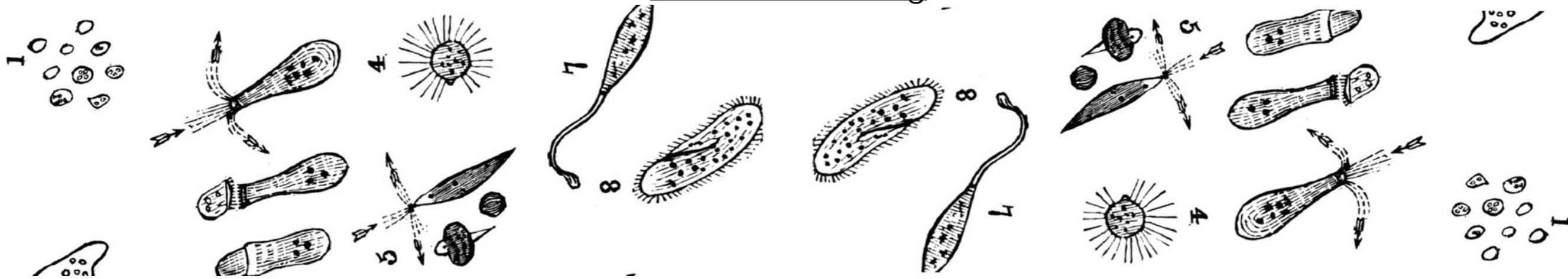


Ensamblaje y anotación

RSG-ISCB
27 de agosto de 2016
Eduardo Castro-Nallar, PhD
www.castrolab.org



Lo que vamos a ver hoy

- Una introducción a los pasos involucrados y cómo funcionan los algoritmos
- Práctica con datos reales

Lo que no va a ocurrir hoy

- No se van a volver expertos
- No vamos a discutir la mejor forma de hacer algo
- No vamos a cubrir todas las estrategias que existen ni las últimas

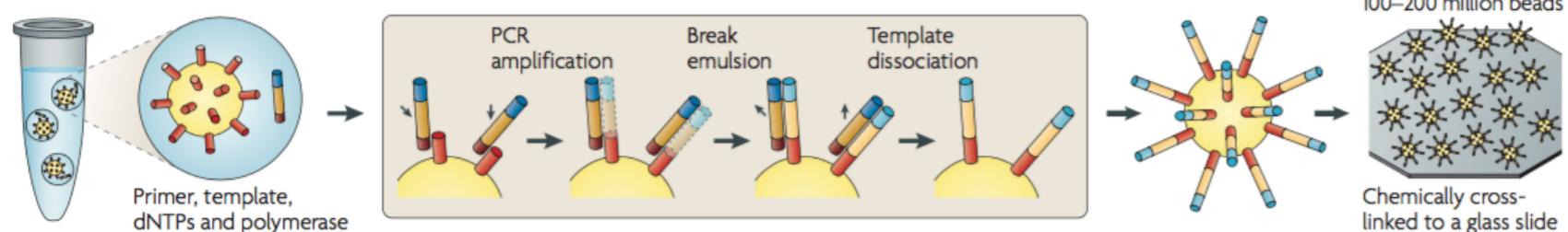
¿Por qué queríamos secuenciar un genoma bacteriano?

- Aislado medioambiental que hace algo interesante
- Aislado patogénico
- Entender evolución de un grupo en particular
- Desarrollar métodos de detección

¿Cómo secuenciamos hoy?

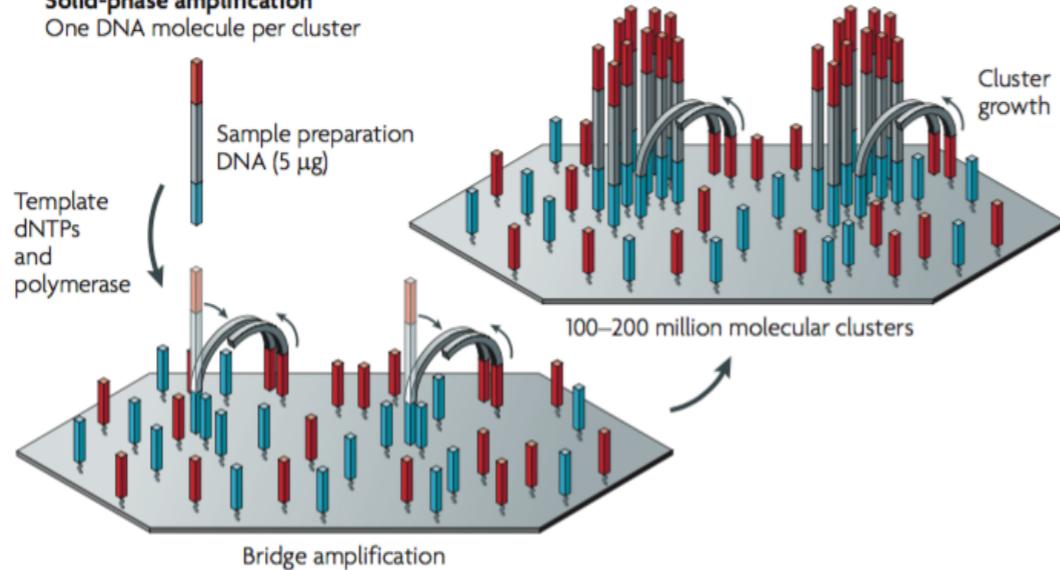
a Roche/454, Life/APG, Polonator Emulsion PCR

One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion

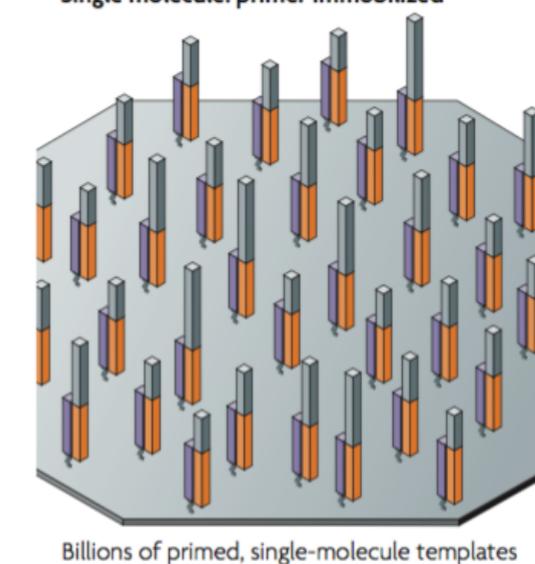


b Illumina/Solexa Solid-phase amplification

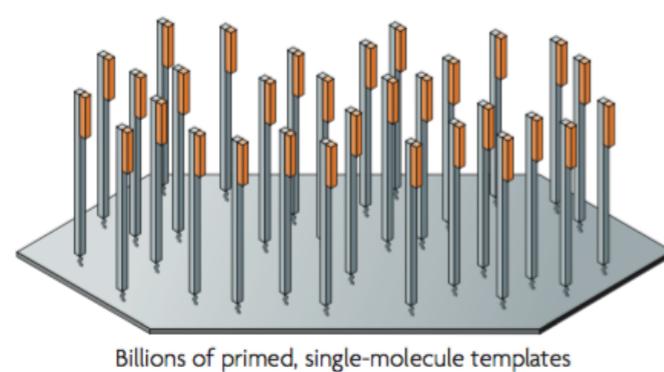
One DNA molecule per cluster



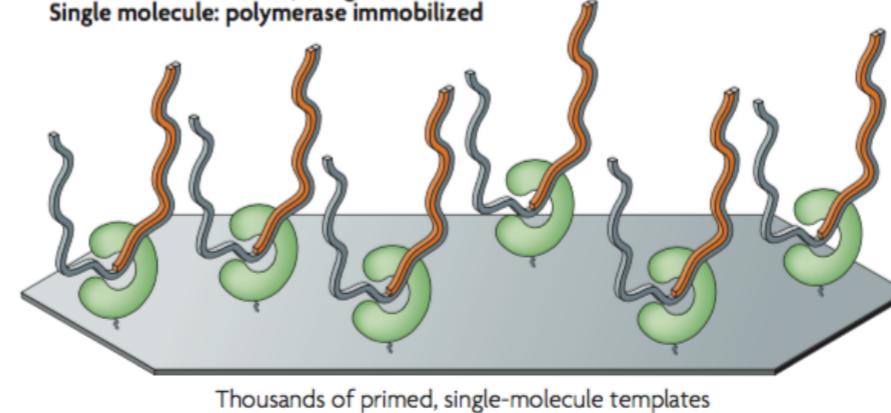
c Helicos BioSciences: one-pass sequencing Single molecule: primer immobilized



