Secuenciación del genoma de bacterias: ensamblado y anotación

Isabel Cuesta
Unidad de Bioinformática
17-21 Junio 2019, 7ª Edición
Programa Formación Continua, ISCIII

Ensamblado

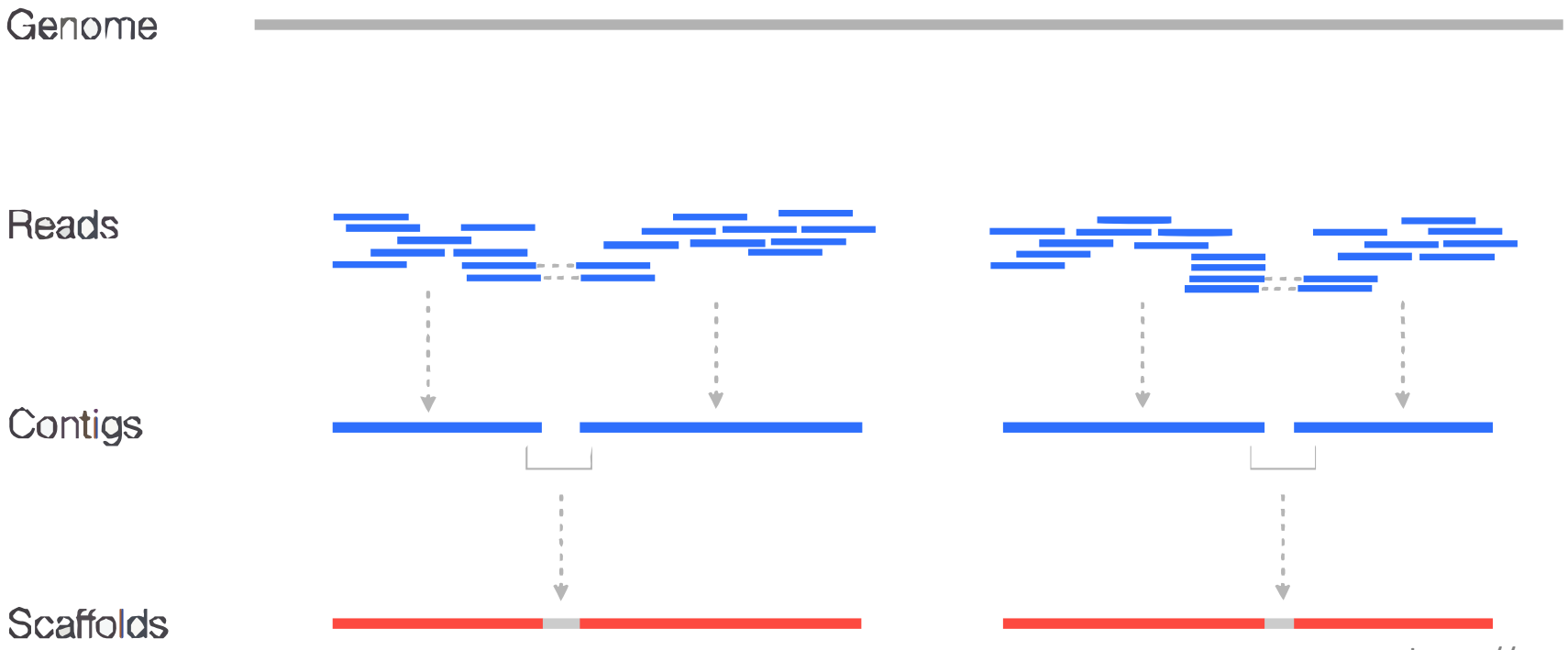
Reconstruir la **secuencia de DNA original** a partir de **lecturas** o secuencias de mucho menos tamaño.

- ***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 maypr tamaño posible.

