CLASE 11

Mapeo / alineamiento de reads a genoma de referencia

DBT 792
GENÉTICA Y GENÓMICA EN PRODUCCIÓN ANIMAL

Profesor
Dr. José Gallardo

PLAN DE LA CLASE

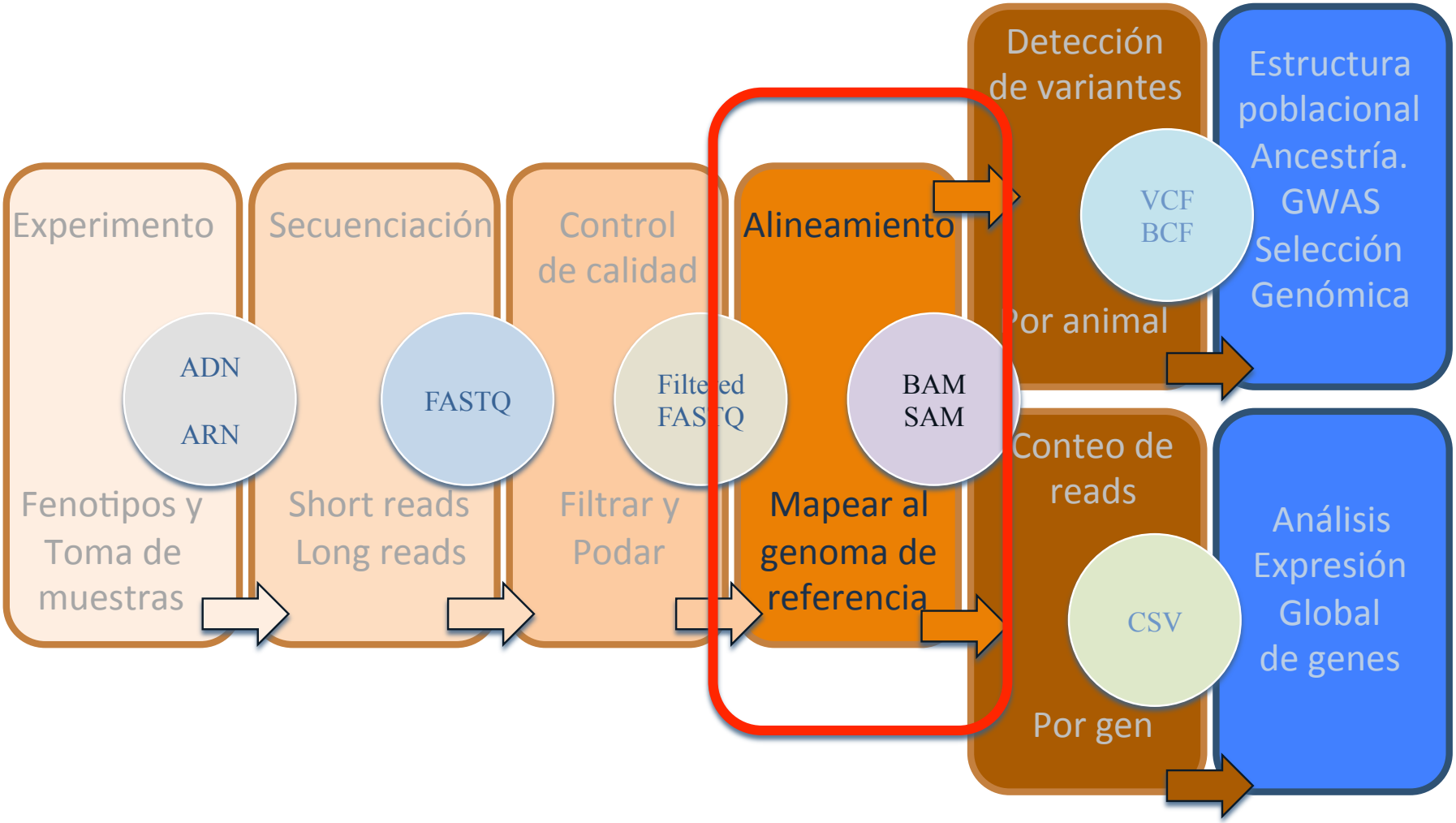
Introducción

- Repaso flujo de trabajo proyecto genómica aplicada.
- Secuencias paired-end y alineamiento.
- Formatos de alineamiento SAM / BAM
- Indexar secuencia del genoma de referencia para optimizar alineamiento: Software Bowtie 2.
- Alineamiento de reads a la referencia: Software Bowtie 2.
- Estadísticas básicas y visualización del alineamiento.