d Helicos BioSciences: two-pass sequencing Single molecule: template immobilized

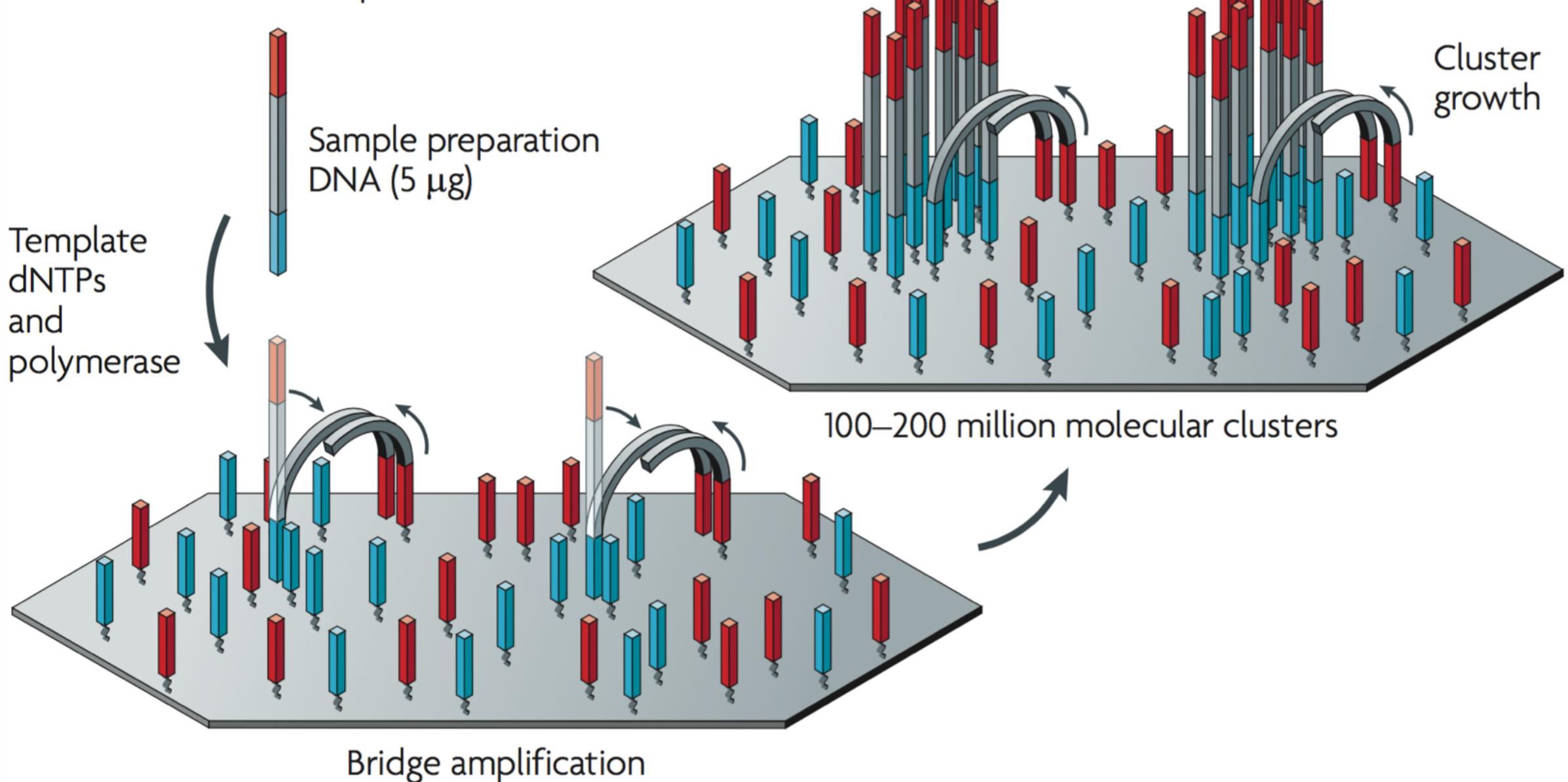


e Pacific Biosciences, Life/Visigen, LI-COR Biosciences Single molecule: polymerase immobilized



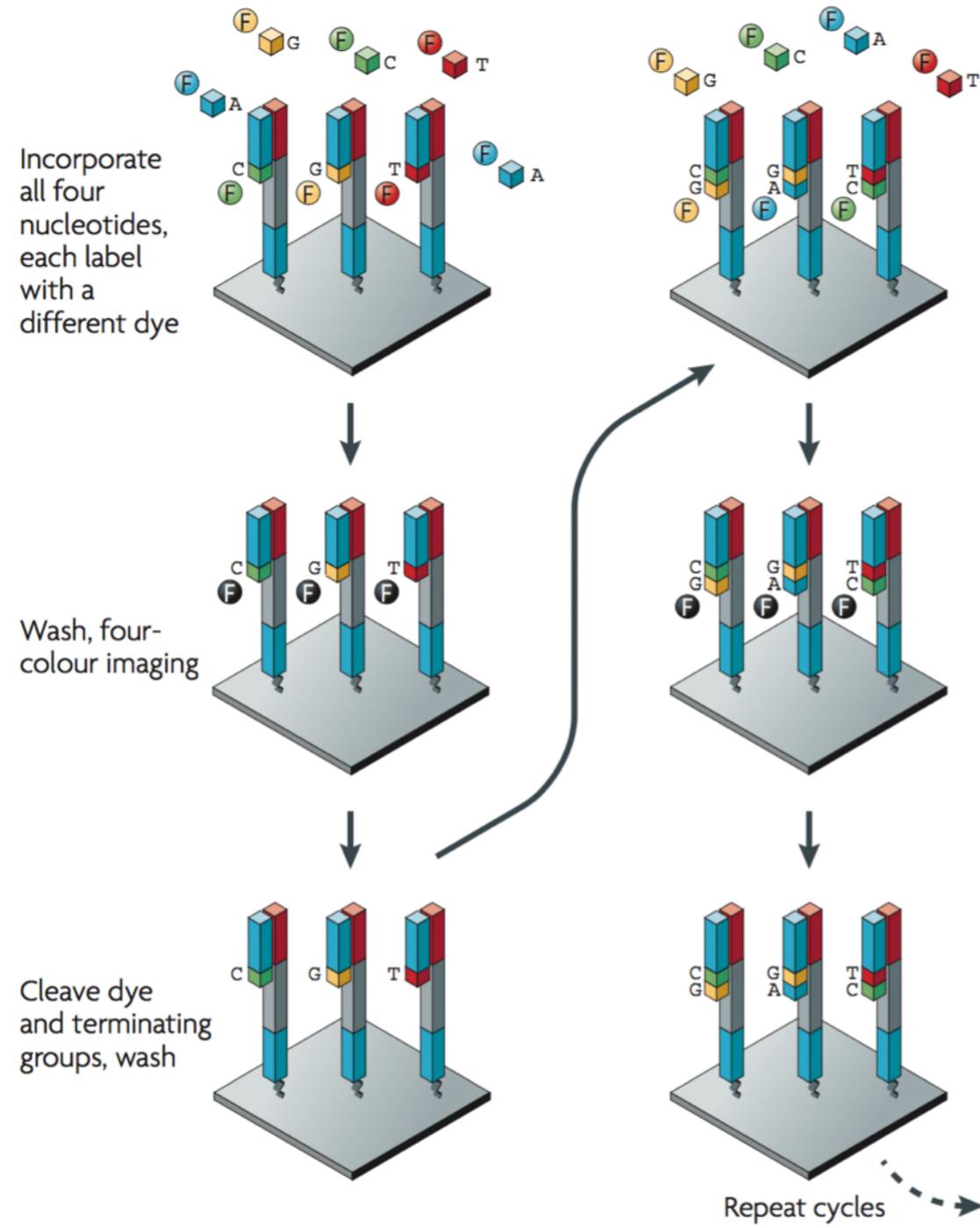
Illumina - secuenciamiento por síntesis

b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster



Illumina - secuenciamiento por síntesis

a Illumina/Solexa — Reversible terminators



Illumina - secuenciamiento por síntesis

- Genera fragmentos de DNA cortos, 75 a 300 bp
- Normalmente se secuencias fragmentos en ambas direcciones, Paired-end
- Baja tasa de error
- En el modo más barato puede costar 100 dólares por muestra

