INSTITUTO DE ECOLOGIA UNAM

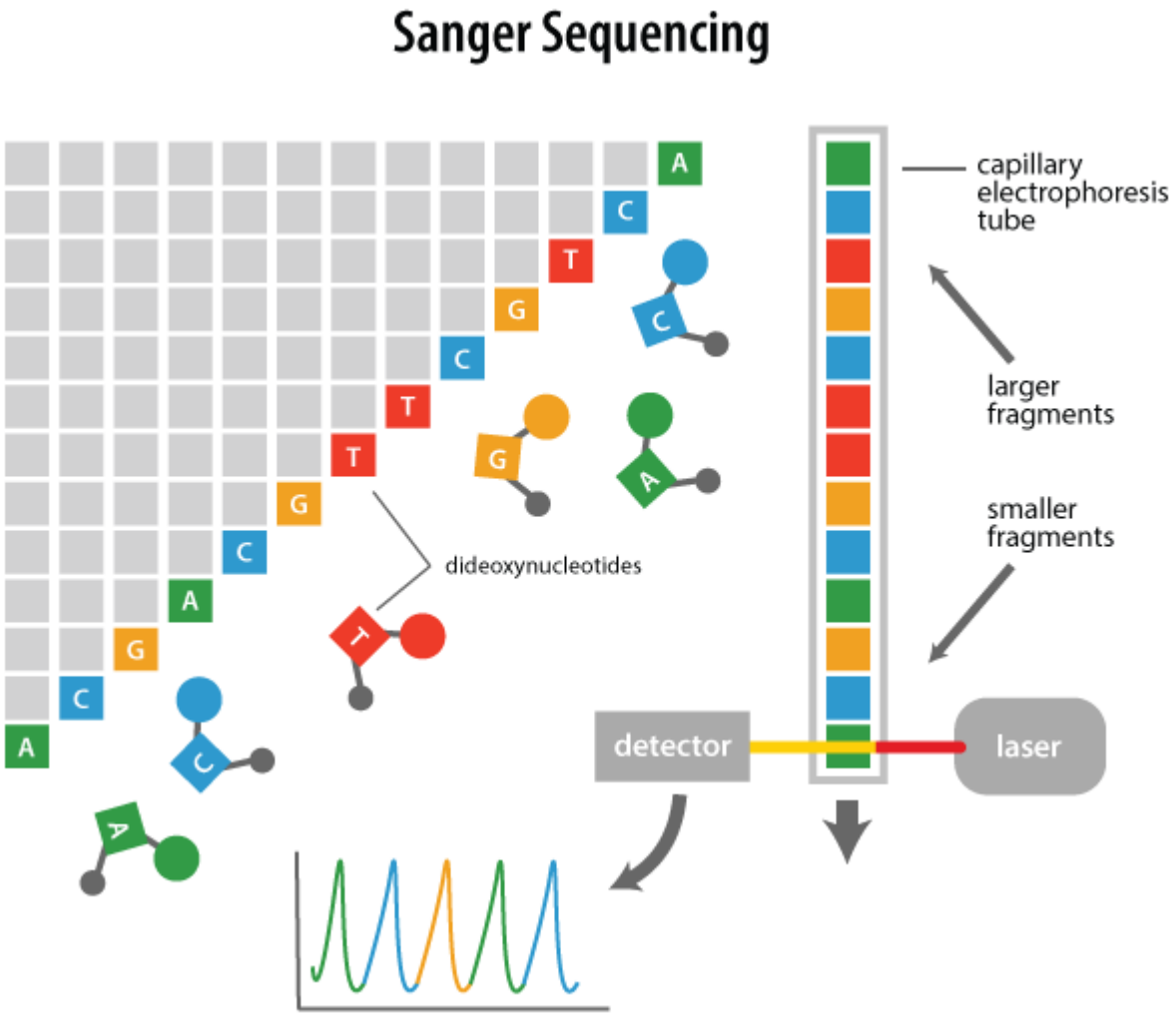
Procesamiento de datos genómicos

Biól. Manuel García Ulloa Gámiz

manuel.gug@hotmail.com

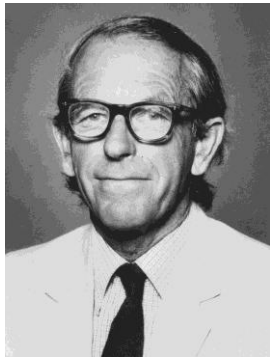
Génesis (1975)

Secuenciación de Sanger

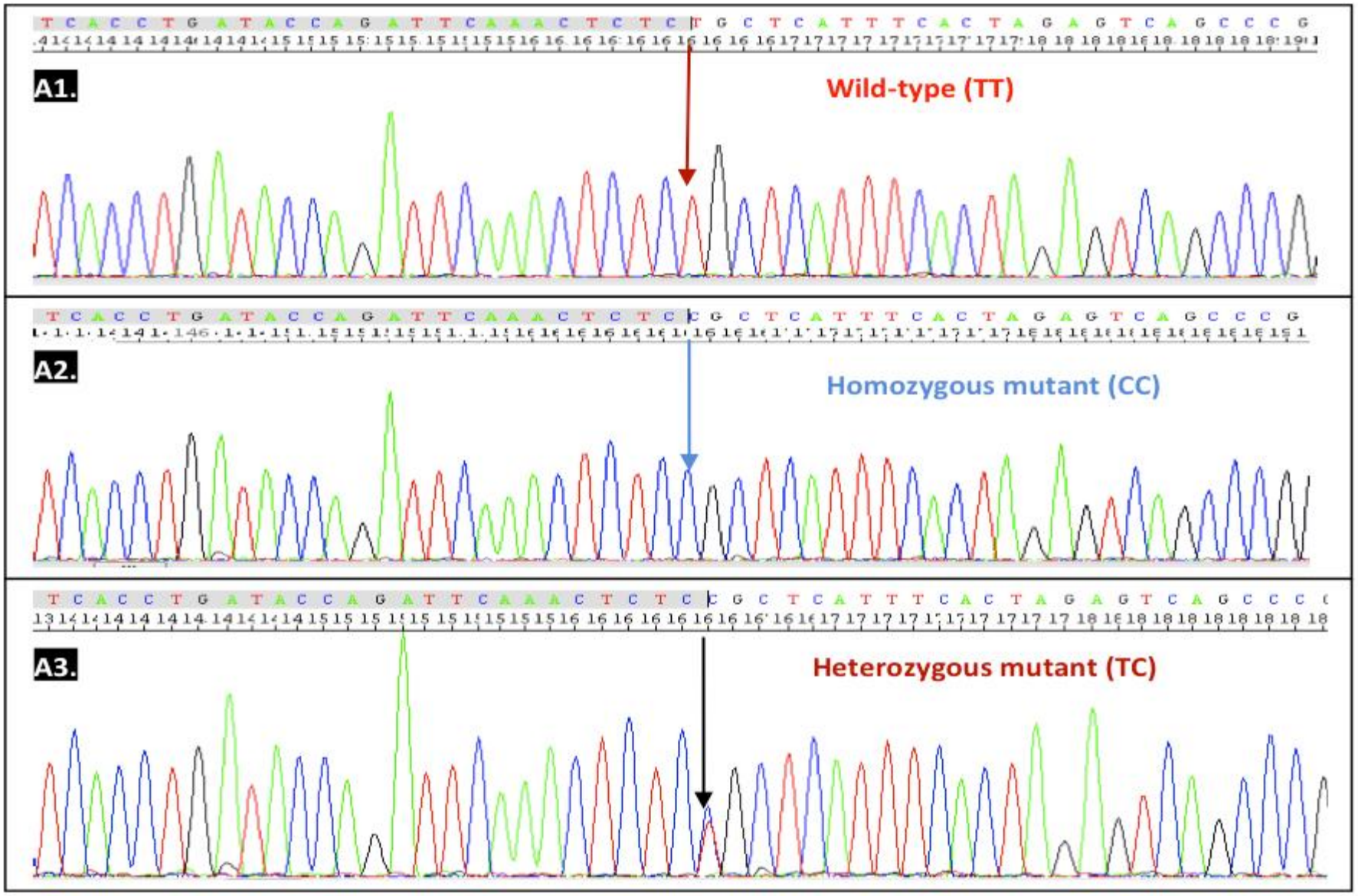


Ventajas:
-Alta precisión, ideal para secuencias cortas (conocidas)