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

Ensamblado: contig y scaffold

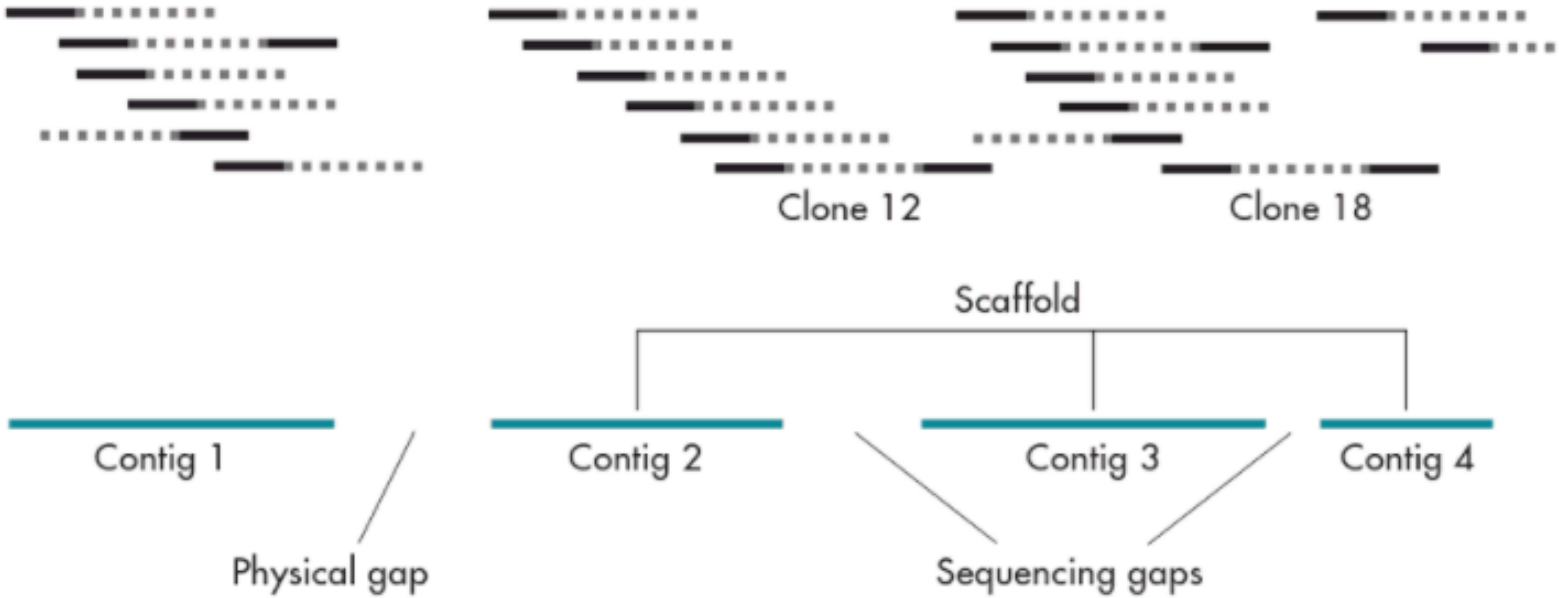
- **Contig**: secuencia continua del genoma formada por lecturas solapantes
- **Scaffold**: dos o más contigs unido por información de longitudes conocidas (pair-end, mate pair, referencia)



