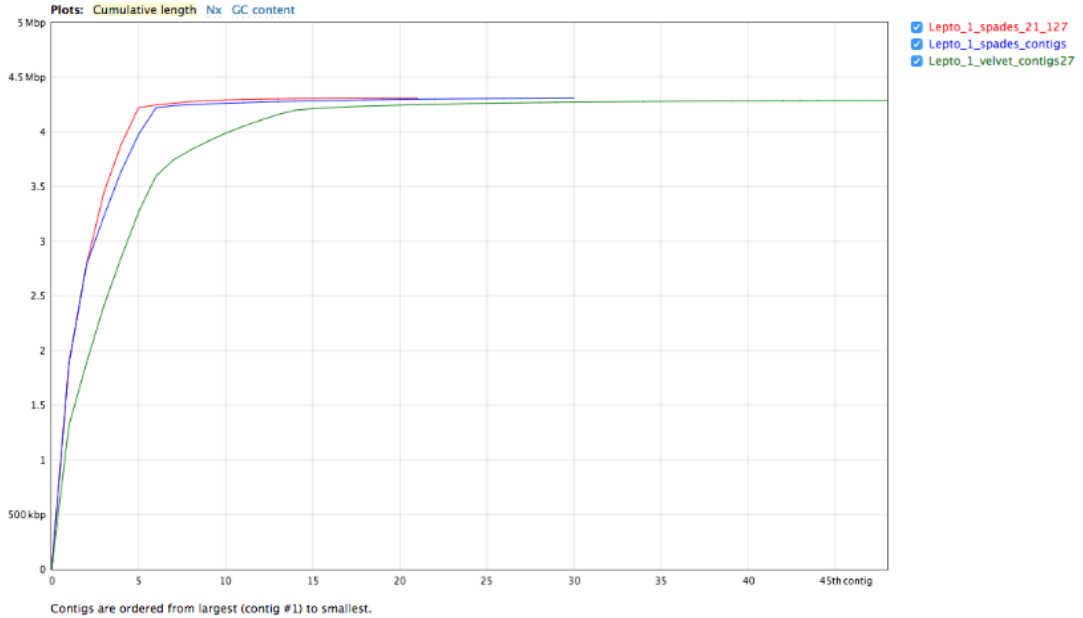
Figure 2.4 Data processing stages involved in microbial next-generation sequencing projects.

Comparativa entre Ensambladores

Ensamblaje de Leptospira - Spades *vs* Velvet

QUAST

Quality Assessment Tool for Genome Assemblies by CAB



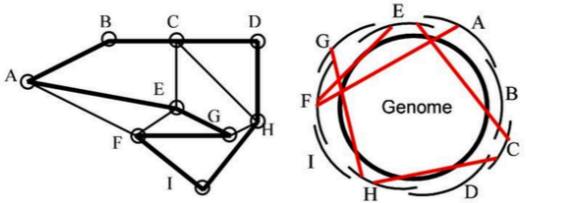
Principales Algoritmos de Ensamblaje

Proceso de **reconstrucción una secuencia** de ADN "original" de un organismo a partir de secuencias cortas *(short reads)*

Overlap Graph

Each read forms a *Node*

Edge exists between two nodes if the reads



Red/Thin: False Overlaps

Algorithm:

Step 1: Removing redundant edges, classify edges as required/optional

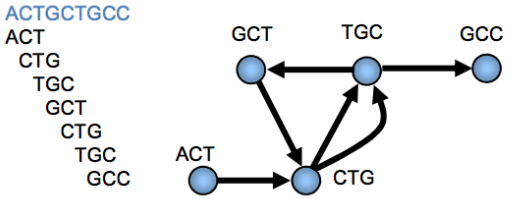
Step 2: Find the shortest walk which includes all required edges

Examples: Newbler, Celera, Arachne

De Bruijn Graph

Each read forms a *Node*

Edge exists between two nodes if the reads overlap



Algorithm:

Step 1: Construct kmer hash table

Step 2: Build de Bruijn graph

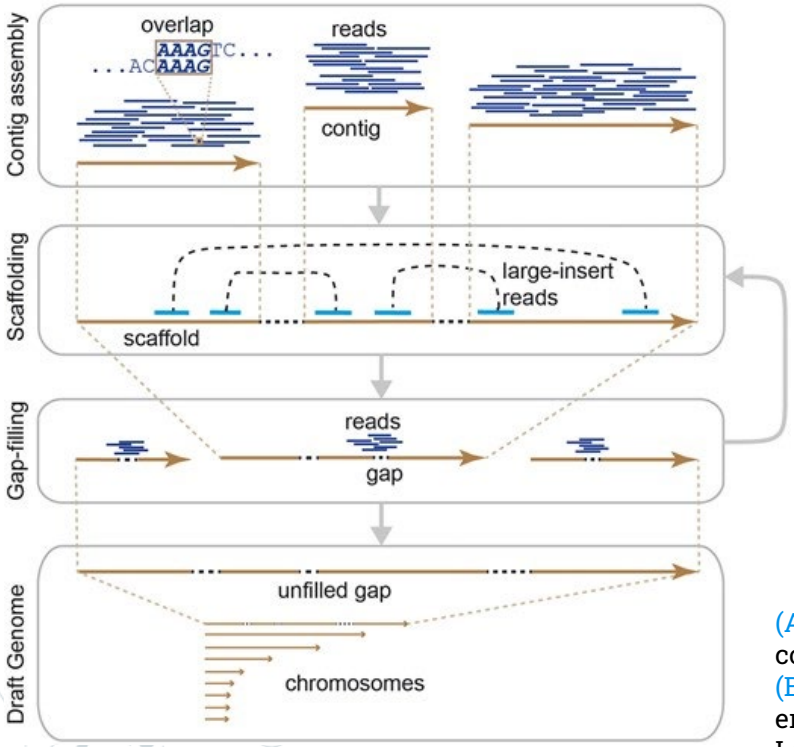
Step 3: Simplify the graph and search Eulerian path

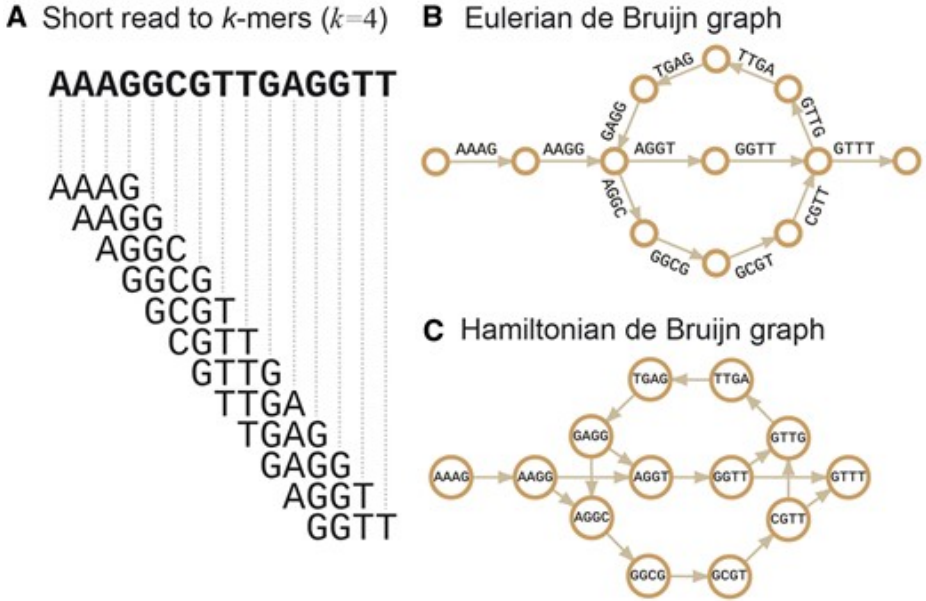
Examples: Euler, Velvet, Allpaths, Abyss, SOAPdenovo

Sohn & Nam, (2018). *The present and future of de novo whole-genome assembly.* Briefings in Bioinformatics

Ensamblaje

Proceso de **reconstrucción una secuencia** de ADN "original" de un organismo a partir de secuencias cortas *(short reads)*





(A) En el enfoque del gráfico de Bruijn, las lecturas cortas se dividen en k-mers cortos antes de construir los gráficos de Bruijn.
(B) En el enfoque hamiltoniano, los k-mers son los nodos, mientras que en el enfoque euleriano son las aristas.
Los k-mers están conectados a sus vecinos por medio de prefijos y sufijos (k-1) superpuestos.