Desventajas:
-Cara
-Lenta



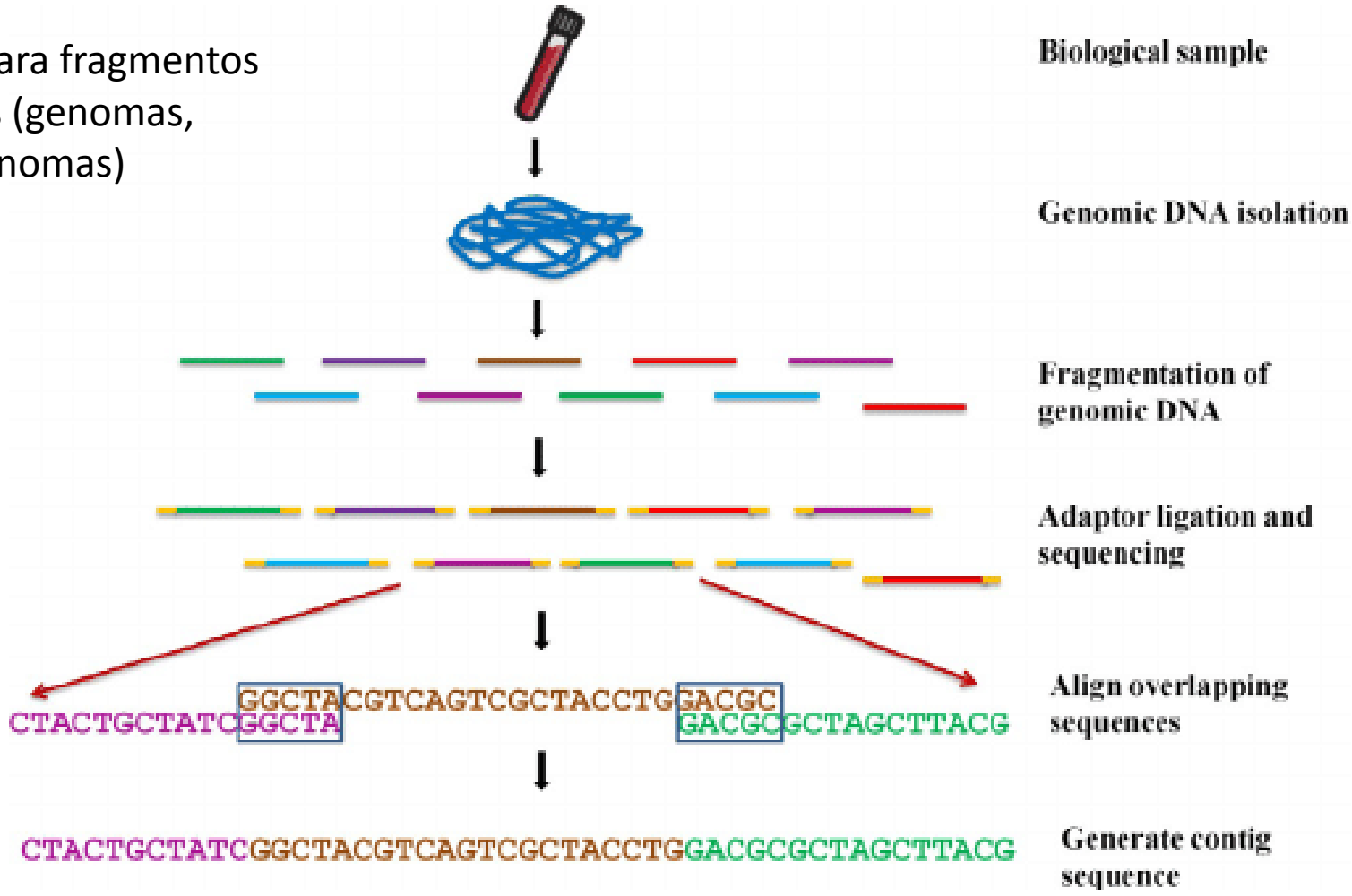
Mutaciones puntuales



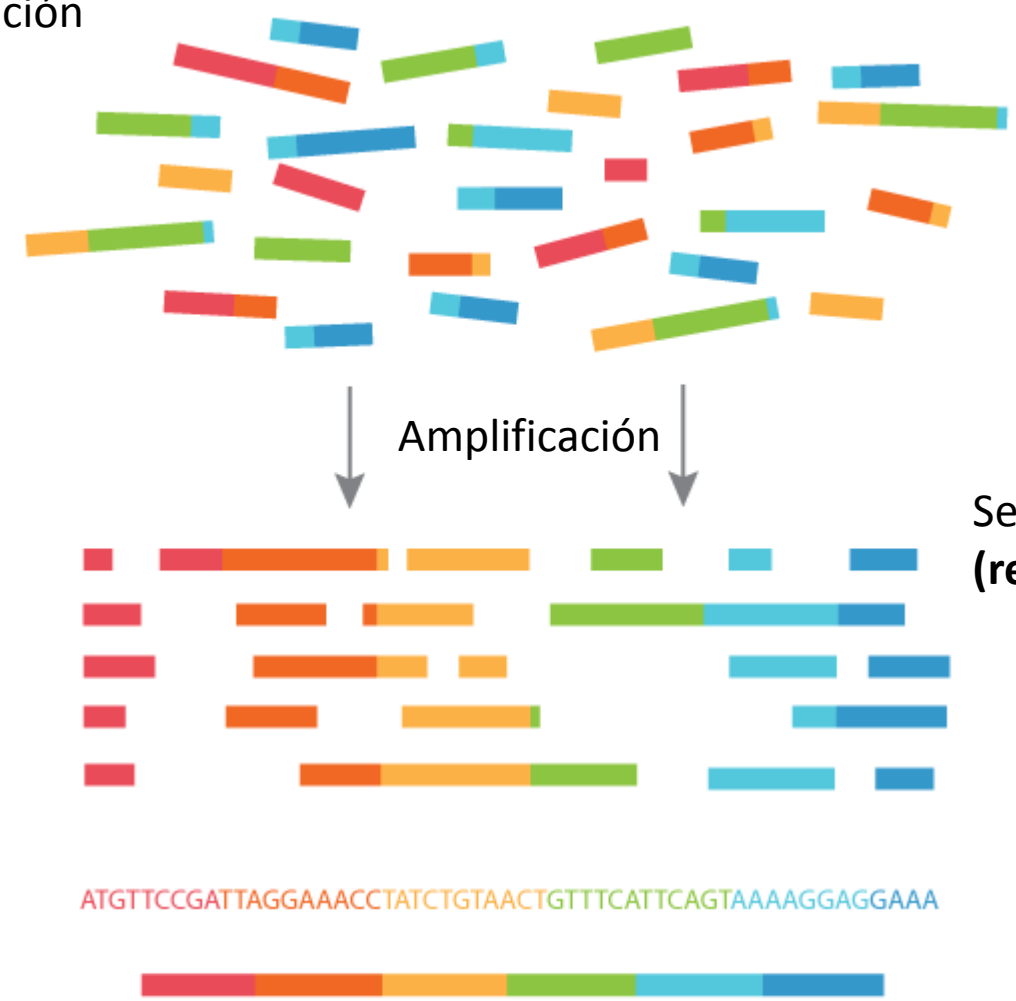
2da generación (1990’s - 2011)

Sequenciación de siguiente generación (NGS)

-Rápida
-Barata
-Ideal para fragmentos grandes (genomas, metagenomas)



Fragmentación



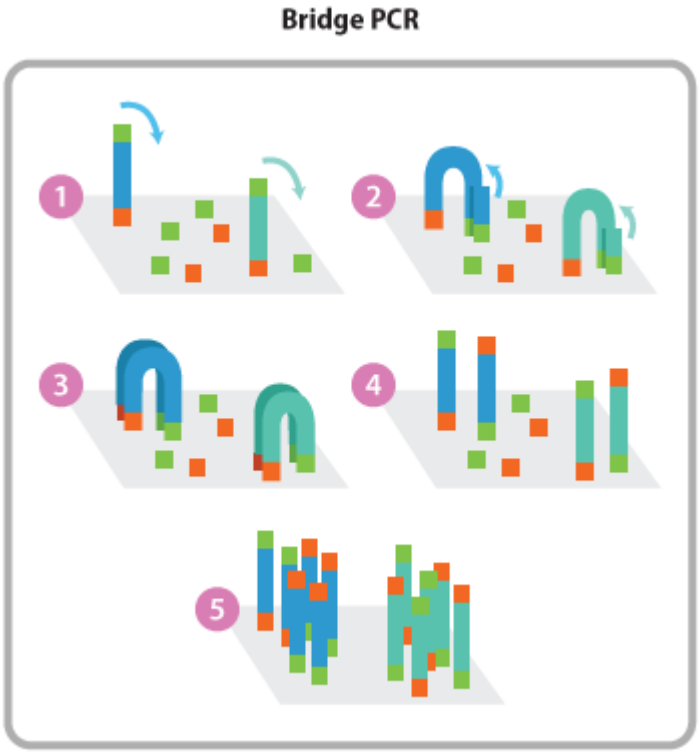
Secuenciación
(reads)

Ensamble
(contigs, scaffolds)

Elementos comunes entre tecnologías NGS

- Preparación de librerías
  - Fragmentación de librerías
  - Ligación de adaptadores (linker + primer + barcode)
  - Posicionamiento en superficie

