“Introducción al ensamble de genomas"

Michael Torres

LANGEBIO - CINVESTAV

alan.torres@cinvestav.mx

SERVICIOS GENÓMICOS

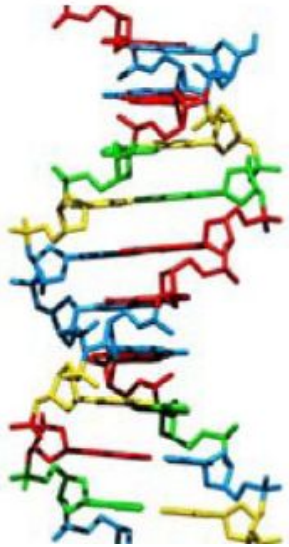
LANGEBIO

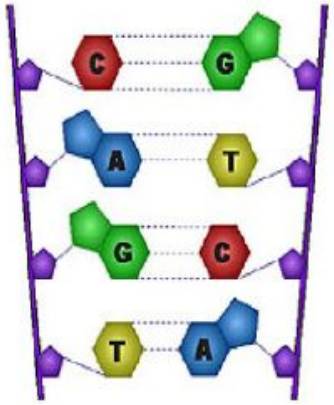
Contenido

1. **Tecnologías Genómicas: Secuenciamiento, fundamentos y aplicaciones de NGS.**
2. **Algoritmos de ensamble.**
3. **Control de calidad de ensamble.**
4. **Hands-on.**

Secuenciamiento de ADN

- Es un proceso que nos permite determinar el orden exacto de las bases A, T, C y G en un fragmento de ADN.





...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT...

Secuenciamiento de ADN: Historia

Secuenciamiento de ADN: Historia

A brief history of genomics

1971	Wu & Taylor determine the first ever DNA sequence (all 12 bp of it!)
1977	Sanger *et al.* sequence the first ever (DNA-based) virus genome - 5,375 bp
1995	First complete bacterial genome sequence (*Haemophilus influenzae*) - 1.83 Mb
1996	First complete eukaryotic genome (*Saccharomyces cerevisiae*) - 12 Mb
1998	First animal genome (*Caenorhabditis elegans*) - 100 Mb

It took 18 years before we knew the structure of DNA before anyone could sequence it. First DNA sequence was from the end of a bacteriophage lambda virus (written in a 20 page paper). First genome was actually an RNA viral genome determined in 1975 by Fiers et al. The 1980's and 1990's saw the start of widespread DNA sequencing for genes of interest in species of interest. Moving to eukaryotic genome sequencing means determining multiple chromosomes, and tackling bigger repeats (more assembly problems).

Secuenciamiento de ADN: Historia

- Secuenciamiento basado en gel de electroforésis
  - 300~500 bases de una hebra de ADN Métodos de terminación de la cadena: Método de Sanger (Sanger et al, 1975 Premio Nobel en Química)
- El primer secuenciamiento de ADN completo de un genoma
  - Bacteriofago phi X 174 de 5,375 nucleótidos (Sanger et al, 1977)
- **Primeros proyectos Shotgun**
  - 6,569 bp DNA mitocondrial humano (Anderson et al, 1981)
  - 48,502 bp secuencia nucleotídica del Bacteriofago λ (Sanger et al, 1982)
  - 172,282 bp secuencia del virus EpsteinBarr (Baer et al, 1984)

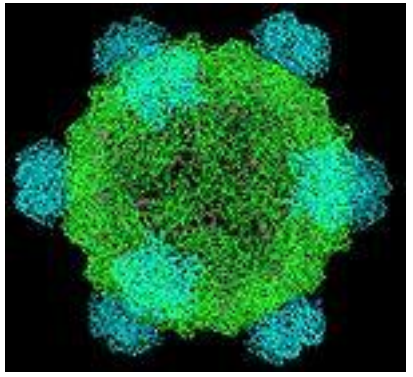
Technical Note: Sequencing

illumina

Using a PhiX Control for HiSeq® Sequencing Runs

A low-concentration spike-in of Illumina PhiX Control v3 provides quality and calibration controls.

http://es.wikipedia.org/wiki/Phi-X174

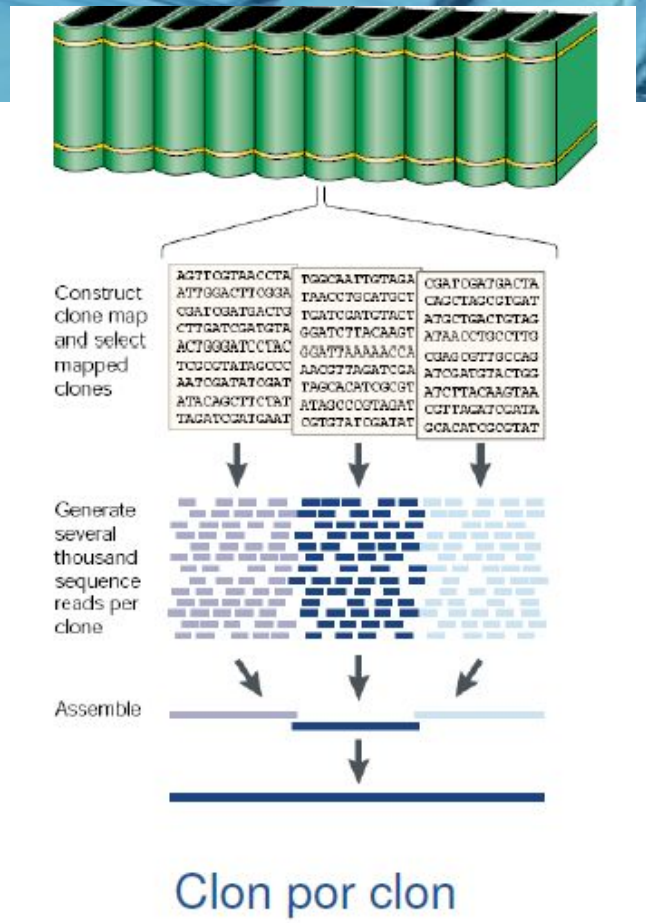


Secuenciamiento de ADN: Tipos de secuenciamiento.

- Secuenciamiento jerárquico:
  Construcción de bibliotecas BAC's.
  Mapa físico requerido.
- Secuenciamiento "Shotgun":
  Fraccionamiento y secuenciamiento directo de ADN.
  No hay necesidad de bibliotecas BAC's

Secuenciamiento de ADN: Tipos de secuenciamiento.

- Secuenciamiento jerárquico:



Whole genome

BAC library ¿?