Sohn & Nam, (2018). *The present and future of de novo whole-genome assembly.* Briefings in Bioinformatics

Ensamblador

Proceso de **reconstrucción una secuencia** de ADN "original" de un organismo a partir de secuencias cortas *(short reads)*

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 5, 2012
© Mary Ann Liebert, Inc.
Pp. 455–477
DOI: 10.1089/cmb.2012.0021

Original Articles

SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing

ANTON BANKEVICH,[1,2] SERGEY NURK,[1,2] DMITRY ANTIPOV,[1] ALEXEY A. GUREVICH,[1] MIKHAIL DVORKIN,[1] ALEXANDER S. KULIKOV,[1,3] VALERY M. LESIN,[1] SERGEY I. NIKOLENKO,[1,3] SON PHAM,[4] ANDREY D. PRJIBELSKI,[1] ALEXEY V. PYSHKIN,[1] ALEXANDER V. SIROTKIN,[1] NIKOLAY VYAHHI,[1] GLENN TESLER,[5] MAX A. ALEKSEYEV,[1,6] and PAVEL A. PEVZNER[1,4]

20.000 Citas

Sohn & Nam, (2018). *The present and future of de novo whole-genome assembly*. Briefings in Bioinformatics

Ensambladores

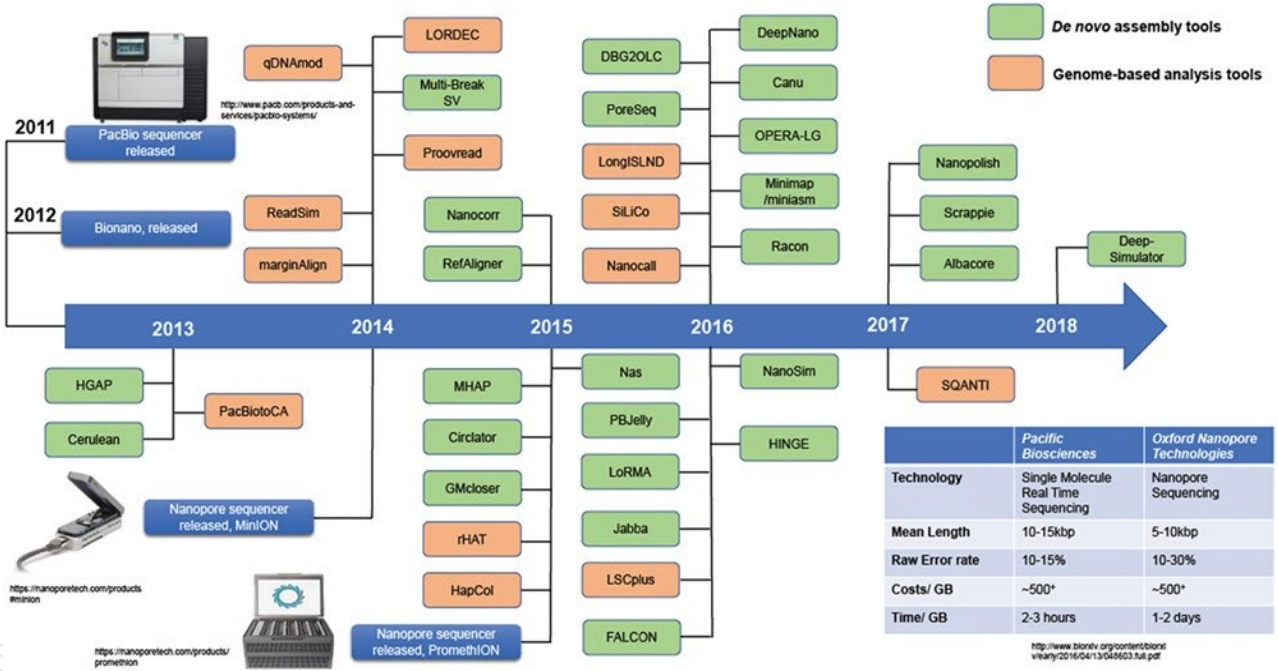
Más empleados en genómica bacteriana

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
SPADES	0.0	0.0	0.0	0.5	3.1	11.3	30.8	45.4	49.4	49.1	54.8
CLC	1.5	9.6	8.8	4.9	9.9	22.8	17.0	19.6	10.6	7.3	6.6
VELVET	34.6	9.5	21.1	20.5	7.4	11.3	13.0	6.0	6.1	6.3	3.0
ALLPATH	0.0	1.1	7.9	35.4	49.6	4.4	0.2	1.3	0.8	0.4	0.0
HGAP	0.0	0.0	0.0	0.7	2.2	5.7	7.7	7.8	7.7	7.8	5.7
NEWBLER	55.2	64.0	25.2	13.3	8.9	9.8	7.6	2.7	2.8	2.4	2.9
SOAP	0.3	3.0	7.4	6.2	3.1	5.7	4.8	2.8	8.6	7.2	8.1
A5	0.0	0.0	0.0	1.0	0.6	6.0	2.0	2.1	5.2	4.1	0.5
ABYSS	0.0	0.2	0.7	1.3	3.2	14.2	5.3	1.8	2.6	0.5	2.7
CELERA	4.8	12.7	27.8	12.6	1.3	2.7	1.2	0.9	0.9	0.3	0.3
PLATANUS	0.0	0.0	0.0	0.0	0.0	0.4	0.1	1.5	0.5	4.3	0.1
UNICYCLER	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.3	4.1	10.3
MIRA	1.0	0.7	1.5	1.7	1.5	1.6	2.6	1.4	0.3	0.2	0.1
MASURCA	0.0	0.0	0.0	0.4	6.5	0.2	1.1	0.2	0.2	0.4	0.2
IDBA	0.0	0.0	0.1	0.1	2.3	4.2	0.8	1.0	0.2	0.1	1.0
CANU	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.7	1.5	1.6	2.1
PHRED/PHRAP/CONSED	1.3	1.6	1.4	1.6	1.9	1.7	0.2	0.1	0.1	0.0	0.0
GENEIOUS	0.0	0.1	0.0	0.2	0.2	0.5	0.6	0.6	0.4	0.6	0.1
RAY	0.0	0.0	0.1	0.5	0.2	2.0	0.2	0.9	0.1	0.0	0.7
DNASTAR	0.0	0.0	0.7	0.3	0.2	0.9	0.4	0.6	0.2	0.0	0.5
FALCON	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.3
SKESA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0

Segerman, (2020). *The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments.* Frontiers