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

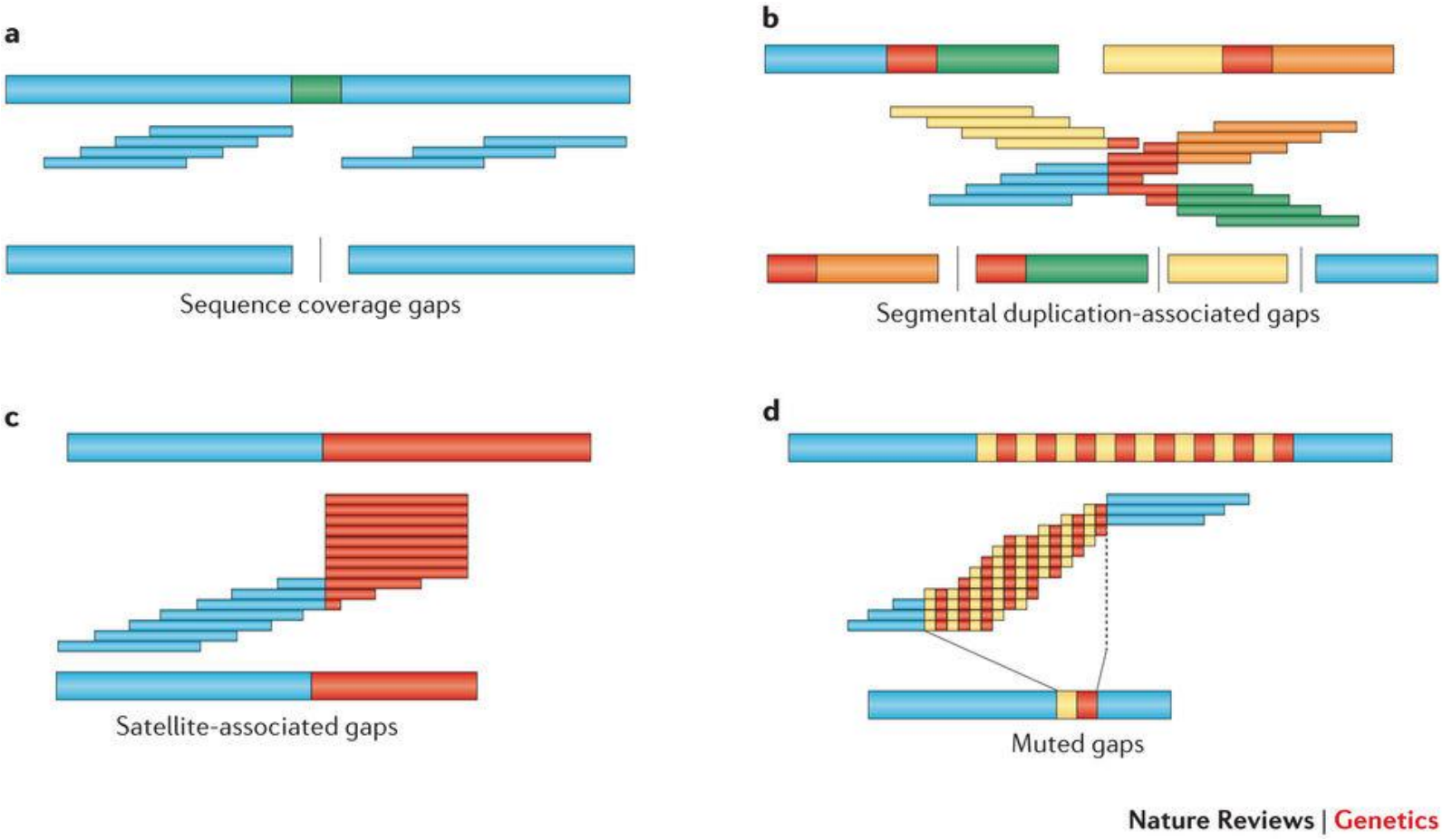
Ensamblado: gaps

- **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*.