- Secuenciación
  - Amplificación
  - Lectura de secuencias

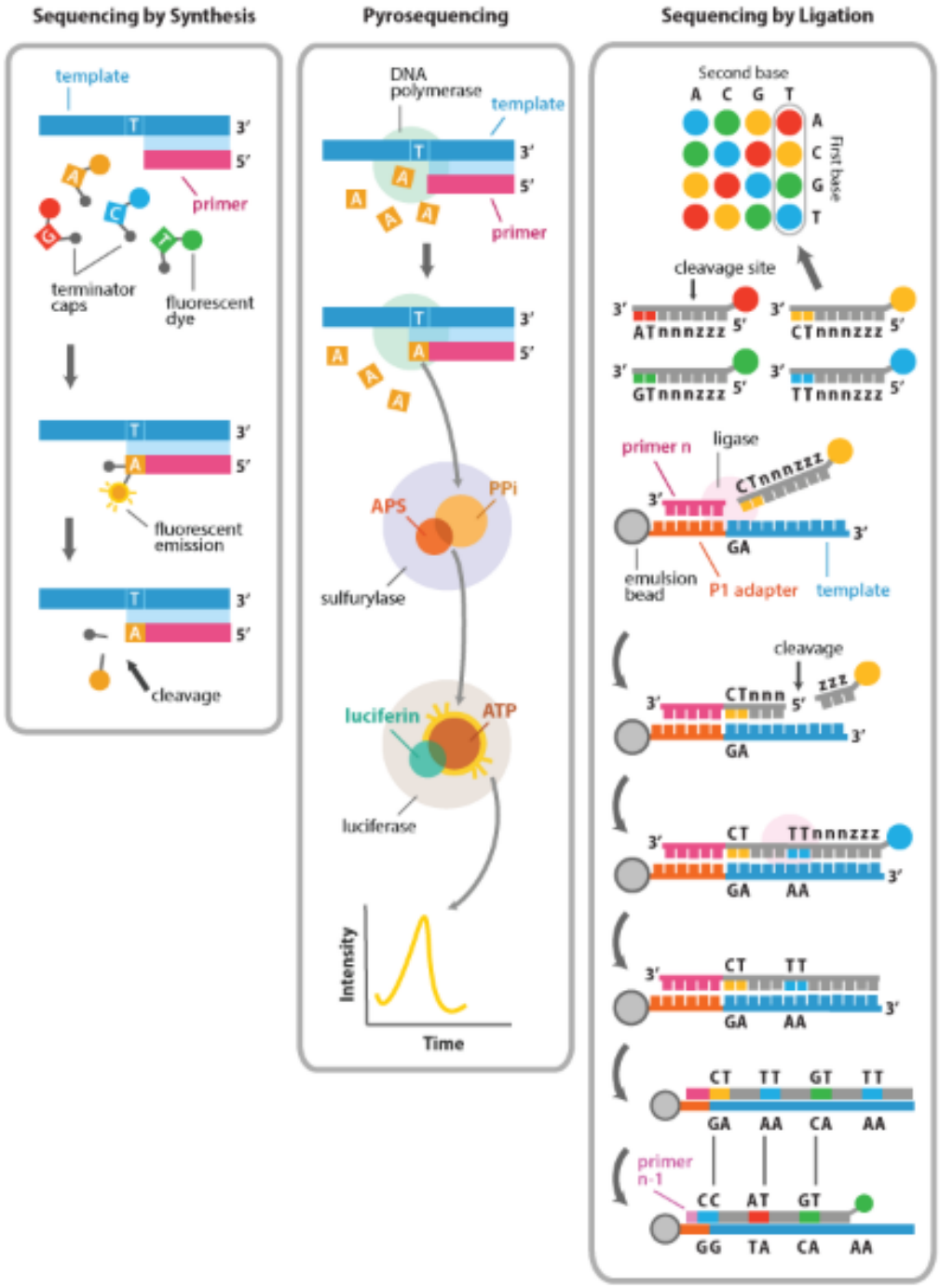


TIPOS DE SECUENCIACIÓN

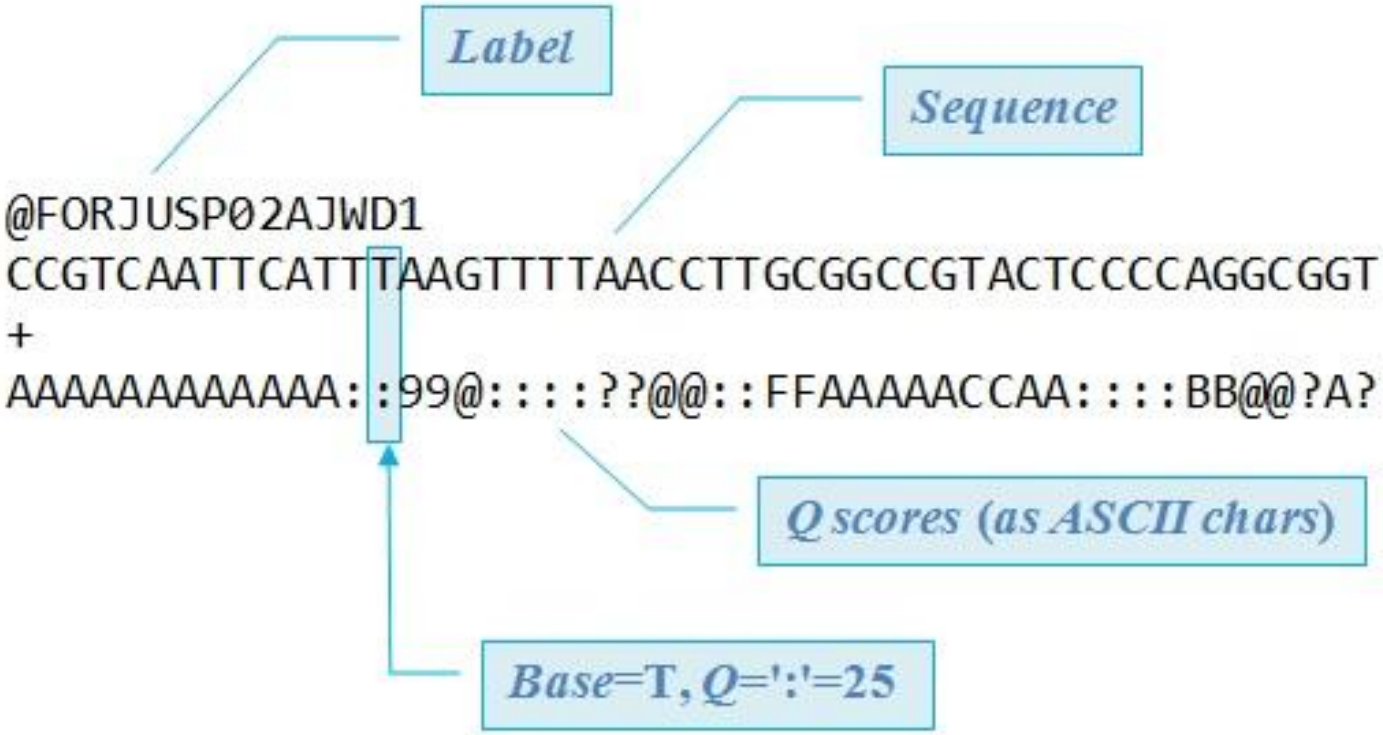
a) Por Síntesis
   -cons: tasa de error aumenta con el largo de la secuencia debido a la remoción incompleta de fluoróforo.

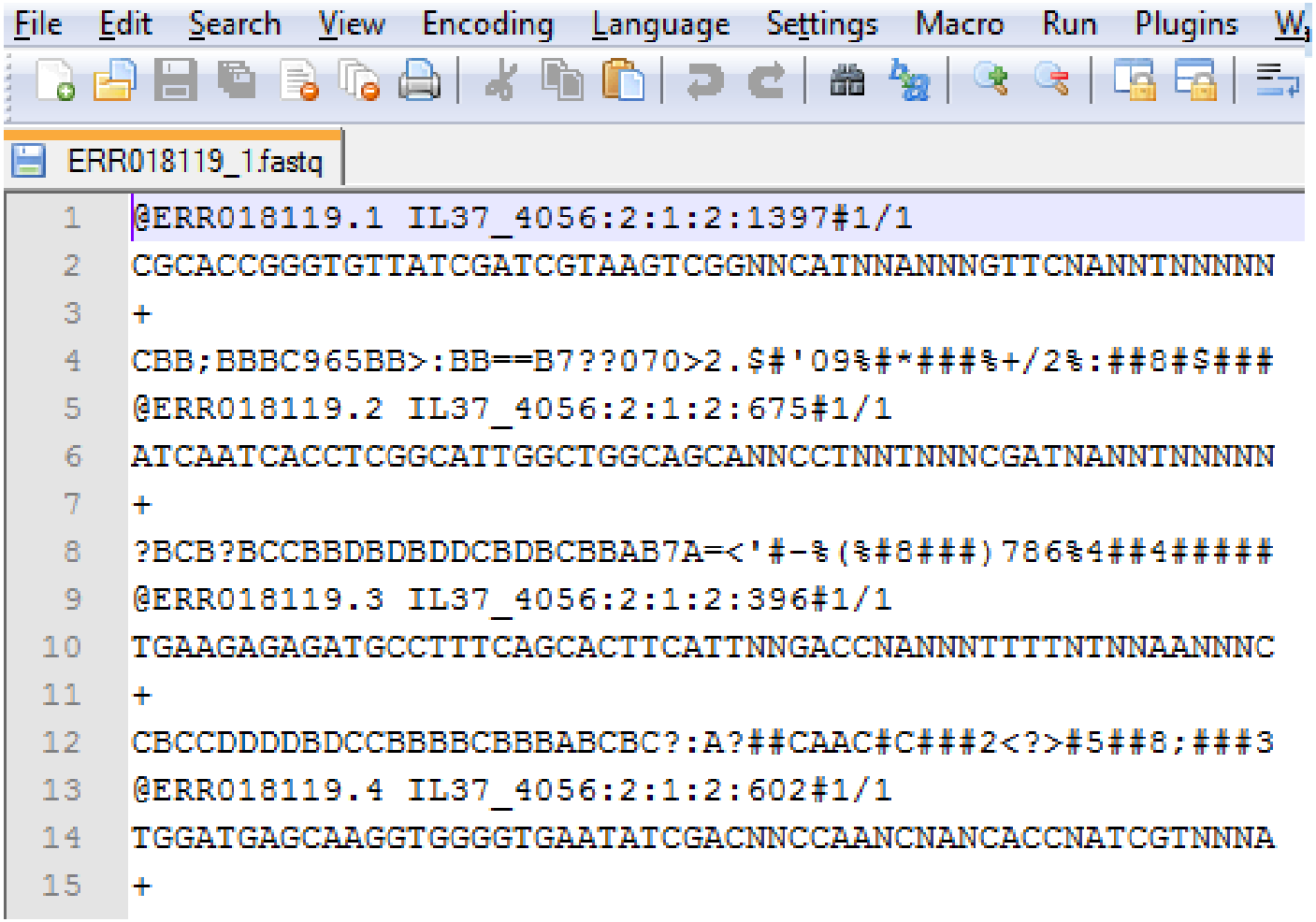
b) Pirosecuenciación
   -cons: cara, alta tasa de error en secuencias con 6 nt seguidos iguales.

c) Por ligación
   -cons: reads muy cortos (máximo 75pb)



- Generación de datos
  - Formato FASTQ




File Edit Search View Encoding Language Settings Macro Run Plugins
ERR018119_1.fastq
1 @ERR018119.1 IL37_4056:2:1:2:1397#1/1
2 CGCACCGGGTGTTATCGATCGTAAGTCGGNNCATNNANNNGTTCNANNTNNNNN
3 +
4 CBB;BBBC965BB>:BB==B7??070>2.$#'09%#*###%+/2%:##8#$###
5 @ERR018119.2 IL37_4056:2:1:2:675#1/1
6 ATCAATCACCTCGGCATTGGCTGGCAGCANNCCTNNTNNNCGATNANNTNNNNN
7 +
8 ?BCB?BCCBBDBDBDDCBDBCBBAB7A=<'#-%(%#8###)786%4##4#####
9 @ERR018119.3 IL37_4056:2:1:2:396#1/1
10 TGAAGAGAGATGCCTTTCAGCACTTCATTNNGACCNANNNTTTTNTNNAANNNC
11 +
12 CBCCDDDDBDCCBBBBCBBBABCBC?:A?##CAAC#C###2<?>#5##8;###3
13 @ERR018119.4 IL37_4056:2:1:2:602#1/1
14 TGGATGAGCAAGGTGGGGTGAATATCGACNNCCAANCNANCACCNATCGTNNNA
15 +

Comparación de tecnologías

- Cobertura: cantidad de veces que un nucleótido particular promedio es secuenciado.