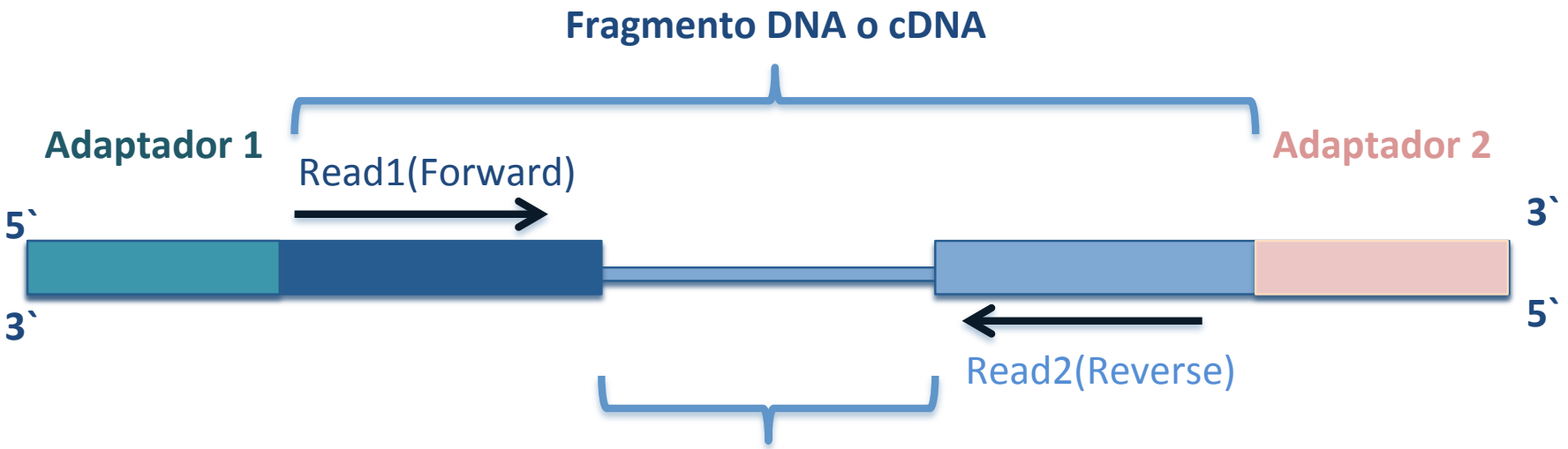
Práctica

- Linux para genómica: Mapeo y alineamiento a genoma de referencia.