Sequencing

Assembly

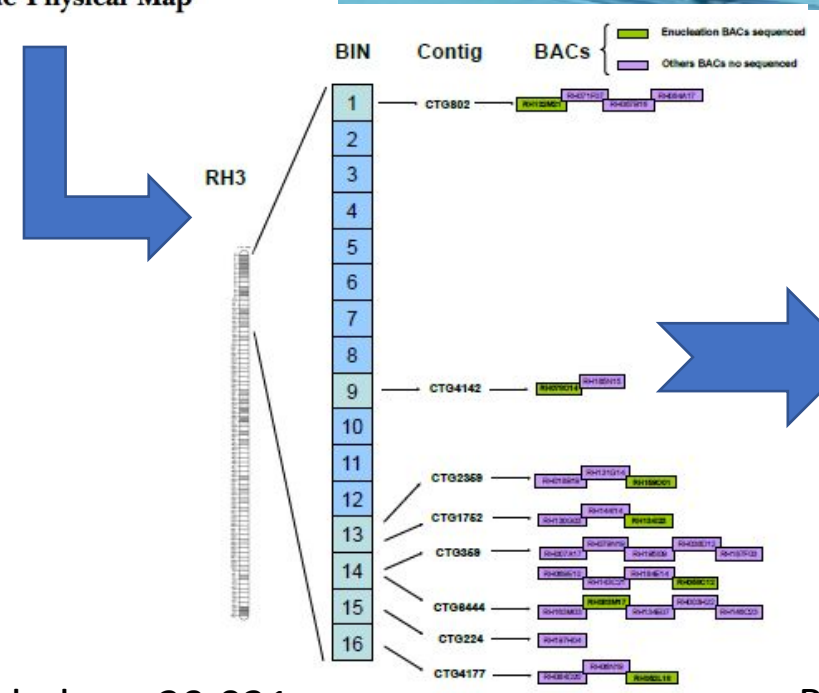
BAC: Bacterial Artificial Chromosome(fragmento de ADN de 150,000 a 200,000 nt)

Secuenciamiento de ADN: Tipos de secuenciamiento

- Potato Genome Sequencing Project:

Construction of a 10,000-Marker Ultradense Genetic Recombination Map of Potato: Providing a Framework for Accelerated Gene Isolation and a Genomewide Physical Map

Año 2006: BAC library of 78,000 clones

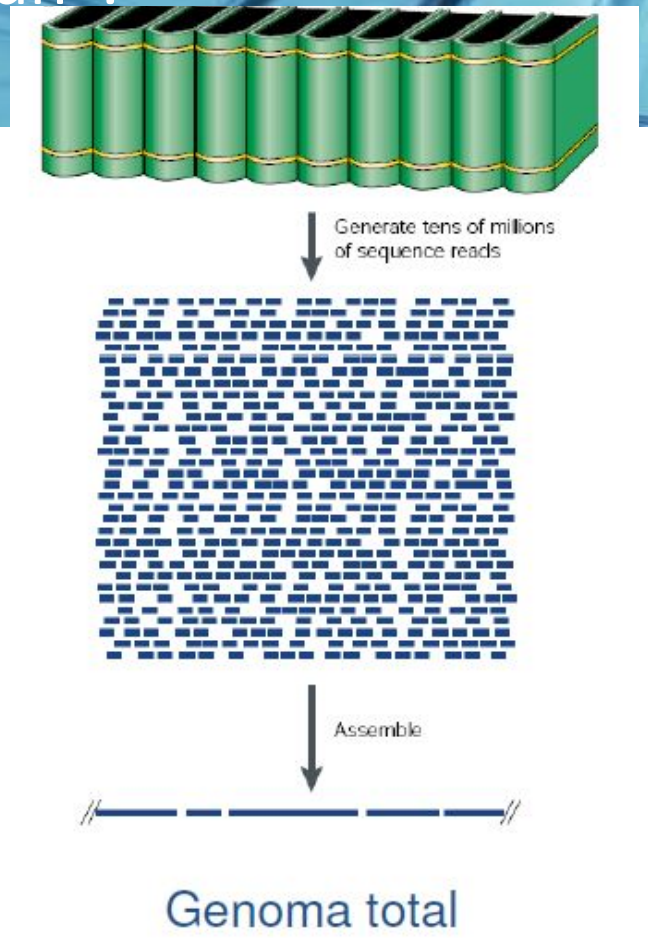


727mb ensamblados y 39,031 genes codificantes. Illumina MP y 454 ensamblaje híbrido.

Published: 10 July 2011
doi:10.1038/nature10158

Secuenciamiento de ADN: Tipos de secuenciamiento.

- Secuenciamiento “Shotgun”:



Whole genome

Sequencing

Assembly

Secuenciamiento de ADN: Tipos de secuenciamiento.

VENTAJAS	DESVENTAJAS
Jerárquico: Genoma final casi completo y ordenado.	-Generación de bibliotecas de cromosomas artificiales (BAC's library). -Generación de mapa físico para el correcto reordenamiento de BAC's.
Shotgun: Genoma borrador generado en mucho menos tiempo que el jerárquico, además este método es de muy bajo costo en comparación al secuenciamiento jerárquico.	-Secuencias generadas al azar requieren de un gran poder computacional para el ensamblaje de contigs. -Por lo general debe complementarse con otros métodos de secuenciamiento (ensamblaje híbrido o de mapeo baseo en EST's)

Plataformas de secuenciación masiva

Diferentes plataformas y tecnologías han sido usadas para secuenciar el ADN. Estas difieren en:

- **Longitud de las reads.**
- **Frecuencia y tipos de errores en su secuencia.**
- **Rendimiento: cantidad de datos de secuencia generados.**
- **Precio.**

Plataformas de secuenciación masiva

- **Illumina**
- **Pacific Biosciences (PacBio)**
- **Ion Torrent**
- **454**
- **Oxford Nanopore**

Secuenciamiento de ADN: Tecnologías y "output" del "shotgun"

High-end sequencing- Platform†	Sequencing chemistry	Read lengths/ through put	Run time	Template prep	Application
Roche 454 -Titanium FLX	Pyrosequencing	400 bp 400 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa -HiSeq 2000	Reversible terminator chemistry	2×100bp 600 GB/ run (dual cell)	11.5 days	Solid-phase	Human WGS, exome seq, RNA-seq, Methylation
ABI/LifeTechnology-SOLiD 5550XL	Sequencing by ligation	2×60bp 15 GB/day	8 days	Emulsion PCR	Human WGS, exome seq, RNA-seq, Methylation
HelicosBiotechnologies	Reversible Terminator chemistry	25-55 bp 28 GB/run (avg)	>1 GB/hour	Single molecule	Human WGS, exome seq, RNA-seq, Methylation
Roche 454- GS Junior	Pyrosequencing	400 bp 50 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa- MiSeq	Reversible terminator chemistry	2×150bp 1.0-1.4 Gb	26 hours	Solid-phase	Microbial discovery, Exome seq, Targeted capture
ABI/ Lifetechnology- Iontorrent	H+ Ion sensitive transistor	320 Mb/run	8 hours*	Emulsion PCR	Microbial discovery, Exome seq, Targeted capture

*Sample preparation – 6 hours, sequencing time – 2 hours, †Data shown here represent the highest figures currently available on the company website and is highly likely to change by the time this article is published

Tecnologías genómicas

- Plataformas:

	454 FLX +	HiSeq 2000	PacBio *RS*	Ion Torrent 316
Company	Roche (USA)	Illumina (USA)	Pacific Biosciences (USA)	Life Technologies (USA)
Sequencing method	Synthesis (pyrosequencing)	Synthesis (cyclic reversible terminator)	Realtime sequencing	Synthesis (H^+ detection) on the chip
Amplification	emPCR	BridgePCR	None	emPCR
Run time	23 h	11 days (dual flow cell)	0.5 - 2 h	2 h
Reads Mb/run	1,000	540,000-600,000	5 - 10	>100
Reagent cost/run	$6,200	~$20,000	$110 - 900	$750
Reagent cost/Mb	$7	>$0.04	$11 - 180	<$7.5
Read length	500-1,000 (mode 700)	2*100 (paired-end reads)	860 - 1,100	>200
Primary errors	Indel	Substitution	CG deletion	Indel
Pros	Long read length	Highest throughput and lowest cost per Mb	Longest read length, no amplification error	Low cost per sample
Cons	High capital cost and high cost per Mb	High capital cost and high computation needs	Error rates, comparatively small outputs, high cost per Mb	High cost per Mb

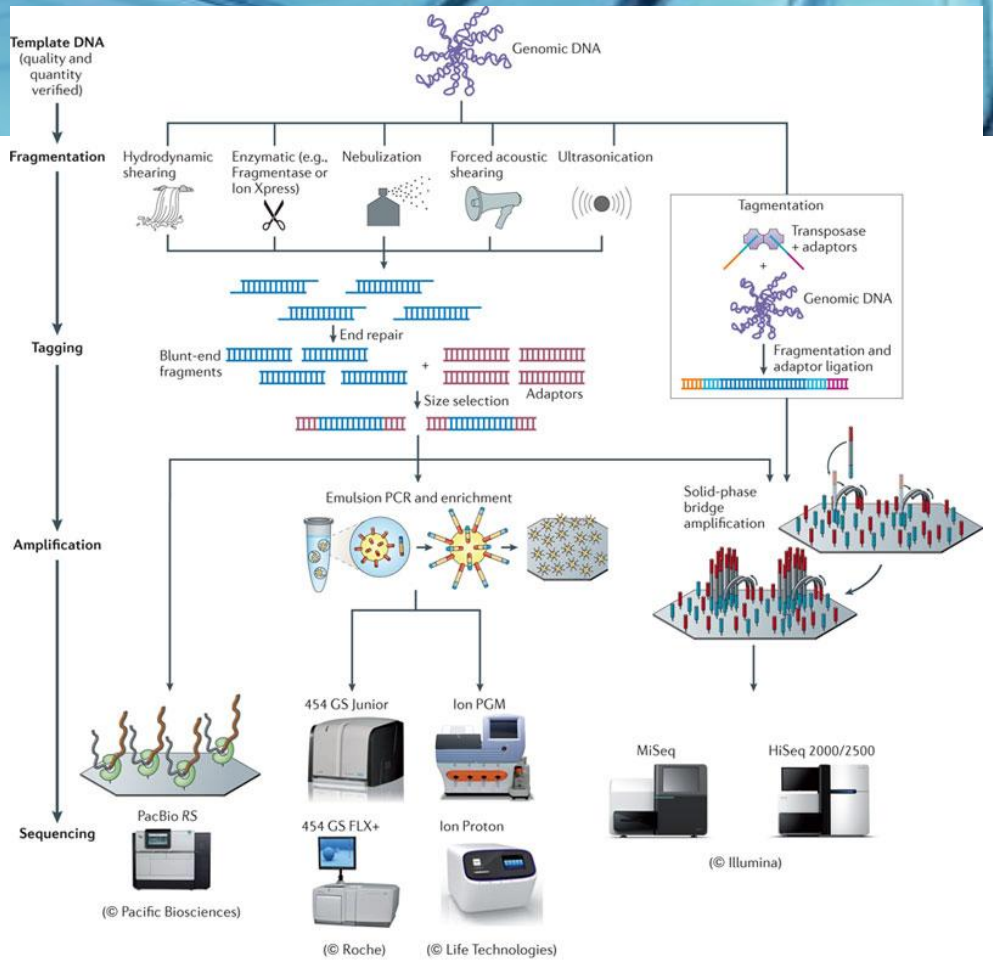
Tecnologías genómicas

- Plataformas:

Sequencing instruments	Method	Read length (bp)	Error (%); primary error	Output (million reads/ run)	Remarks
454	Emulsion PCR and Pyrosequencing	700	1; indels	1	Longer read but comparatively high cost/Mb data and high error rate compared to MiSeq®/HiSeq™. Ok for marker gene-based metagenomic sequencing
Ion PGM	Emulsion PCR and semiconductor Sequencing	400	≈1; indels	0.4–0.5 (314 chip)	Faster and reasonably longer read than MiSeq/HiSeq but low throughput than to MiSeq/HiSeq. Also too much hands-on time and high error rate associated with indels, OK for marker gene-based metagenomic sequencing but strict size selection is a problem
				2–3 (316 chip)	Same as 314 chip
				4–5.5 (318 chip)	Same as 314 chip
Ion proton	Emulsion PCR and semiconductor sequencing	200	≈1, indels	60–80 (PI chip)	Higher throughput than Ion PGM but lower throughput than HiSeq, shorter reads than Ion PGM/MiSeq/HiSeq/PacBio
MiSeq	Bridge amplification and reversible dye terminator sequencing	300+300	≈0.1; substitution	25	High throughput, low error rate, and comparable read length in paired-end sequencing make this platform best among all for marker gene-based metagenomic sequencing
HiSeq 2500	Bridge amplification and reversible dye terminator sequencing	250+250	≈0.1; substitution	600 (rapid run v2 kit)	Same as MiSeq but with much higher throughput and slightly shorter read length. Good for shotgun Meta-omic sequencing
		125+125		4000	
PacBio® RS II	Single molecule real-time (SMRT) sequencing	8500 bp	≈11; indels	0.05 (per smart cell per run)	No amplification step, longest read among all sequencing platforms (up to 20 kb) but with low throughput and very high error rate in a single pass. Good for shotgun sequence assembly in meta-omic