Ensambladores

Más empleados en Secuencias de Tercera Generación

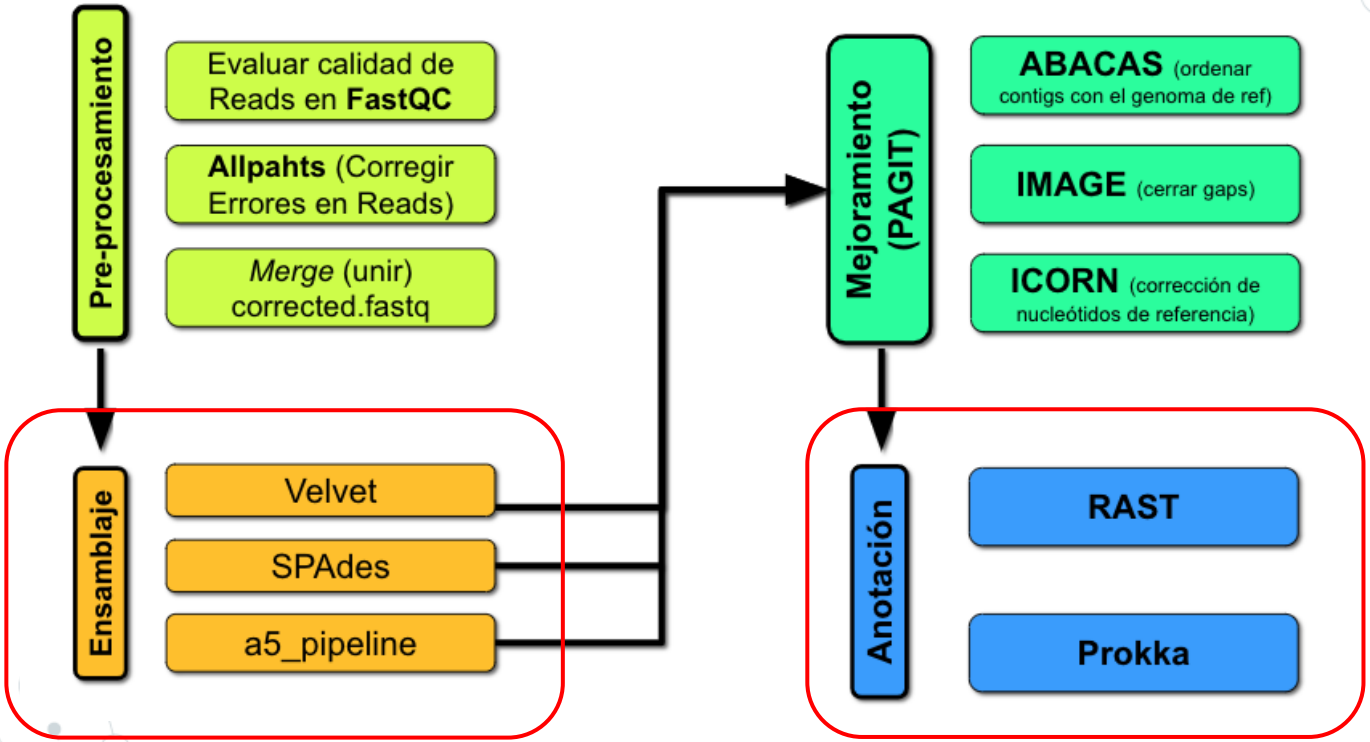


	Pacific Biosciences	Oxford Nanopore Technologies
Technology	Single Molecule Real Time Sequencing	Nanopore Sequencing
Mean Length	10-15kbp	5-10kbp
Raw Error rate	10-15%	10-30%
Costs/ GB	~500*	~500*
Time/ GB	2-3 hours	1-2 days

Wee *et al.* (2019). *The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing.* Briefings in Functional Genomics

Flujo de Trabajo

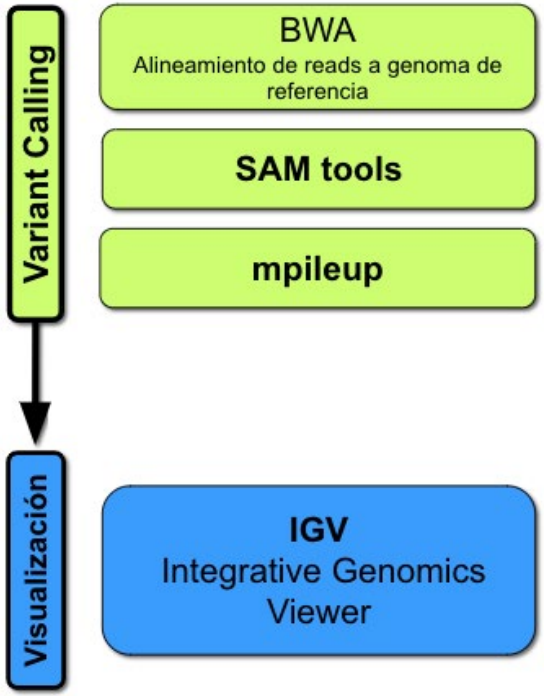
Ejemplos del botón "rojo"

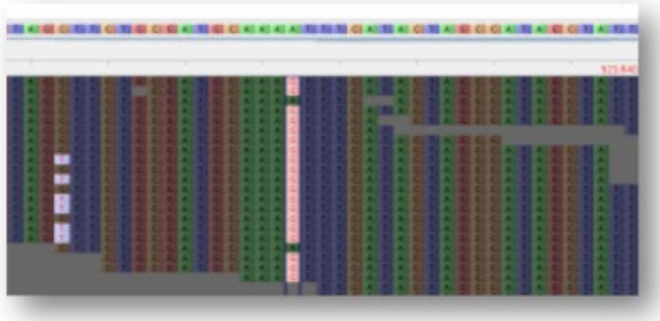


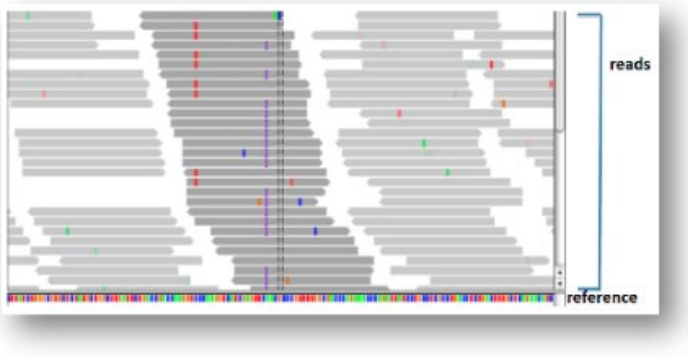
https://nanoporetech.com/nanopore-sequencing-data-analysis // https://www.ensembl.org/info/website/upload/gff.html

Flujo de Trabajo

Ejemplos del botón “rojo”