ATGACTGCTGA

TGACTGCTGATTG Cobertura según A= **5x**

GACTGCTGATTGAACT

TGCTGATTGAACT

GATTGAACTATGGTGA

Pirosecuenciación

Roche	GS Junior	GS Junior Plus	GS FLX+ System	
			GS FLX Titanium XL+	GS FLX Titanium XLR70
Human	0	0	0	0
Mouse	0	0	0	0
Arabidopsis thaliana	0	1	5	3
E-Coli	8	15	151	97

Secuenciación por síntesis

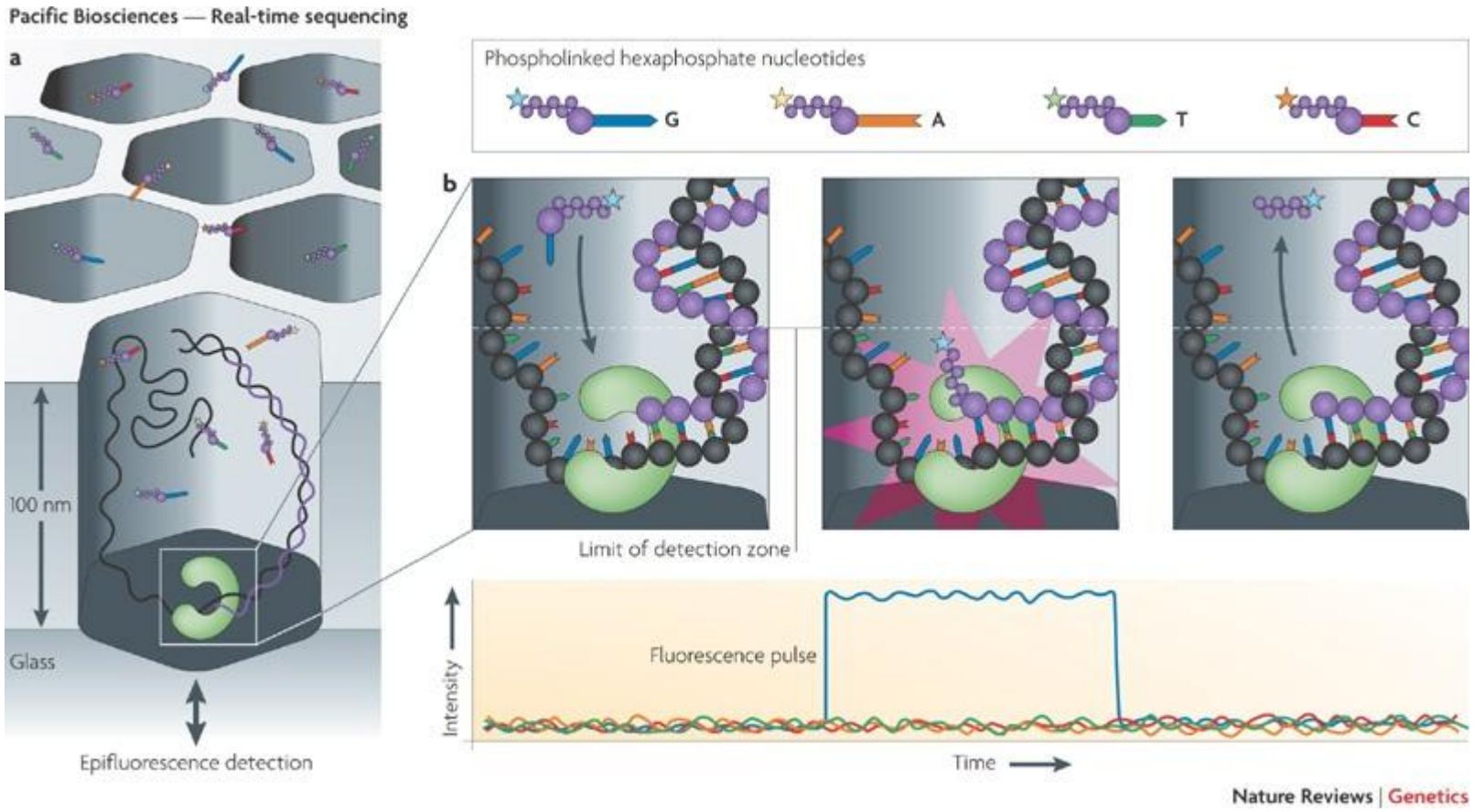
Illumina	MiSeq	NextSeq 500		HiSeq 2500		HiSeq 3000	HiSeq 4000
Human	5	12	36	91	303	227	455
Mouse	5	14	43	107	357	268	536
Arabidopsis thaliana	111	289	889	2,222	7,407	5,556	11,111
E-Coli	3,233	8,407	25,866	64,666	215,553	161,665	323,330

Secuenciación por ligación

Applied Biosystems	Genetic Analyzer V2.0	
	5500W System	5500xl W System
Human	48	97
Mouse	57	114
Arabidopsis thaliana	1,185	2,370
E-Coli	34,489	68,977

3ra generación (2011 -)

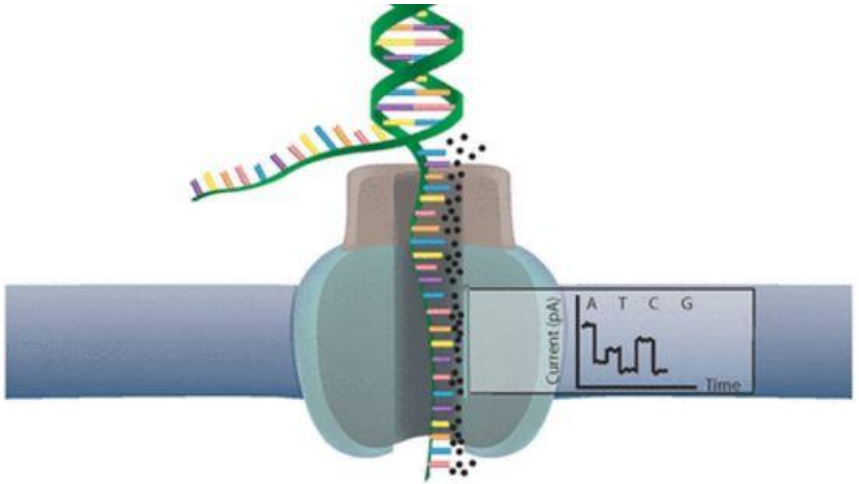
Secuenciación SMRT (single molecule real time)



- Genera secuencias muchas muy largas (hasta 50,000 secuencias de 30kb)
- Detecta metilación
- Alta tasa de error (hasta 10%)

Secuenciación Nanopore

- Aún en desarrollo
- Secuencias de hasta 79kb
- Tasa de error de 15-40%...



Ejercicio:
Proyecto típico de genómica

Pasos

- 1) Limpieza de datos
- 2) Ensamble de genoma
- 3) Control de calidad
- 4) Anotación de genoma
- 5) Filogenia

