

Ensamblaje de genomas y predicción de genes*

Bioinformática para biotecnología BIT120

22 agosto 2017

Eduardo Castro, PhD

www.castrolab.org

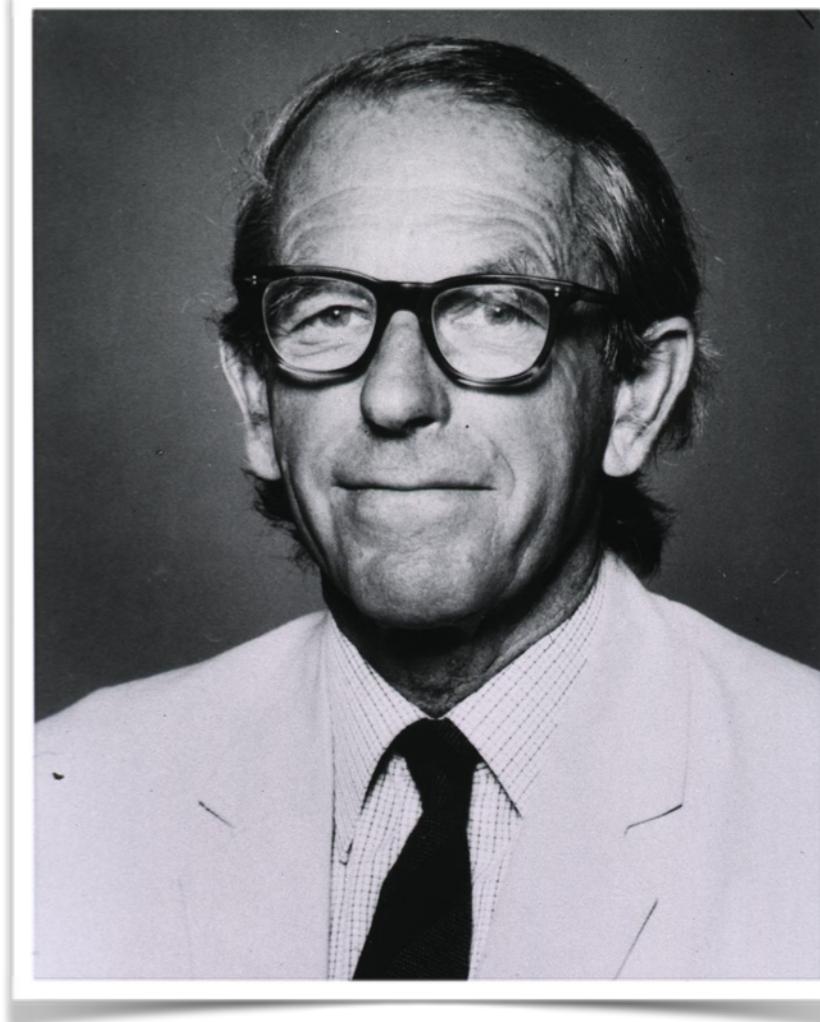
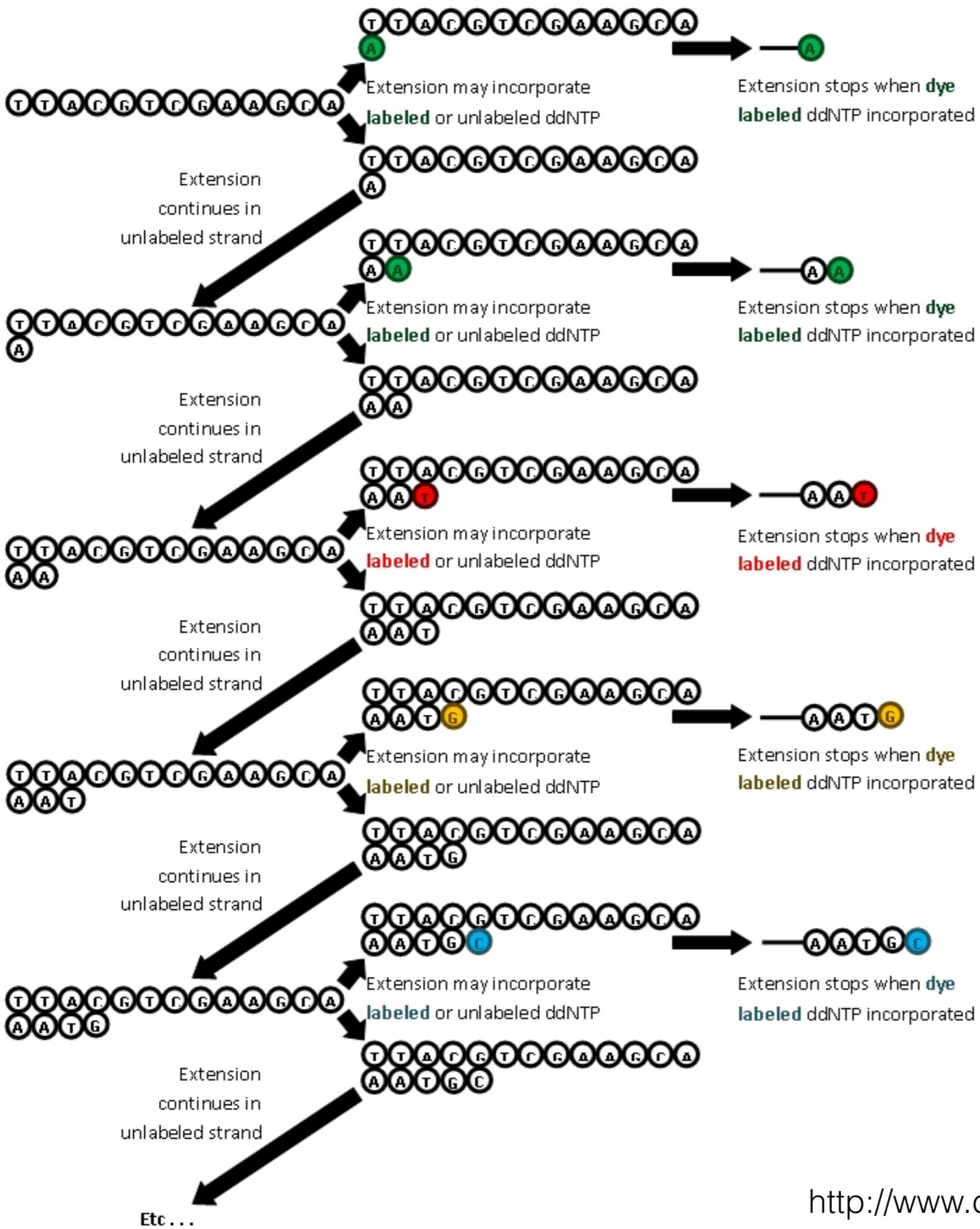
*un poco de secuenciamiento masivo

¿Por qué necesitamos ensamblar un genoma?

- ¿Qué es lo que tenemos que ensamblar?
- ¿Podemos obtener un genoma completo sin necesidad de ensamblar algo?

¿Cómo secuenciamos los genomas?

- Reacción similar al PCR
- Se usan polimerasas modificadas. Principio biológico es la replicación
- Por terminación de cadena naciente, por síntesis, etc.



- 1958 estructura de proteínas, especialmente insulina
- 1980 determinación de secuencia de bases en ácidos nucleicos

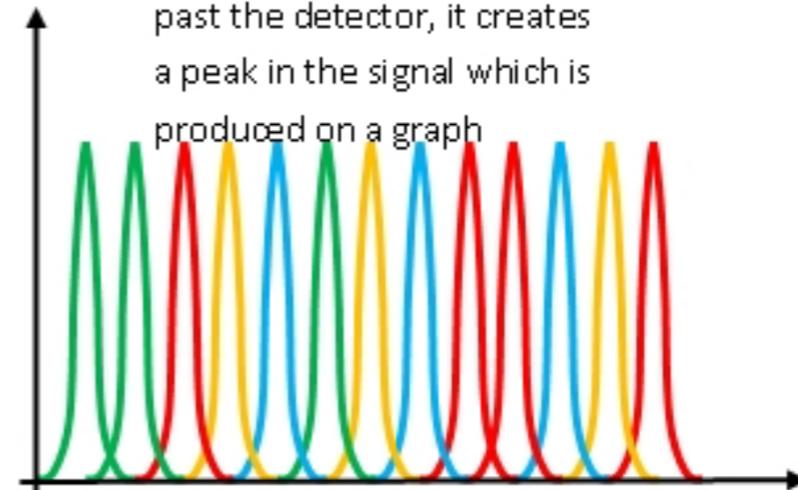
DNA Fragments with Dye Terminators
(Smaller fragments pass through the capillary first)



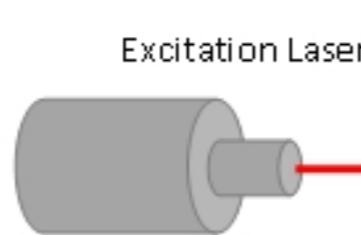
Capillary tube



As each band of colour
(caused by collections of
dye terminated fragments
of the same size) moves
past the detector, it creates
a peak in the signal which is
produced on a graph