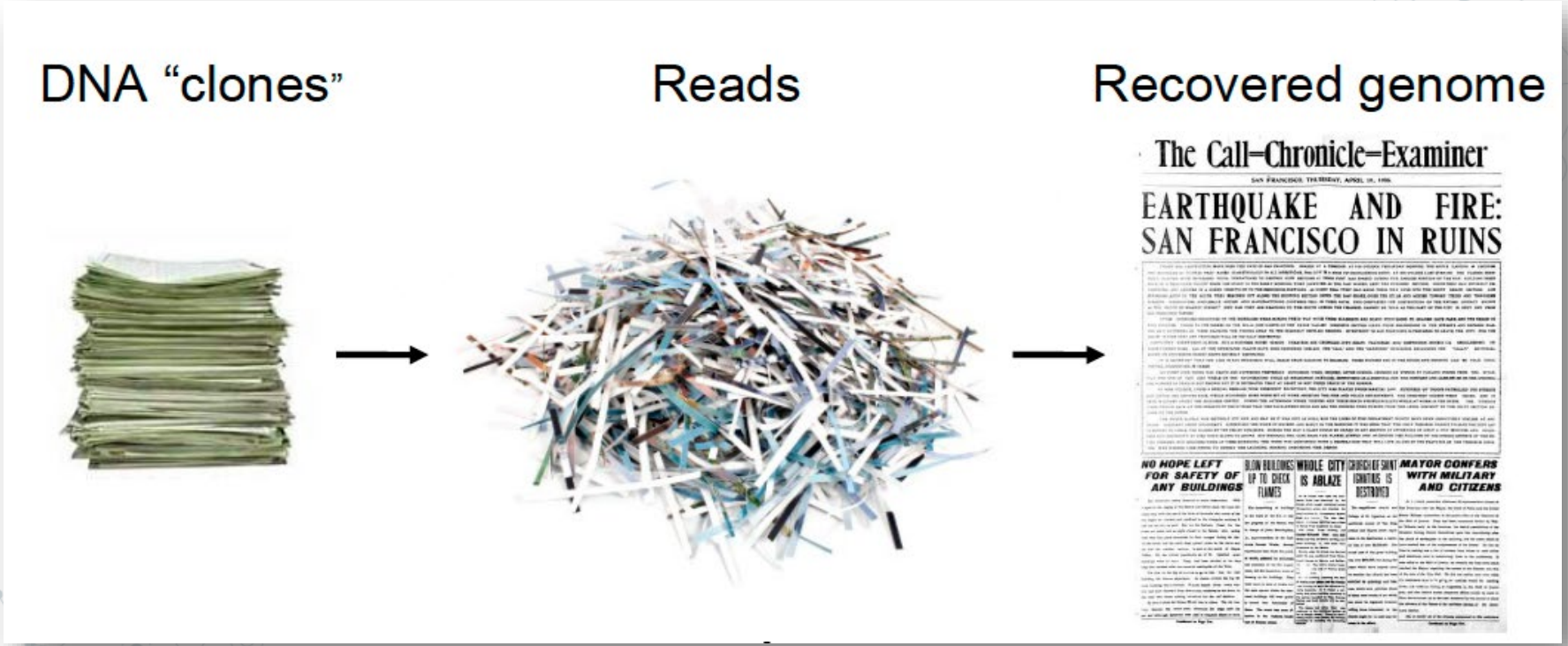




https://nanoporetech.com/nanopore-sequencing-data-analysis // https://www.ensembl.org/info/website/upload/gff.html

Ensamblaje

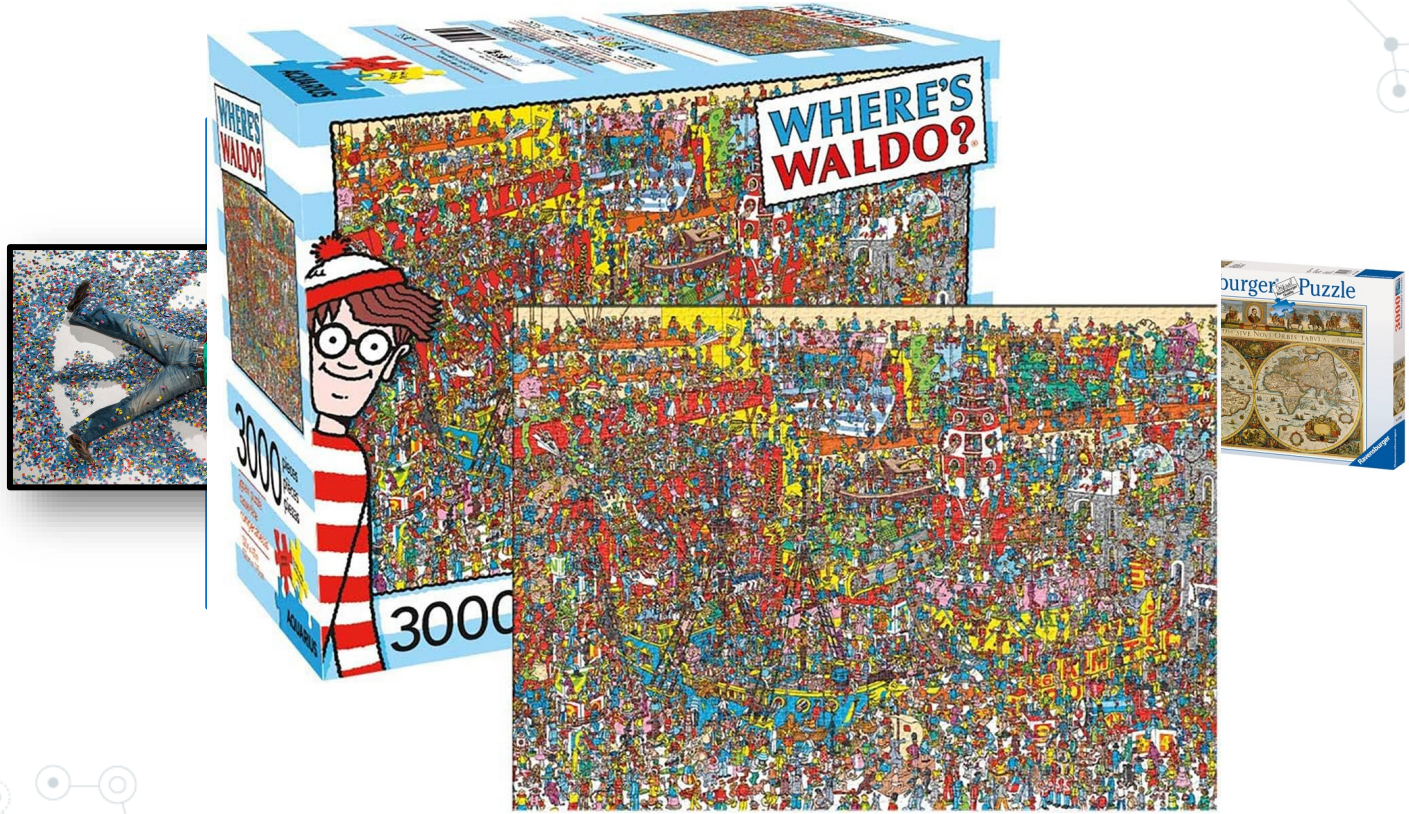
Proceso de **reconstrucción una secuencia** de ADN “original” de un organismo a partir de secuencias cortas *(reads)*



Seemann, (2010). *Genome Assembly Strategies*.

Estrategias de Ensamblaje

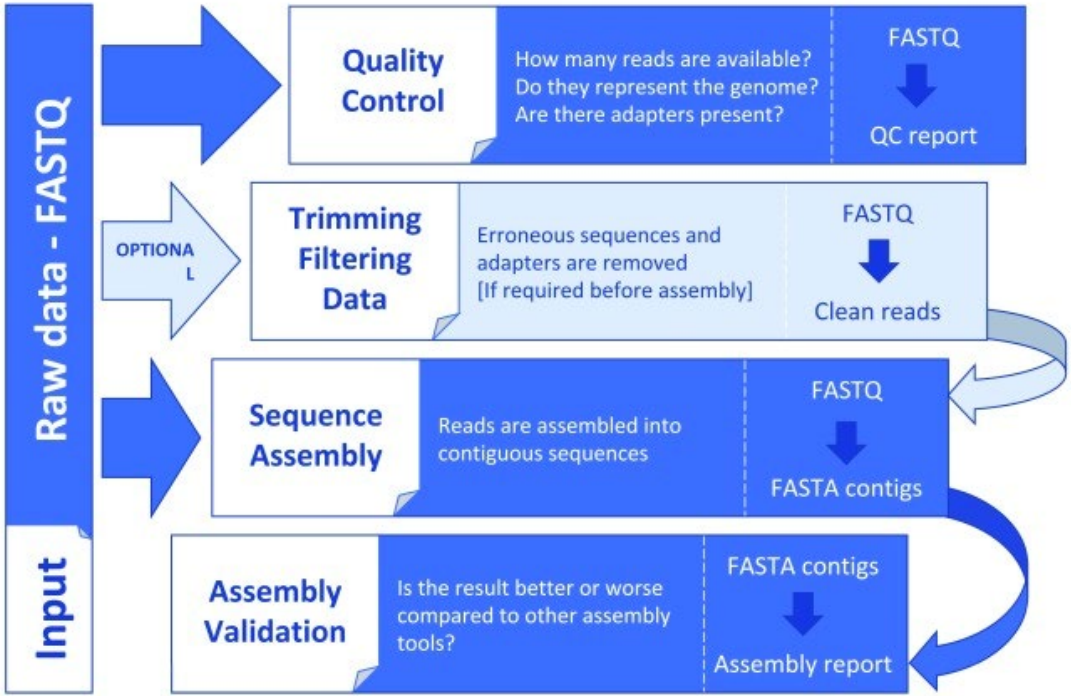
De novo y con referencia



Seemann, (2010). *Genome Assembly Strategies.*

Ensamblaje

Workflow para realizar un "buen" ensamblaje



https://f1000research.com/articles/7-148/v1

Mejorando el Ensamblaje

PROTOCOL

A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs

Martin T Swain[1,2], Isheng J Tsai[1], Samual A Assefa[1], Chris Newbold[1,3], Matthew Berriman[1] & Thomas D Otto[1]