Illumina - secuenciamiento por síntesis

	 MinSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million [†]	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	<ul style="list-style-type: none"> • MinSeq System for low-throughput targeted DNA and RNA sequencing 	<ul style="list-style-type: none"> • MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics 	<ul style="list-style-type: none"> • NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays 	<ul style="list-style-type: none"> • HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics 	<ul style="list-style-type: none"> • HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

Tipos de “reads” o lecturas

Single Read



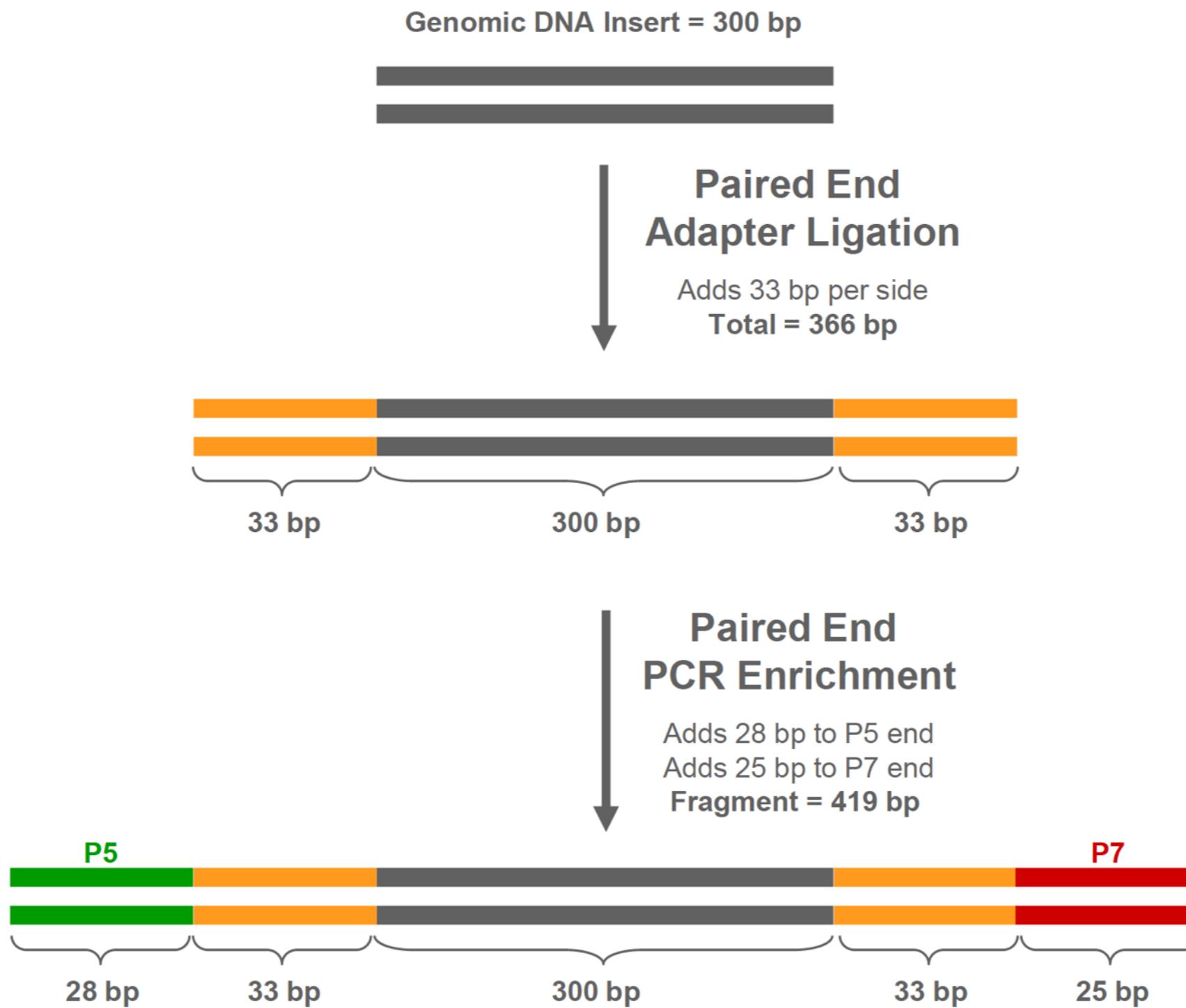
Paired Read



Indexed Paired Read

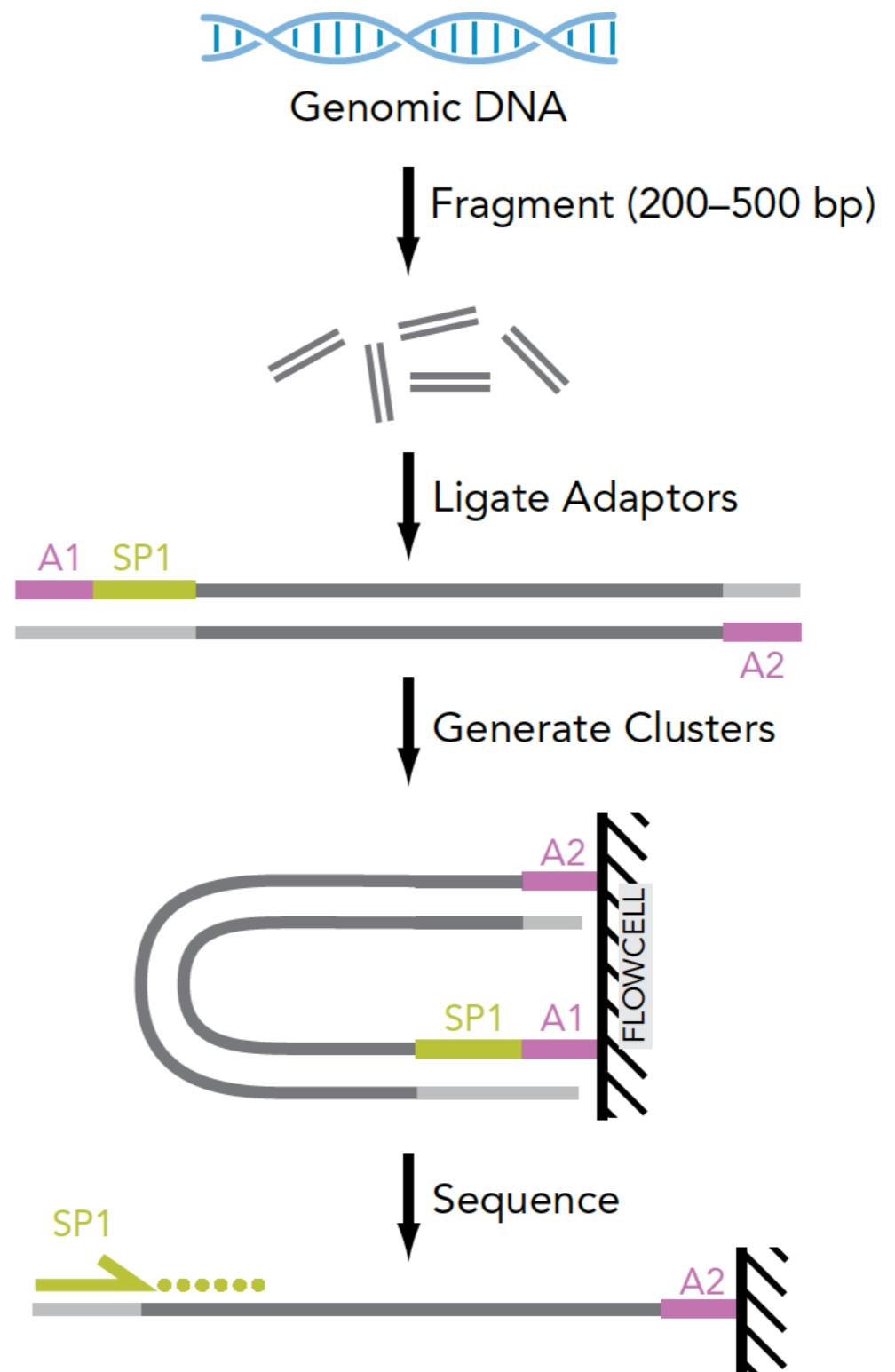


Inserto y secuencia útil



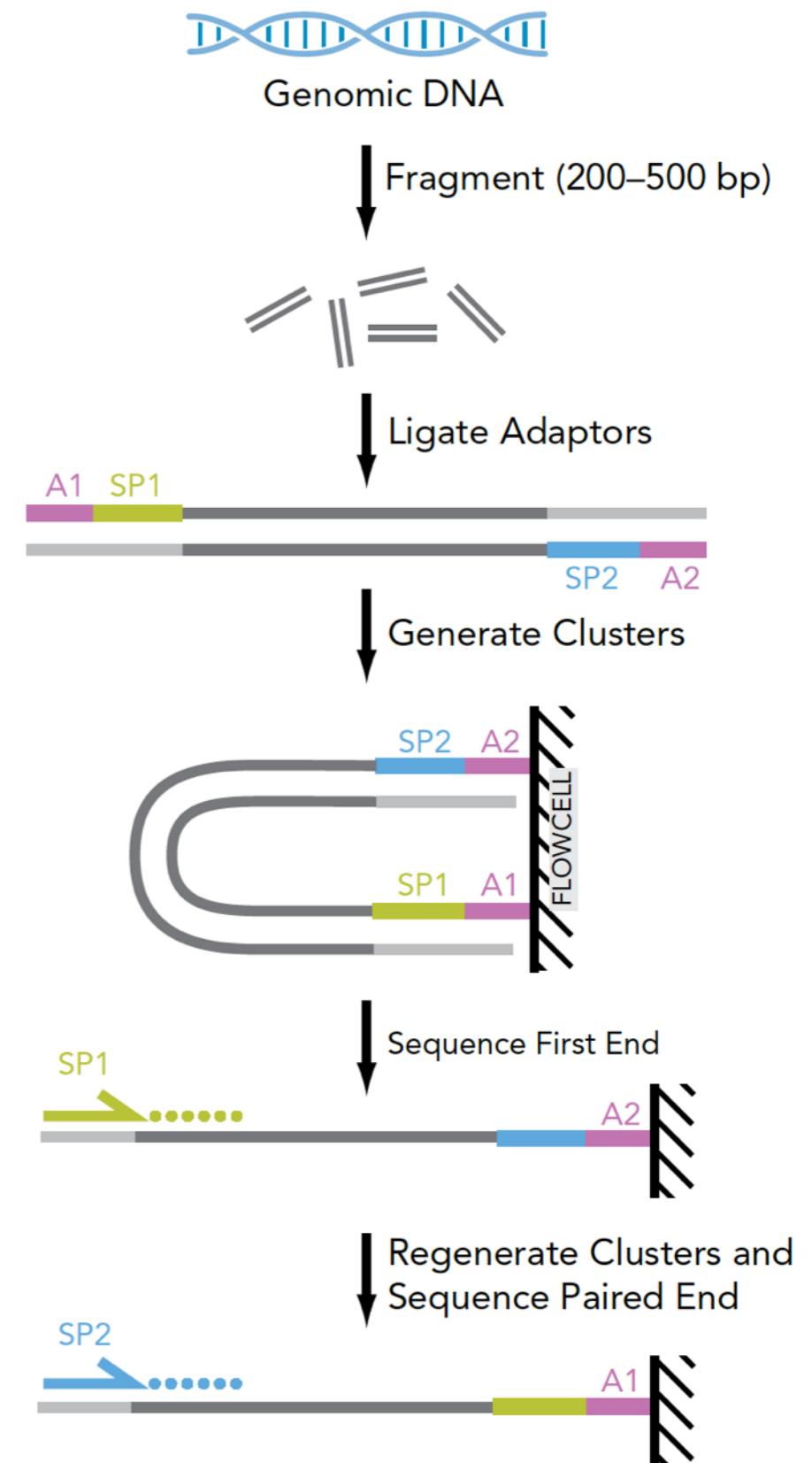
Single-end