FLUJO DE TRABAJO DE UN PROYECTO DE GENÓMICA APLICADA



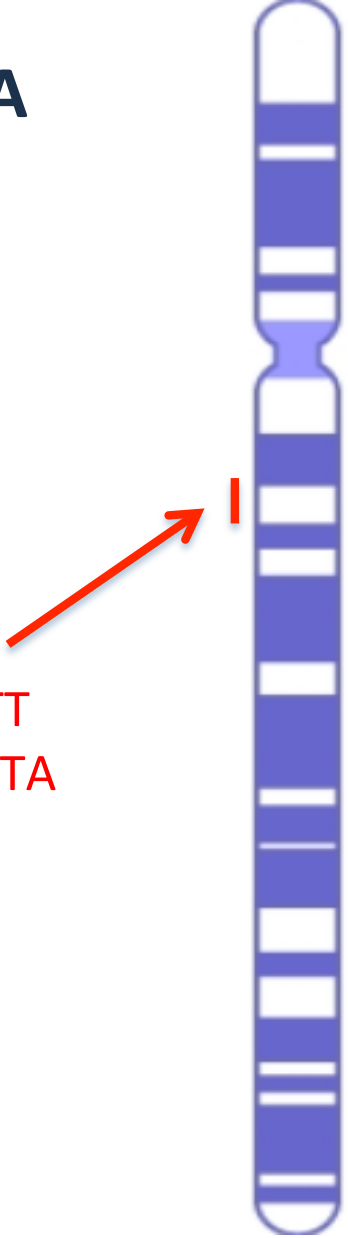
SECUENCIAS PAIRED-END



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

ALINEAMIENTO A UNA REFERENCIA

Principal problema a resolver: Para cada read, varias decenas de milllones en el forward y reverse, el alineador debe determinar el punto de origen o localización más probable en la referencia. La referencia puede contener en algunos casos billones de nucleótidos distribuidos en distintos cromosomas.



>gnl|SRA|**SRR5585856**.159293651.1 **159293651** (Biological)
CATTCNTTTTGAAGTTTTATTTTTTGNNTAAAATATTAATTGATCTCATTT
TGCTCAATATTCATTNGTCAACACTCATCATCAGTTTGCTGTCGCGGCTTA
AGAAGACAGAATGTGAAAAACACAGCAGATGAAATACAA

Salmo salar
RefSeq assembly accession: GCF_000233375.1
Total Length genoma Salmón: 2,966,873,538
Total Length Ssa8: 26,434,011

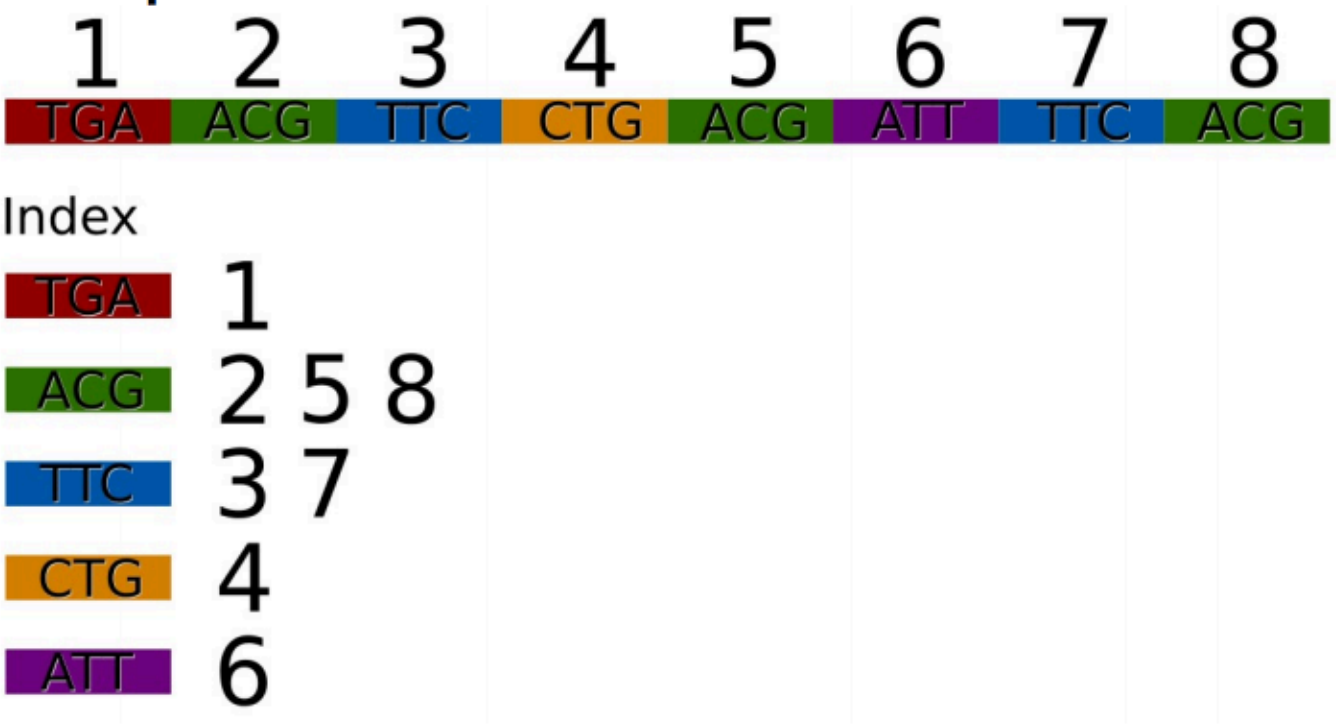
ALINEAMIENTO UNO A UNO: DEMASIADO LENTO.

