Bioinformatica

Para realizar la siguiente practica por favor use el servidor Galaxy en el servidor de la union europea, disponible en el siguiente link, https://usegalaxy.eu (por favor no use https://usegalaxy.org, porque puede presentar incovenientes durante la practica al usar algunas herramientas), el sistema pedirá que se registre nuevamente si no se habia registrado antes.

Practica de Ensamblaje de genomas

El ensamblaje de genomas es el proceso mediante el cual se reconstruye un genoma completo a partir de fragmentos de ADN obtenidos a través de técnicas de secuenciación. En términos más sencillos, se trata de ensamblar pequeñas piezas de información genética (lecturas de secuencias de ADN) en un genoma completo.

Cuando se realiza la secuenciación de un genoma, la tecnología moderna divide el ADN en fragmentos más pequeños, que se leen y generan secuencias de bases (A, T, C, G). Estas lecturas deben ser organizadas y alineadas correctamente para crear una representación precisa y continua del genoma original. El proceso de ensamblaje puede llevarse a cabo con diferentes enfoques, como:

- **Ensamblaje de novo**: Se realiza cuando no se tiene un genoma de referencia, y el objetivo es reconstruir el genoma desde cero a partir de las lecturas sin ninguna guía externa.
- **Ensamblaje de referencia**: Se realiza cuando se tiene un genoma de referencia conocido, y las lecturas se alinean y ensamblan sobre este genoma para obtener la secuencia específica del organismo en estudio.

Para esta practica, tenemos un conjunto de lecturas de una bacteria *Staphylococcus aureus* imaginaria con un genoma en miniatura (197.394 pb). Nuestro conjunto de lecturas de la cepa mutante se secuenció con el método de escopeta de genoma completo, utilizando un instrumento de secuenciación de ADN Illumina. A partir de estas lecturas, nos gustaría reconstruir nuestra bacteria imaginaria *Staphylococcus aureus* mediante un ensamblaje de novo de un conjunto de lecturas cortas utilizando el ensamblador Velvet. Velvet es uno de los numerosos ensambladores de novo que utilizan conjuntos de lecturas cortas como entrada (por ejemplo, Illumina Reads). El método de ensamblaje se basa en la manipulación de gráficos de de Bruijn, mediante la eliminación de errores y la simplificación de regiones repetidas.

Las lecturas han sido secuenciadas a partir de una bacteria imaginaria *Staphylococcus aureus* utilizando un instrumento de secuenciación de ADN Illumina. Obtuvimos los 2 archivos que importamos (`mutant_R1` y `mutant_R2`)

1. Cree y nombre un nuevo historial para este tutorial. En la parte superior derecha de sus pantallas encotnraran un simbolo (+) donde podran crear una nueva historia y dando click en el simbolo de lapiz podra editar el nombre

Importa desde Zenodo o desde la biblioteca de datos los archivos. Para esto de click en `Upload` en la parte superior izquierda de las pantallas. Alli luego mantenga la pestaña `regular` y en la parte inferior del lado derecho de click en `Paste/Fecth data`. Luego copie los link de abajo y continuar. Cierre la ventana y luego en la parte derecha podra ver que se estan cargando los archivos, espere a que este en verde. Esto indica que ya esta listo. Naranja indica que esta en proceso y rojo que no se pudo realizar y debe repetir el cargue.

```
https://zenodo.org/record/582600/files/mutant_R1.fastq
https://zenodo.org/record/582600/files/mutant_R2.fastq
https://zenodo.org/record/582600/files/wildtype.fna
```

2. Inspeccione el contenido de un conjunto de datos

   Pregunta

   a. ¿Cuáles son las cuatro características principales de un archivo FASTQ?

   b. ¿Cuál es la principal diferencia entre un archivo FASTQ y un archivo FASTA?