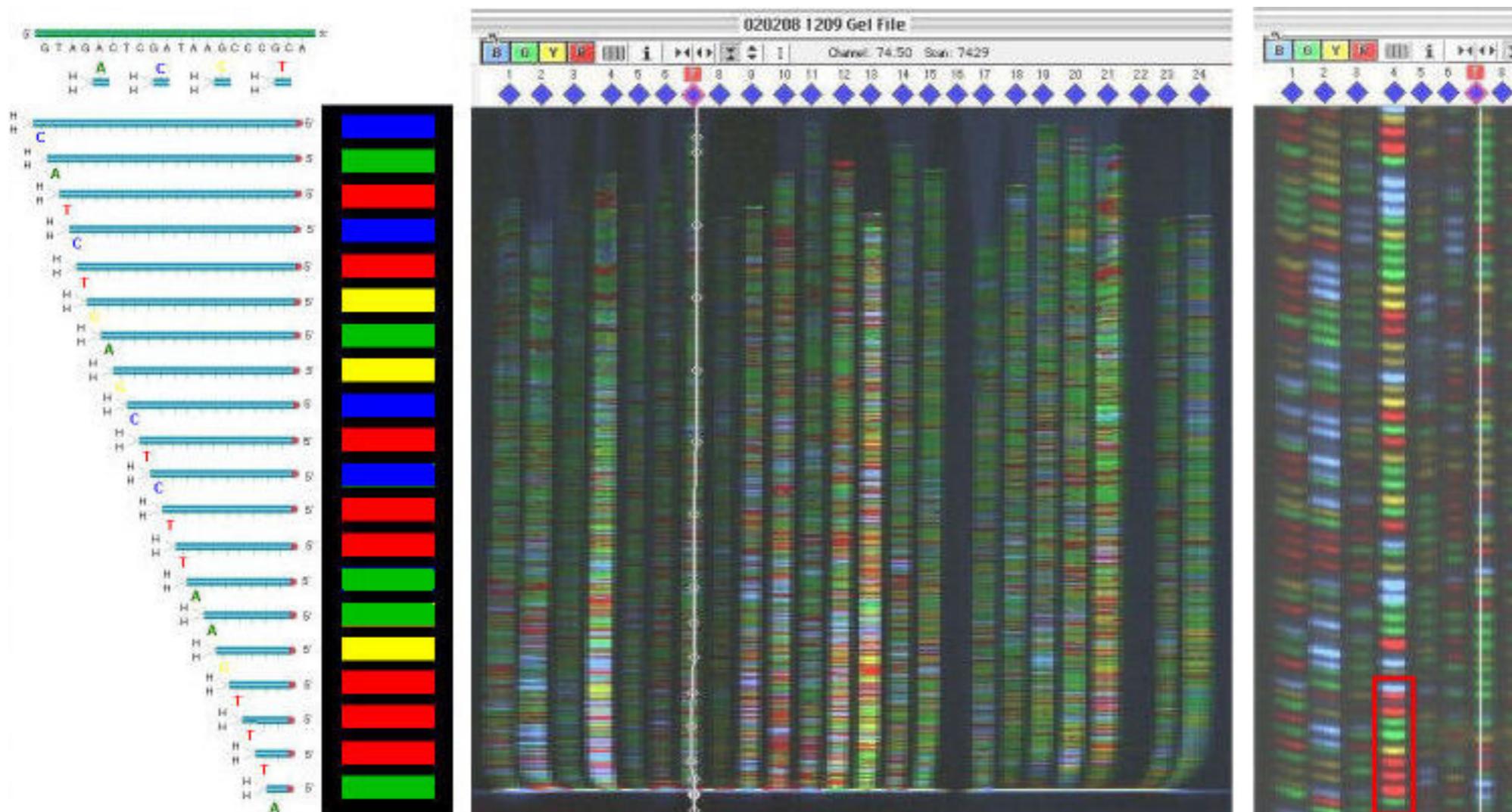


Sequence output



Detector

Imagen real de un gel de secuenciamento



...Transformada a un chromatograma



https://www.mun.ca/biology/scarr/377_Chromatogram.html

Requisitos del método de Sanger

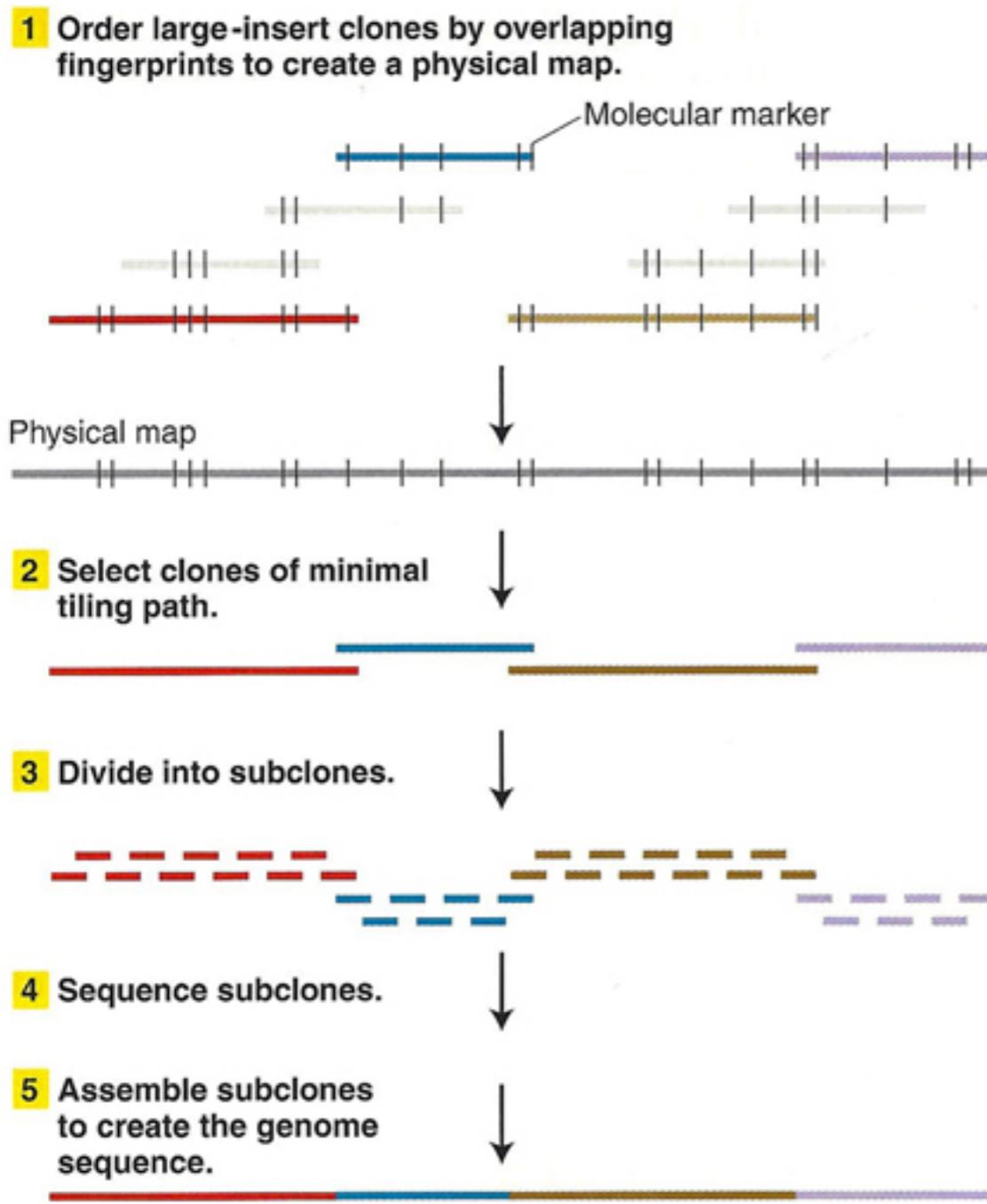
- DNA tiene que tener un primer; 1000 bp a la vez
- Lento y laborioso —> clonar en BACs, YACs
- Dos estrategias —> Directo o clon por clon y aleatorio o shotgun

Características Sanger

First generation (Sanger) sequencing

throughput	50-100kb, 96 sequences per run
read length	0.5-1.1 kbp
accuracy	high quality bases - 99%: ~900bp very high quality bases - 99.9%: ~600bp 99.999%: 400-500bp
price per raw base	~400k€/Gb

Secuenciamiento por clones o basado en mapas

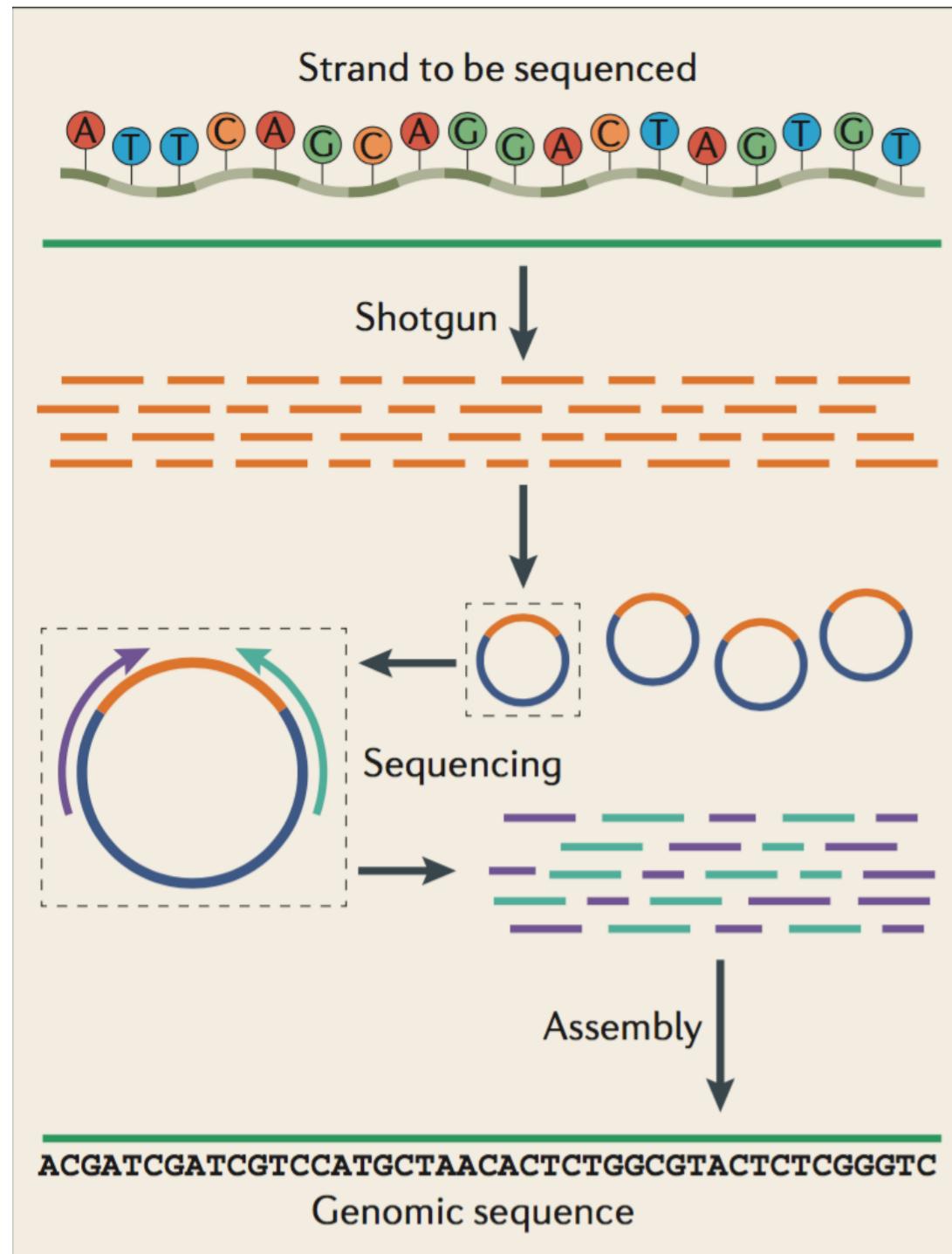


- Crear mapa físico
- Seleccionar clones con trayectoria minima
- Dividir en subclones
- Secuenciar subclones
- Ensamblar

Limitaciones

- Súper laborioso
- Toma mucho tiempo, recursos y personal especializado
- Caro → \$3 mil millones USD proyecto genoma humano NIH

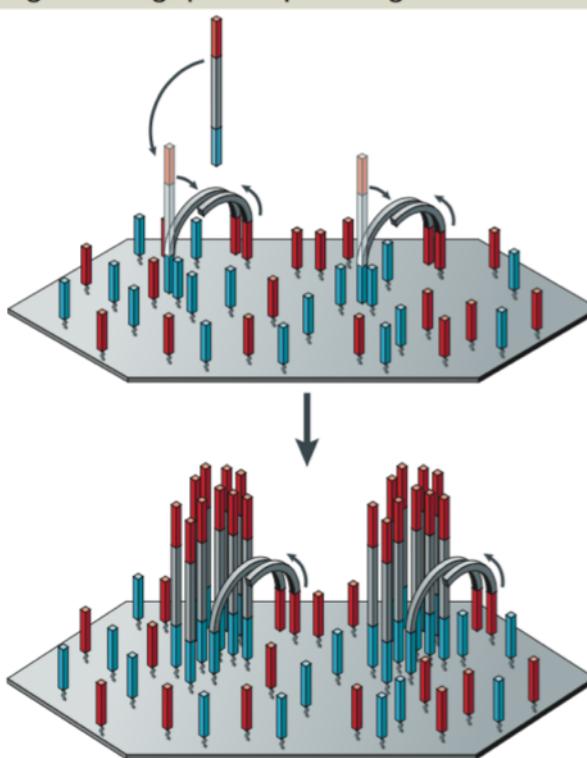
Whole-genome shotgun: la primera revolución



- Más fácil que clone-by-clone
- Creada por J Craig Venter
- Se transforma en la estrategia dominante
- Pone por primera vez la “carga” en el análisis post-secuenciación
- Armar el puzzle después de que el experimento ha concluido
- Menos caro: \$300,000,000 USD

High-throughput sequencing: la segunda revolución

The Second Revolution
High-throughput sequencing



454 sequencing

- Sequencing by synthesis
- Amplified templates generated *in vitro*
- High accuracy outside homopolymers but short read lengths

For example, 454 GS FLX+ (Roche)