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)**

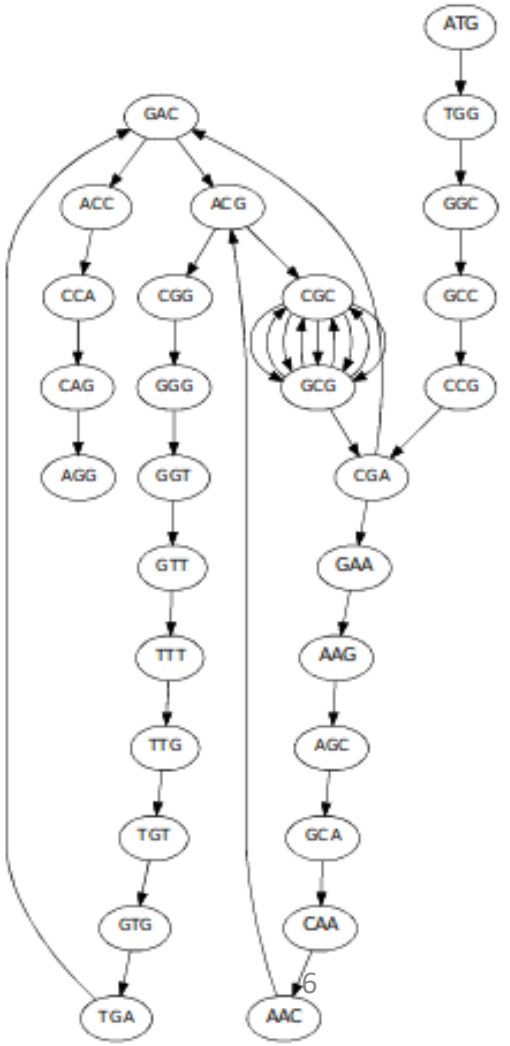
Grafos de Brujin: Elaboración de un grafo de k-mers (fragmentos de secuencia de longitud fija) donde se representan todos los solapamientos entre k-mers. Se unen nodos, burbujas y selección del mejor camino hasta un grafo irreducible del que se obtienen los contigs.

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).