Evaluar las lecturas de entrada

Antes de realizar cualquier ensamblaje, las primeras preguntas que debe plantearse sobre las lecturas de entrada son:

- ¿Cuál es la cobertura de mi genoma?
- ¿Qué calidad tiene mi conjunto de lecturas?
- ¿Necesito solicitar una nueva secuencia?
- ¿Es adecuada para el análisis que necesito hacer?

Evaluaremos las lecturas de entrada con la herramienta **FastQC**. Esta herramienta ejecuta una serie de pruebas estándar en su conjunto de lecturas y devuelve un informe relativamente fácil de interpretar. La utilizaremos para evaluar la calidad de nuestros archivos FASTQ y combinar los resultados con MultiQC.

3. **FastQC** (Galaxy version 0.73+galaxy0) con los siguientes parametros:

   - "Raw read data from your current history"*: `mutant_R1.fastq` and `mutant_R2.fastq`

4. **MultiQC** (Galaxy version 1.11+galaxy1) con los siguientes parametros:

   - "Results: Which tool was used to generate logs?": `FastQC`
   - Click "Insert FastQC output"
     - "Type of FastQC output?":`multiple datasets, select the raw data files from FastQC`

```
MultiQC genera una página web que combina informes para FastQC en ambos
conjuntos de datos. Incluye estos gráficos y tablas:

 - Estadísticas generales

Necesitamos conocer los datos para nuestro análisis. En
particular, necesitamos conocer las longitudes de lectura, ya
que es importante para establecer el tamaño máximo de k-mer para
un ensamblaje. Para obtener la longitud de las secuencias:

a. Busque la salida MultiQC que es una página web y haga clic
para verla

b. La primera tabla muestra las estadísticas generales de los
archivos de lectura de entrada.

c. En la parte superior de esta tabla, haga clic en Configurar
columnas

d. Asegúrese de que la casilla junto a Longitud está marcada
```

```
e. Cierre la ventana

f. Esta tabla debería mostrar ahora una columna para las
longitudes de lectura

Pregunta

¿Qué longitud tienen las secuencias?

¿Cuál es la cobertura media del genoma, teniendo en cuenta que
nuestra bacteria imaginaria \*Staphylococcus aureus\* tiene un
genoma de 197.394 pb?

 - Histogramas de calidad de la secuencia

Las caídas en la calidad cerca del principio, el medio o el
final de las lecturas pueden determinar los métodos y parámetros
de recorte/limpieza que deben utilizarse, o pueden indicar
problemas técnicos con el proceso de secuenciación/la ejecución
de la máquina.

Pregunta

¿Qué representa el eje y?

¿Por qué disminuye la puntuación de calidad a lo largo de las
lecturas?

 - Contenido GC por secuencia

Los organismos con alto GC tienden a no ensamblarse bien y
pueden tener una distribución desigual de la cobertura de
lectura.

 - Contenido de N por base

La presencia de un gran número de Ns en las lecturas puede
indicar un proceso de secuenciación de baja calidad. Deberá
recortar estas lecturas para eliminar los Ns.

 - Contenido de k-mer

La presencia de k-mers muy recurrentes puede indicar
contaminación de las lecturas con códigos de barras o secuencias
adaptadoras.
```

Ensamblar lecturas con Velvet

Ahora queremos ensamblar nuestras lecturas para encontrar la secuencia de nuestra bacteria imaginaria *Staphylococcus aureus*. Realizaremos un ensamblaje de novo de las lecturas en secuencias largas contiguas utilizando el ensamblador de lecturas cortas Velvet.

El primer paso del ensamblador es construir un grafo de Bruijn. Para ello, dividirá nuestras lecturas en k-mers, es decir, fragmentos de longitud k. Velvet requiere que el usuario introduzca un valor de k (tamaño

k-mer) para el proceso de ensamblaje. Los k-mers pequeños darán mayor conectividad, pero los k-mers grandes darán mayor especificidad.