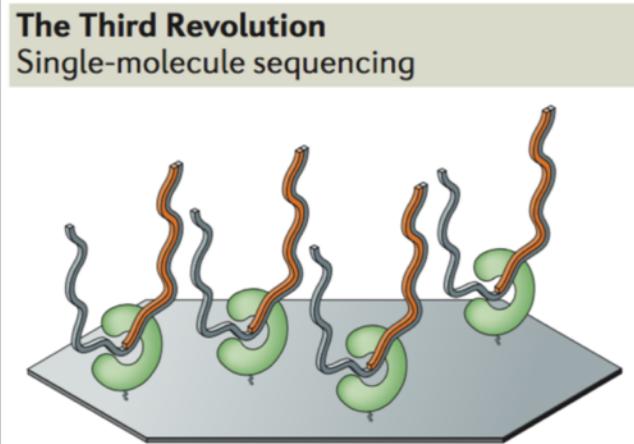
Illumina sequencing

- Sequencing by synthesis
- Amplified templates generated *in vitro*
- High accuracy but short read lengths

For example, MiSeq (Illumina)

- 2005
- No hay terminación temprana de la cadena naciente
- Secuenciamiento por síntesis
- Alto rendimiento —> 1 M a 400 M de fragmentos o “reads”
- Mayor tasa de error que Sanger pero no importa

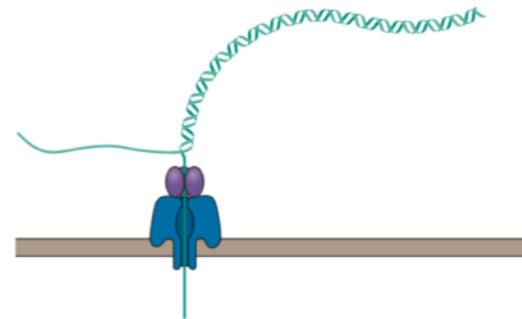
Tercera revolución? Single-molecule sequencing



Pac Bio SMRT sequencing

- Sequencing by synthesis
- Single-molecule templates
- Low accuracy but long read lengths

For example, PacBio RS
(Pacific Biosciences)



Oxford Nanopore sequencing

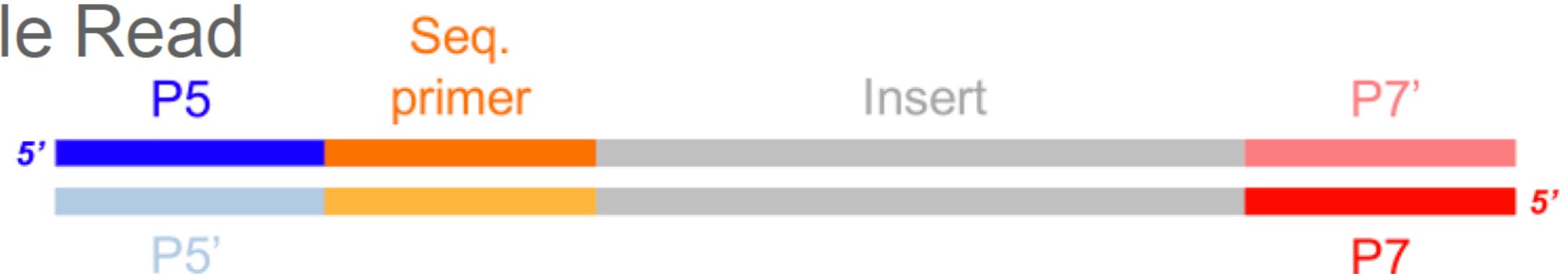
- Nanopore sequencing
- Single-molecule templates
- Low accuracy but long read lengths

For example, MinION
(Oxford Nanopore)

- 2009
- PacBio SMRT → Secuenciamiento por síntesis
- Produce fragmentos largos de hasta 200 kbp
- Mayor tasa de error que Illumina o 454
- No hay amplificación
- Puede capturar señales epigenéticas