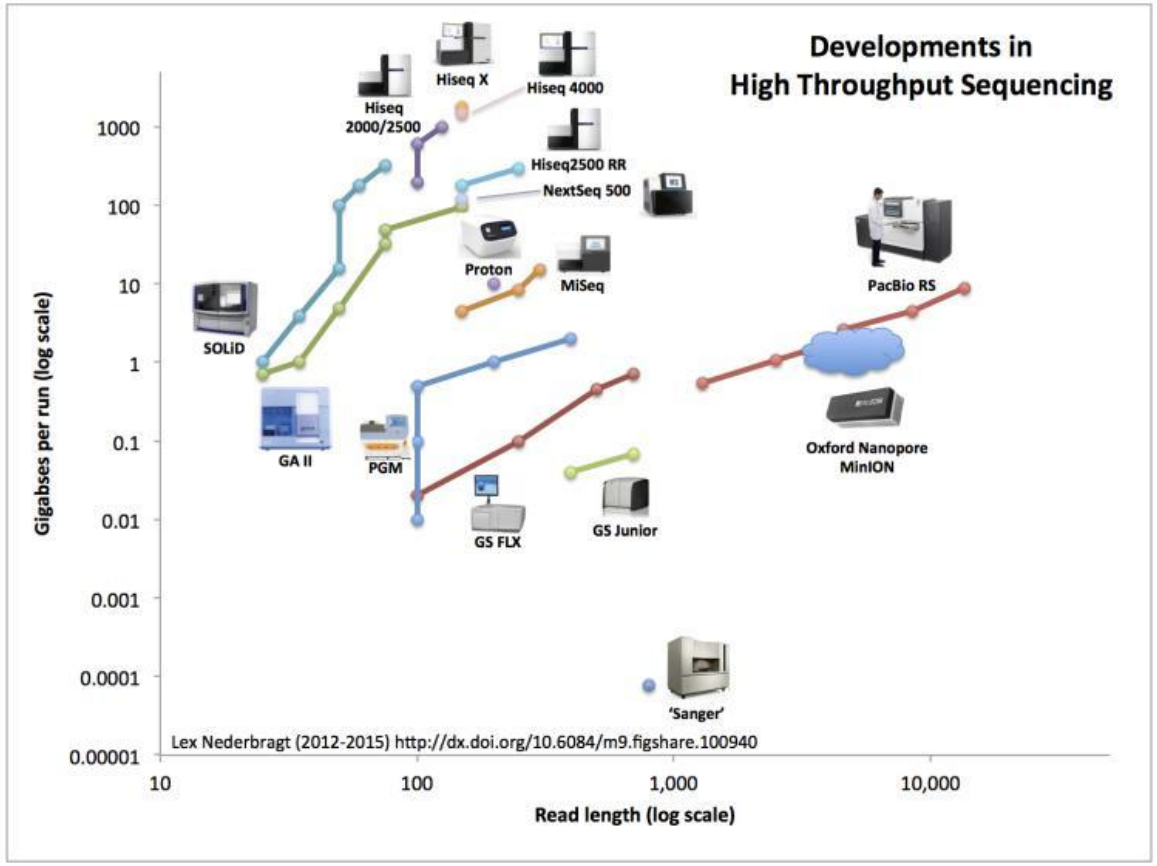
Tecnologías genómicas

- Plataformas:



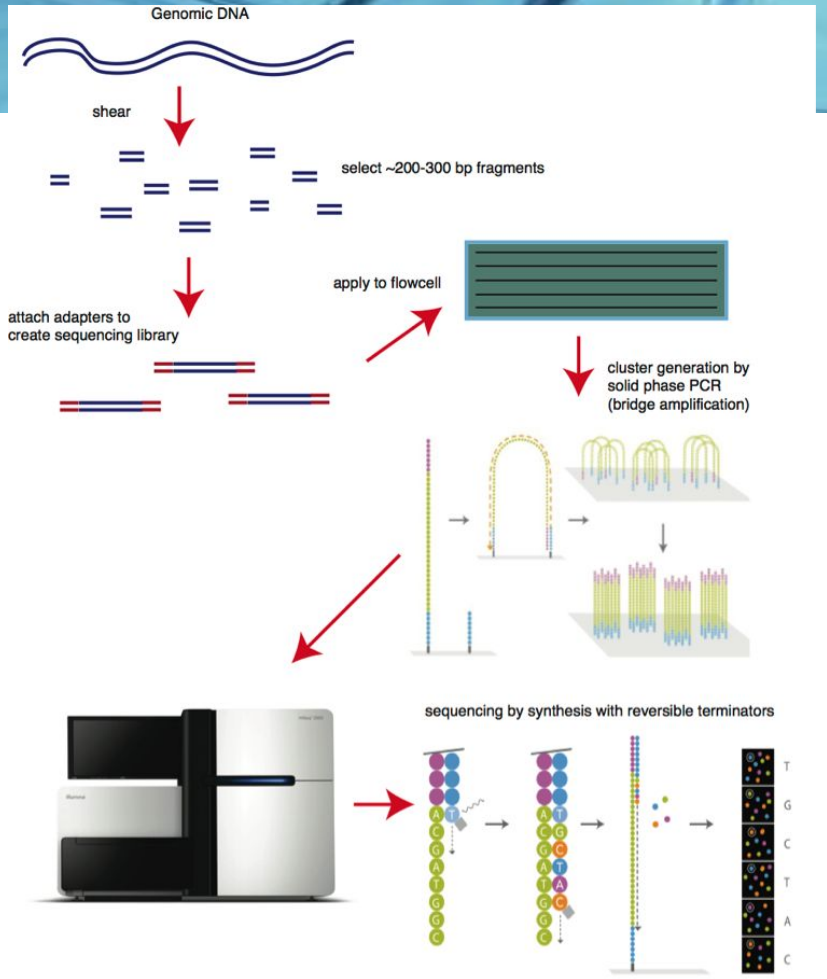
Tecnologías genómicas

- Plataformas:

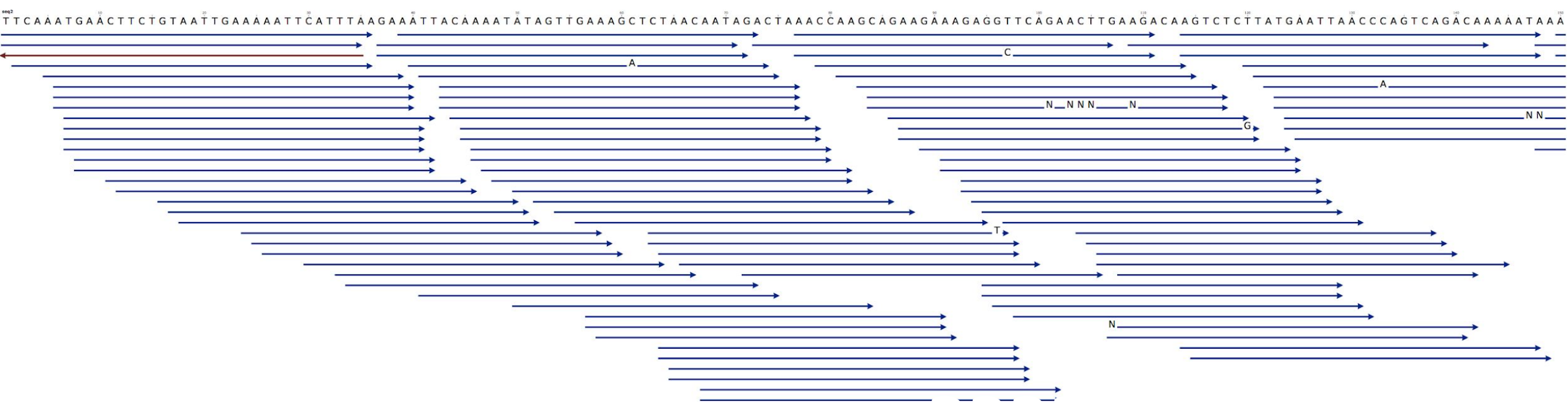


Tecnologías genómicas

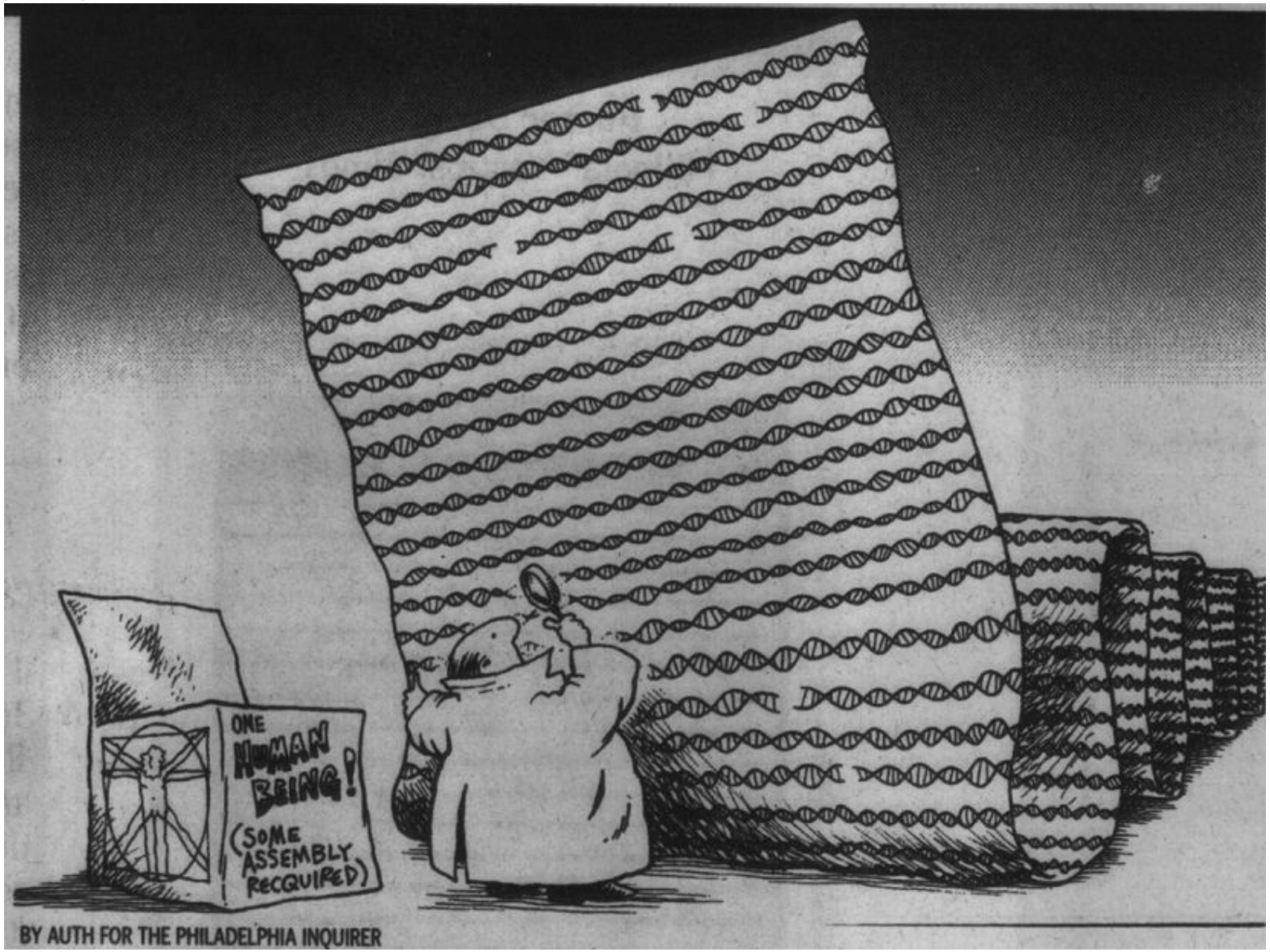
- Plataforma Illumina:



Secuenciamiento de ADN: Tecnologías y "output" del "shotgun"



¿Y ahora?

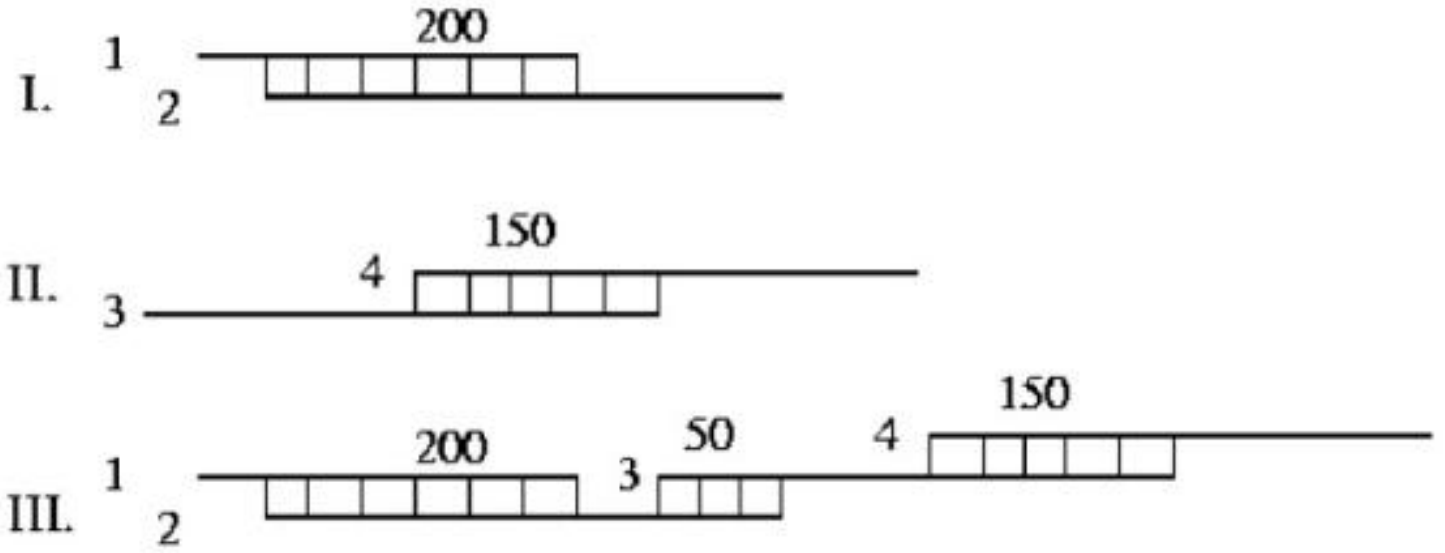

ONE HUMAN BEING!
(SOME ASSEMBLY RECQUIRED)
BY AUTH FOR THE PHILADELPHIA INQUIRER

Algoritmos de ensamblaje

- Paradigmas de ensamblaje
  - Algoritmo “Greedy”.
  - Algoritmo “OLC graphs”.
  - Algoritmo “Bruijn graphs”.