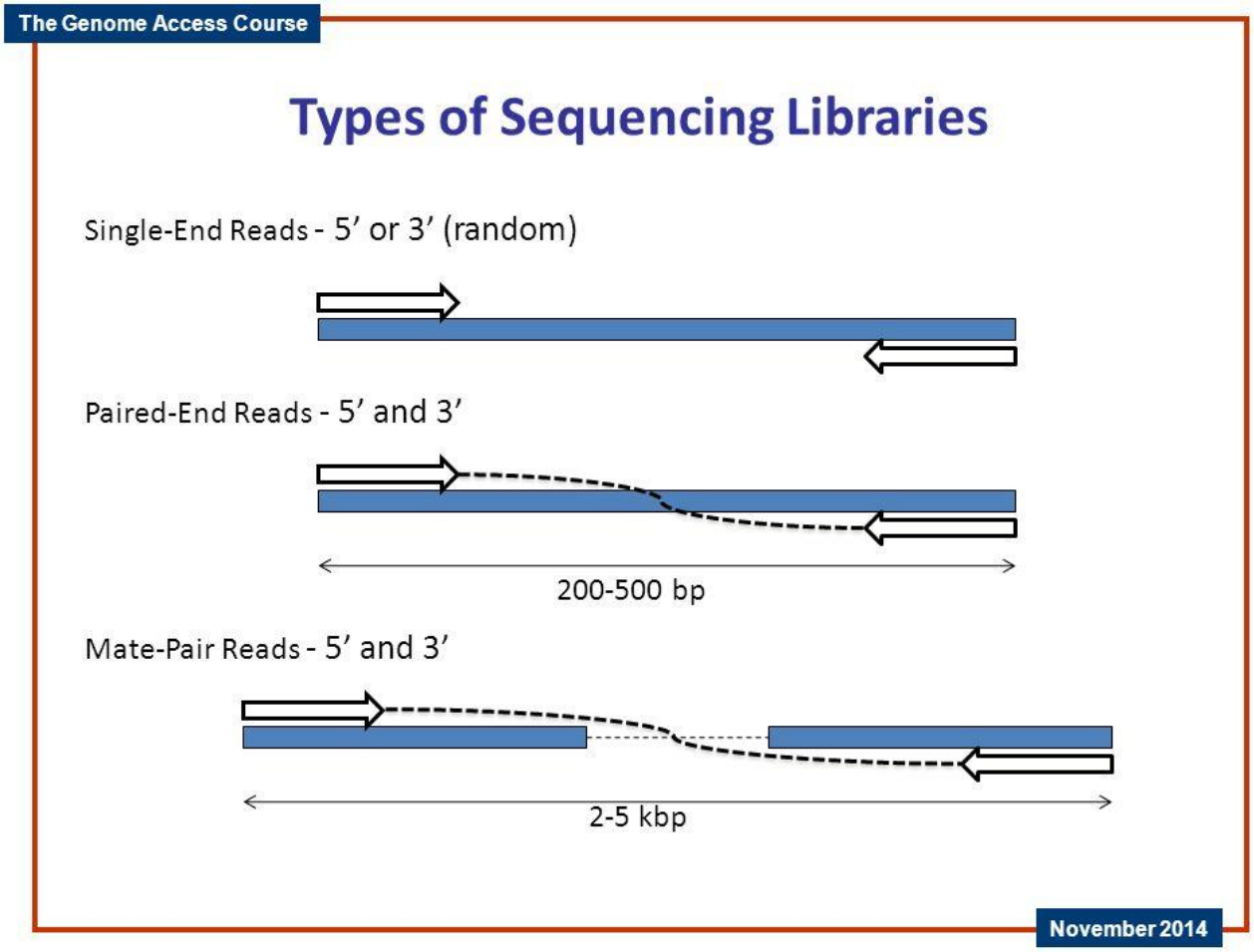
Datos

- Secuenciación del genoma de *Carsonella ruddii* strain BT (genoma bacteriano más pequeño: 150-170 kb, ~180 ORFS)
- Tipo de datos: Illumina paired-end 2x300 (**secuenciación por síntesis**)
- Carpeta “DATOS”:
  - reads_B1_2900x300bp_0S_0I_0D_0U_0N_1.fq
  - reads_B1_2900x300bp_0S_0I_0D_0U_0N_2.fq

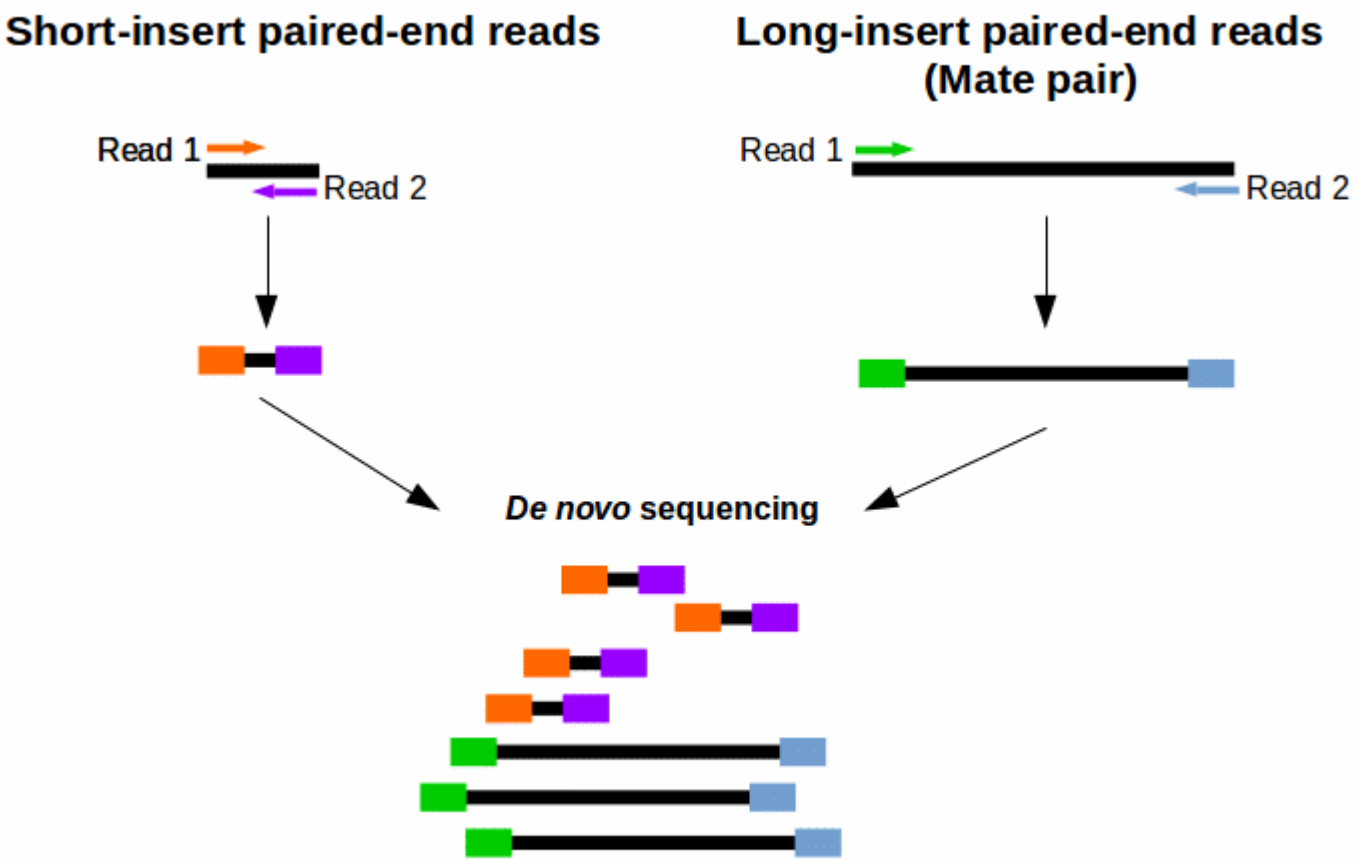
Mate-pair vs paired-end (Illumina)

Paired-end
-Mayor cobertura y exactitud
-Cortos (máx 500pb)
-Problemas con secuencias repetitivas
✔Polimorfismos, indels

Mate-pair
-Menor cobertura y exactitud
-Largos (máx 5kb)
-Buenos con secuencias repetitivos
✔Variación estructural, rearreglos