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



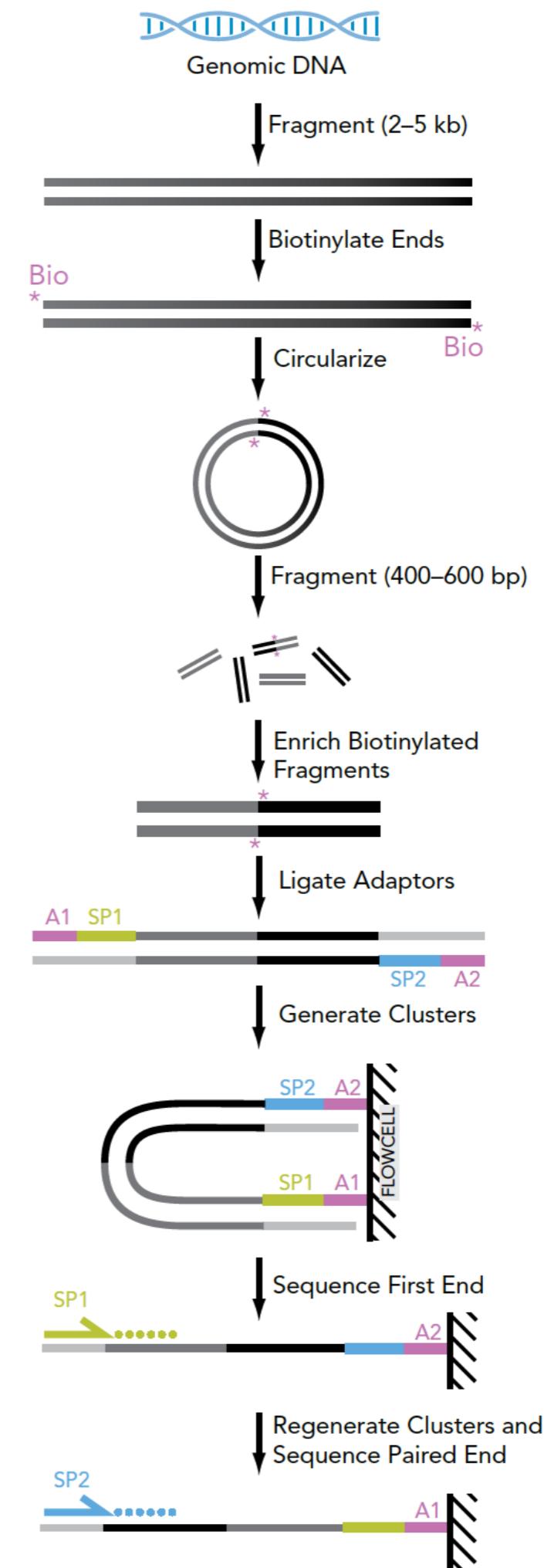
Paired-end

- 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

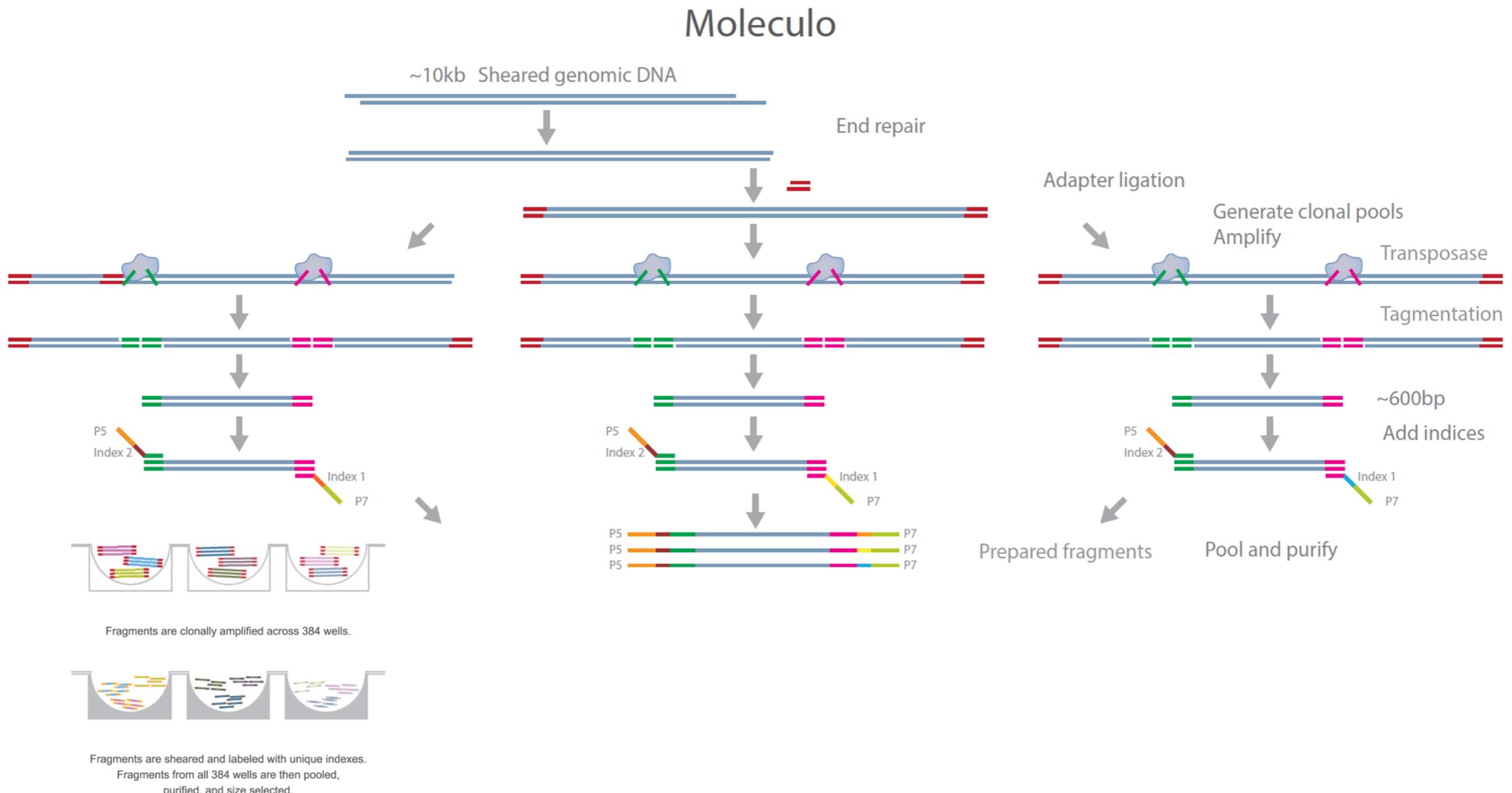


Mate-pairs

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



Artificial long reads



genomas complejos, “phasing” de alelos, finalizar genomas

Resultado de la secuenciación



Métodos de ensamblaje

- Overlay-layout-consensus
- Encontrar regiones que se superponen entre reads
- Escala con el cuadrado del número de reads → muy lento para Illumina/Ion Torrent
- Finalmente se escoge la secuencia consenso más probable

Métodos de ensamblaje

X: CTCGGCCCTAGG
||| | | | |
Y: GGCTCTAGGCC

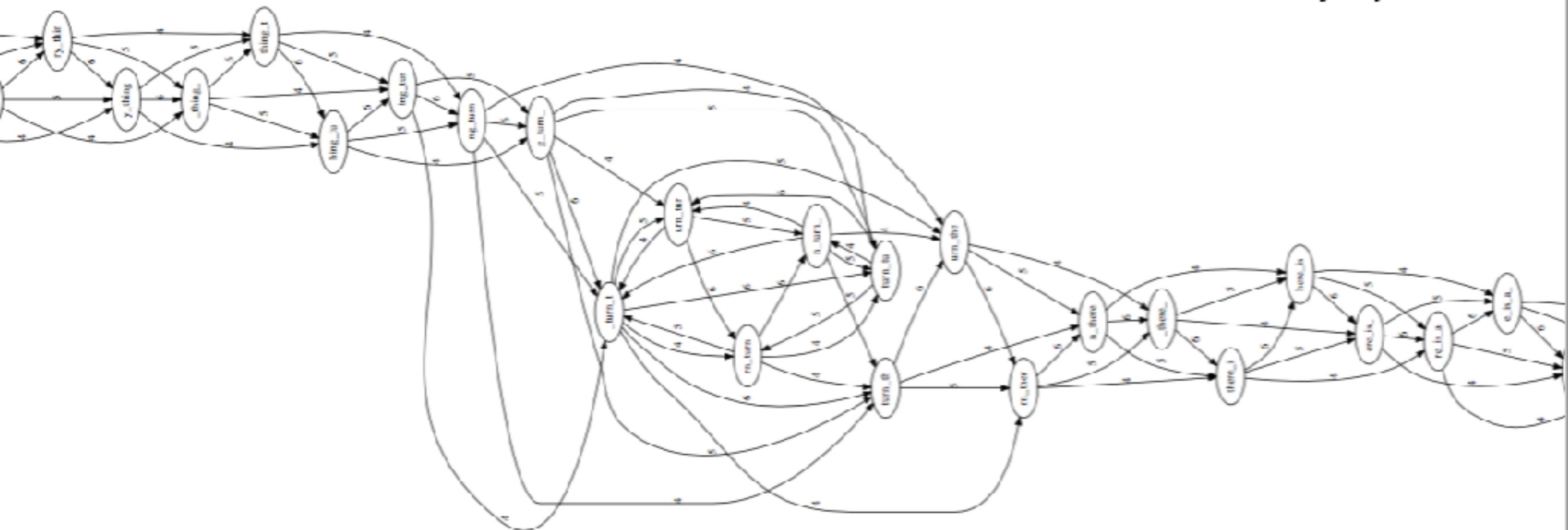
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

↓ ↓ ↓ ↓

TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote



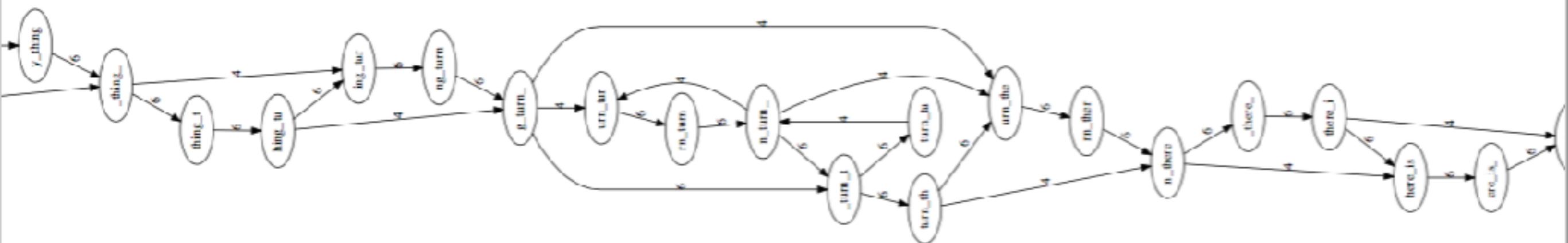
Métodos de ensamblaje