- Naive approach:
  - Evaluate every location on the reference

- Too slow for billions of reads on a big reference

PASO CLAVE 1: CREAR UN ÍNDICE ACELERA EL PROCESO

- Speed up with the creation of a reference index

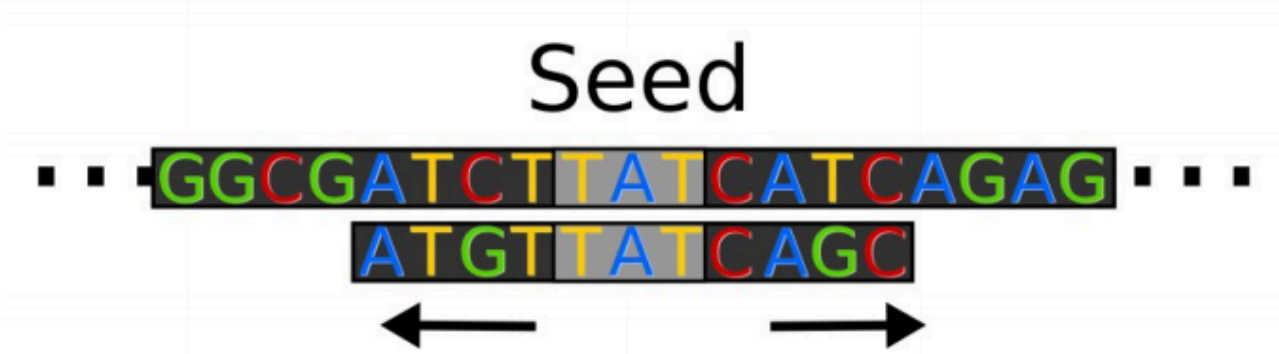


- Fast lookup table for subsequences in reference

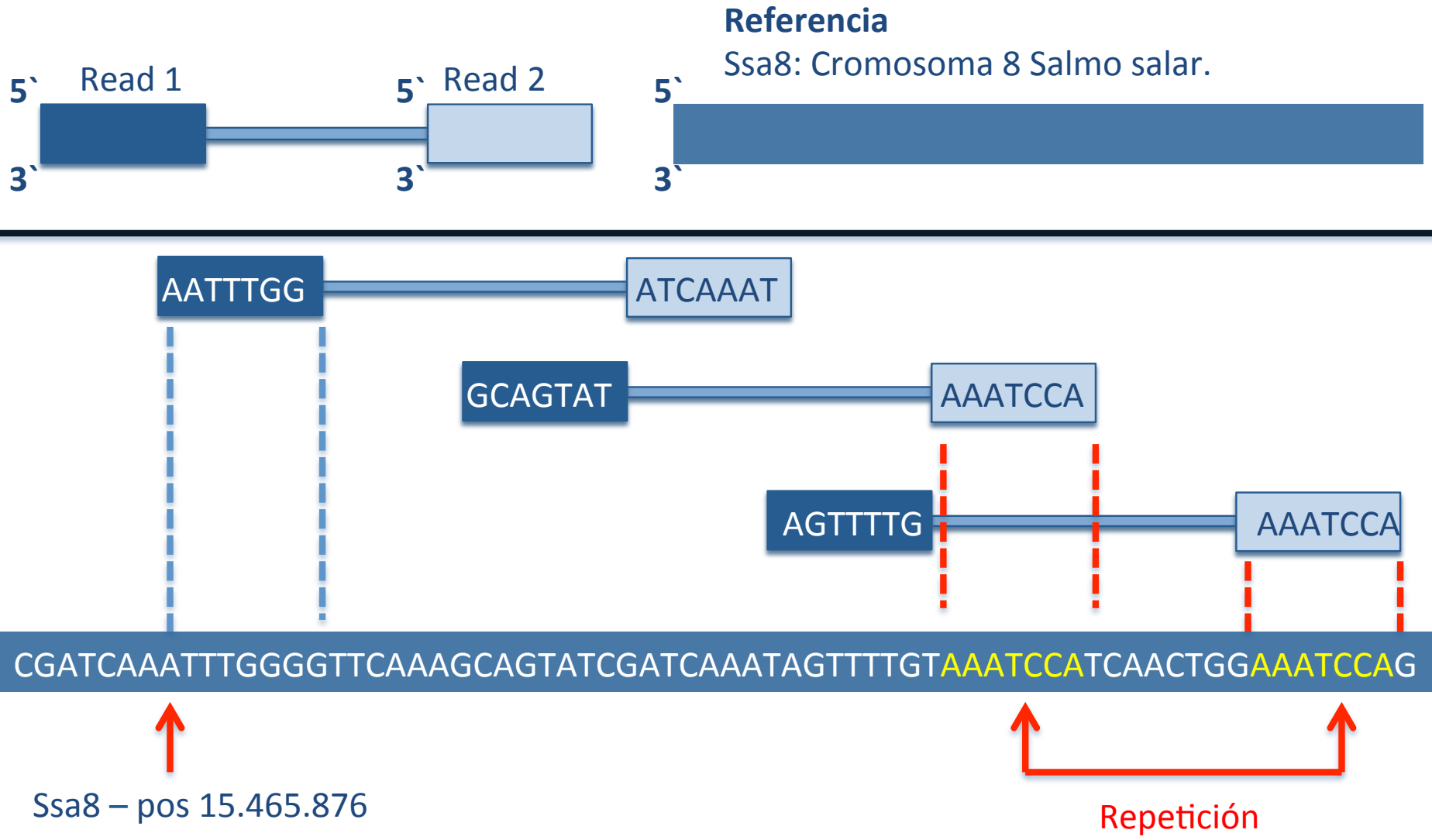
PASO CLAVE 2: SEMBRAR Y EXTIENDER CADA SEMILLA