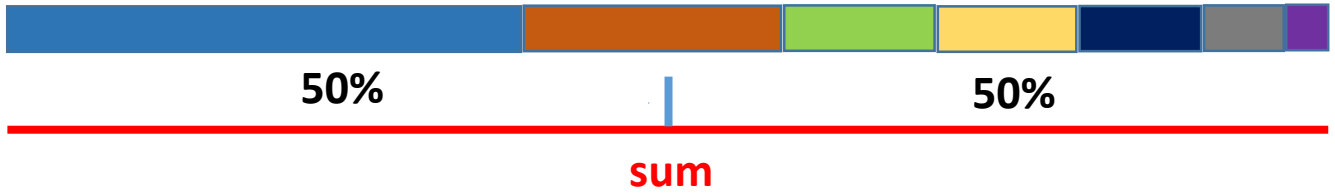
Ej. Assembler SGA, String Graph…

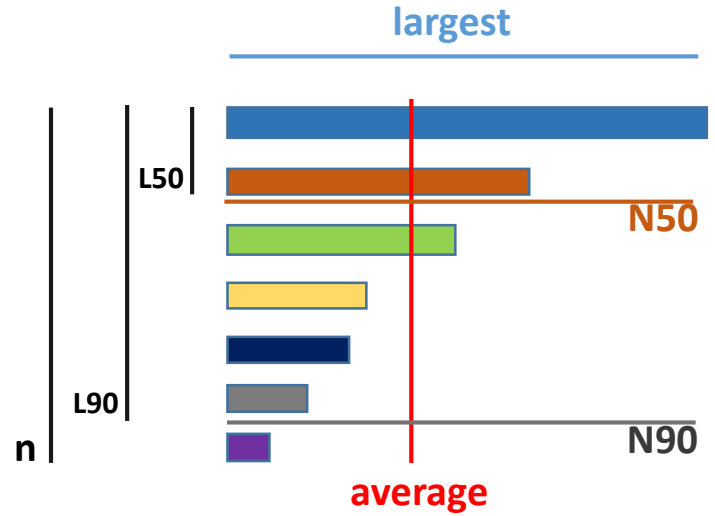


https://medium.com/@han_chen

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

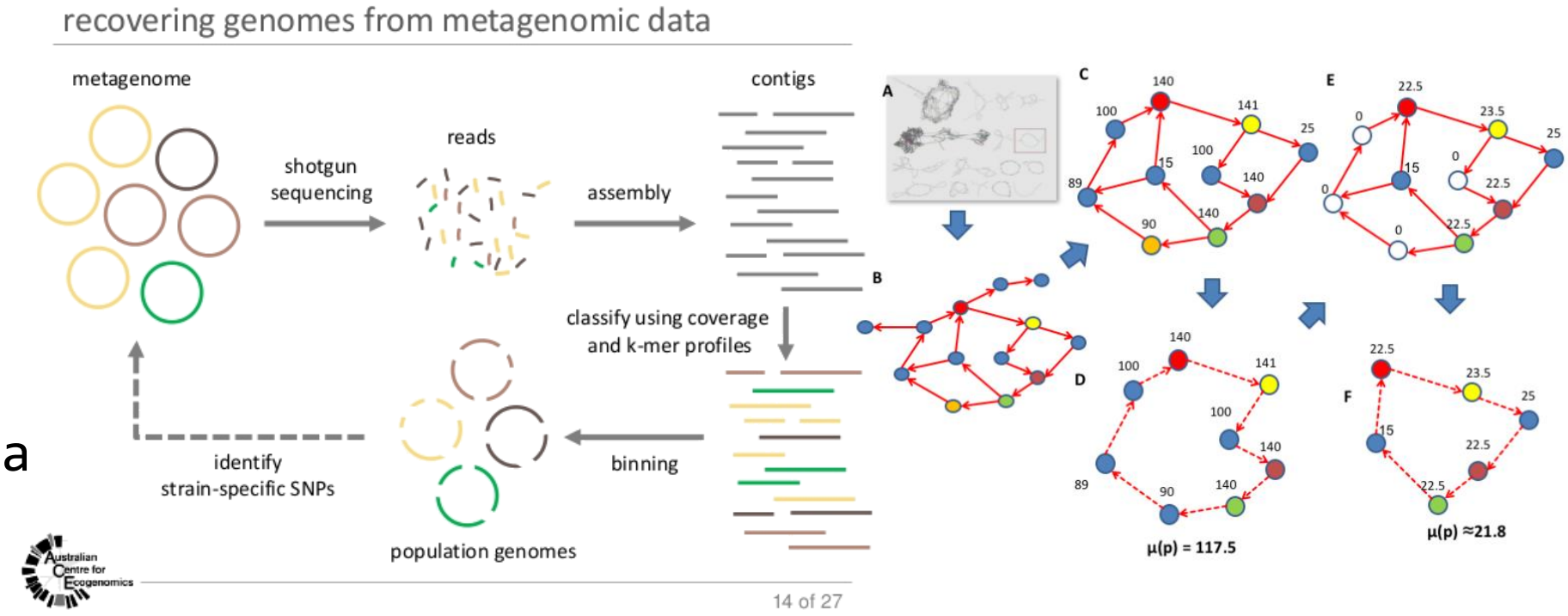
Ensamblado: Ensambladores

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

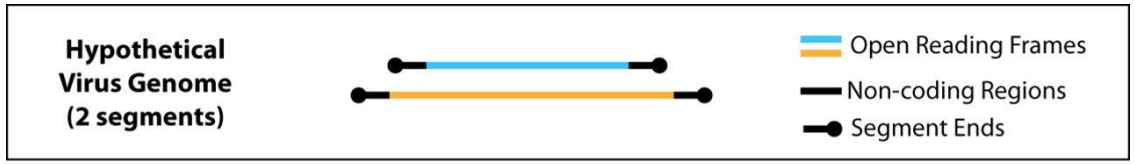
*Licences: OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics

Ensamblado: Ensamblados especiales

- Genomas diploides
- Metagenomas
- Plásmidos
- Transcriptoma



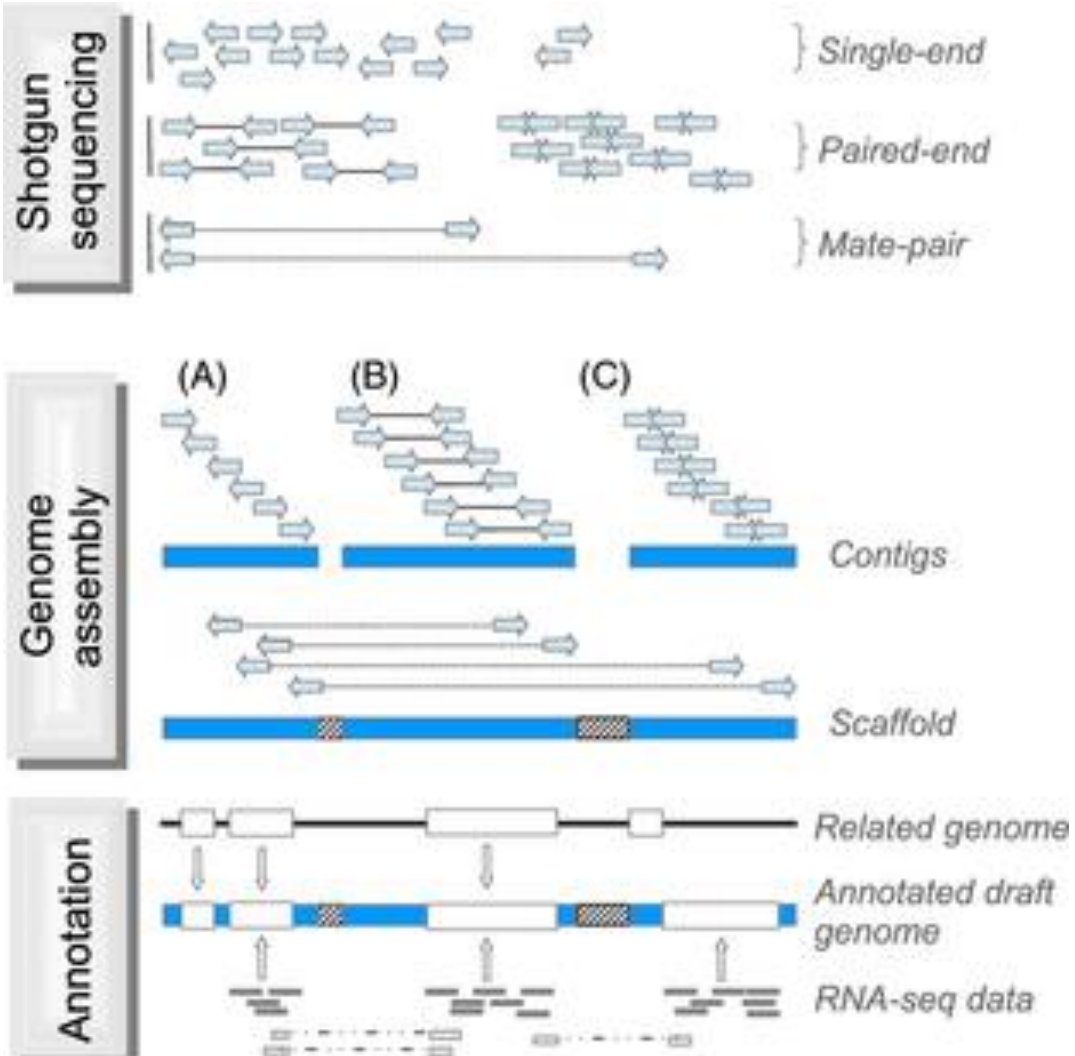
Ensamblado: Categorías



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*.