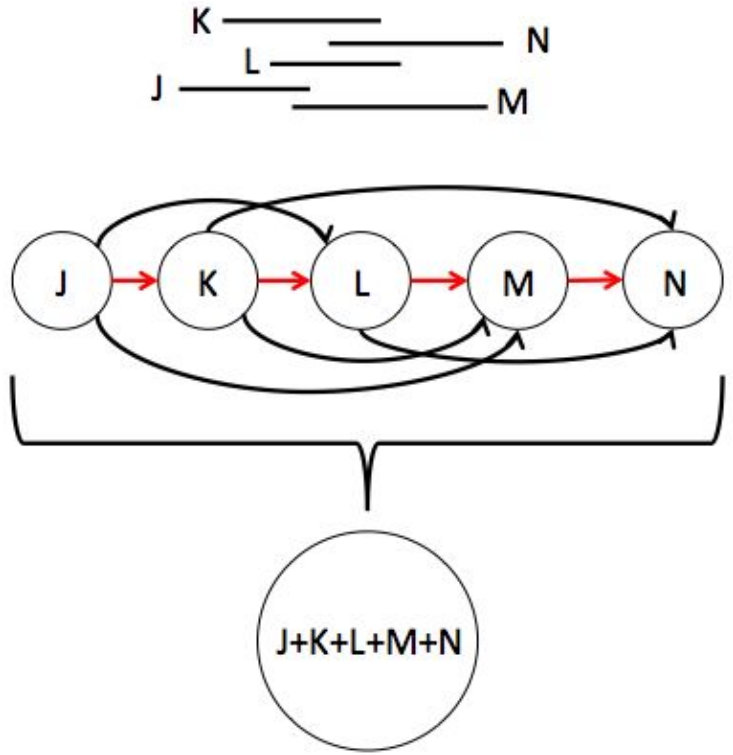
Algoritmos de ensamblaje

- Algoritmo “Greedy”: Busca mejor alineamiento entre 2 secuencias (Best score de alineamiento, i.e Newbler)



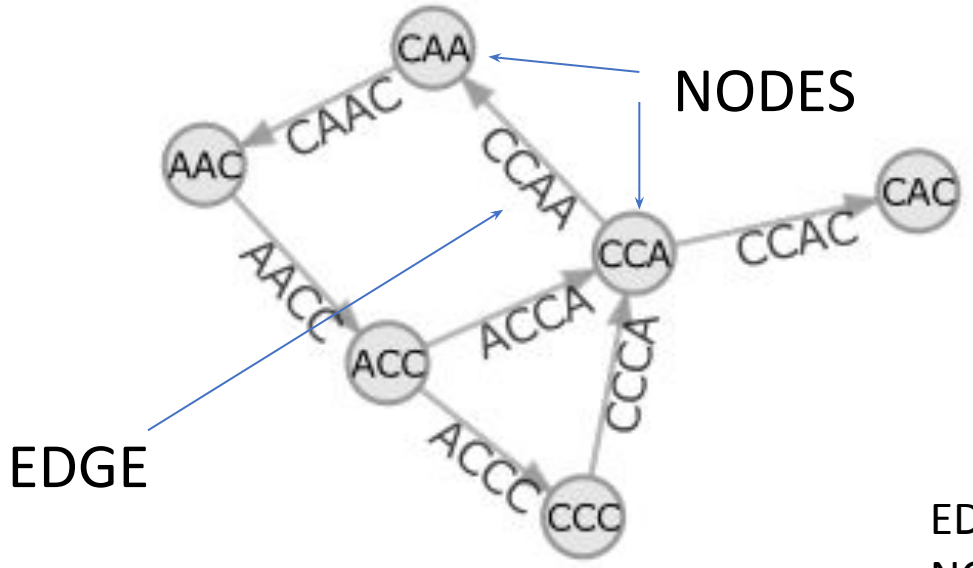
Algoritmo de Overlap-Layout-Consensus (OLC)

- Usado para secuencias Sanger y genomas pequeños y recientemente optimizado para genomas más grande (Celera Assembler)



Algoritmo de Bruijn

- Optimizado para lecturas cortas basado en el uso de k-mers (trozos de secuencia de longitud fija, i.e ABySS, Velvet, SOAPdenovo, etc)



K-MER (K) = 4

-C-C-A-
-C-C-A-A-
----C-A-A-

EDGE = K = N
NODE = K – 1
ALIGN = K – 2

Ensamble de genomas

- Ensamble “De novo”.
- Ensamble por mapeo.

De Novo	Por mapeo
No existe información previa del genoma. Y es necesario para genomas nuevos. Mayor consumo de CPU y memoria.	La lectura se alínea contra una referencia. Usado ampliamente para el estudio del genoma humano. Generan contigs que se ensamblan en supercontigs.

Ensamble "*de novo*"