Tipos de “reads” o lecturas

Single Read



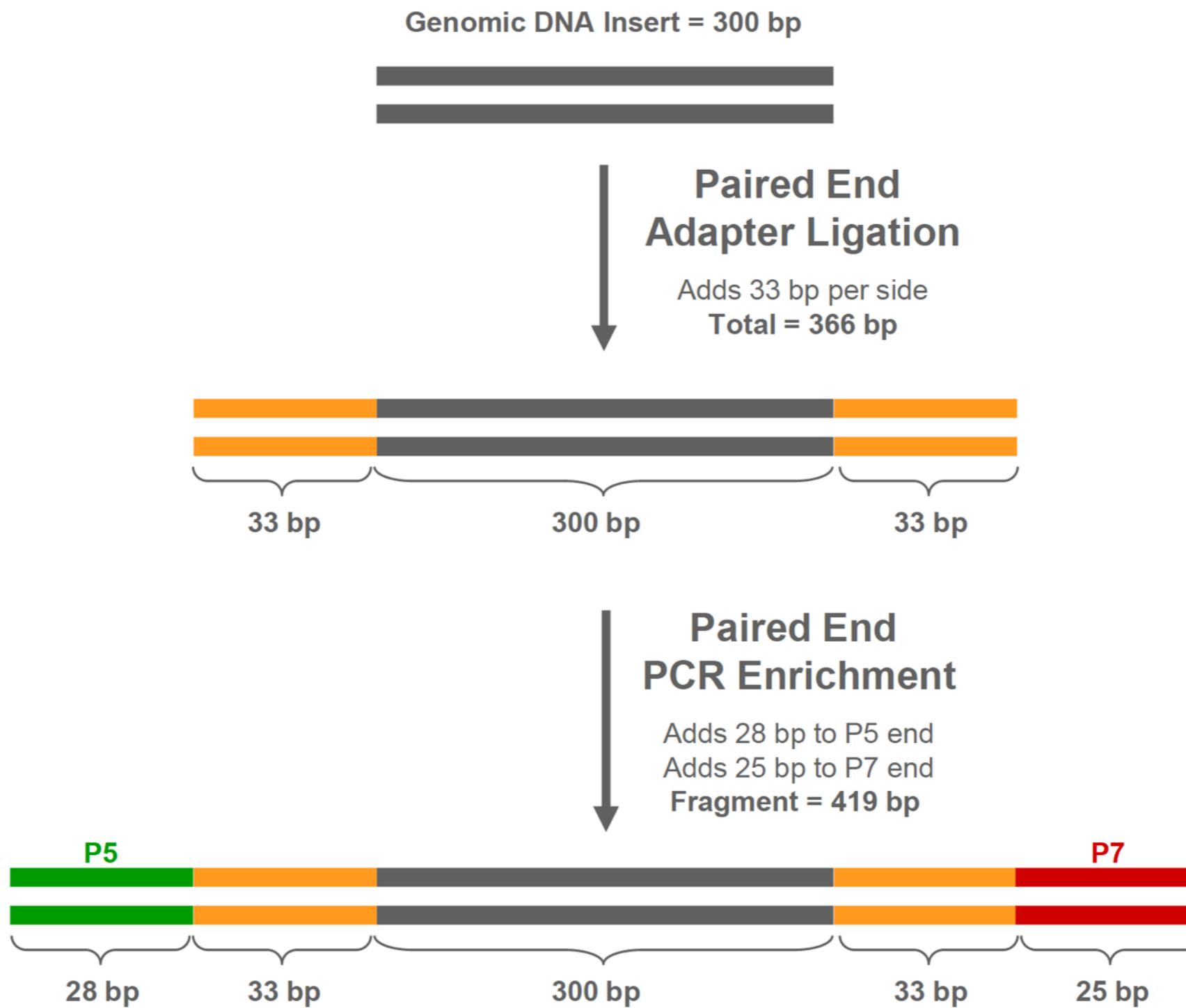
Paired Read



Indexed Paired Read

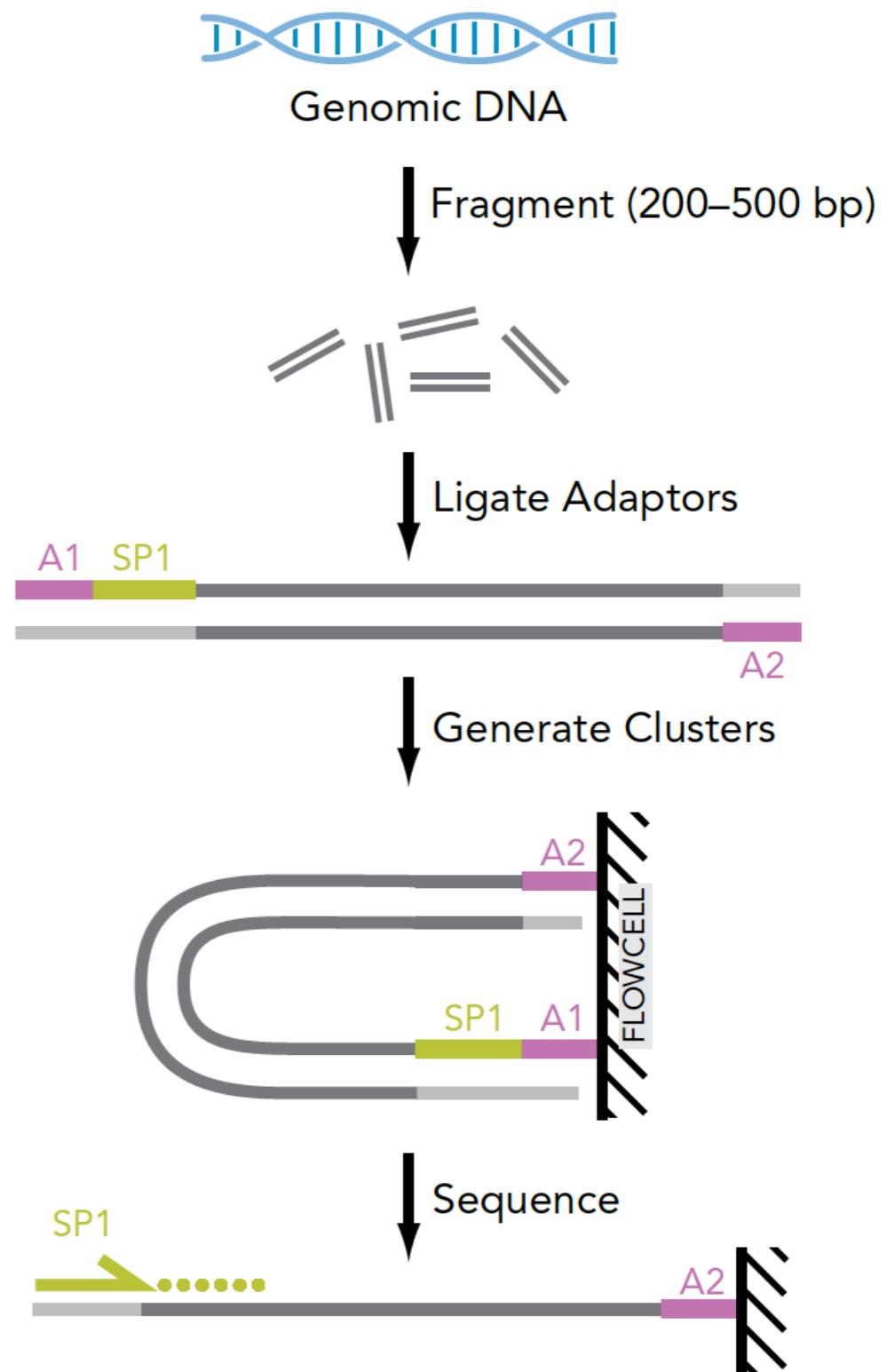


Inserto y secuencia útil



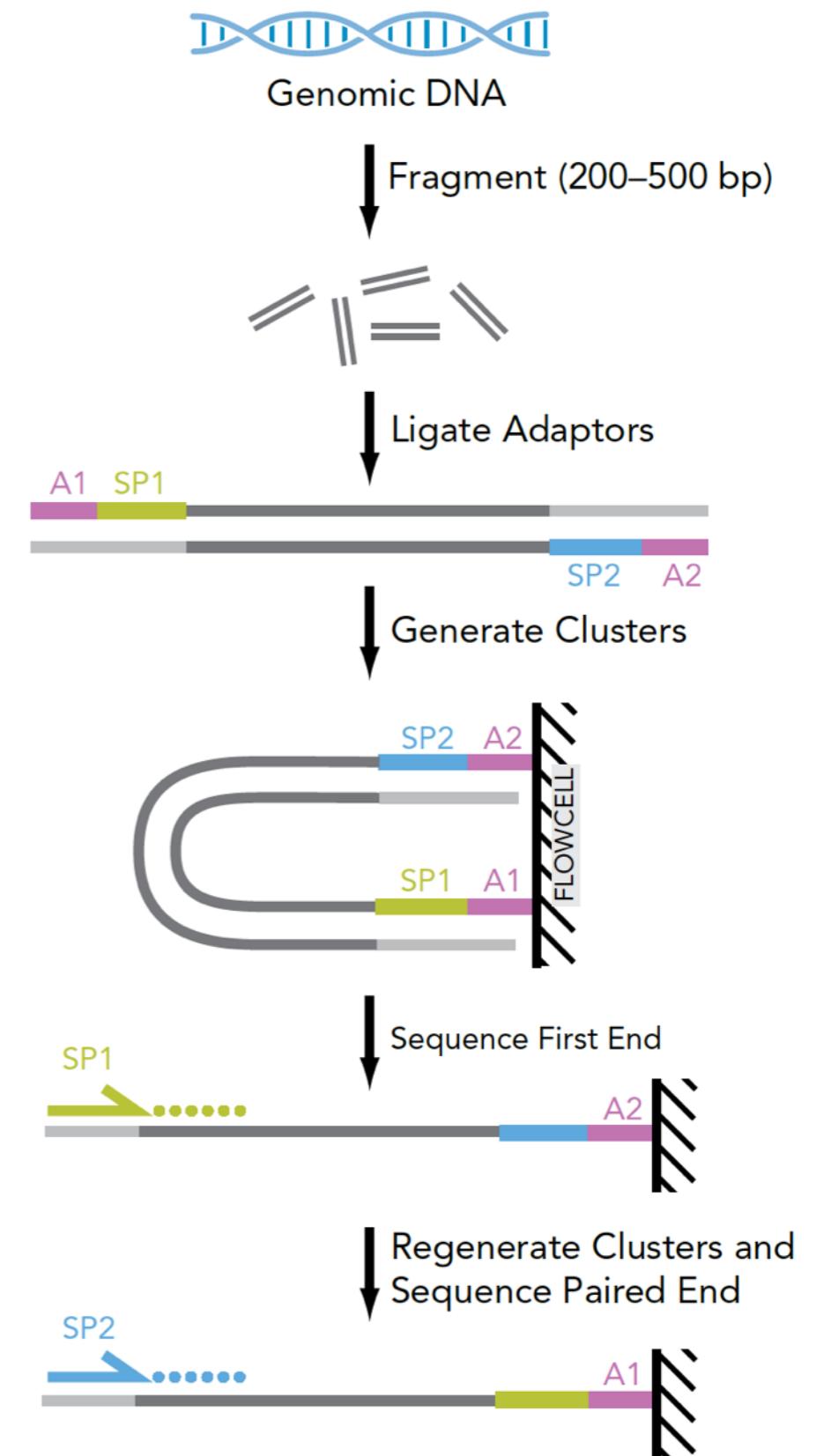
Single-end

- Solo un partidor para secuenciar
- Rápido, más barato
- Descontinuado



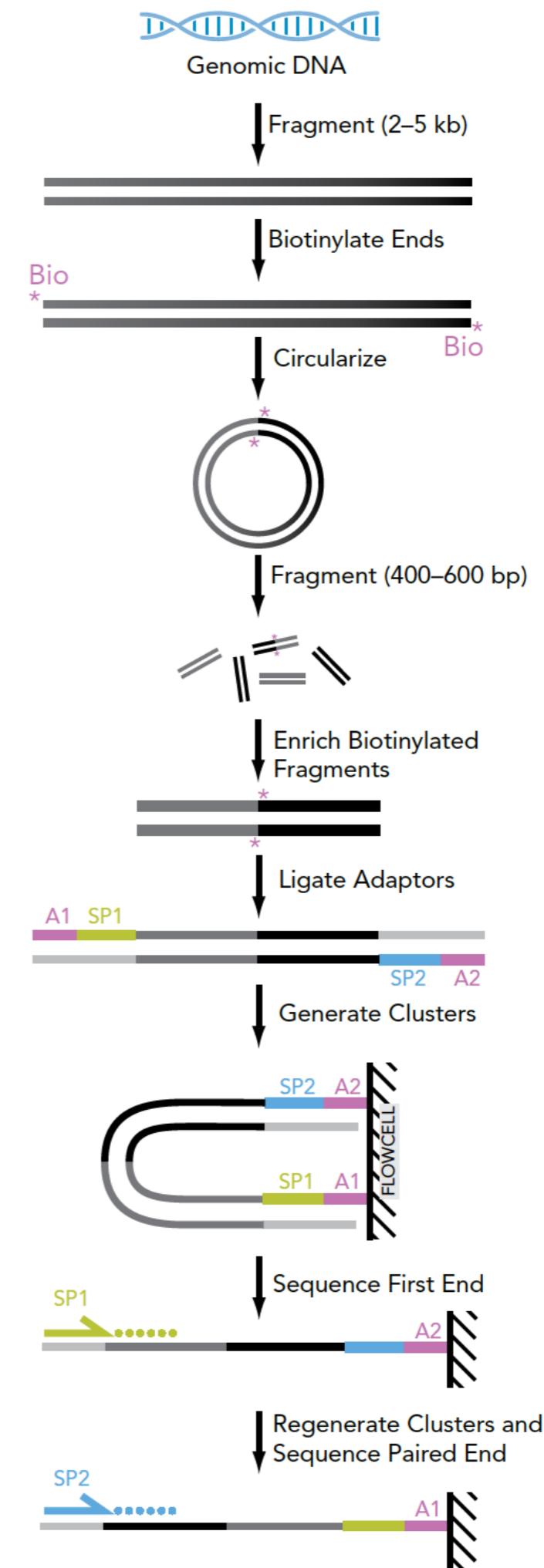
Paired-end

- Se secuencia el mismo inserto dos veces
- Es posible “alargar” el tamaño de la read
- Captura información estructural
- Toma el doble de tiempo, más caro



Mate-pairs

- Información estructural
- Finalizar genomas, genomas de alta calidad
- Resolver genes multicopia, regiones repetitivas



Resultado de la secuenciación



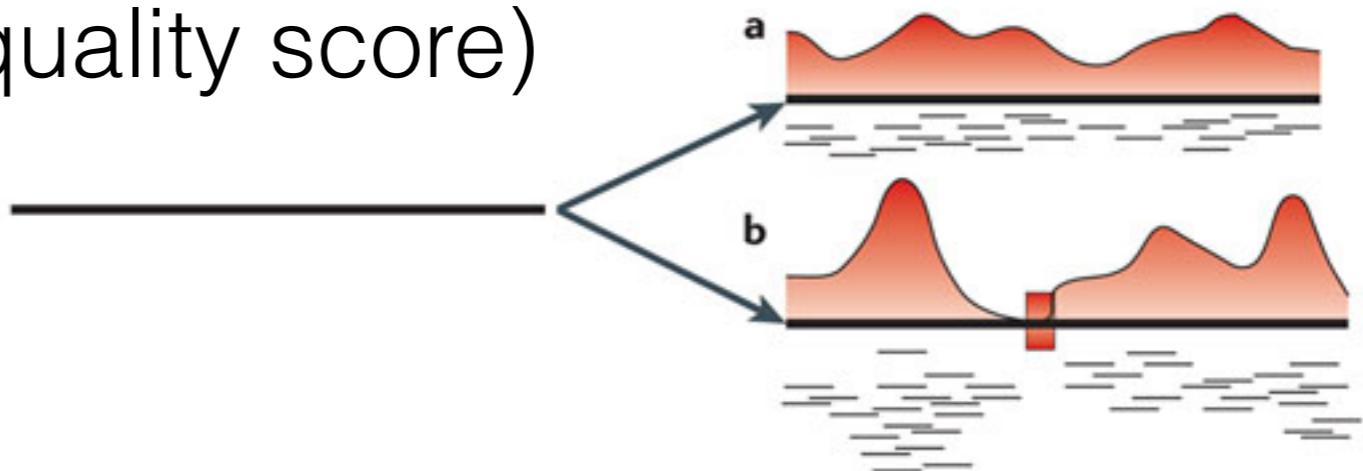
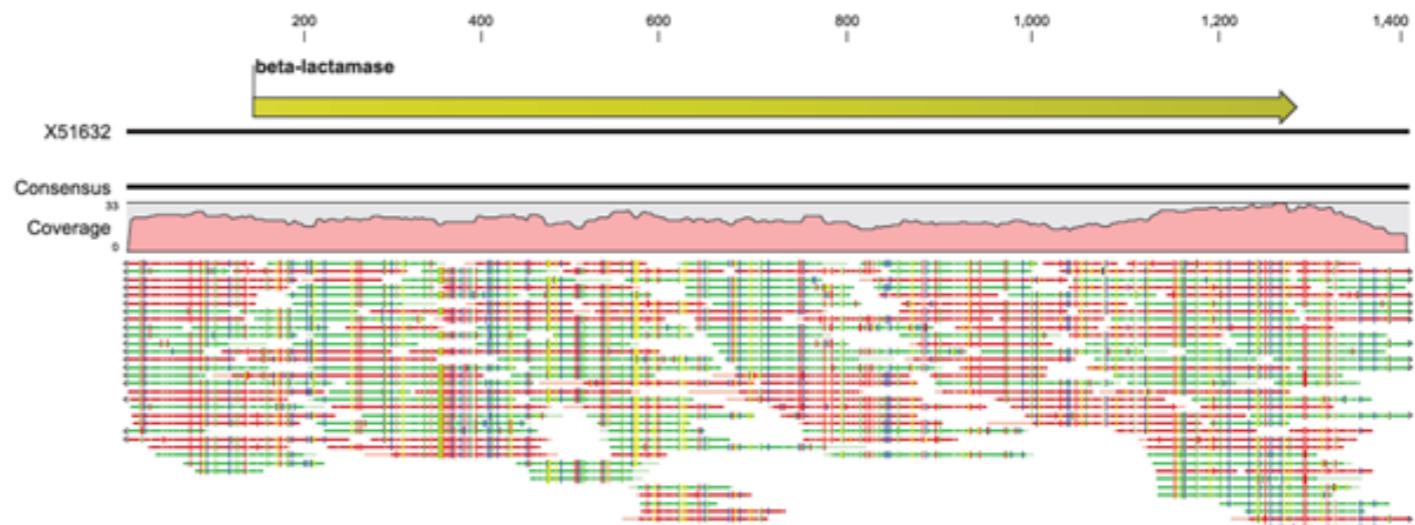
Resultado de la secuenciación

Baño, café, aire fresco
10 minutos de recreo

Estrategias para armar
el rompecabezas

Terminologia

- Coverage
- Reads, contigs, scaffolds
- Profundidad (depth)
- Puntaje de calidad (quality score)