Anotación



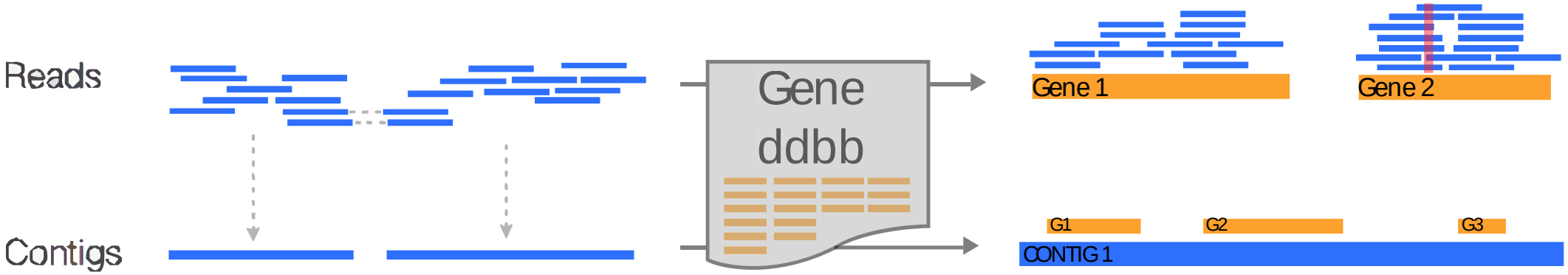
A field guide to whole-genome sequencing, assembly and annotation. Ekblom R, Wolf J