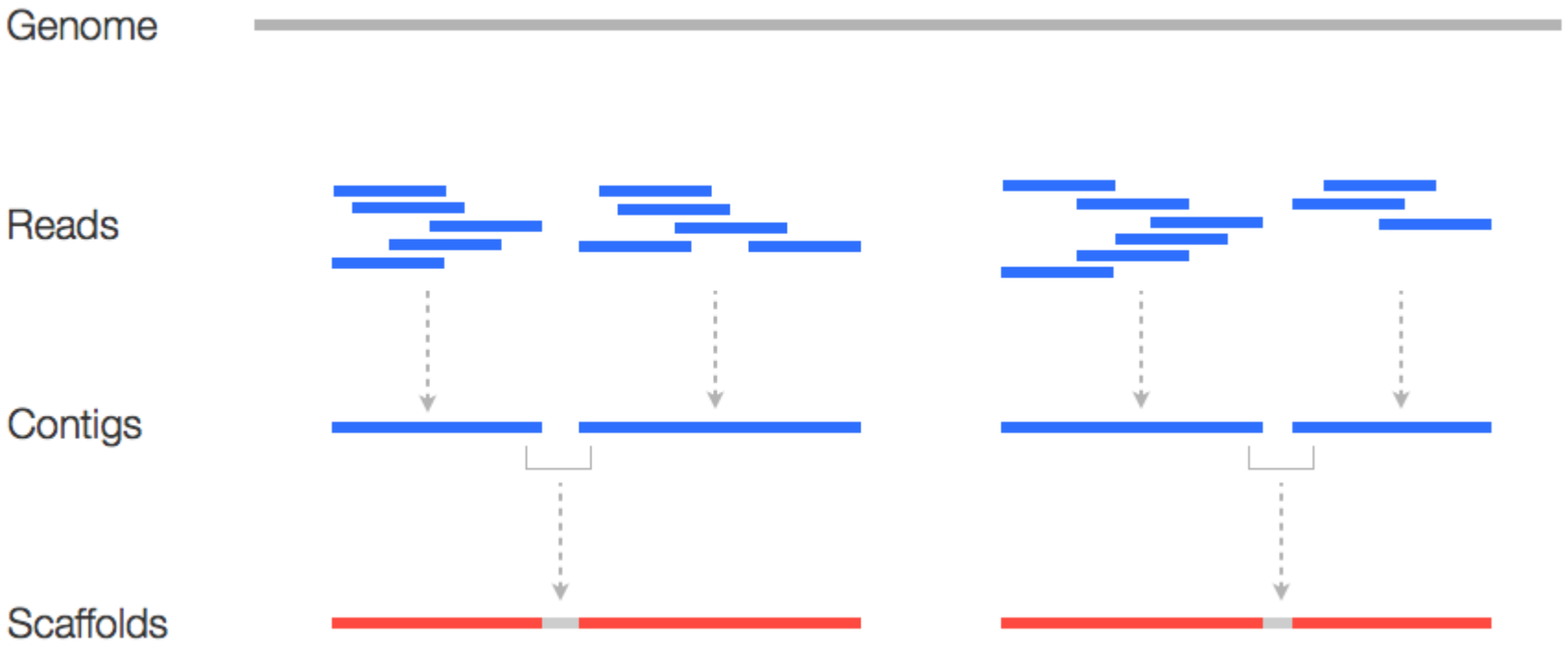


Uso conjunto



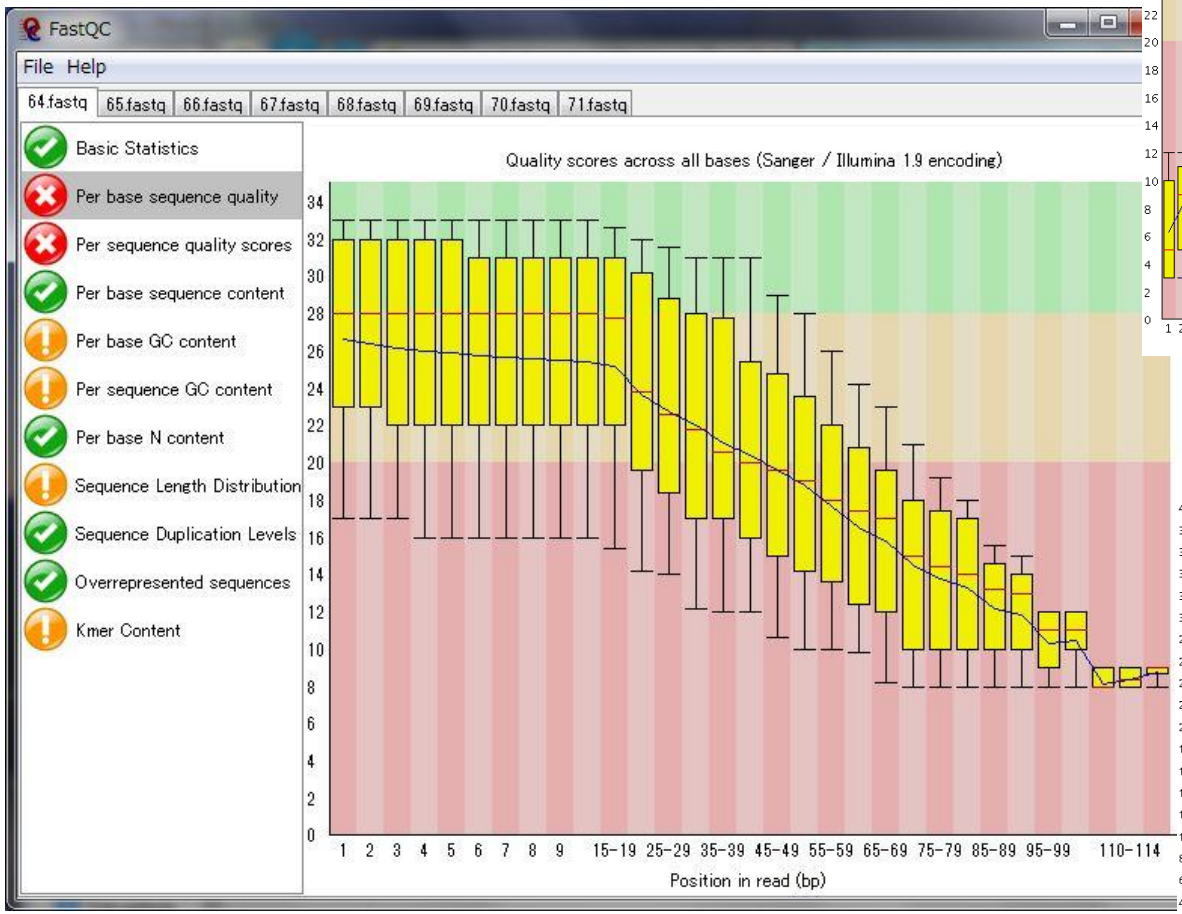
Ensamble de genoma

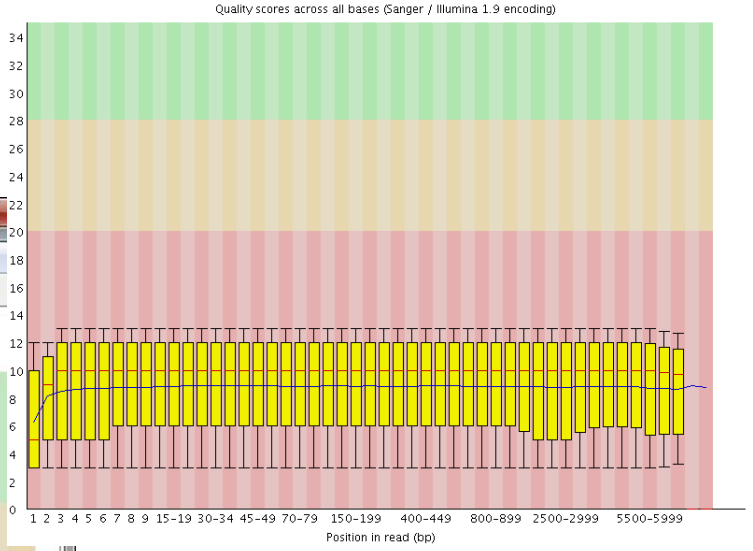
- Ensamblador: programa que organiza reads de tal manera que la secuencia de la cual fueron obtenidos sea reconstruida (lo mejor posible dadas las condiciones iniciales de los datos).

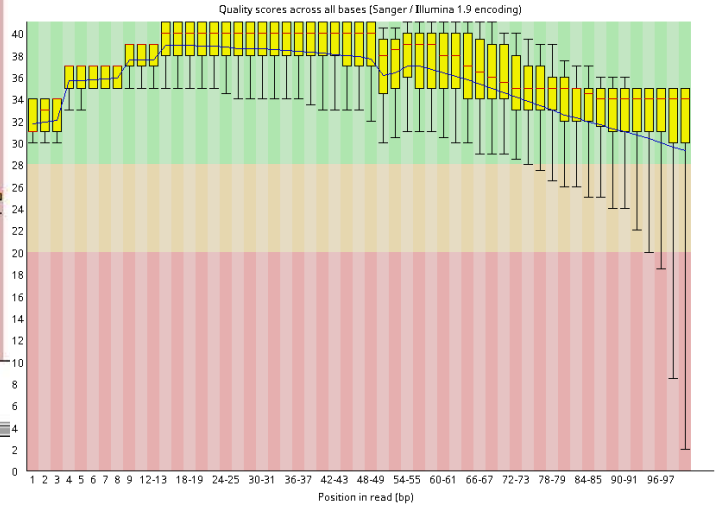


¿Cómo están mis datos?

- $ sudo apt-get install fastqc
- $ fastqc







Limpieza de datos crudos

- http://www.usadellab.org/cms/?page=trimmomatic
  - Version 0.38: binary

- $ java -jar trimmomatic-0.38.jar PE -phred33 ../DATOS/reads/reads_B1_2900x300bp_0S_0I_0D_0U_0N_2.fq ../DATOS/reads/reads_B1_2900x300bp_0S_0I_0D_0U_0N_2.fq NOMBRE_1_paired.fq NOMBRE_1_unpaired.fq NOMBRE_2_paired.fq NOMBRE_2_unpaired.fq ILLUMINACLIP:/home/mint/Downloads/Trimmomatic-0.38/adapters/TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50

- $ scp NOMBRE*.fq ohta@132.248.49.136:/home/ohta/Desktop/CURSO_BIOINFO/GENOMICA/2_ensamble/SPAdes-3.13.0-Linux/bin

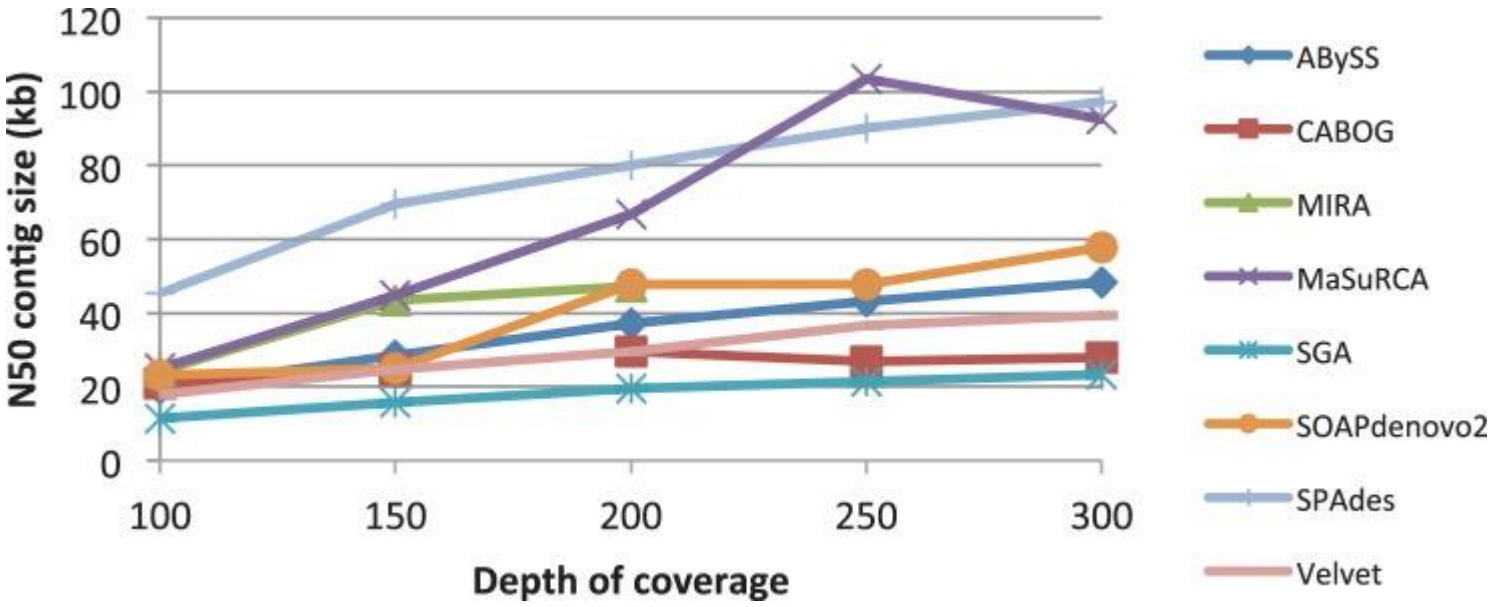
OXFORD JOURNALS Bioinformatics (Oxford, England)

Bioinformatics. 2013 Jul 15; 29(14): 1718–1725.
Published online 2013 May 10. doi: 10.1093/bioinformatics/btt273

PMCID: PMC3702249
PMID: 23665771

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

Tanja Magoc,[1] Stephan Pabinger,[1,2] Stefan Canzar,[1] Xinyue Liu,[3] Qi Su,[3] Daniela Puiu,[1] Luke J. Tallon,[3] and Steven L. Salzberg[1,*]



OXFORD JOURNALS Bioinformatics (Oxford, England)

Bioinformatics. 2013 Jul 15; 29(14): 1718–1725.
Published online 2013 May 10. doi: 10.1093/bioinformatics/btt273
PMCID: PMC3702249
PMID: 23665771

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

Tanja Magoc,[1] Stephan Pabinger,[1,2] Stefan Canzar,[1] Xinyue Liu,[3] Qi Su,[3] Daniela Puiu,[1] Luke J. Tallon,[3] and Steven L. Salzberg[1,*]