5. **FASTQ interlacer** (Galaxy version 1.2.0.1+galaxy0) con los siguientes parametros:

    - "Type of paired-end datasets": `2 separate datasets`
    - "Left-hand mates": `mutant_R1.fastq`
    - "Right-hand mates": `mutant_R2.fastq`

Actualmente, nuestras lecturas emparejadas están en 2 archivos (uno con las lecturas hacia delante y otro con las lecturas hacia atrás), pero Velvet sólo necesita un archivo, en el que cada lectura esté junto a su lectura emparejada. En otras palabras, si las lecturas están indexadas desde 0, entonces las lecturas 0 y 1 están emparejadas, 2 y 3, 4 y 5, etc. Antes de hacer el ensamblaje propiamente dicho, tenemos que preparar los archivos combinándolos.

6. **velveth** (Galaxy version 1.2.10.3) con los siguientes parametros:

    - "Hash Length": `29`
    - "Input Files"
        - Click on + "Input Files"
            * In "1: Input Files"
                · "Choose the input type": `interleaved paired end`
                · "read type": `shortPaired reads param-files`
                · "Dataset": `pairs output of *FASTQ interlacer*`

La herramienta toma nuestras lecturas y las descompone en k-mers.

7. **velvetg** (Galaxy version 1.2.10.2) con los siguientes parametros:

    - param-files "Velvet Dataset": `outputs of velveth`
    - "Using Paired Reads": `Yes`

Esta última herramienta realiza realmente el ensamblaje.

Se generan cinco archivos. Veremos el archivo **contigs** y el archivo **stats**:

- El archivo **contigs**

Este archivo contiene las secuencias de los contigs. En la cabecera de cada contig, se añade un poco de información:

la longitud k-mer (llamada «length»): Para el valor de k elegido en el ensamblaje, una medida de cuántos k-mers se solapan (en 1 pb cada solapamiento) para dar esta longitud la cobertura de k-mers (denominada «cobertura»): Para el valor de k elegido en el ensamblaje, una medida de cuántos k-mers se solapan cada posición de base (en el ensamblaje).

- El archivo de **stats**

Se trata de un archivo tabular que proporciona para cada contig las longitudes de k-mer, las coberturas de k-mer y otras medidas. Tenga en cuenta que sus resultados pueden diferir del ejemplo de la imagen siguiente.

Recopilar algunas estadísticas sobre los contigs

Pregunta ¿Cuántos contigs se han construido? ¿Cuál es la longitud media, mínima y máxima de los contigs?

Esta tabla es limitada, pero ahora recopilaremos estadísticas más básicas sobre nuestro ensamblaje.

8. **Quast** (Galaxy version 5.2.0+galaxy1) con los siquientes parametros:

    - "Assembly mode": `Individual assembly (1 contig file per sample)`
    - "Use customized names?": `No`
    - "Contigs/scaffolds file": `contigs output of velvetg`
    - "Type of assembly": `Genome`
    - "Use a reference genome?": `Yes`
    - "Reference genome": `wildtype.fna`
    - "Type of organism": `Prokaryotes`
    - "Lower Threshold": `500`
    - "Advanced options: Comma-separated list of contig length thresholds": `0,1000`

Preguntas

¿Cuántos contigs se han construido?

¿Qué proporción del genoma de referencia representan?

¿Cuántos montajes erróneos se han encontrado?

¿Ha introducido el ensamblaje desajustes e indels?

¿Qué son N50 y L50?

¿Existe un sesgo en el porcentaje de GC inducido por el ensamblaje?

Trabajo para analizar

Corran nuevamente pero en el paso 6

6. *velveth* (Galaxy version 1.2.10.3) con los mismos parametros de antes pero prueben

- "Hash Length": un valor entre 31 y 101

Practica de Anotación de genomas

La anotación genómica consiste en describir la estructura y función de los componentes del genoma, prediciéndolos, analizándolos e interpretándolos para extraer su significado biológico y comprender los procesos biológicos en los que participan. Entre otras cosas, identifica la localización de los genes y todas las regiones codificantes de un genoma (anotación estructural) y determina qué hacen esos genes (anotación funcional).