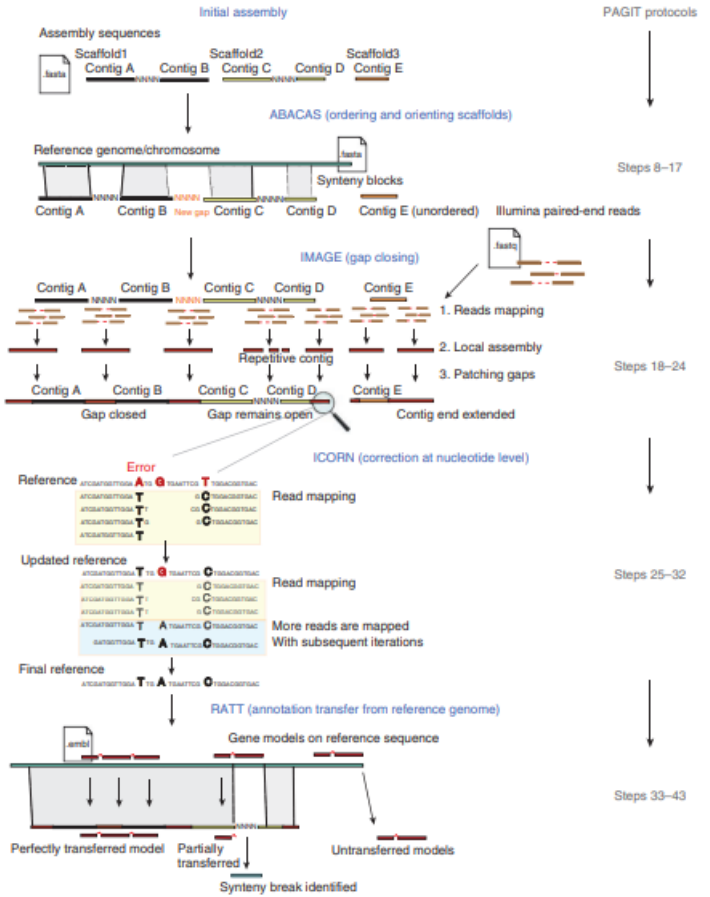
[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [2]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Penglais Campus, Aberystwyth, UK. [3]Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. Correspondence should be addressed to T.D.O. (tdo@sanger.ac.uk).

Published online 7 June 2012; doi:10.1038/nprot.2012.068

Tool | Analysis | Annotation | Sequence data processing | Archived

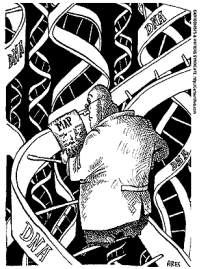
PAGIT

Tools to generate automatically high quality sequence by ordering contigs, closing gaps, correcting sequence errors and transferring annotation.



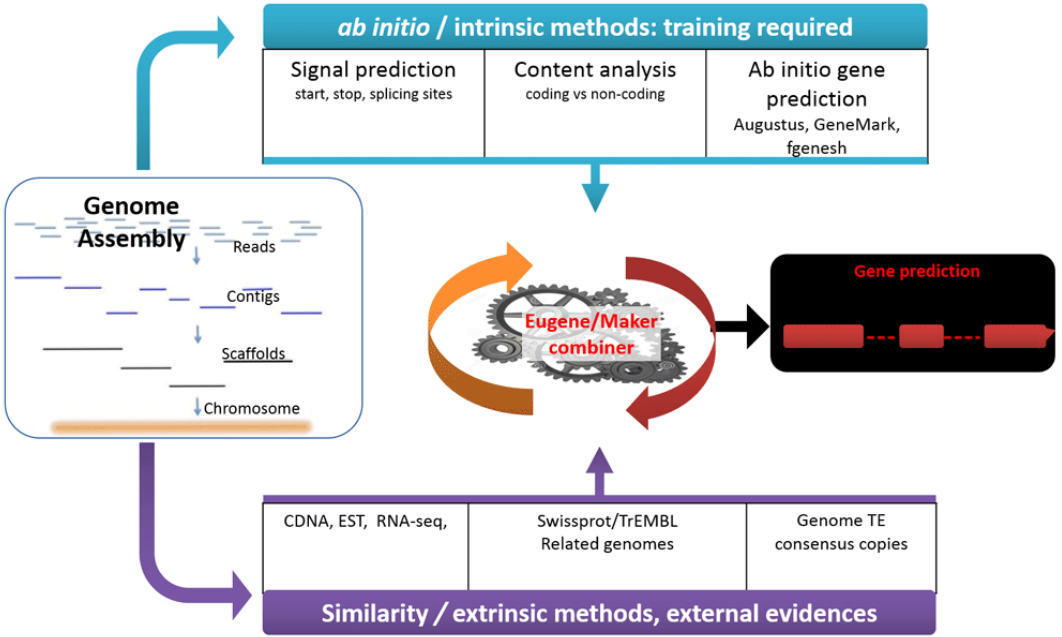
Anotacion

Proceso de identificación de características de interés en un genoma.



Incluye:

- ❑ Genes
- ❑ RNA no codificantes
- ❑ Operones
- ❑ Nuevos genes?

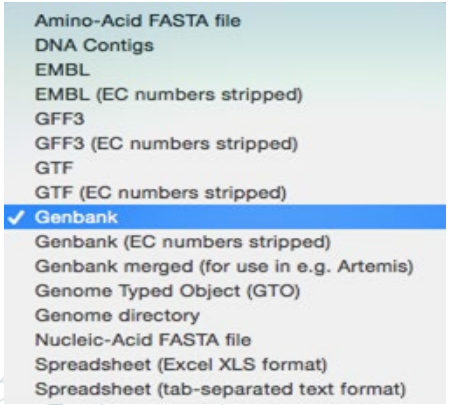


La precisión en la anotación genómica es importante y, en algunas veces crítica, en la interpretación biológica *downstream*

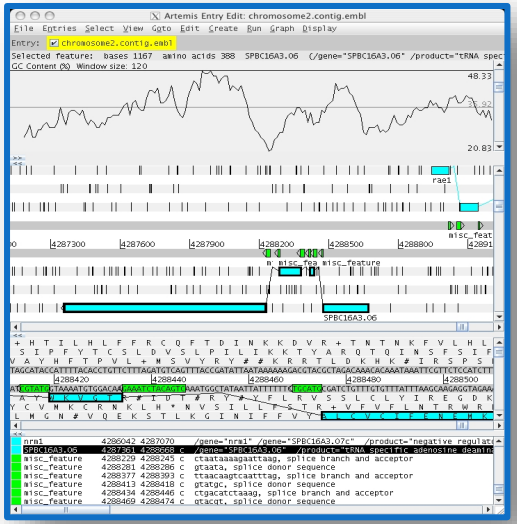
Anotacion

Proceso de identificación de características de interés en un genoma.