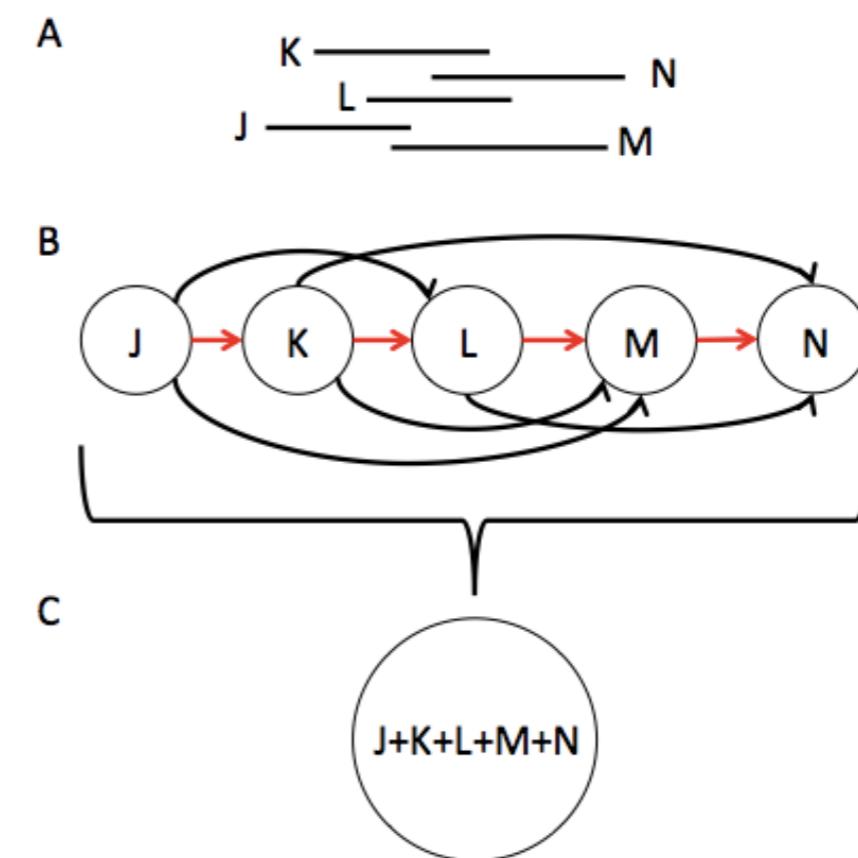
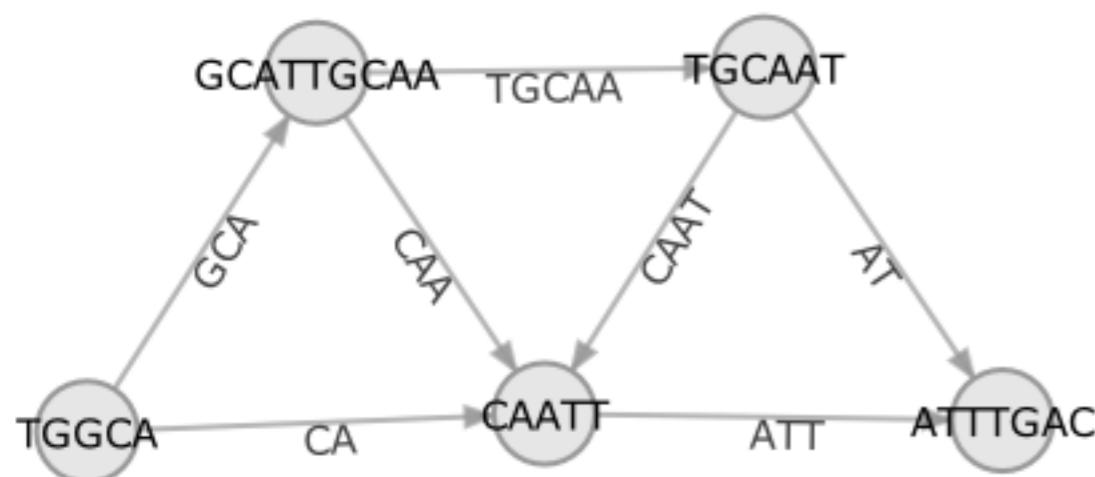


De novo

- Utilizar las reads por sí solas para reconstruir el genoma
- Dos estrategias: Overlay-layout-consensus y De Bruijn graphs

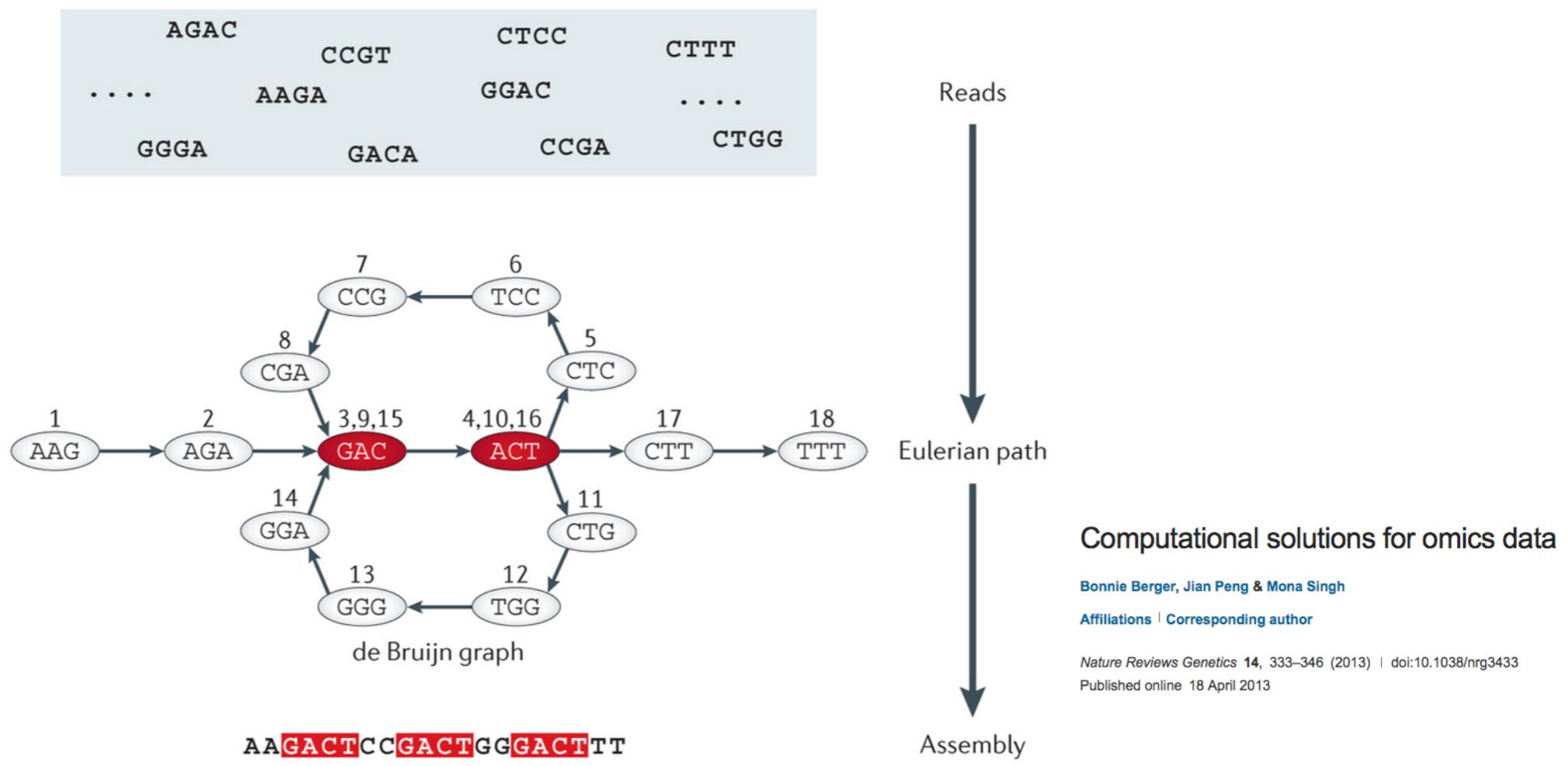
Overlay-layout-consensus

- Sobreponer reads por identidad de secuencia, unir reads sobrepuertas y encontrar un camino, formar un consenso



De Bruijn graphs

- Fragmentar reads en pedazos de longitud K (llamados kmers), generar un gráfico sobrelapando kmers. Finalmente se forma una secuencia al trazar un camino donde cada kmer se visita una vez



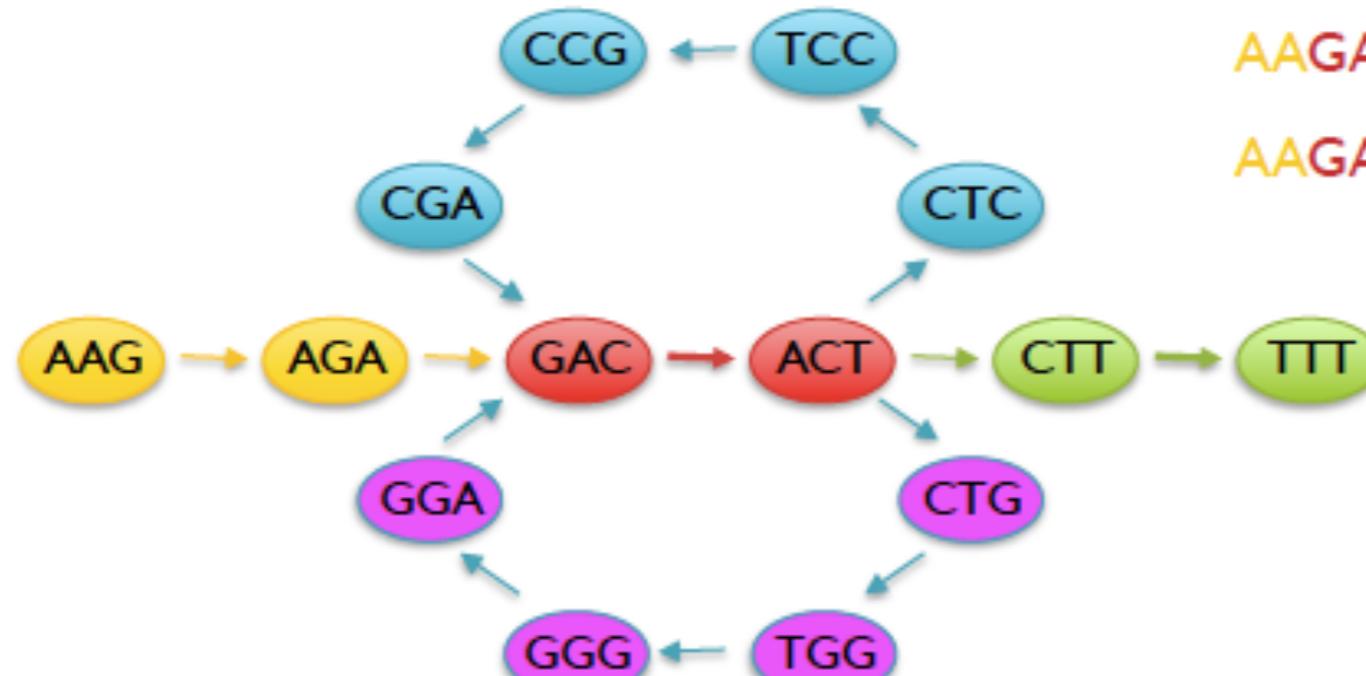
De Bruijn graphs

- Más de una solución para el mismo gráfico

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
...

de Bruijn Graph



Potential Genomes

AAGACTCCGACTGGGACTTT
AAGACTGGGACTCCGACTTT

Consideraciones con ensamblaje de novo

- DBG - valor de k, repeticiones más largas que k o que reads.
Muchos errores con repeticiones
- DBG - mejor con reads cortas, aunque reads cada vez más largas
- DBG - requieren mucha memoria RAM, e.g., 140 GB a 2 TB
- OLC - lento, semanas en supercomputadora
- OLC - requiere calcular todas las combinaciones de reads
- OLC - errores cuando datos tienen mucha profundidad