X: CTCGGCCCTAGG
||| | | | |
Y: GGCTCTAGGCC

TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGTGATGGCGTAA CTA
↓
TAGATTACACAGATTACTGACTTGTGATGGCGTAA CTA

Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote



Métodos de ensamblaje

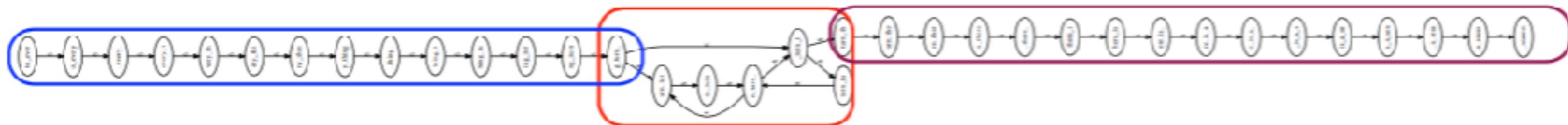
X: CTCGGCCCTAGG
||| |||||
Y: GGCTCTAGGCC

TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

↓ ↓ ↓ ↓
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote



Métodos de ensamblaje - Alternativa

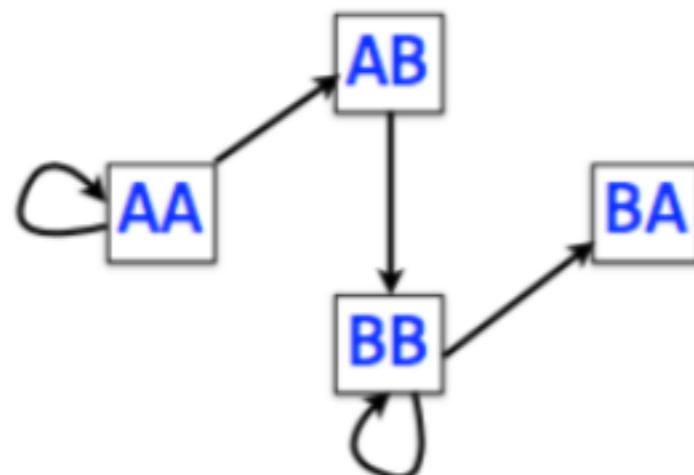
- De Bruijn graphs
- Trabaja sobre kmers no reads
- Escala linealmente con el número de reads (más rápido)
- Requiere memoria RAM proporcional al tamaño del genoma no de la librería
- Downside: no resuelve repeticiones muy bien

Métodos de ensamblaje - Alternativa

Take each length-3 input string and split it into two overlapping substrings of length 2. Call these the *left* and *right* 2-mers.