- **RAST: Rapid Annotation using Subsystem Technology,** es un servicio totalmente automatizado para anotar genomas bacterianos y de arqueas.



- Formato del output: *.gff
- Se visualiza en: Artemis (Genome Browser and Annotation Tool)



Anotacion

Proceso de identificación de características de interés en un genoma.

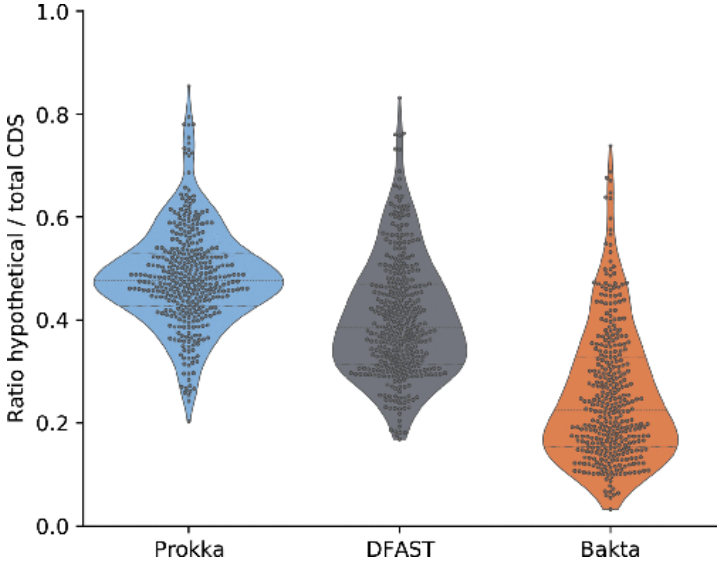
- **BAKTA** nueva herramienta de línea de comandos para la anotación automatizada y estandarizada de genomas bacterianos que tiene como objetivo un equilibrio entre el **rendimiento del tiempo de ejecución y las anotaciones completas.**
  - tRNA,
  - tmRNA,
  - rRNA,
  - ncRNA
  - genes,
  - CRISPR,
  - CDS.

MICROBIAL GENOMICS
Volume 7, Issue 11
Research Article | Open Access

Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification

Find out more about Bakta, the motivation, challenges and applications, here.

Oliver Schwengers[1], Lukas Jelonek[1], Marius Alfred Dieckmann[1], Sebastian Beyvers[1], Jochen Blom[1], Alexander Goesmann[1]

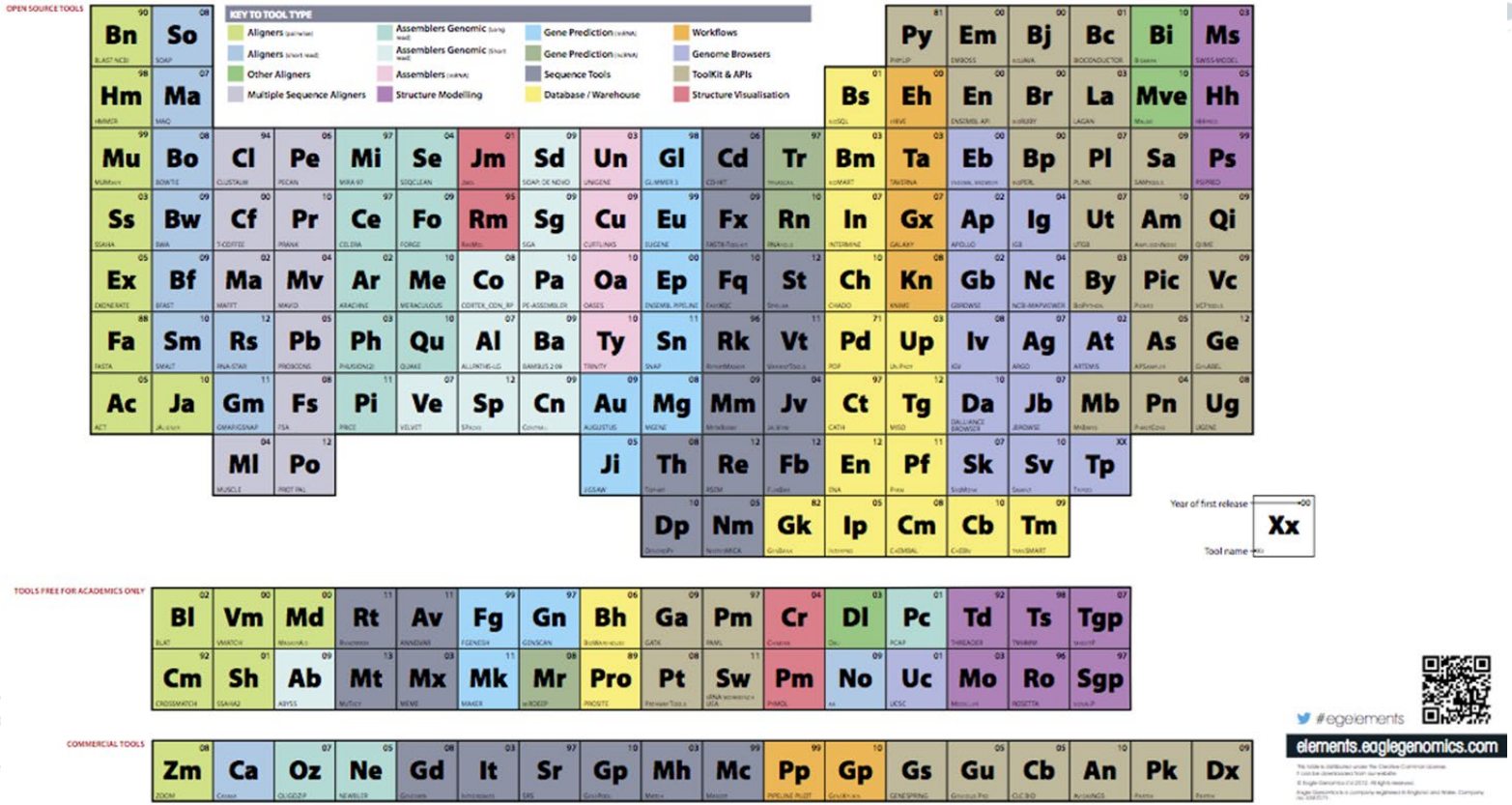


Proporción de secuencias de proteínas anotadas como proteínas hipotéticas. 362 genomas del GenBank que comprenden especies de géneros no definidos.

Schwengers O., *et al.* (2021). *Bakta.* Microbial Genomics

La tabla periódica....bioinformática

Un ecosistema de programas



Ensamblaje - Epílogo

Un par de preguntas importantes al ensamblar un genoma

1. ¿Cuál es el mejor ensamblador que se puede utilizar?

Depende.

Diferentes ensambladores tendrán un rendimiento diferente para diferentes genomas. Factores como:

- tamaño del genoma,
- Las secuencias repetidas,
- el contenido de GC y otros pueden influir en el rendimiento de los ensambladores.

2. ¿Cuándo termina mi ensamblaje?

Actualmente, la respuesta a esta pregunta es Nunca.

Como ejemplo: el Genoma Humano es el genoma mejor estudiado del mundo, con miles de individuos secuenciados y millones de dólares gastados. Aun así, hasta un 20% del Genoma Humano sigue sin ser ensamblado a pesar de los recientes avances.

Significa que su ensamblaje estará terminado cuando pueda responder a las preguntas que desea formular.