Consideraciones con ensamblaje de novo

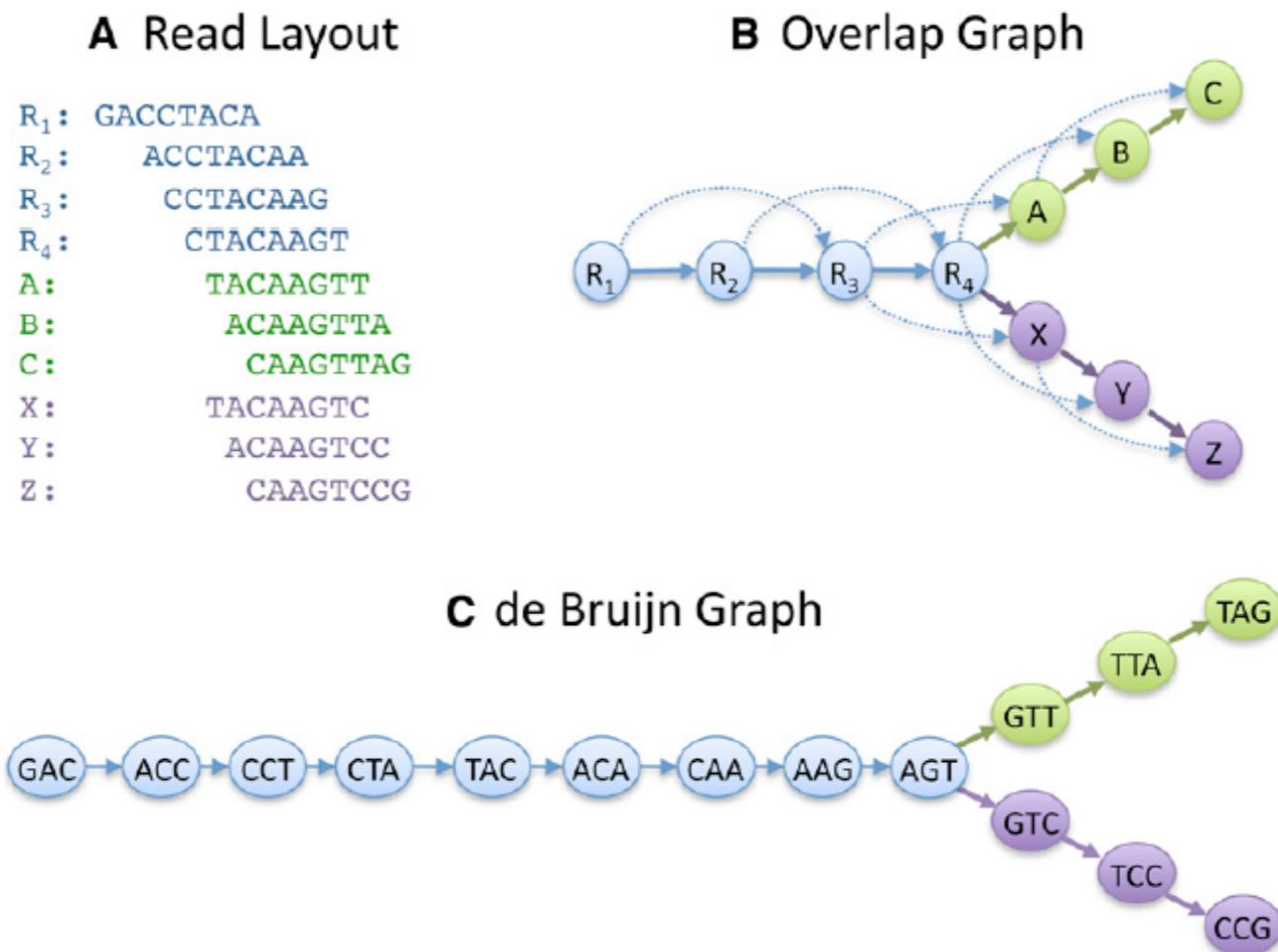


Figure 2. Differences between an overlap graph and a de Bruijn graph for assembly. Based on the set of 10 8-bp reads (A), we can build an overlap graph (B) in which each read is a node, and overlaps >5 bp are indicated by directed edges. Transitive overlaps, which are implied by other longer overlaps, are shown as dotted edges. In a de Bruin graph (C), a node is created for every k-mer in all the reads; here the k-mer size is 3. Edges are drawn between every pair of successive k-mers in a read, where the k-mers overlap by $k - 1$ bases. In both approaches, repeat sequences create a fork in the graph. Note here we have only considered the forward orientation of each sequence to simplify the figure.

Assembly of large genomes using second-generation sequencing

Michael C. Schatz, Arthur L. Delcher and Steven L. Salzberg

Genome Res. published online May 27, 2010
Access the most recent version at doi:10.1101/gr.101360.109

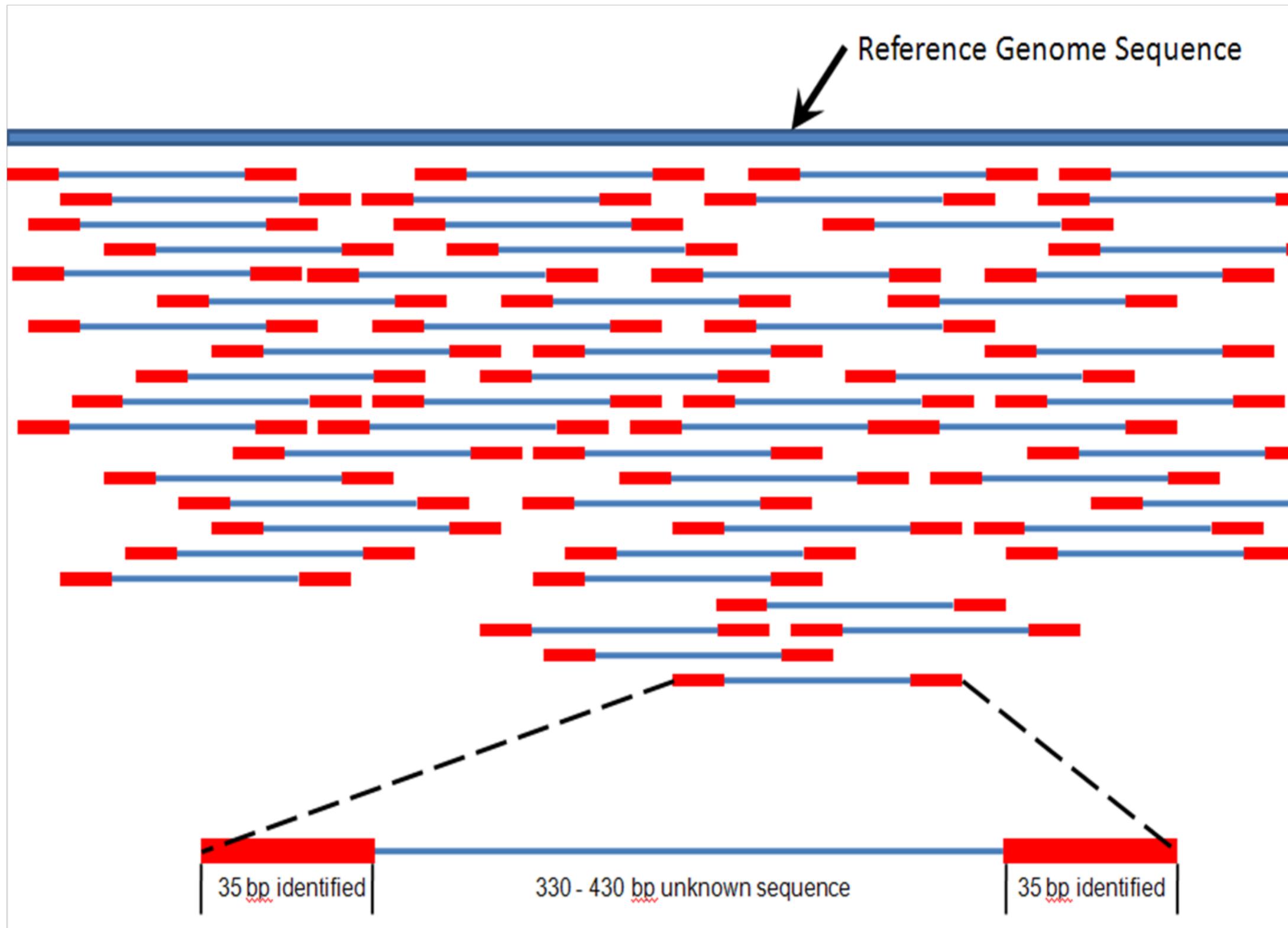
Consideraciones con ensamblaje de novo

- Mientras más largas las reads mejor es el ensamblaje, menos ambigüedad
- Se necesita mucho coverage
- El resultado está fragmentado
- Evaluar la calidad —> N50, mediana, media
- N50 = después de ordenar contigs, se divide la distribución de bases por la mitad, la longitud del contig donde esto ocurre es el N50

Por referencias

- Útil para estudios de resecuenciamiento, e.g., UK10K, GenomeTrakr
- Se usa un genoma ensamblado para “mapear” reads
- Computacionalmente más fácil que *de novo*
- Reads cortas pueden mapear en varias partes en la referencia
- Limita conocer la estructura de genomas nuevos, restringe reads a la referencia

Por referencias

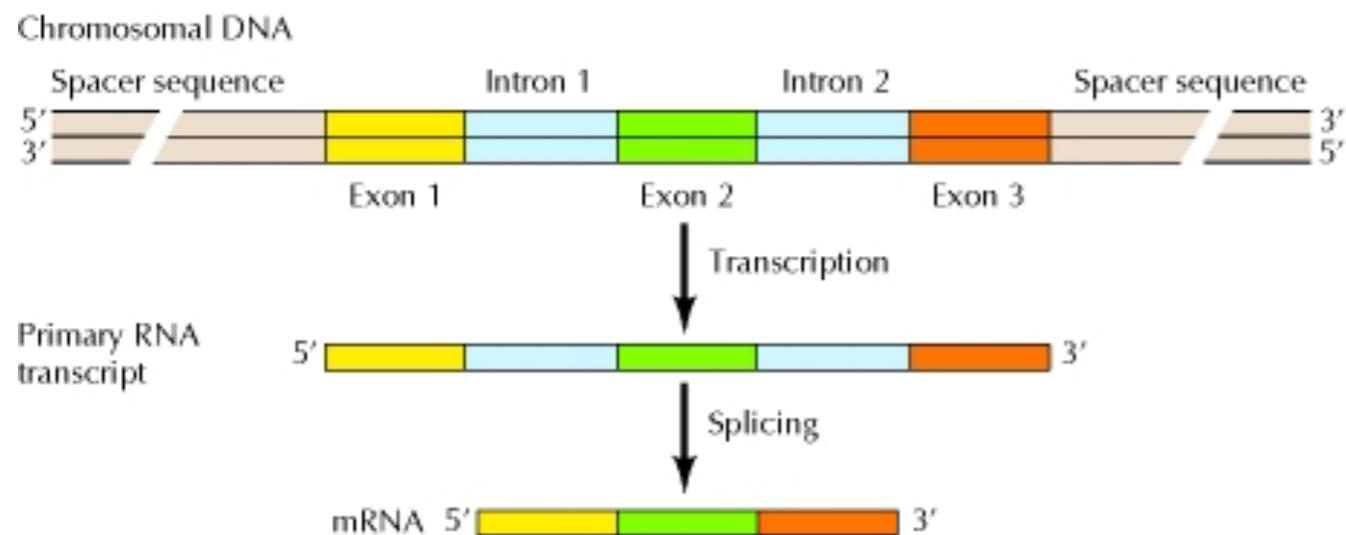


¿Qué obtenemos al final del ensamblaje?