Let 2-mers be nodes in a new graph. Draw a directed edge from each left 2-mer to corresponding right 2-mer:



Each edge in this graph corresponds to a length-3 input string

Una cosa más antes de comenzar

Conceptos

- Read
- Contig
- Repeat
- kmer
- Sequence quality
- Scaffolding
- Coverage

Manos a la obra!

- Vamos a seleccionar datos crudos, recién salidos del secuenciador para ensamblarlos usando SPAdes (De Bruijn graphs)

Muchos sabores y colores

- MIRA, Soap-denovo, Velvet, Abyss y un largo etcétera
- A través de simulaciones y uso de datos reales se pueden evaluar las ventajas relativas de los ensambladores

Muchos sabores y colores

- MIRA, Soap-denovo, Velvet, Abyss y un largo etcétera
- A través de simulaciones y uso de datos reales se pueden evaluar las ventajas relativas de los ensambladores

Bioinformatics. 2013 Jul 15;29(14):1718-25. doi: 10.1093/bioinformatics/btt273. Epub 2013 May 10.

GAGE-B: an evaluation of genome assemblers for bacterial organisms.

Magoc T¹, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL.

Author information

¹Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21025, USA.

<http://www.ncbi.nlm.nih.gov/pubmed/23665771>

Genome Res. 2011 Dec;21(12):2224-41. doi: 10.1101/gr.126599.111. Epub 2011 Sep 16.

Assemblathon 1: a competitive assessment of de novo short read assembly methods.

Earl D¹, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, Yang SP, Wu W, Chou WC, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P, Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev D, Ramirez-Gonzalez R, Leggett R, MacLean D, Xia F, Luo R, Li Z, Xie Y, Liu B, Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Yin S, Sharpe T, Hall G, Kersey PJ, Durbin R, Jackman SD, Chapman JA, Huang X, DeRisi JL, Caccamo M, Li Y, Jaffe DB, Green RE, Haussler D, Korf I, Paten B.

<http://www.ncbi.nlm.nih.gov/pubmed/21926179>

Manos a la obra!

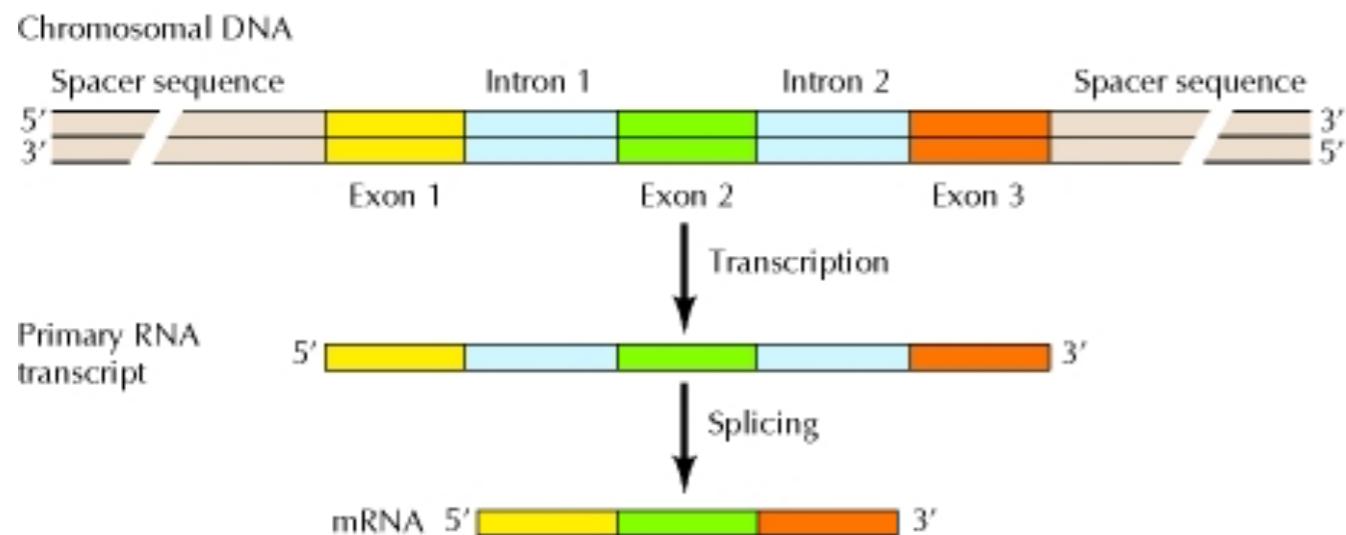
- Datos para trabajar:
- Opción I, *Pseudomonas fluorescens*
- Opción II, *Escherichia coli*
- Disco duro o vínculo

Obtener las reads

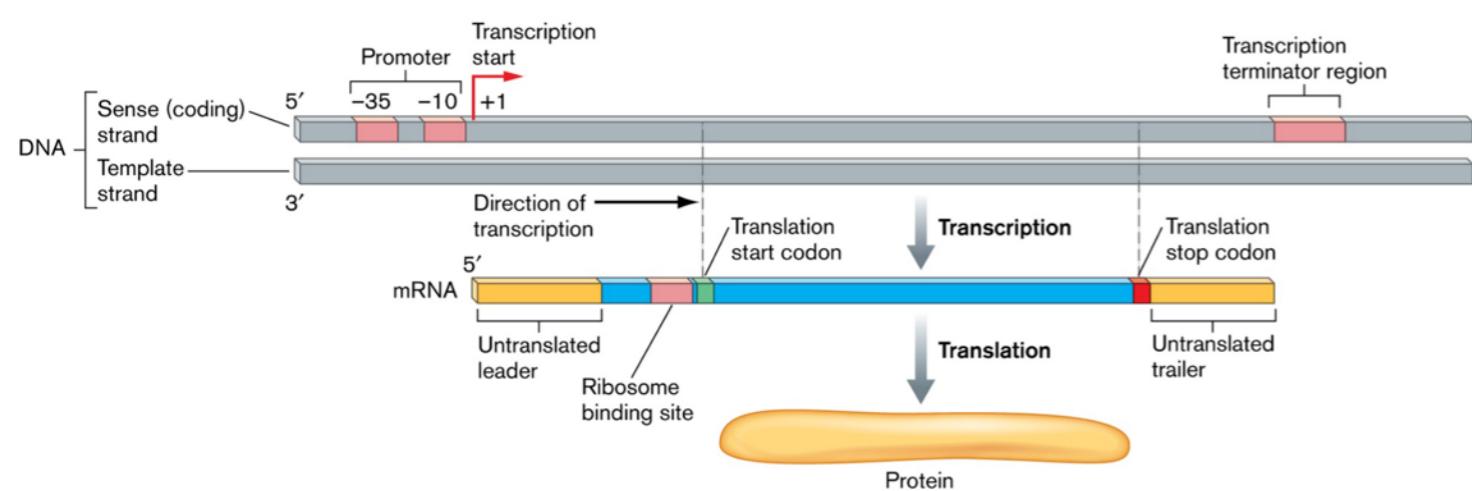
- Opción I, *Pseudomonas fluorescens*
- <http://tinyurl.com/jn7nukh>
- Opción II, *Escherichia coli*
- <http://tinyurl.com/zu32w3j>