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

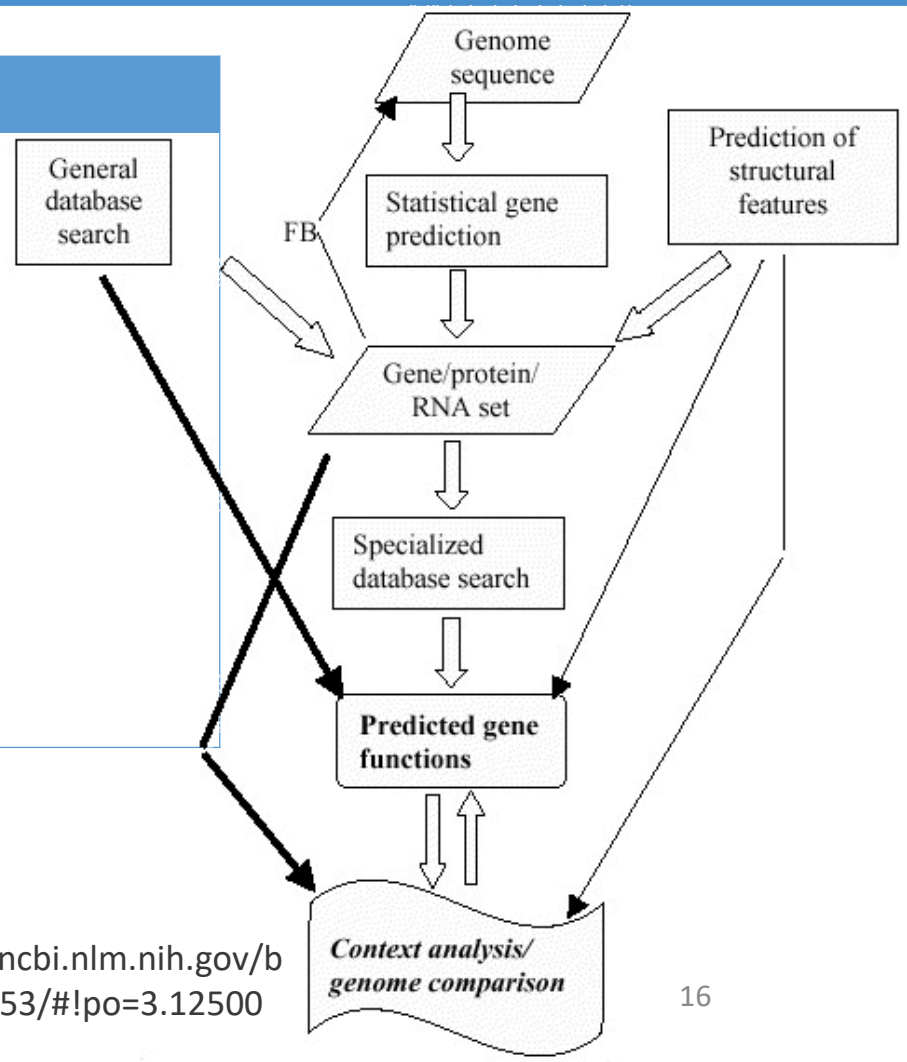
Anotación funcional

- 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

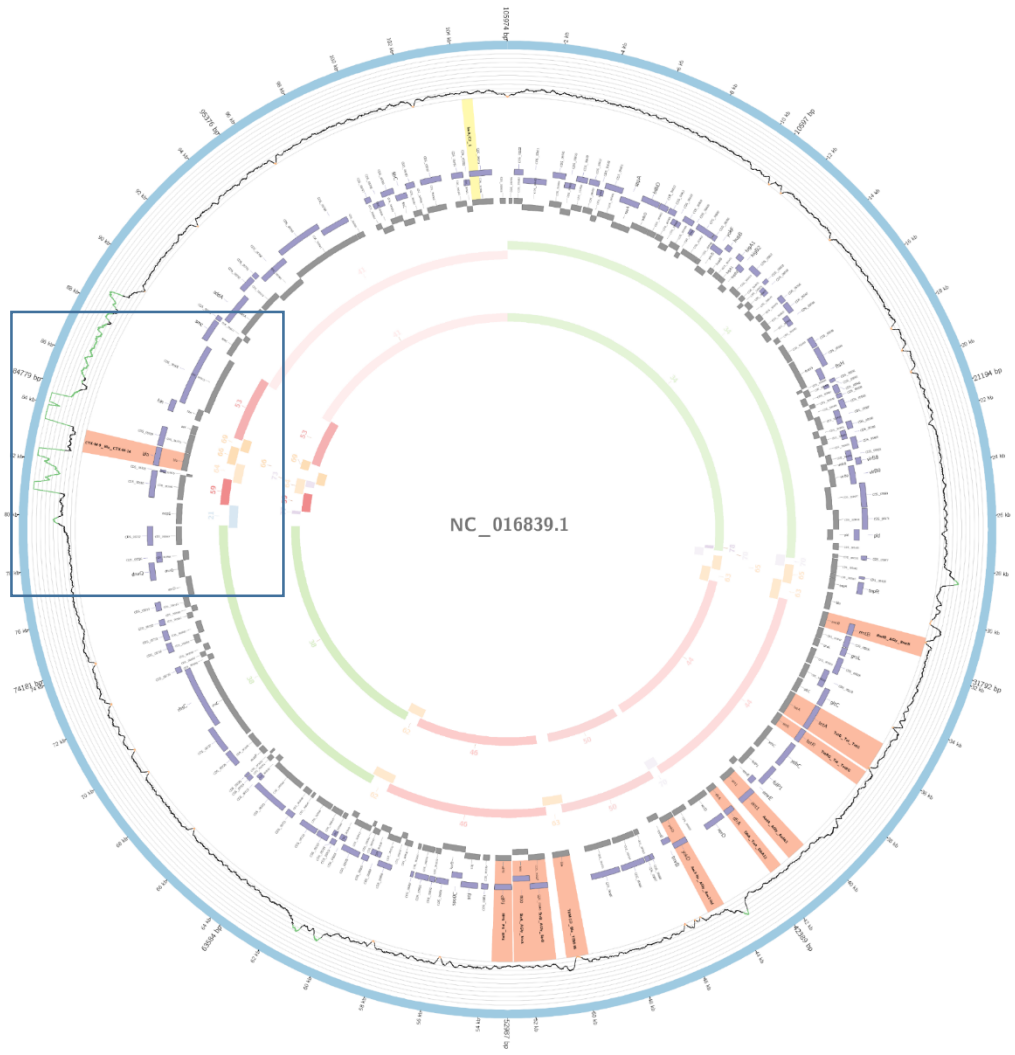
Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen et al. , 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen et al. , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA
BLAST+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

PlasmidID


NC_016839.1


84779 bp
84 kb
82 kb
80 kb
78 kb
CTX-M-9_Bla_CTX-M-14
bla
hin
dnaQ
mcpS