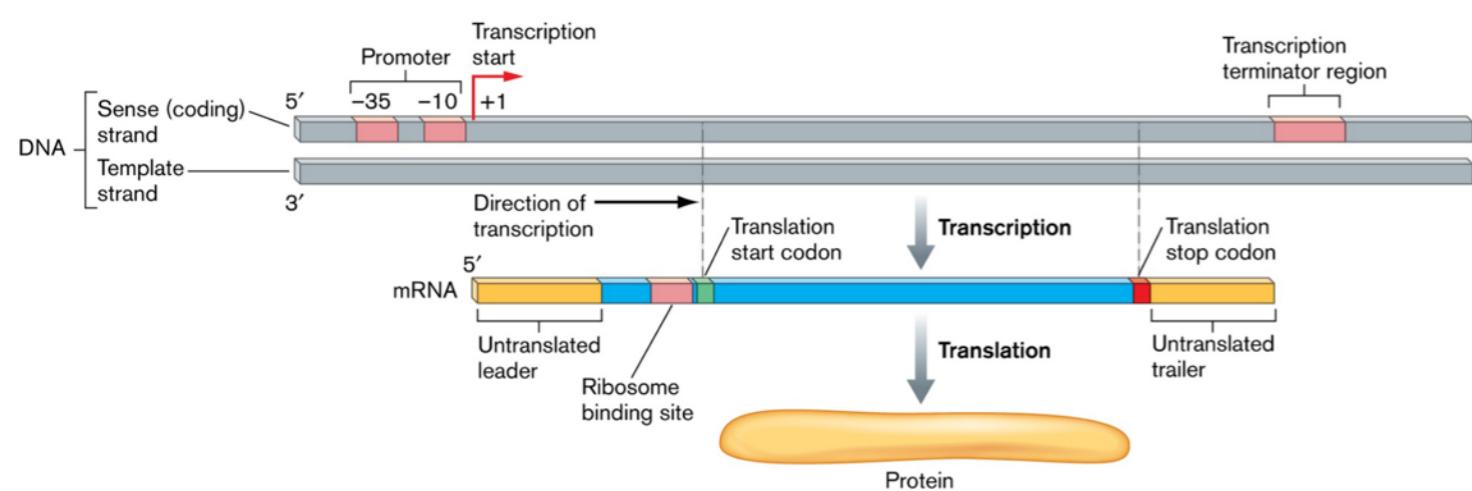
- Contigs o scaffolds
- Difícilmente se recupera el genoma completo, i.e., cromosomas lineales o circulares
- 100 contigs para bacterias es común
- “Finalizar” o “cerrar” es más caro y laborioso

Predicción de genes

Estructura de genes

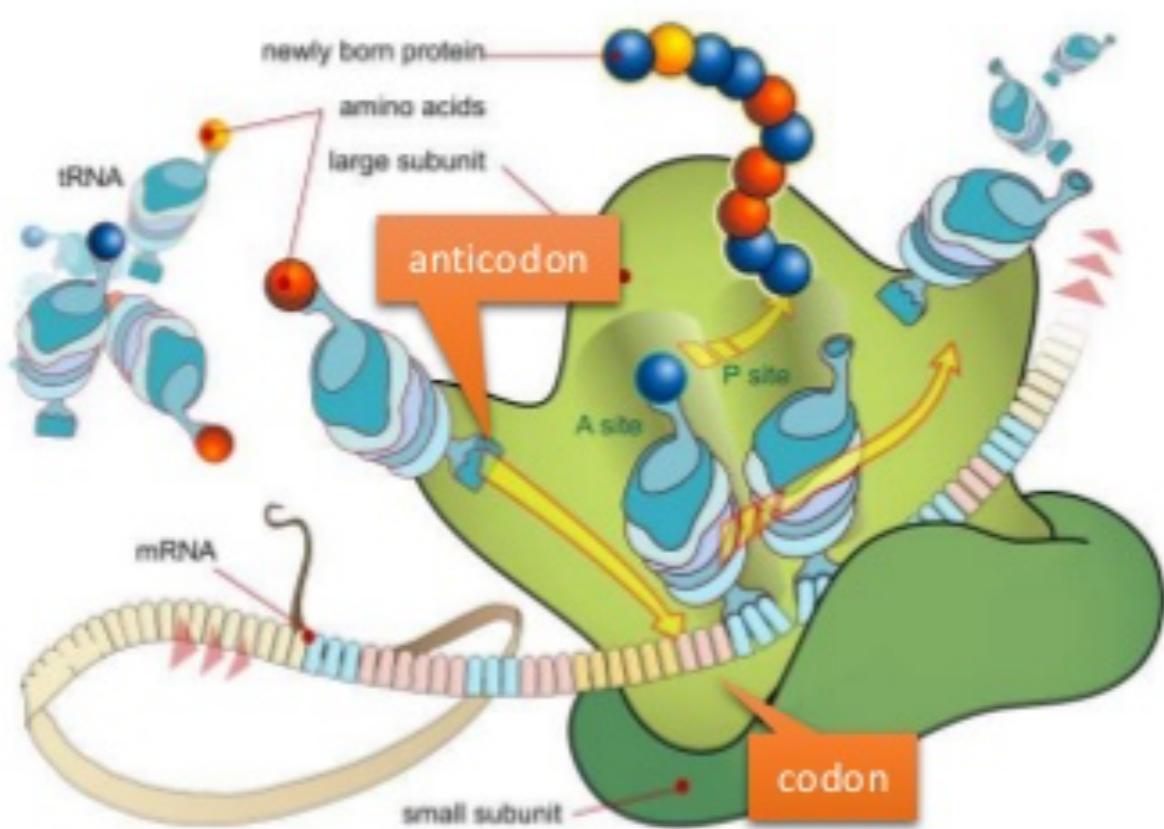


- ADN eucarionte envuelto en histonas, resulta en patrones repetitivos. Promotores están cerca de estos sitios



- Prokaryotes no tienen intrones y regiones promotoras y codones de inicio están conservados
- Ambos difieren en uso de codones

Predicción de genes



- Uso de codones es especie específico
- Regiones funcionales como promotores, sitios de splicing, inicio de la traducción varian por especie

Dos metodologías clásicas

- ***Ab initio* o intrínsecos** —> solo a partir de la secuencia de DNA, busca señales inequívocas de la presencia de un gen o región de interés, e.g., codones de inicio/término, sitios de unión de factores de transcripción
- **Extrínsecos o por homología/evidencia** —> búsquedas en bases de datos curadas de proteínas, mRNAs o transcriptomas.

Ab initio

- **Procariontes** —> más estudiados, se sabe qué buscar y genomas presentan cierta regularidad
 - ORF largos flanqueados por codones de inicio y término. Virtualmente no hay secuencias intergénicas
- **Eucariontes** —> sabemos menos, altamente variables. Sitios de unión para colas de poliA, islas CpG. Intrones y secuencias intergénicas + splicing alternativo lo hacen más complicado
 - Ventaja = intrones son más ricos en A/T que en exones

Predicción de genes

- Modelos génicos
- Coordenadas de inicio y término de elementos genéticos
- En eucariontes, no hay exones sobrelapantes, exones deben estar en el mismo marco de lectura, al juntar dos exones no se debe formar un codón de término

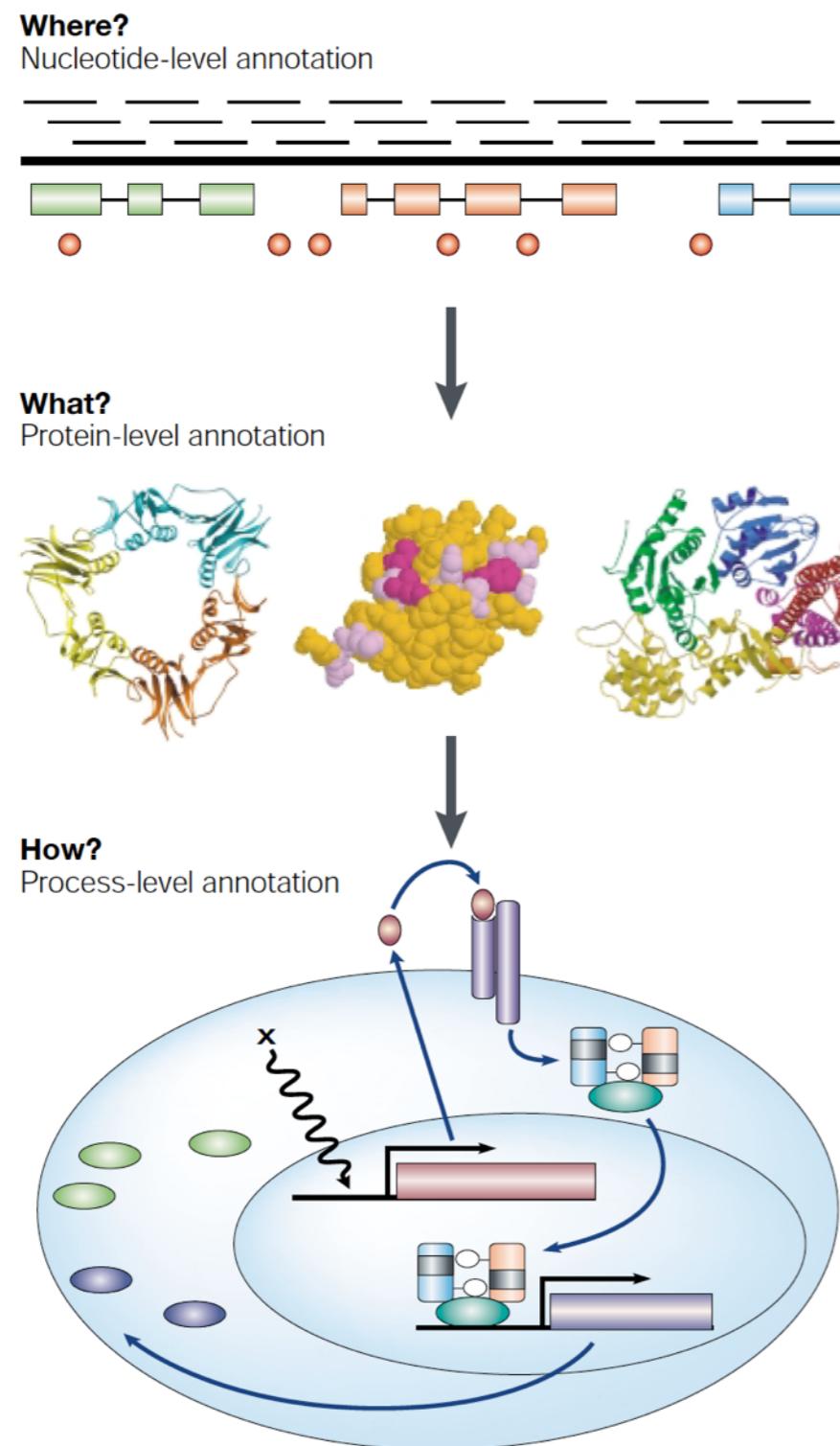
¿Qué tan bien funciona?

- **Procariontes** —> 50-70% por homología, resto *ab initio*. Difícil en genes que se superponen
- **Eucariontes** —> 40% por homología, resto *ab initio*. Refinación con RNASeq.
- Actualmente siempre se usa una combinación de distintos métodos y bases de datos para lograr mejores modelos génicos

Predicción de genes parte de “anotación genómica”

- Una secuencia por si sola no tiene mucho valor
- Es necesario asignar límites dentro de una secuencia para definir donde yacen elementos funcionales del genoma, e.g., genes, rRNAs, tRNAs, lncRNAs, promotores, sitios de unión de proteínas, etc.