Parte II:Predicción de genes

Estructura de genes

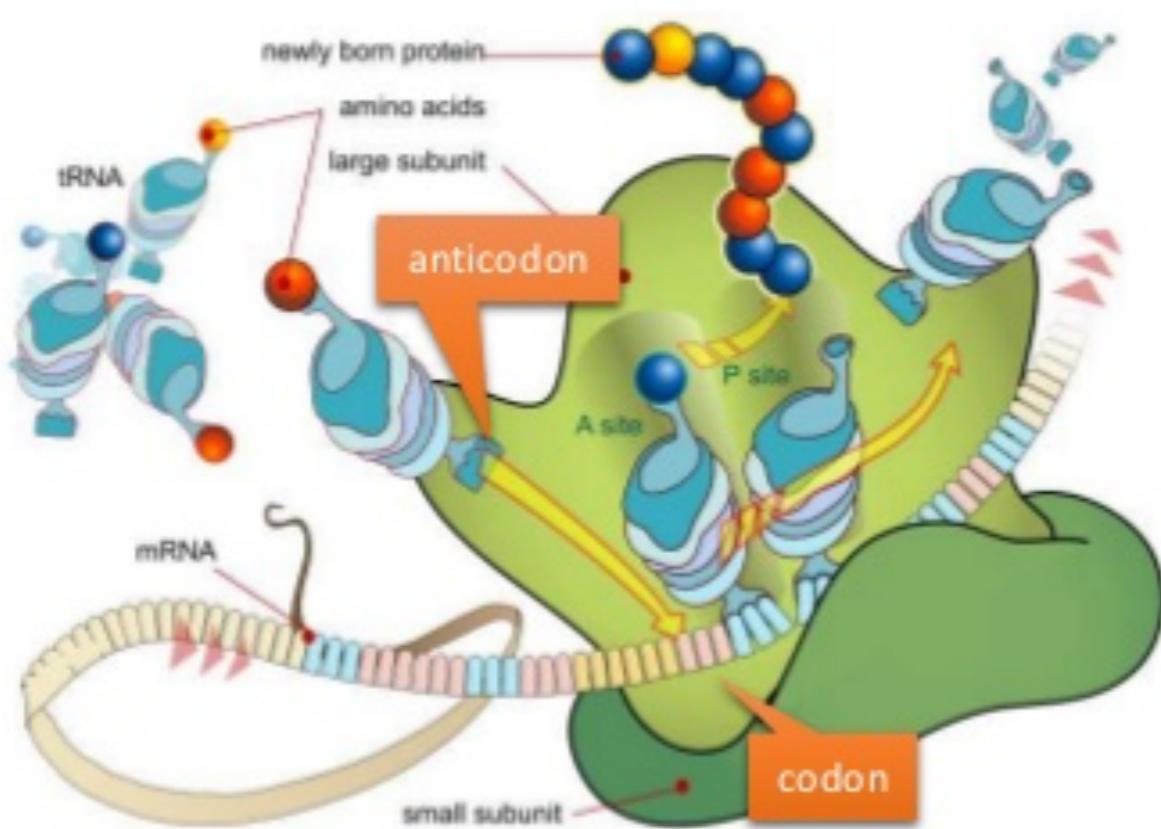


- ADN eucarionte envuelto en histonas, resulta en patrones repetitivos. Promotores están cerca de estos sitios



- 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

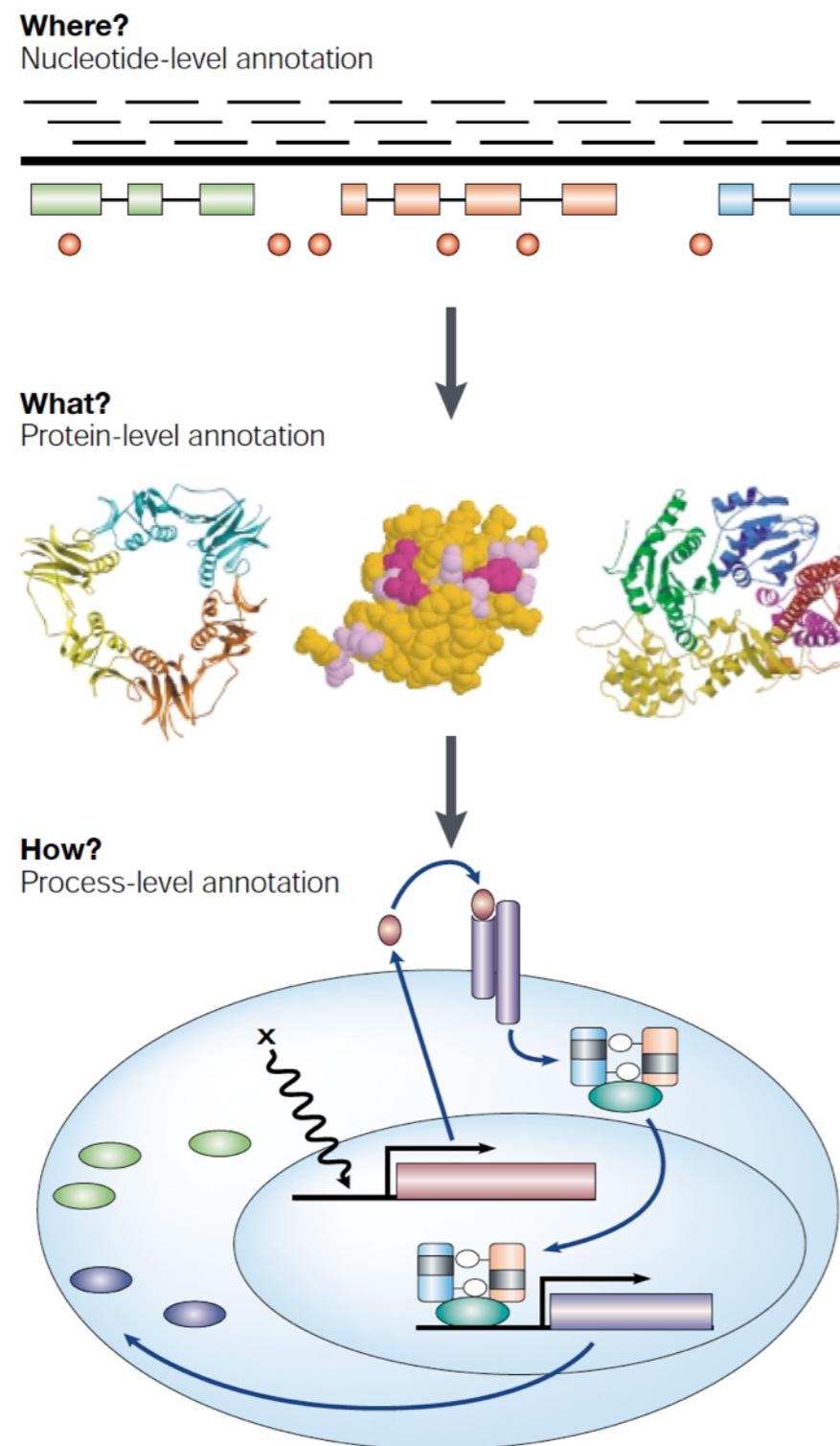
¿Qué tan bien funciona?

