Taller ENSAMBLAJE POR REFERENCIA

Estudiante: Francisco Salamanca

En este taller prenderás a alinear lecturas genómicas reales de Mycoplasma genitalium contra su genoma de referencia utilizando herramientas estándar en bioinformática como BWA y Samtools. Este proceso, conocido como ensamblaje por referencia, es clave en análisis de variantes, resecuenciación y estudios comparativos.

Objetivo:

Este taller tiene como propósito que usted desarrolle las habilidades necesarias para:

- Identificar y descargar un genoma de referencia desde NCBI.
- LLevar a cabo y comprender el proceso de indexación genómica mediante BWA.
- Alinear lecturas (reads) usando un genoma de referencia.
- Convertir, ordenar y filtrar alineamientos mediante Samtools.
- Obtener y reconocer las carcterísticas de un archivo de ensamblaje en formato BAM.

Dataset:

- Organismo: Mycoplasma genitalium.
- Datos: Lecturas Illumina MiSeq 2x150 bp (pareadas).
- Accesión: ERR486840 .
- Genoma de referencia de Mycoplasma genitalium (~580 kb)

Introducción

Mycoplasma genitalium Mycoplasma genitalium es una bacteria de pequeño tamaño perteneciente a la clase Mollicutes. Fue identificada por primera vez en la década de 1980 y es conocida por tener uno de los genomas más pequeños de todos los organismos autosuficientes (~580 kbp), lo que la convierte en un modelo interesante para estudios de biología mínima y genómica sintética. Desde el punto de vista clínico, M. genitalium es un patógeno de transmisión sexual (ETS) que puede causar infecciones tanto en hombres como en mujeres. En hombres, se asocia comúnmente con uretritis no

gonocócica, mientras que en mujeres puede provocar cervicitis, enfermedad inflamatoria pélvica, y se ha vinculado con complicaciones reproductivas.

Algunas características relevantes de M. genitalium son:

Carece de pared celular, lo que la hace intrínsecamente resistente a antibióticos β-lactámicos como la penicilina.

Crecimiento lento y difícil en cultivo, lo que complica su diagnóstico clínico.

Alta tasa de resistencia a antibióticos, especialmente a macrólidos (como la azitromicina) y fluoroquinolonas, lo que plantea un desafío en su tratamiento.

Su importancia clínica ha ido en aumento en los últimos años, siendo reconocida como una ETS emergente y motivo de preocupación por su resistencia creciente y su impacto en la salud sexual y reproductiva.

Desarrollo del taller

Parte 0: Copia de Archivos

Enlace los archivos originales en su espacio de trabajo creando un enlace simbòlico de los mismos:

ln -s /datos/resources/examples/rawdata/mycobacterium/ERR486840_1.fastq ERR486840_1.fastq

ln -s /datos/resources/examples/rawdata/mycobacterium/ERR486840_2.fastq ERR486840_2.fastq

In [ ]:
```
#Crear accesos directos a los datos de ejemplo de Mycobacterium
tuberculosis

!ln -s   /datos/resources/examples/rawdata/mycobacterium/ERR486840_1.fastq
ERR486840_1.fastq

!ln -s   /datos/resources/examples/rawdata/mycobacterium/ERR486840_2.fastq
ERR486840_2.fastq
```

Parte 1: Descargar el genoma de referencia

Un genoma de referencia es una secuencia genómica conocida que se usa como modelo o "molde" para guiar el alineamiento de lecturas nuevas. En el caso de Mycoplasma genitalium, su genoma es pequeño (~580 kb) y está completamente secuenciado, lo que lo hace ideal para análisis de referencia.

Determinar cual genoma de referencia usar es una tarea fundamental en el proceso de ensamblaje y no siempre es un proceso directo que sea fácil de llevar a cabo, pues requiere estar seguro de elegir el mejor genoma posible para llevar a acabo el alineamiento lo cual requiere de un conocimiento mínimo de la biología del organismo a ensamblar.

Para este taller obtenga el genoma de referencia del NCBI en el enlace proveido:

https://www.ncbi.nlm.nih.gov/datasets/taxonomy/2097/

Responda las preguntas asociadas:

https://docs.google.com/forms/d/e/1FAIpQLSfisjTZ-N1TSEKa3V7Jdu7-s24sthkH4s4SPxu4flvUY_achQ/viewform?hr_submission=ChkI9uvrh-QBEhAIkoK8vOQWEgcI08PbkqsUEAA

In [ ]:
```
# Se usara el archivo GCF, debido a que tiene una curaduria previa.
# El archivo GCA es el genoma completo sin curaduria previa.
# NOTA: En adelante el genoma de referencia será llamado
mycoplasma_ref.fna
```

Pregunta 1: ¿Cuales de estas son características correctas del genoma de referencia de Mycoplasma genitalium usado en el taller?

R/ Para responder esta pregunta se revisan los archivos assembly_data_report.jsonl, el cual contiene las estadisticas detalladas del ensamblaje. Al igual que el archivo data_summary.tsv.

Contig N50: 580.1Kb; Genome Coverage de 533X, Strain G37

Parte 2: Indexación del genoma con BWA

El proceso de indexación es fundamental en bioinformática y es usado en casi todos los procesos de aceleración de alineamientos, como en BLAST y en ensamblajes de genoma completo. En este caso, antes de poder alinear lecturas, BWA necesita preparar (indexar) el genoma para búsquedas rápidas y eficientes. Este proceso genera estructuras de datos internas que permiten alinear millones de lecturas sin recorrer el genoma secuencia por secuencia.

¿Qué es BWA? BWA (Burrows-Wheeler Aligner) es un alineador de lecturas cortas contra genomas de referencia. Es rápido, eficiente y ampliamente usado en pipelines genómicos.