- Find all possible alignment positions
  - Called seeds

- Evaluate every seed

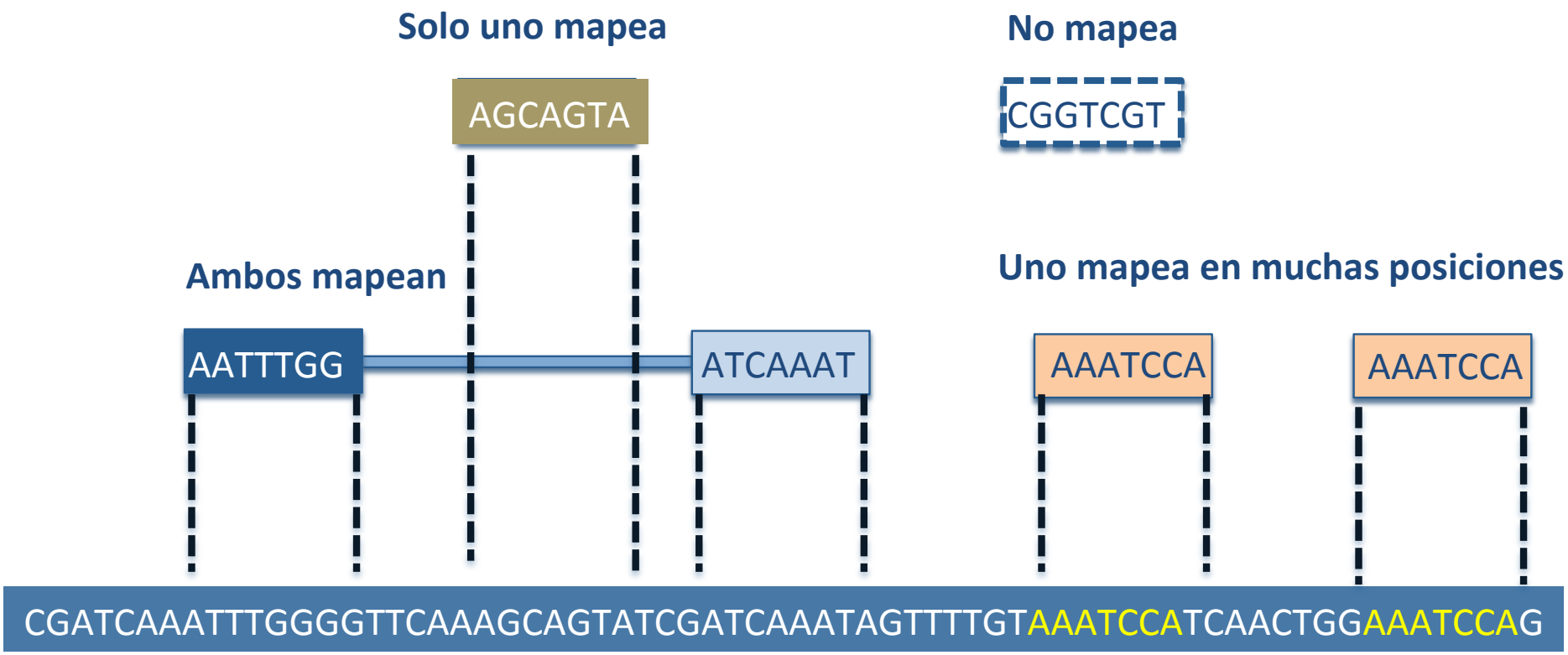