Processed reads

Denovo algorithm

contigs_are_ ntigu ences_
re_conti us_se es_of_DNA
guous_sequen

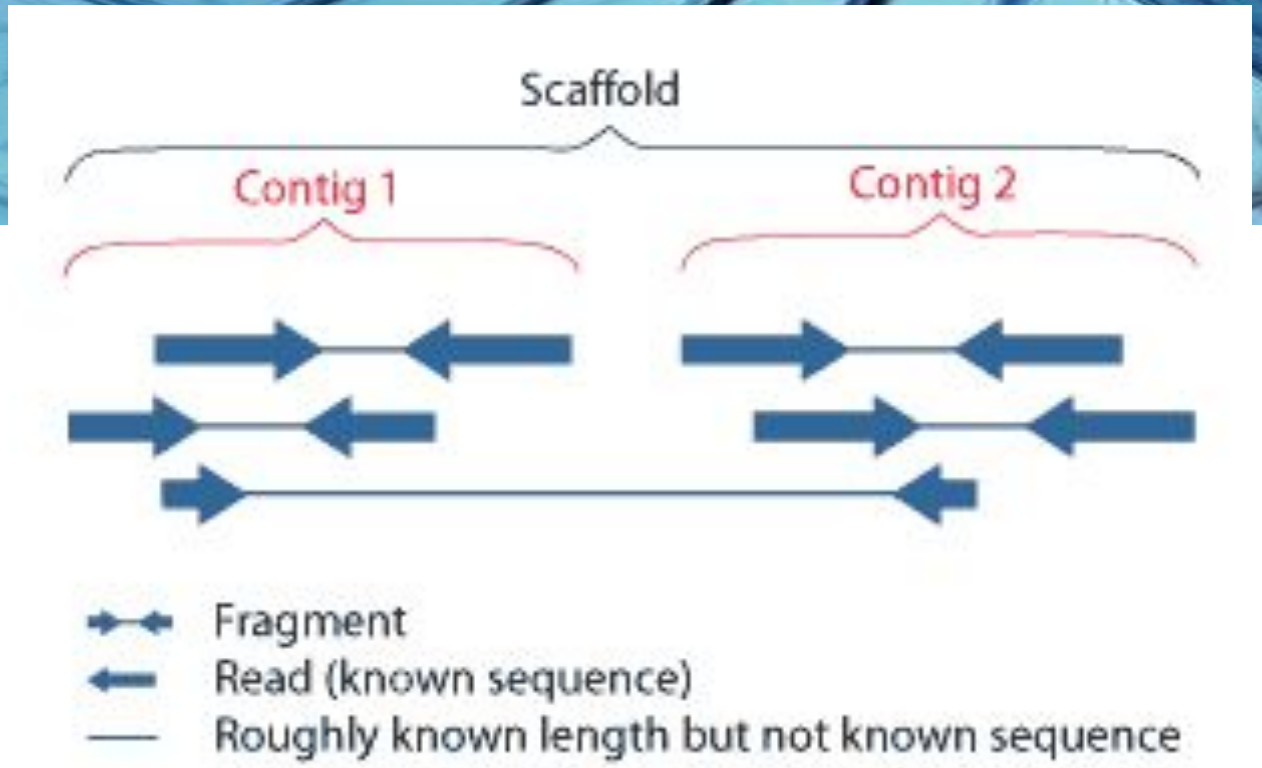
contigs_are_contiguous_sequences_of_DNA

Contig

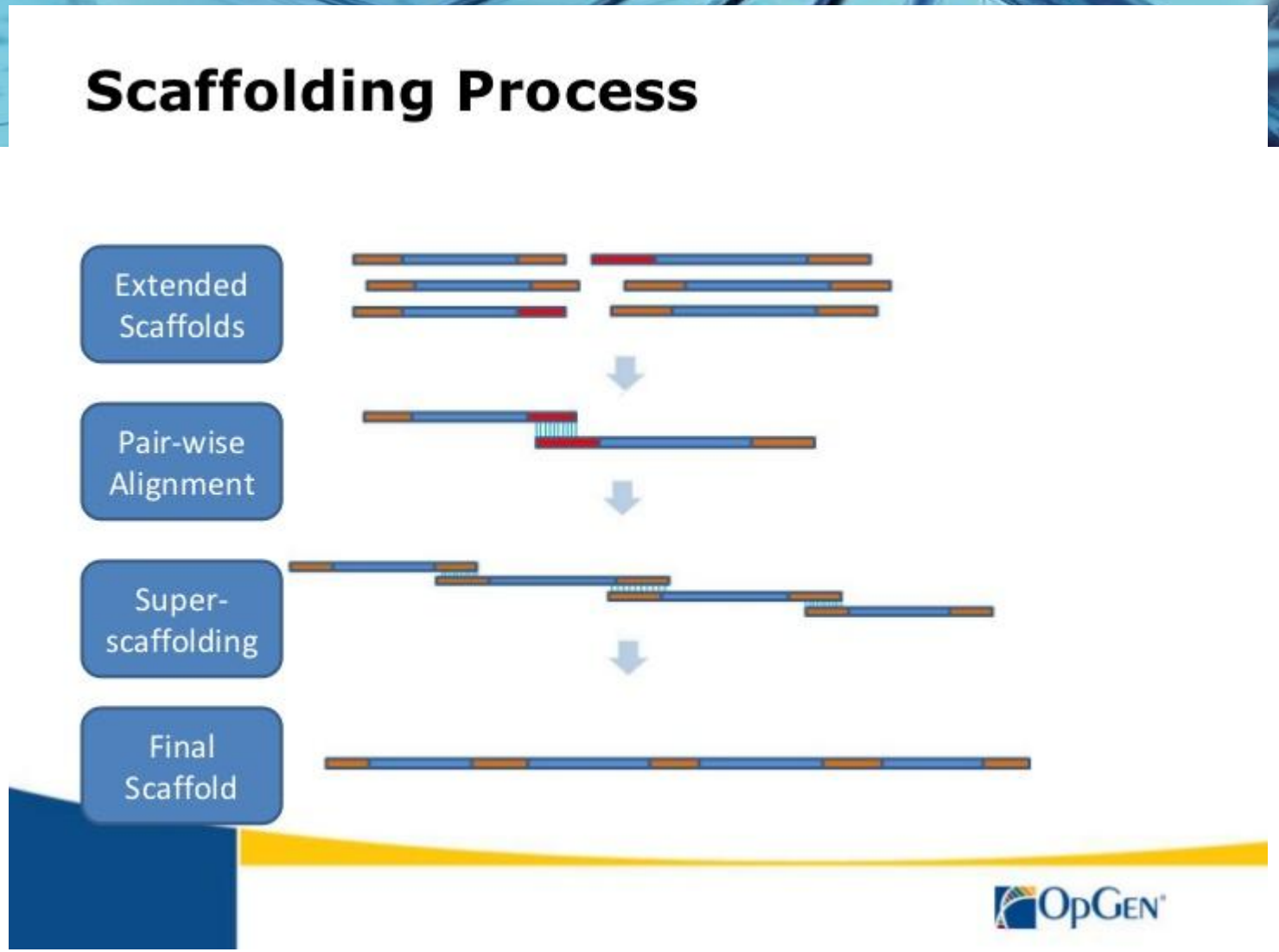
Algoritmos de ensamblaje: Ensambladores disponibles

Nombre	Tipo	Método	Plataformas	Autor
SSAKE	de novo	Greedy	Solexa	Warren, R. et al.
SHARCGS	de novo	Greedy	Solexa	Dohm et al.
VCAKE	de novo	Greedy	Solexa	Jeck W. et al.
Newbler	de novo	Greedy/OLC	454, Sanger	454/Roche
Celera Assembler	de novo	OLC	Sanger	Myers G. et al.
Arachne	de novo	OLC	454, Solexa	Batzoglou S. et al.
CAP	de novo	OLC	454, Solexa	Kolehmainen et al.
PCAP	de novo	OLC	454, Solexa	Kolehmainen et al.
CABOG	de novo	OLC	Sanger, 454, Solexa	Miller G. et al.
Euler	de novo	DBG	Sanger, 454	Pevzner P. et al.
Velvet	de novo	DBG	Sanger, 454, Solexa, SOLiD	Zerbino D. et al.
ABySS	de novo	DBG	Solexa, SOLiD	Simpson J. et al.
AllPaths	de novo	DBG	Solexa	Butler J. et al.
SOAPdenovo	de novo	DBG	Solexa	Li R. et al.
Bowtie	mapping	BWT	454, Solexa, SOLiD	Langmead B. et al.

Ensamble *"de novo"*

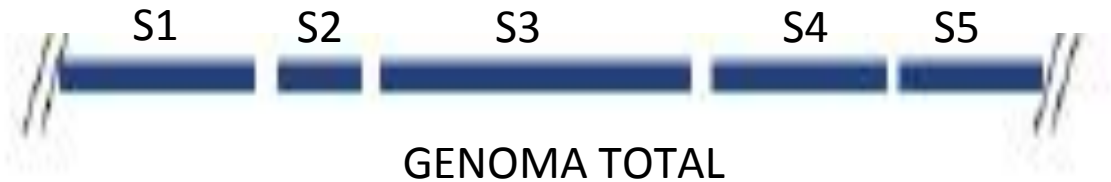


sample “*de novo*”