**N50 en kb

Assembler	Species assembled						
	HiSeq (100 bp) reads			MiSeq (250 bp) reads			
	R.sphaeroides	*M.abscessus*	*V.cholerae*	*B.cereus*	*R.sphaeroides*	*M.abscessus*	*V.cholerae*
ABySS	13.0	115.7	93.0	130.6	21.4	68.5	60.3
CABOG	11.2	78.2	48.8	150.5	30.5	8.3	32.5
MIRA	17.7	129.2	87.1	100.0	15.4	75.0	108.7
MaSuRCA	**176.8**	**194.0**	**236.4**	**246.7**	**130.7**	36.2	71.6
SGA	12.1	27.9	23.4	25.5	9.1	12.8	27.3
SOAPdenovo	10.5	147.2	106.5	**246.3**	33.5	113.3	65.5
SPAdes	83.5	147.9	77.1	103.7	118.1	**215.4**	**246.6**
Velvet	13.1	60.3	39.5	24.5	24.2	41.5	67.1

SPAdes

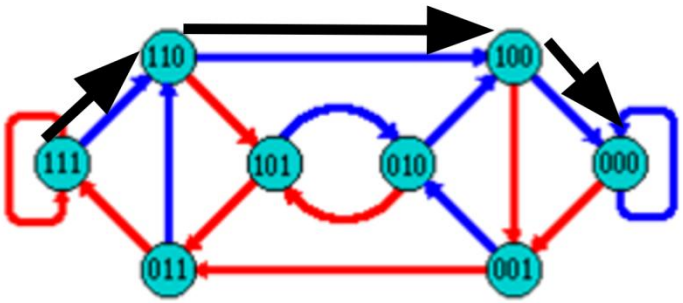
SPAdes Assembler 3.0 BETA
ALGORITHMIC BIOLOGY LAB

- Algoritmo basado en gráficas de Bruijn
  - Ruptura de reads en k-meros (secuencias cortas de k longitud, debe ser número impar)
  - Sobrelape de k-1 entre k-meros
  - Búsqueda de camino más corto

e.g. 111000 = 111 -> 110 -> 100 -> 000

```
1 1 1
  1 1 0
    1 0 0
      0 0 0
-----------
1 1 1 0 0 0
```

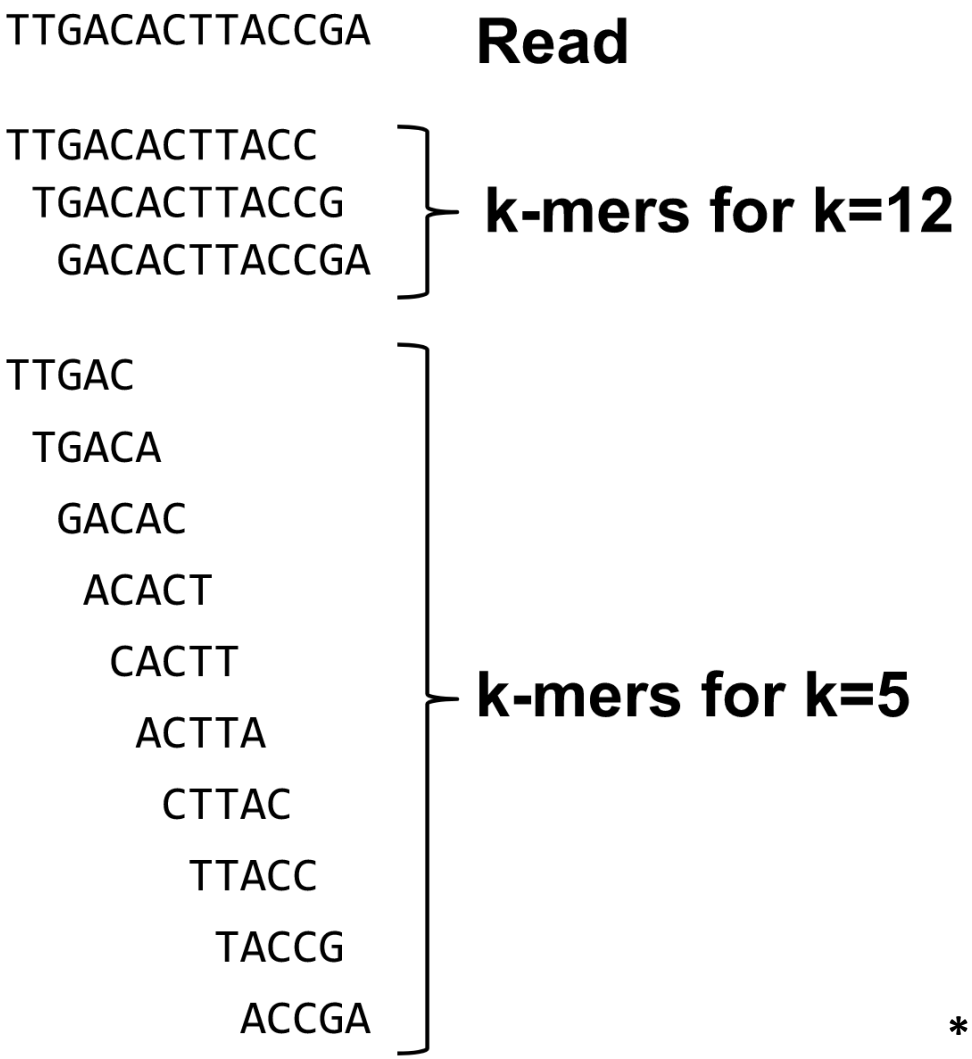
Gráfica para: **111000**
k = 3



Para qué usamos k-meros?

- Para poder usar gráficas de Bruijn es necesario que los reads sobrelapen L-1 pb. Sin embargo:
  - Por sí mismos, no todos los reads sobrelaparán de la misma manera debido a:
    - Errores de secuenciación
    - “Agujeros” de cobertura
  - No todos los reads tendrán la misma longitud, dependiendo de la tecnología utilizada y el paso de limpieza.

Diferencias entre tamaño de k-meros



K-meros grandes:
-gráfica con **menos conexiones**
-**mejor** oportunidad de resolver secuencias repetitivas
-**menor cobertura** k-mérica

K-meros pequeños:
-gráfica con **más conexiones**
-**menor** oportunidad de resolver secuencias repetitivas
-**mayor cobertura** k-mérica

***Mayor cobertura = menos errores**

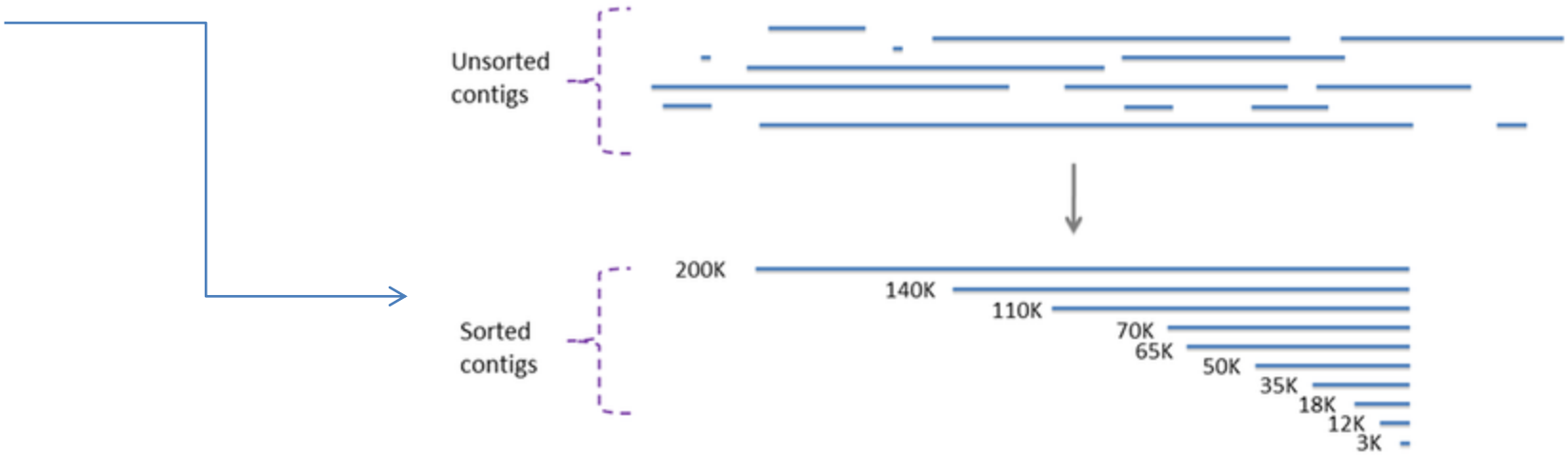
SPAdes

- Desde Ohta:
- $ cd Desktop/CURSO_BIOINFO/GENOMICA/2_ensamble/SPAdes-3.13.0-Linux/bin
- $ cat NOMBRE_*_unpaired.fq > NOMBRE_unpaired_all.fq
- $ python spades.py -1 NOMBRE_1_paired.fq -2 NOMBRE_2_paired.fq -s _NOMBRE_unpaired_all.fq -t 2 -o NOMBRE_ensamble
- $ cd NOMBRE_ensamble
- mv contigs.fasta NOMBRE_contigs.fasta

Control de calidad

- Tamaño de ensamble
- %GC
- Número de contigs
- Contig más grande

- **N50**



Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

50% total contig length= 703K x 50%= 351.5K

∵200K+140K +110K> 351.5K，∴N50= 110K

QUAST: Quality Assessment Tool for Genome Assemblies

- $ python quast.py -o NOMBRE_quast ../../2_ensamble/SPAdes-3.13.0-Linux/bin/NOMBRE_PRUEBA/NOMBRE_contigs.fasta
- $ cd NOMBRE_quast
- $ less report.txt

Anotación

- Búsqueda de marcos de lectura abiertos (ORF's).
- Predicción de genes:
  - Proteína = ATG……*STOP*
  - Sitios/dominios funcionales
- Uso de base de datos como referencia