Realice la indexación usando el siguiente comando y responda las preguntas asociadas en el formulario presentado anteriormente.

In [ ]:
```
#!bwa index mycoplasma_ref.fna

"""(base) fjosesala@Franciscos-MacBook-Air GCF_040556925.1 % bwa index
GCF_mycoplasma_ref.fna
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.08 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forward-only FASTA... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0.02 sec
[main] Version: 0.7.19-r1273
[main] CMD: bwa index GCF_mycoplasma_ref.fna
[main] Real time: 0.113 sec; CPU: 0.133 sec
(base) fjosesala@Franciscos-MacBook-Air GCF_040556925.1 % """
```

Pregunta 2: Salmon y Bowtie son ensambladores populares que se usan muchas veces como alternativas a BWA ya que permiten también realizar ensamblajes de genoma completo.

R/ FALSO, pues en el caso de bowtie, son alineadores (mapean contra un genoma de referencia, no ensambla desde 0). En el caso de Salmon, es una herramienta que permite secuenciar la expresion genica (transcriptomas), a partir de RNA-seq.

Pregunta 3: En el proceso de indexación del genoma de referencia usando BWA se crearon 5 arhivos con las siguientes extensiones:

R/ amb, ann, .bwt, .pac y .sa

Parte 3: Alineamiento de las lecturas al genoma

bwa mem es un "subcomando" de BWA para lecturas tipo Illumina (paired-end). Usa el algoritmo de Burrows-Wheeler para buscar coincidencias entre las lecturas y la referencia. Este es básicamente el comando que nos permite ubicar la posición de cada uno de los reads en el genoma.

In [ ]:
```
#Correr bwa mem para llevar a cabo el alineamiento:

!bwa mem GCF_mycoplasma_ref.fna ERR486840_1.fastq ERR486840_2.fastq >
alineamiento.sam
```

Examine el archivo SAM obtenido. ¿Que peso tiene este archivo? Asegurese de tener claridad acerca de qué es un archivo SAM y cual es su importancia en Bioinformática, particularmente en el ensamblaje de genomas.

R/ Un Archivo SAM, se usa para almacenar las alineaciones de secuencias de lectura corta (reads), contra un genoma de referencia. Se estructura con un header, que contiene metadatos y una seccion de alineamiento, donde cada fila representa una secuencia de lectura.

Pregunta 4: El archivo SAM obtenido consta de alrededor de cuantas secuencias alineadas?

R/ Más de 780 mil secuencias alineadas

In [ ]:
```
!grep -v '^@' alineamiento.sam | wc

"""782390"""

!ls -lh alineamiento.sam

"""326.8 MB"""
```

Parte 4: Conversión y procesamiento del archivo de alineamiento

El formato SAM (Sequence Alignment/Map) es un archivo de texto plano que contiene cada uno de los alineamientos de las lecturas al genoma de referencia. Es legible por humanos, pero muy pesado y lento de procesar. Por esta razón una práctica común es la de convertir este archivo SAM a un formato más ligero y compacto sin perdidad de información denominado BAM.

In [ ]:
```
!samtools view -S -b alineamiento.sam > alineamiento.bam

!ls -lh alineamiento.bam

"""51.6 MB"""
```

¿Nota alguna diferencia de tamaño? ¿Para qué sirven las opciones -S y -b?

Si, El archivo BAM pesa mucho menos (51.6mb).

- b: Es una opción que le indica a samtools que la salida debe ser en formato BAM.
- S: Es una opción que le indica a samtools que la entrada está en formato SAM.

Parte 5: Ordenar el archivo BAM

Muchos programas, como samtools index o visualizadores como IGV, requieren que los alineamientos estén ordenados de acuerdo a las coordenadas del genoma. Esto también mejora el acceso aleatorio y la eficiencia en análisis posteriores. Por esta razón el ordenamiento de los alineamientos es clave para acelerar cualquier proceso posterior pues se pasa de tener los alineamientos en el archivo SAM y BAM tal y como se fue realizando el alineamiento, a ordenarlos de acuerdo a su posición en el genoma.

In [ ]:
```
!samtools sort alineamiento.bam -o alineamiento_sorted.bam
```

Parte 6: Indexar el archivo BAM

Aunque no es estrictamente necesario, la indexacion del archivo BAM es un proceso rutinario, ya que es requerido por muchos programas de visualización. La indexación genera un archivo .bai que permite:

- Acceso rápido a regiones específicas del genoma.
- Visualización eficiente en programas como IGV.
- Cálculos rápidos de cobertura, variantes, etc.

In [ ]:
```
!samtools index alineamiento_sorted.bam
```

Parte 7: Visualización del ensamblaje mediante Tablet

Aunque el archivo BAM contiene toda la información de las lecturas alineadas al genoma, es difícil de interpretar directamente desde la línea de comandos. Por eso, utilizamos una herramienta gráfica (Tablet) que permitirá:

- Ver cómo se alinean las lecturas una a una sobre la referencia.
- Identificar regiones de baja cobertura, duplicaciones o errores sistemáticos.
- Validar visualmente variantes y errores de secuenciación.
- Comprender de forma intuitiva la relación entre datos crudos y el genoma de referencia.

Pregunta 5: Captura de pantalla de la cobertura de sus reads en el genoma de referencia.

 No description has been provided for this image

Identifique regiones de alta o baja cobertura. Cree que existen SNPs?

R/ Si, se hallan ajustando el parametro variants

Hay fallas de secuenciación en algunos reads?

R/ No parece haber fallas de secuenciacion, pues no se percibe cambios bruscos en los reads ni gaps extremos.