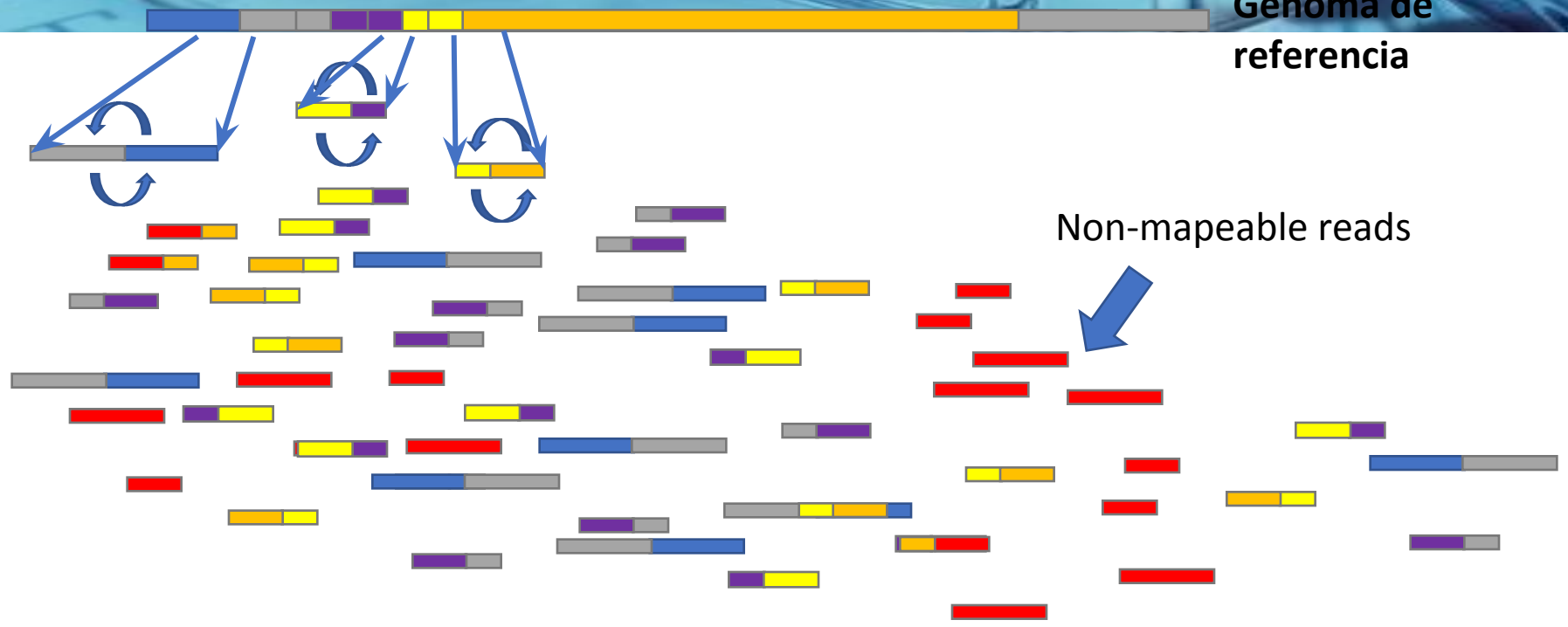
Ensamble *"de novo"*

Reads	• Reads procesados
Contig	• Sobrelapamiento de reads procesados
Scaffold	• Sobrelapamiento de contigs
S-Scaffold	• Sobrelapamiento de scaffolds



Ensamble por mapeo

Influenza
A/Singapore/C2011.803/2011(H3N2)



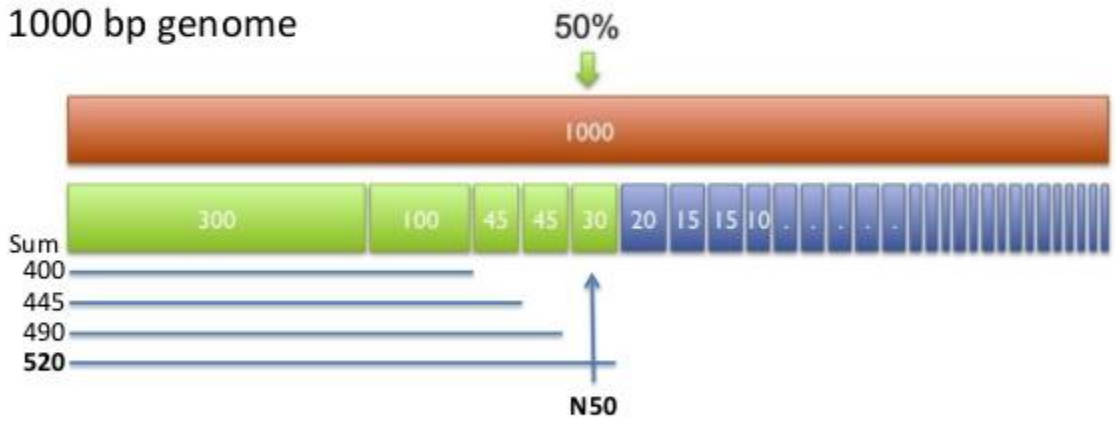
Ensamble de genomas y control de calidad

- Las estadísticas a considerar para el control de calidad son (QUAST):
  - N50.
  - Número de contigs ó scaffolds.
  - Longitud de contig o scaffold más largo.
  - Longitud combinada de todos contigs ó scaffolds.
  - % CEGs (conserved core eukaryotic genes) mapeados.

Ensamble de genomas y control de calidad: N50

N50

50% of the genome is in contigs as large as the N50 value



Courtesy of Michael Schatz, CSHL

HANDS-ON

SSH desde terminal usando el siguiente usuario

ssh -X alumno01@atziri.mazorkita.labsergen.langebio.cinvestav.mx

Crear carpeta con tu apellido

i.e: mkdir TORRES

copiar lecturas desde la carpeta “data”

cp data/* TORRES/

HANDS-ON

Ingresar a la carpeta que le corresponda:

i.e : cd TORRES/

Ejecutar fastqc

fastqc -f fastq *.gz

Crear carpetas para almacenar secuencias rasuradas

mkdir PAIRED
mkdir UNPAIRED

HANDS-ON

Ejecutar trimmomatic:

java -jar /opt/trinityrnaseq-Trinity-v2.4.0/trinity-plugins/Trimmomatic-0.36/trimmomatic.jar PE -threads 12 -phred33 6_S4_R1_paired.fastq.gz 6_S4_R2_paired.fastq.gz PAIRED/6_S4_R1_paired.fastq.gz UNPAIRED/6_S4_R1_unpaired.fastq.gz PAIRED/6_S4_R2_paired.fastq.gz UNPAIRED/6_S4_R2_unpaired.fastq.gz ILLUMINACLIP:/opt/trinityrnaseq-Trinity-v2.4.0/trinity-plugins/Trimmomatic-0.36/adapters/all_adapters.fasta:2:30:10 MINLEN:50 AVGQUAL:28

Measure	RAW	PROCESSED
Filename	6_S4_R1_paired.fastq.gz	6_S4_R1_paired.fastq.gz
File type	Conventional base calls	Conventional base calls
Encoding	Sanger / Illumina 1.9	Sanger / Illumina 1.9
Total Sequences	264498	171167
Sequences flagged as poor quality	0	0
Sequence length	35-285	50-285
%GC	39	39

HANDS-ON

Ejecutar spades, a5_pipeline y cap3:

spades.py -o spades -t 4 -1 mutant_R1.fastq.gz -2 mutant_R2.fastq.gz

/opt/a5_miseq_linux_20160825/bin/a5_pipeline.pl mutant_R1.fastq.gz mutant_R2.fastq.gz --threads=4 A5-assembly-mutant

Transformar archivos fastq a fasta

/d2p10tb/databases/db/viromescan/seqtk seq -a mutant_R1.fastq.gz > mutant_R1.fasta

/d2p10tb/databases/db/viromescan/seqtk seq -a mutant_R2.fastq.gz > mutant_R2.fasta

cat mutant_R1.fasta mutant_R2.fasta > mutant_R1-2.fasta

/opt/iassembler/iAssembler-v1.3.2.x64/bin/cap3 mutant_R1-2.fasta

HANDS-ON

Ejecutar QUAST:

opt/miniconda2/lib/python2.7/site-packages/quast-5.0.2-py2.7.egg/EGG-INFO/scripts/quast.py -o QUAST-3-assemblers -r wildtype.fna mutant_R1-2.fasta.cap.contigs spades-assembly/scaffolds.fasta A5_assembly/A5-assembly-mutant.contigs.fasta