- Anotadores populares:
  - PROKKA
  - Augustus
  - Glimmer

PROKKA

- $ $HOME/prokka/bin/prokka --outdir NOMBRE_prokka ../2_ensamble/SPAdes-3.13.0-Linux/bin/OUTPUT_PRUEBA/NOMBRE_contigs.fasta
- $ less PROKKA*.txt
- $ less PROKKA*.faa
- $ grep “phos” PROKKA*.faa

Filogenia de genomas

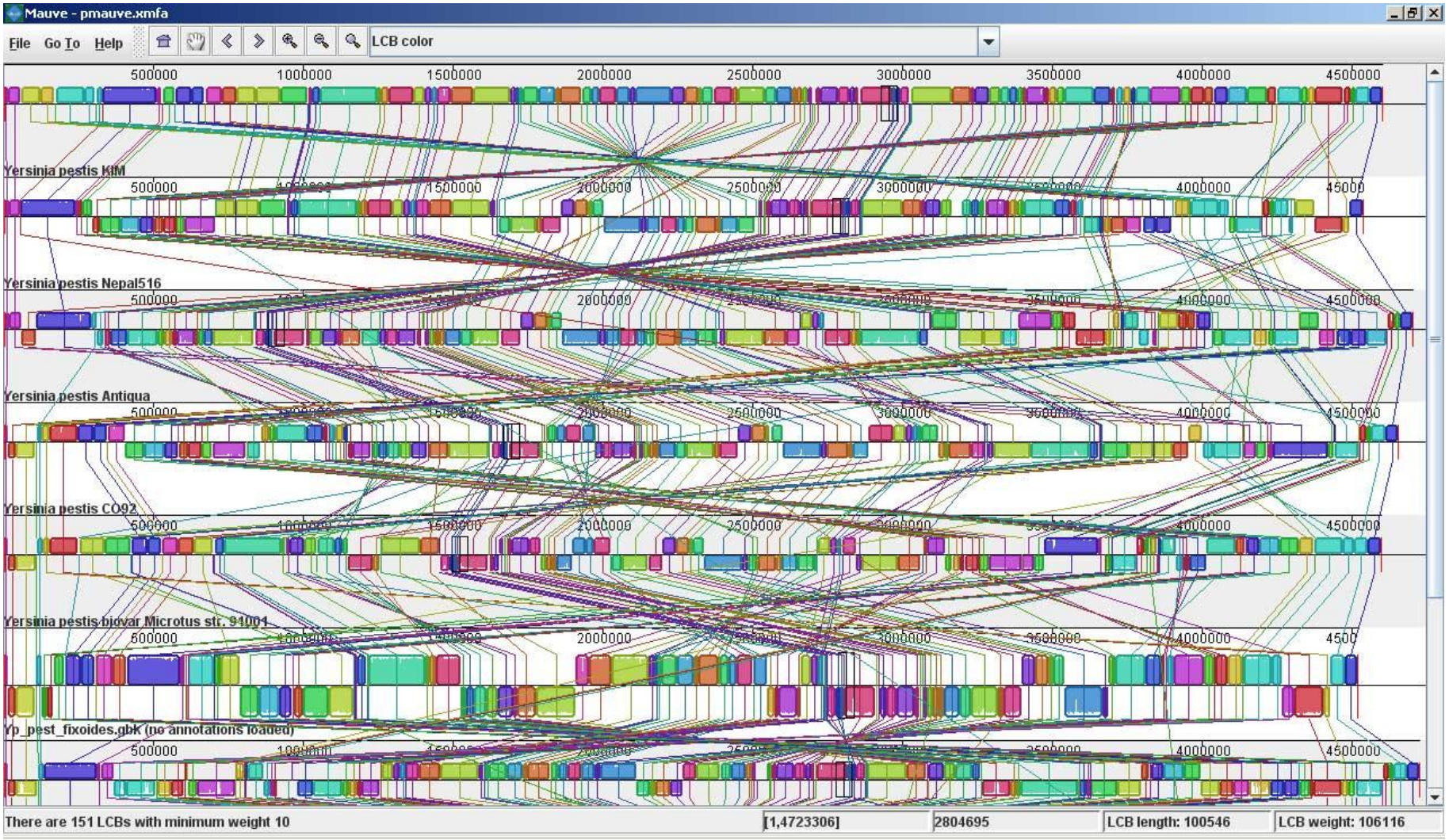
- Alineamiento de secuencias
- Remoción de genoma accesorio (la filogenia se hace con el core)
- Filogenia

Alineamiento de genomas y extracción de genoma core con progressiveMauve

- $ progressiveMauve --output=NOMBRE_mauve.xmfa ../2_ensamble/SPAdes-3.13.0-Linux/bin/OUTPUT_PRUEBA/NOMBRE_contigs.fasta ../DATOS/para_filogenia/*

- $ stripSubsetLCBs NOMBRE_mauve.xmfa NOMBRE_mauve.xmfa.bbcols NOMBRE_core.xmfa 500 6

- $ perl xmfa2fasta.pl --file NOMBRE_core.xmfa > NOMBRE_core.fasta


Mauve - pmauve.xmfa
File
Go To
Help
LCB color
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
Yersinia pestis KIM
Yersinia pestis Nepal516
Yersinia pestis Antiqua
Yersinia pestis CO92
Yersinia pestis biovar Microtus str. 91001
Yp_pest_fixoides.gbk (no annotations loaded)
There are 151 LCBs with minimum weight 10
[1,4723306]
2804695
LCB length: 100546
LCB weight: 106116

Filogenia con FastTree

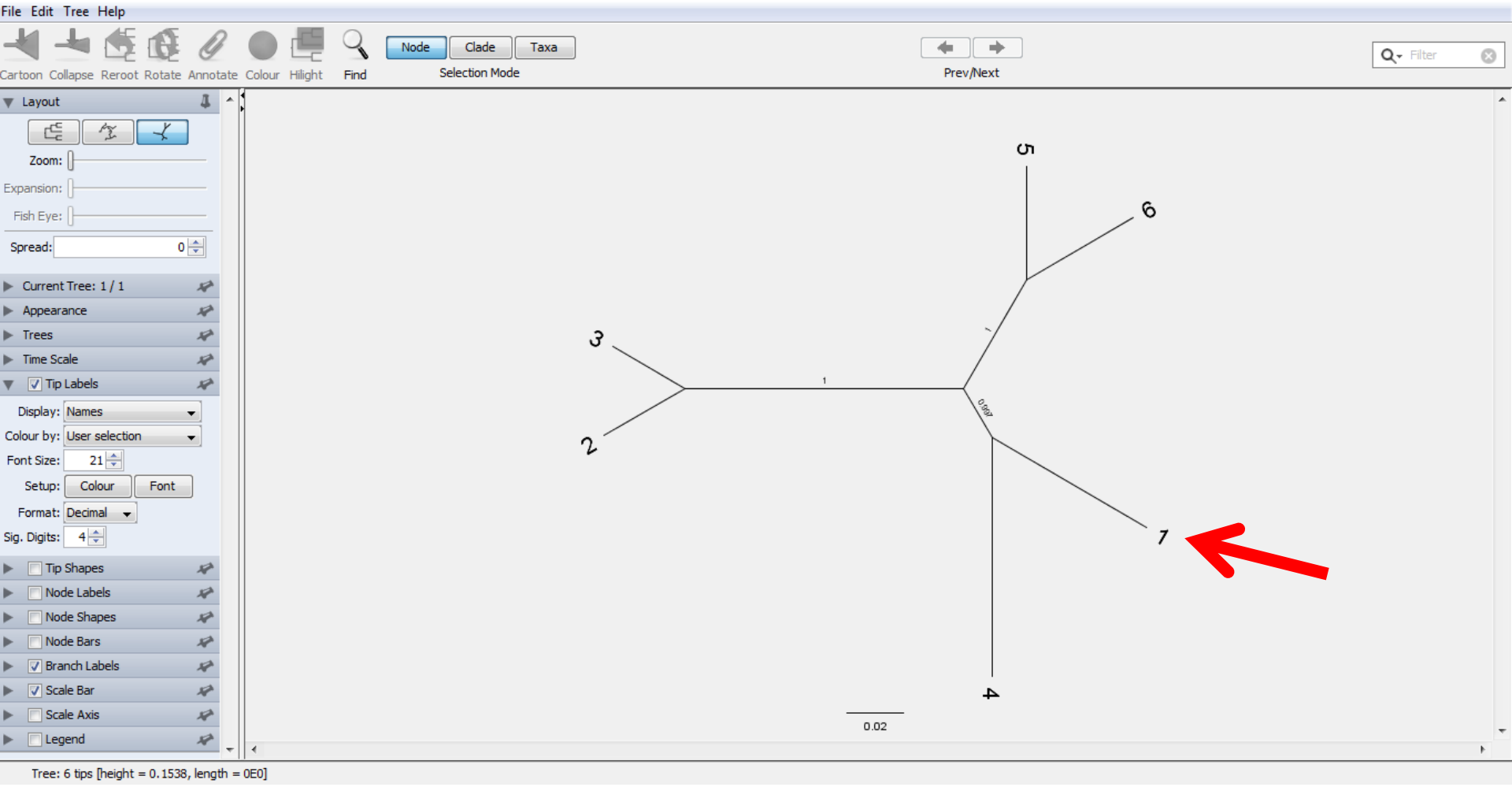
- $ ./FastTreeMP -nt -gtr < NOMBRE_core.fasta > NOMBRE_core.tree

Visualización

- https://github.com/rambaut/figtree/releases

- scp NOMBRE_core.tree a su computadora
- Abrir archivo NOMBRE_core.tree con figtree
- Activar “branch labels” y cambiar a “label”
- Cambiar tipo de filogenia a estrella

Filogenia de genomas core



Resultados

- Calidad de ensamble
  - Tamaño de ensamble
  - %GC
  - Número de contigs
  - Contig más grande
  - N50
- Anotación
  - Número de proteínas
  - Número de ORF’s
- Filogenia
  - Cercano a referencias?
  - Observaciones