Ensamblaje - Epílogo

Un par de preguntas importantes al ensamblar un genoma

PLOS COMPUTATIONAL BIOLOGY

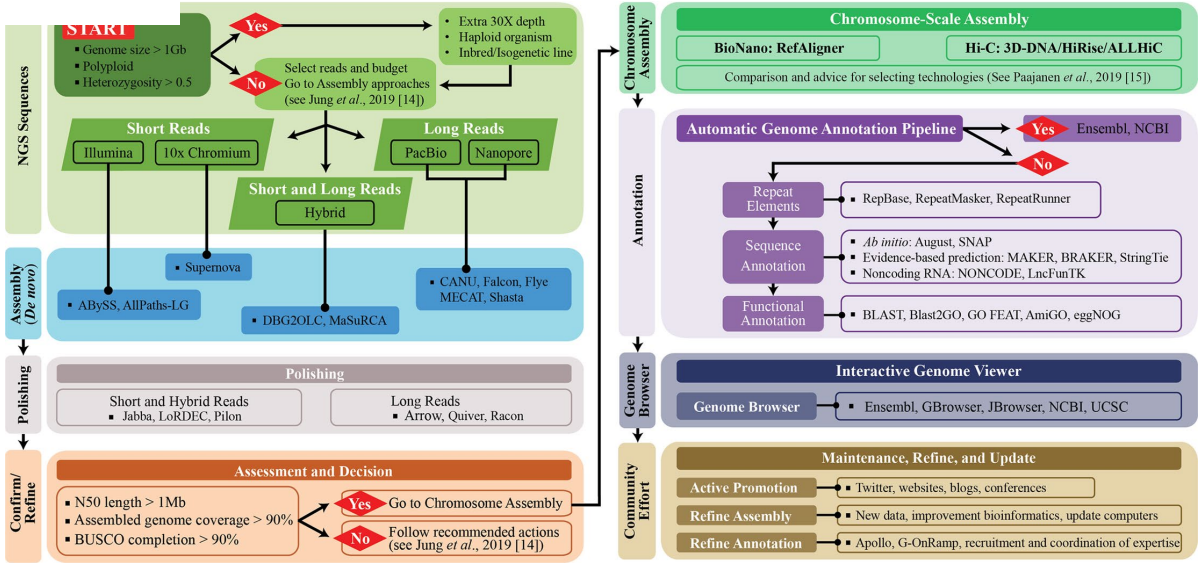
OPEN ACCESS

EDUCATION

Twelve quick steps for genome assembly and annotation in the classroom

Hyungtaek Jung, Tomer Ventura, J. Sook Chung, Woo-Jin Kim, Bo-Hye Nam, Hee Jeong Kong, Young-Ok Kim, Min-Seung Jeon, Seong-il Eyun

Published: November 12, 2020 • https://doi.org/10.1371/journal.pcbi.1008325



Pregunta de investigacion

Estudiando la leptospirosis

Los métodos tradicionales no logran identificar la especie de Leptospira aislada, por lo que se plantea la siguiente pregunta: **¿Empleando la secuenciación de genoma completo se podrá aumentar la resolución e identificar correctamente la especie de Leptospira?**

Antedecentes

Estudiando la leptospirosis

Open Access | Published: 24 April 2003

Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing

Shuang-Xi Ren, Gang Fu, Xiu-Gao Jiang, Rong Zeng, You-Gang Miao, Hai Xu, Yi-Xuan Zhang, Hui Xiong, Gang Lu, Ling-Feng Lu, Hong-Quan Jiang, Jia Jia, Yue-Feng Tu, Ju-Xing Jiang, Wen-Yi Gu, Yue-Qing Zhang, Zhen Cai, Hai-Hui Sheng, Hai-Feng Yin, Yi Zhang, Gen-Feng Zhu, Ma Wan, Hong-Lei Huang, Zhen Qian, Sheng-Yue Wang, Wei Ma, Zhi-Jian Yao, Yan Shen, Bo-Qin Qiang, Qi-Chang Xia, Xiao-Kui Guo, Antoine Danchin, Isabelle Saint Girons, Ronald L. Somerville, Yu-Mei Wen, Man-Hua Shi, Zhu Chen, Jian-Guo Xu & Guo-Ping Zhao ✉ - Show fewer authors

Nature **422**, 888–893(2003) | Cite this article

Antedecentes

Estudiando la leptospirosis

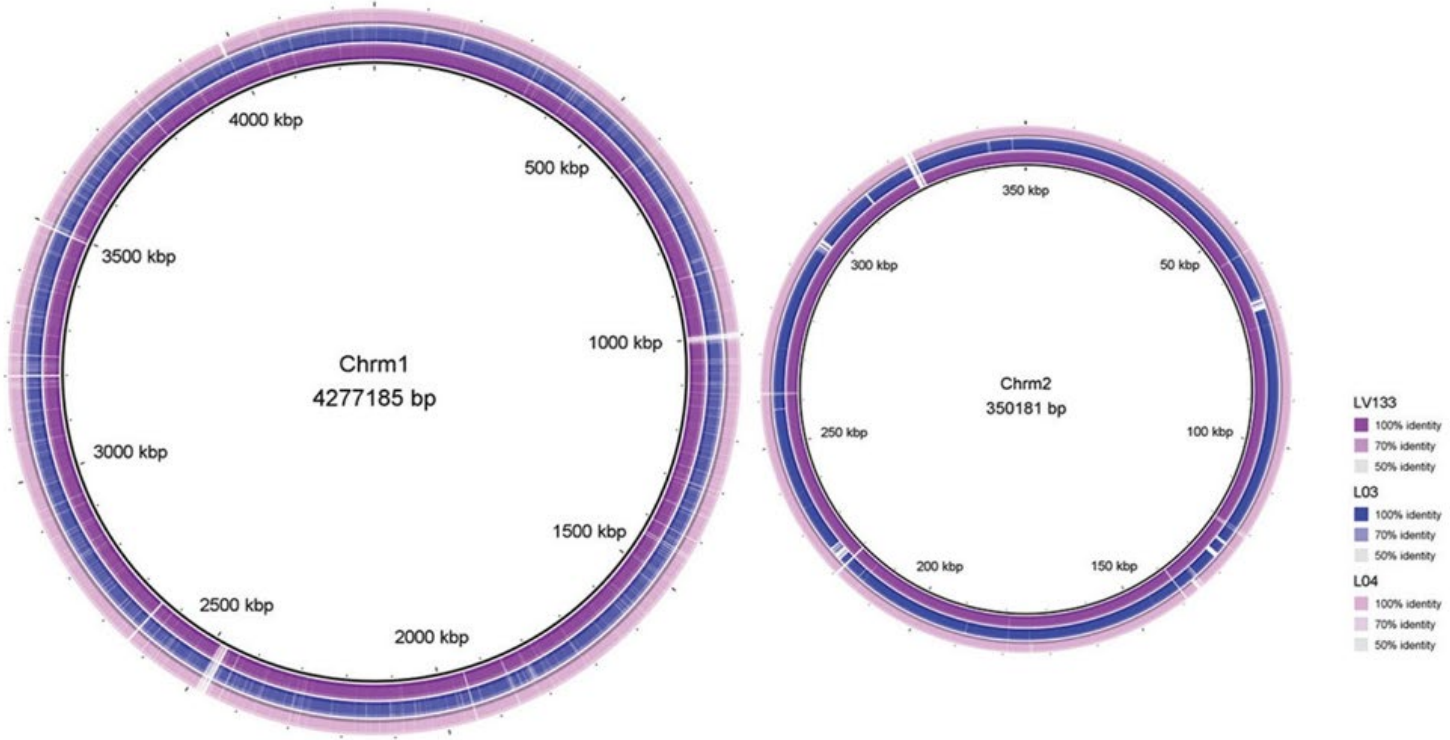


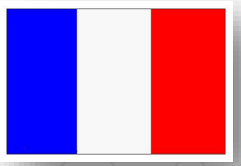
Fig. 1 BLAST ring image generator (BRIG) plot displaying whole genome comparison of *Leptospira interrogans* serovars Copenhageni (internal reference line) and Canicola. http://dx.doi.org/10.1590/0074-02760170119