- **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



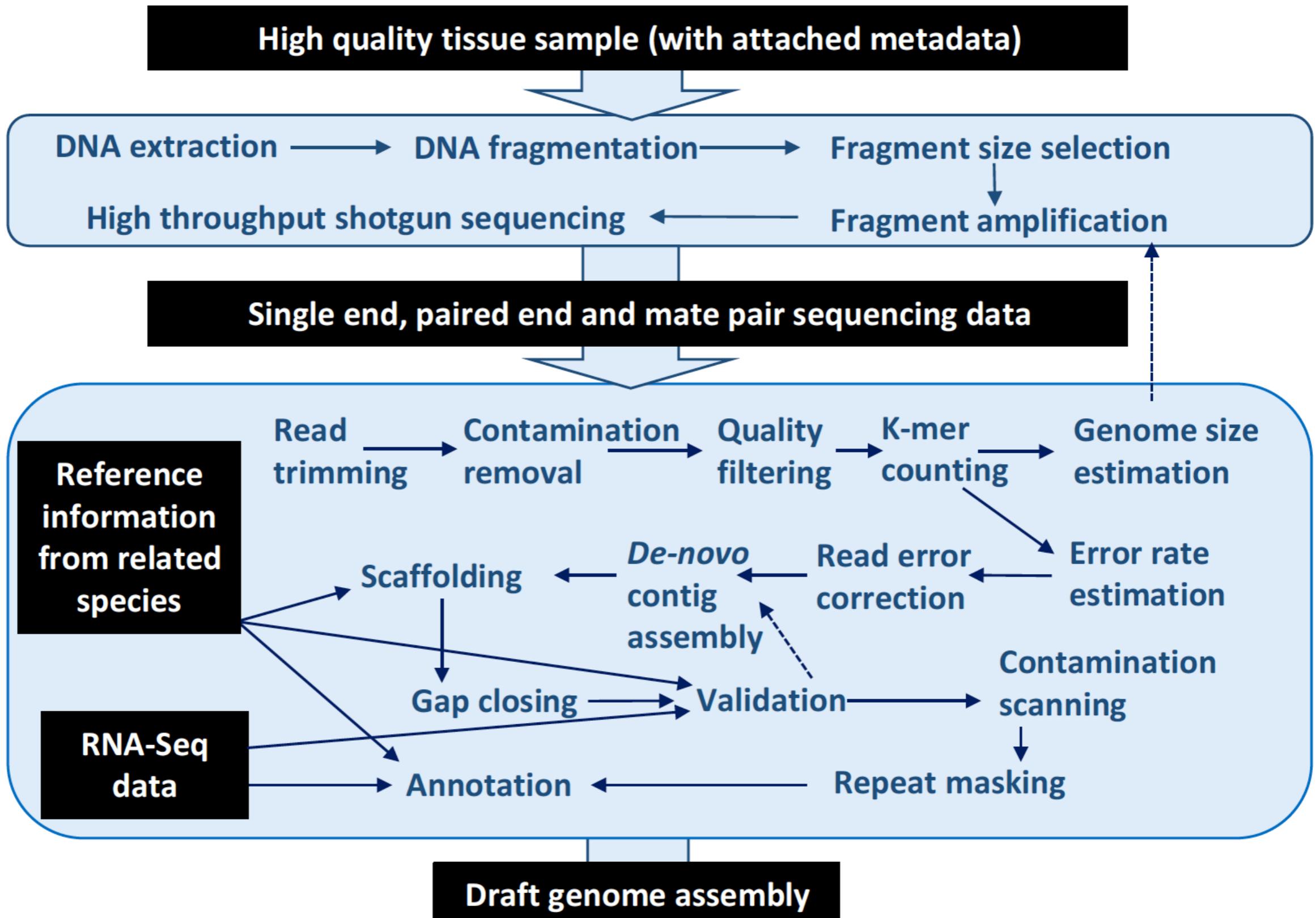
Nature Reviews Genetics **2**, 493-503 (July 2001) | doi:10.1038/35080529

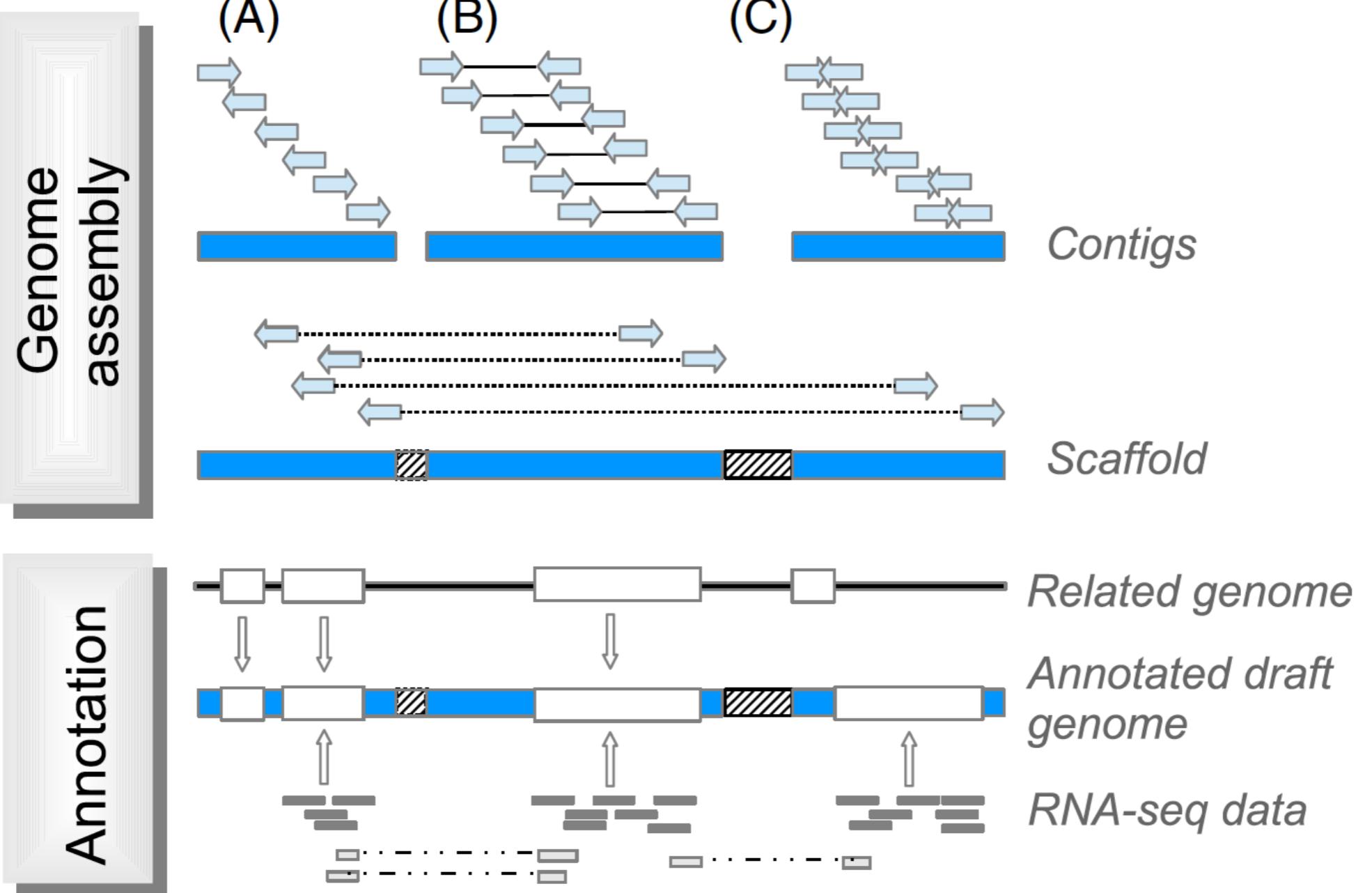
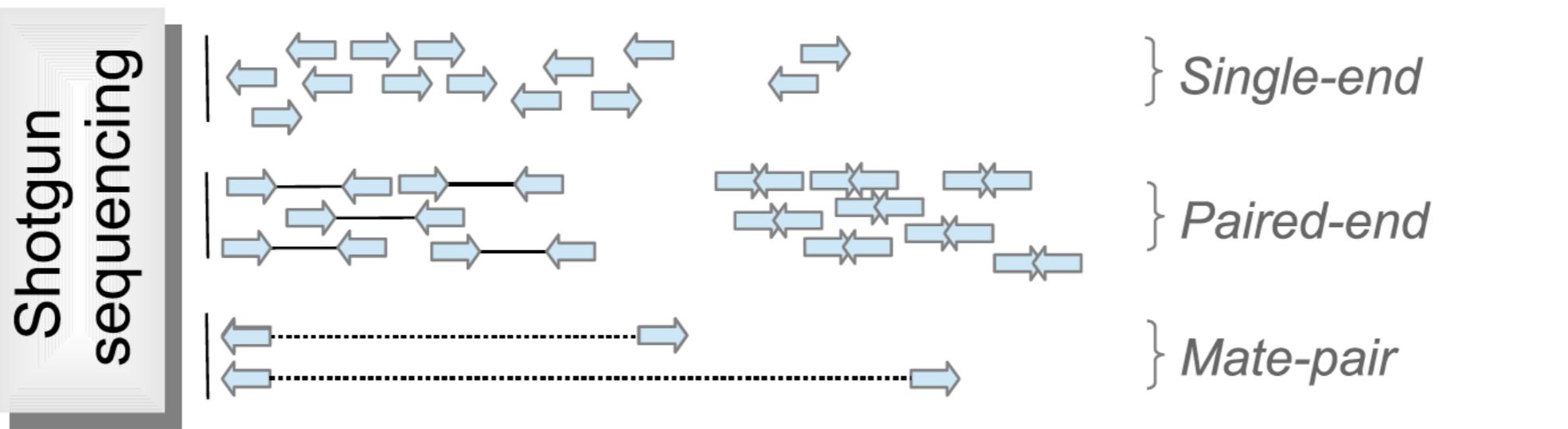
Genome annotation: from sequence to biology

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