MEJORAR EL ALINEAMIENTO CON PAIRED-END READS:



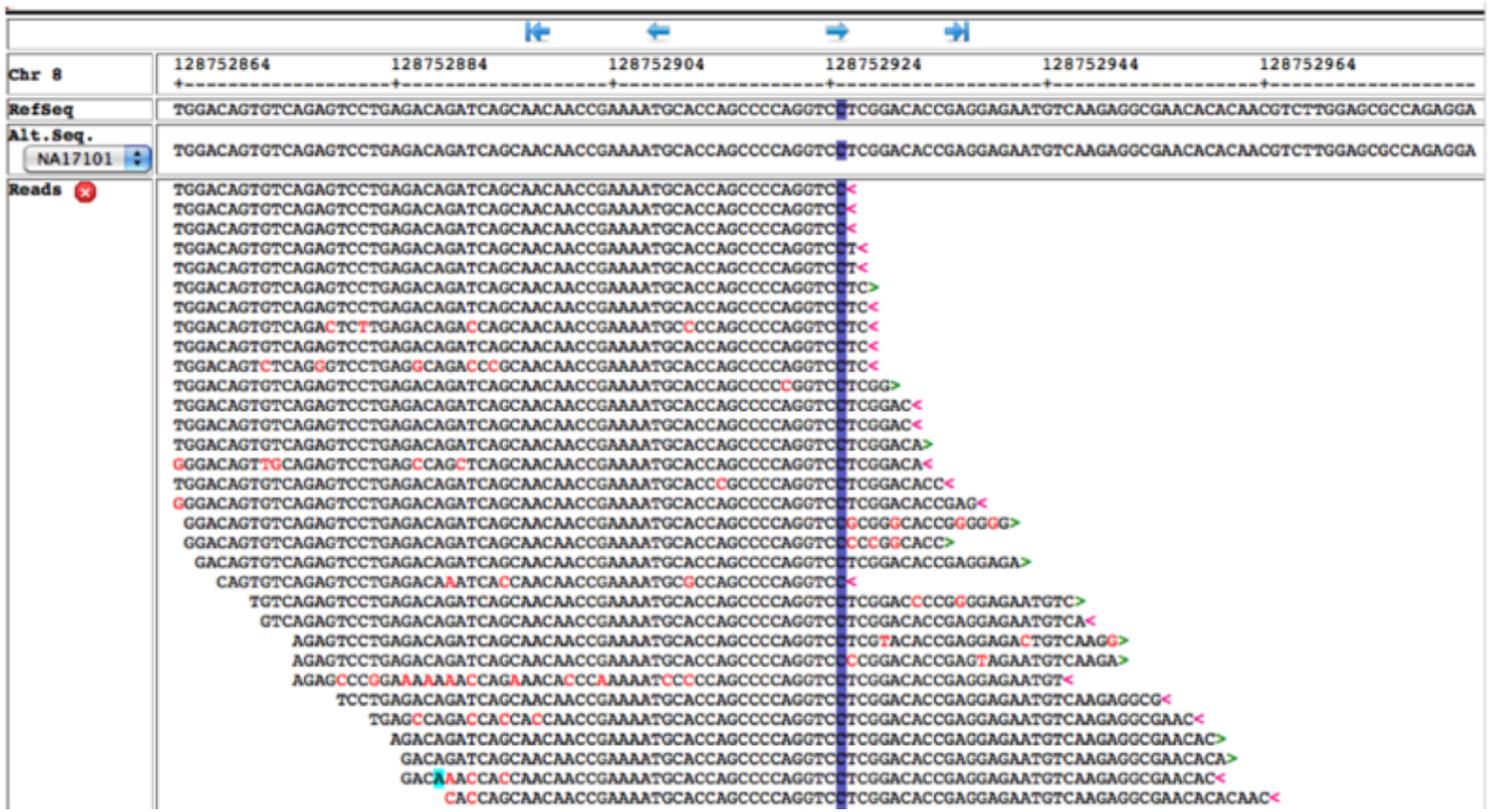
https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