Producción científica

Leptospira venezuelensis

IVIC

Pasteur Montevideo



Pasteur Francia

INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

TAXONOMIC DESCRIPTION

Puche *et al.*, *Int J Syst Evol Microbiol*

DOI 10.1099/ijsem.0.002528

MICROBIOLOGY SOCIETY

Leptospira venezuelensis sp. nov., a new member of the intermediate group isolated from rodents, cattle and humans

Rafael Puche,[1]† Ignacio Ferrés,[2]† Lizeth Caraballo,[3] Yaritza Rangel,[3] Mathieu Picardeau,[4] Howard Takiff[5,*] and Gregorio Iraola[2,*]

Abstract

Three strains, CLM-U50^T, CLM-R50 and IVIC-Bov1, belonging to the genus *Leptospira*, were isolated in Venezuela from a patient with leptospirosis, a domestic rat (*Rattus norvegicus*) and a cow (*Bos taurus*), respectively. The initial characterisation of these strains based on the *rrs* gene (16S rRNA) suggested their designation as a novel species within the 'intermediates' group of the genus *Leptospira*. Further phylogenomic characterisation based on single copy core genes was consistent with their separation into a novel species. The average nucleotide identity between these three strains was >99 %, but below 89 % with respect to any previously described leptospiral species, also supporting their designation as a novel species. Given this evidence, these three isolates were considered to represent a novel species, for which the name *Leptospira venezuelensis* sp. nov. is proposed, with CLM-U50^T (=CIP 111407^T=DSM 105752^T) as the type strain.

Producción científica

Leptospira venezuelensis

Se aislaron tres cepas:

- CLM-U50T aislada de un paciente con leptospirosis de moderada a severa, caracterizada por fiebre, ictericia y altos niveles de enzimas hepáticas, presentando además vómito, mialgia y artralgia, sin embargo, no se diagnosticó fallas renales ni respiratorias.
- CLM-R50 aislado del riñón de una rata (*Rattus norvegicus*) la cual fue capturada en la misma zona donde vivía el paciente con leptospirosis.
- IVIC-Bov1 se aisló de la orina de una vaca (*Bos taurus*) perteneciente a una granja situada a 40 km de la zona donde vivía el paciente.

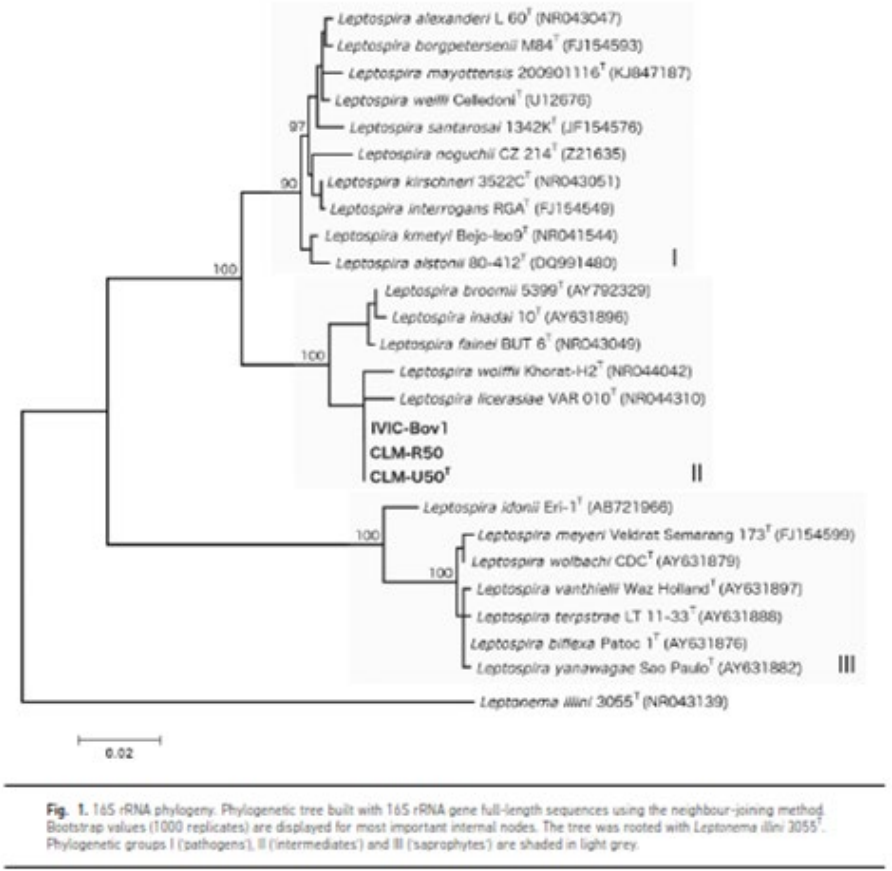
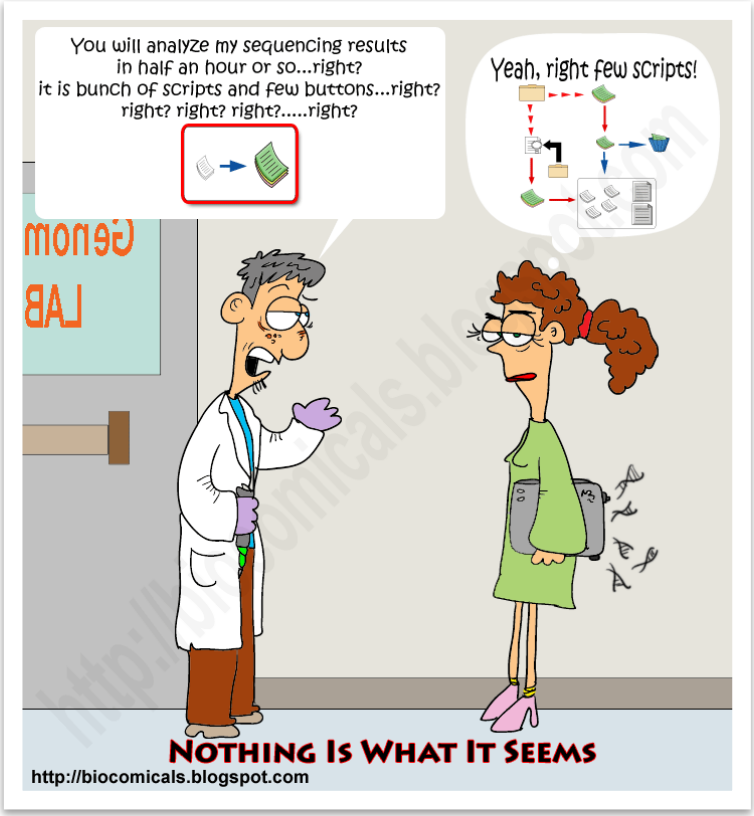


Fig. 1. 16S rRNA phylogeny. Phylogenetic tree built with 16S rRNA gene full-length sequences using the neighbour-joining method. Bootstrap values (1000 replicates) are displayed for most important internal nodes. The tree was rooted with *Leptonema illini* 3055T. Phylogenetic groups I ('pathogens'), II ('intermediates') and III ('saprophytes') are shaded in light grey.

No es tan simple como se aparenta...



...pero no imposible

Muchas preguntas por responder...

Muchas preguntas por hacer...

pregu

preguntas
preguntados
preguntas interesantes
preguntas para ask

"Presionar Enter para buscar"

Gracias!

¿Mas preguntas?

0212-5041622

rafael.puche@pedeciba.edu.uy

Twitter: @rpucheq

Telegram: t.me/BioinformaticaVZLA