Dónde, qué y cómo



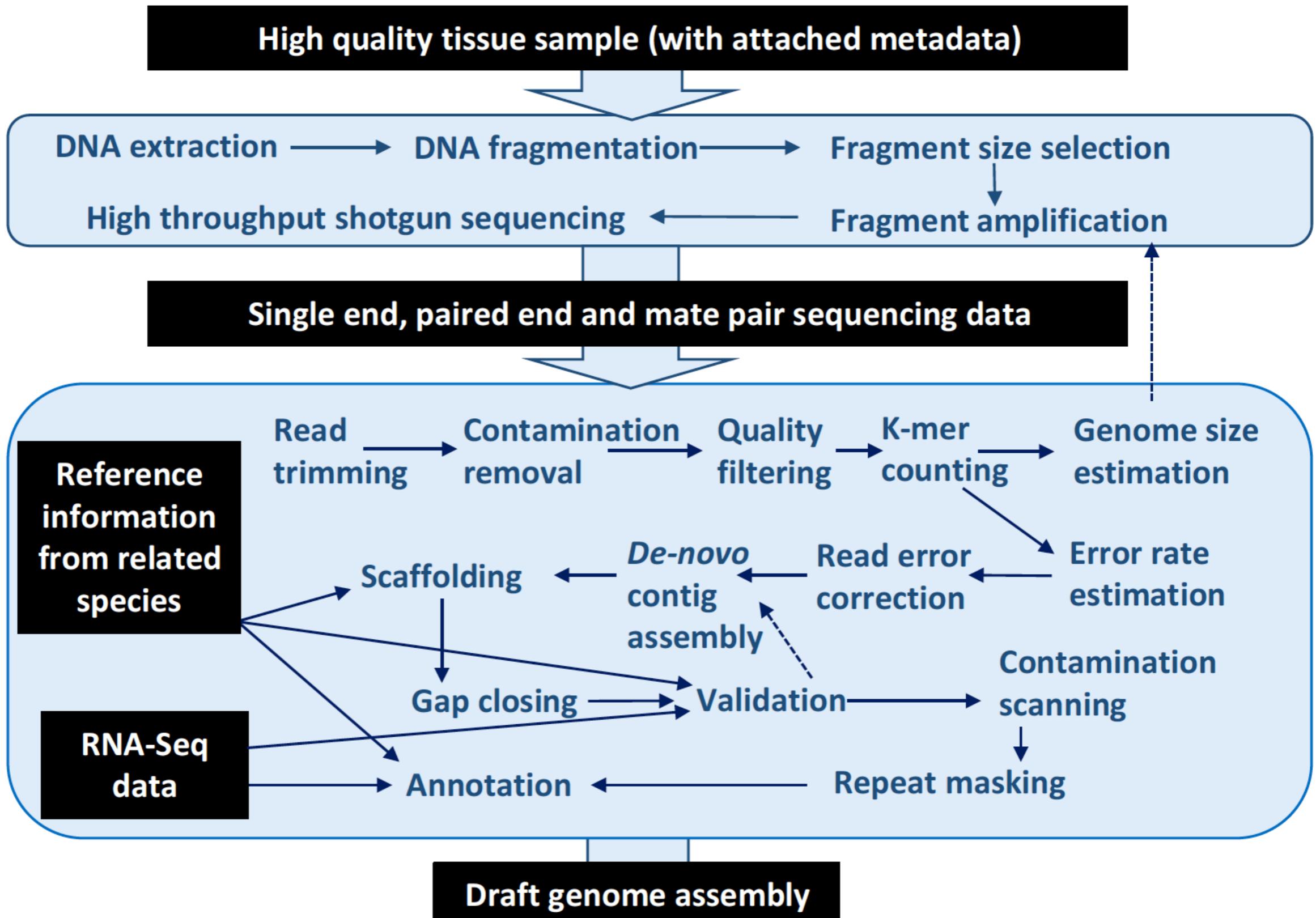
Nature Reviews Genetics **2**, 493-503 (July 2001) | doi:10.1038/35080529

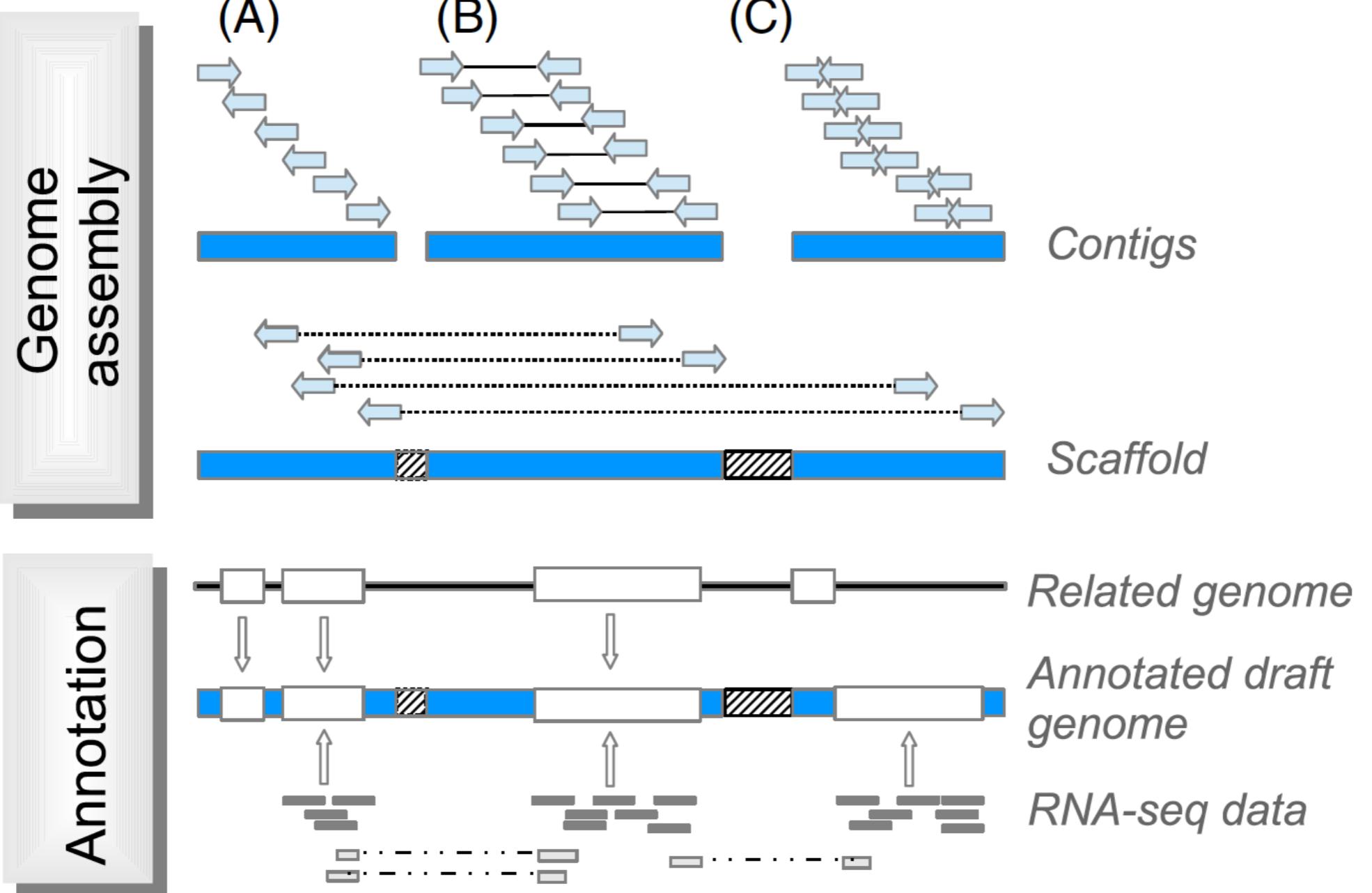
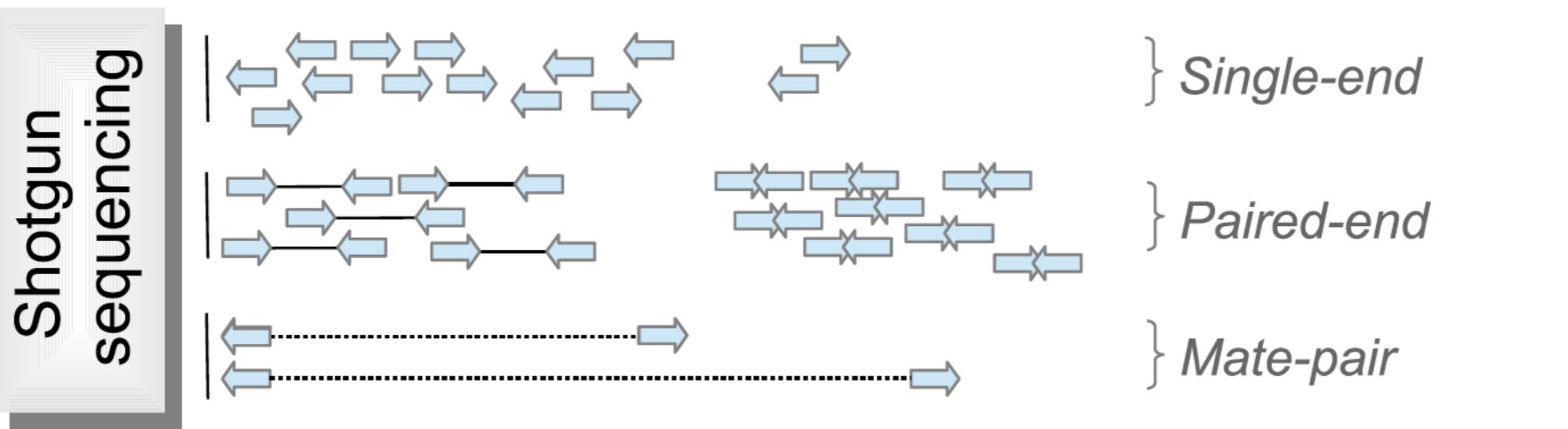
Genome annotation: from sequence to biology

Lincoln Stein

En resumen...

Wet-lab procedures





- Un genoma ensamblado y anotado es un modelo
- Genomas no son estáticos, siempre se pueden mejorar
- Regiones UTR y genes no codificantes son difíciles de predecir
- Genoma humano tiene muchas versiones y parches

Human Genome Assembly Data

Metrics for the current genome assembly

Statistics for the current assembly are available below. Information on tiling path files (TPFs) for the human assembly is available at [TPF Overview](#).

Assembly Statistics for GRCh38.p6 Release date: December 23, 2015 Choose another assembly



- ✓ GRCh38.p6
- GRCh38.p5
- GRCh38.p4
- GRCh38.p3
- GRCh38.p2
- GRCh38.p1
- GRCh38
- GRCh37.p13
- GRCh37.p12
- GRCh37.p11
- GRCh37.p10
- GRCh37.p9
- GRCh37.p8
- GRCh37.p7
- GRCh37.p6
- GRCh37.p5
- GRCh37.p4
- GRCh37.p3
- GRCh37.p2
- GRCh37.p1
- GRCh37
- NCBI36
- NCBI35

[Chromosome Lengths](#)

[Total Lengths](#)

[Ungapped Lengths](#)

[N50s](#)

[Gaps](#)

[Counts](#)

Chromosome lengths are calculated by summing the length of the placed scaffolds and ends.

Primary Assembly

chr	total length	GenBank Accession	RefSeq ID
1	248,956,422	CM000663.2	NC_000001.11
2	242,193,529	CM000664.2	NC_000002.12
3	198,295,559	CM000665.2	NC_000003.12
4	190,214,555	CM000666.2	NC_000004.12
5	181,538,259	CM000667.2	NC_000005.10
6	170,805,979	CM000668.2	NC_000006.12
7	159,345,973	CM000669.2	NC_000007.14
8	145,138,636	CM000670.2	NC_000008.11
9	138,394,717	CM000671.2	NC_000009.12
10	133,797,422	CM000672.2	NC_000010.11
11	135,086,622	CM000673.2	NC_000011.10

Global stats for GRCh38.p6

General Info

Assembly Type	haploid with alt loci
Release Type	patch
Number of Assembly Units	38
Total Bases in Assembly	3,231,297,122
Total Non-N Bases in Assembly	3,069,928,971
Primary Assembly N50	67,794,873
Region Information	
Total number of defined regions	238
Number of Regions with Alternate Loci	178
Number of Regions with Fix Patches	40
Number of Regions with Novel Patches	21
Number of Regions as PAR	4



Identifying bacterial genes and endosymbiont DNA with Glimmer

Arthur L. Delcher^{1,*}, Kirsten A. Bratke², Edwin C. Powers³ and Steven L. Salzberg¹

+ Author Affiliations

*To whom correspondence should be addressed.

Received August 3, 2006.

Revision received December 15, 2006.

Accepted January 14, 2007.

Research

Genome Biology

August 2006, 7:S11

First online: 07 August 2006

Open Access

AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome

Mario Stanke  , Ana Tzvetkova, Burkhard Morgenstern