RESULTADOS DE UN ALINEAMIENTO PE



FORMATO SAM / BAM

SAM: Sequence Alignment Map
BAM: Binary Alignment Map



https://www.samformat.info/sam-format-flag

Alineamiento a una referencia: Software BWA.

BWA (Burrows-Wheeler Aligner): Software que permite realizar un rápido alineamiento de reads a un genoma de referencia. Fue desarrollado por Ben Li y Durbin en 2009. Utiliza la transformación de Burrows–Wheeler y el índice FM para generar un índice sobre el cual buscar una secuencia, esto acelera el proceso de forma significativa.

Algoritmos

BWA-backtrack: diseñado para lecturas de secuencia de Illumina de hasta 100 pb.

BWA-SW y BWA-MEM: Diseñado para secuencias más largas entre 70 pb y 1 Mbp.

Recomendado: Se ha demostrado consistentemente que BWA-MEM es más rápido y más preciso superando BWA-backtrack para lecturas de Illumina de 70-100 pb.

Alineamiento con BWA.

Etapa 1: Indexar

bwa index genome.fasta

Etapa 2: Alieneamiento

bwa mem seq.fasta sample_1.fastq sample_2.fastq > sample.sam

Software Samtools

Conjunto de programas para explorar y analizar datos de secuenciación de alto rendimiento.

Consta de tres repositorios separados:

Samtools
Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format

BCFtools
Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants

HTSlib
A C library for reading/writing high-throughput sequencing data

Software Samtools: Funcionalidades clave.

Luego de obtener nuestro archivo de alineamiento SAM las tareas lógicas a desarrollar son:

1.- Convertir nuestro archivo .sam a .bam usando **samtools view**.
Los archivos .bam son mas livianos y contienen la misma información que los archivos .sam, pero algunas restricciones quizás apliquen.
Archivos de salida: muestra.bam

2.- Ordenar el alineamiento por coordenadas de la referencia usando **samtools sort**.
Archivos de salida: sorted.muestra.bam

3.- Indexar los arhivos "sorted.bam" usando **samtools index** para poder filtrar el alineamiento según la localización de un gen de interes o según la calidad que deseamos visualizar usando **samtools view**.
Archivos de salida: sorted.muestra.bam.bai

Software Samtools: Funcionalidades clave 2da parte.

4.- Obtener estadísticas del alineamiento.

Es posible realizar, analizar y comparar una gran diversidad de parámetros estadísticos conectando mediante tuberías comandos de diferentes software y shell. **Lo mejor es tomar los datos de salida y moverse rápido a gráficas de alto nivel en Python o R.**

Estadística estándar

samtools flagstat file.bam > muestra_stat.txt

Profundidad por locus y promedio de los reads.

samtools depth muestra.bam > muestra_covertura.txt

samtools depth –a muestra.bam | awk '{c++;s+=$3}END{print s/c}'

https://sarahpenir.github.io/bioinformatics/awk/calculating-mapping-stats-from-a-bam-file-using-samtools-and-awk/

OBJETIVOS DEL TRABAJO PRÁCTICO

Esta práctica tiene como propósito:

Realizar el alineamiento de una muestra en formato .fastq a un genoma de referencia.