Para ilustrar el proceso de anotación de un genoma bacteriano, tomamos un ensamblaje de un genoma bacteriano (muestra KUN1163) generado siguiendo un tutorial de ensamblaje de genomas bacterianos a partir de los datos producidos en «Complete Genome Sequences of Eight Methicillin-Resistant Staphylococcus aureus Strains Isolated from Patients in Japan» (Hikichi et al. 2019).

Staphylococcus aureus resistente a la meticilina (MRSA) es un patógeno importante que causa infecciones nosocomiales, y las manifestaciones clínicas de MRSA van desde la colonización asintomática de la mucosa nasal hasta la infección de tejidos blandos y la enfermedad invasiva fulminante. Aquí presentamos las secuencias genómicas completas de ocho cepas de SARM aisladas de pacientes en Japón.

Cargue los datos

Importe el archivo contig desde Zenodo o desde las bibliotecas de datos compartidos Galaxy:

`https://zenodo.org/record/10572227/files/DRR187559_contigs.fasta`

Anotación de contigs

Para anotar los contigs, existen varias herramientas para hacerlo: **Prokka** (Seemann 2014), **Bakta** (Schwengers et al. 2021), etc. Aquí, utilizamos Bakta como recomendado por Torsten Seemann avatar Torsten Seemann como el sucesor de Prokka.

Bakta es una herramienta para la anotación rápida y estandarizada de genomas bacterianos y plásmidos tanto de aislados como de genomas ensamblados de metagenomas (MAGs). Implementa un flujo de trabajo de anotación exhaustivo para genes codificantes y no codificantes (es decir, ARNt, ARNr).

1. En el panel de herramientas busque `Bakta` y ajuste los siguientes parametros:
    - En "Input/Output options":
        - "Select genome in fasta format": DRR187559_contigs.fasta
        - "Bakta database": V5.1_2024-01-19
        - "AMRFinderPlus database": V3.12-2024-05-02.2
    - En "Optional annotation":
        - "Keep original contig header": Yes
    - En "Selection of the output files":
        - "Output files selection":
            1. Annotation file in TSV
            2. Annotation and sequence in GFF3
            3. Feature nucleotide sequences as FASTA
            4. Summary as TXT
            5. Plot of the annotation result as SVG

Responda las siguientes preguntas

En el archivo `Analisis summary`

- ¿Cuantos contigs habian en el input?
- ¿Cuan largo es el draft genome?
- ¿Cuantos CDSs fueron encontrados?
- ¿Cuantas small proteins?
- ¿Que otros componentes fueron encontrados?
- ¿Si compara los resultados obtenidos con aquellos para KUN1163 en la Tabla 1 en Hikichi et al. 2019, que tal va?

En el archivo `nucleotide_sequences`

- ¿Cuantas secuencias hay en el archivo?
- ¿Cuales secuencias hay almacenadas ahi?

En el archivo `annotation_summary`

- ¿Que hay almacenado ahi?

En el archivo `Annotation_and_sequences`

- ¿Que hay almacenado ahi?

En el archivo `SVG`

- ¿Que significan los dos anillos en el centro?
- ¿Que significan los dos anillos de color gris?

Anotación adicional

Plasmidos

Para identificar plásmidos en nuestros contigs, utilizamos PlasmidFinder (Carattoli y Hasman 2020), una herramienta para la identificación y tipificación de secuencias de plásmidos en la secuenciación del genoma completo. Utiliza la base de datos plasmidfinder con cientos de secuencias para predecir el plásmido en los datos.

1. En el panel de herramientas busque `PlasmidFinder` y ajuste los siguientes parametros:
    - En "Input parameters":
        - "Choose a fasta or fastq file": `DRR187559_contigs.fasta`
        - "PlasmidFinder database": utilice la más actualizada

`PlasmidFinder` genera varios resultados:

`raw_results.txt`: Un archivo de texto que contiene la tabla de resultados y las alineaciones `results.tsv`: Un archivo tabular con las siguientes columnas:

```
- Base de datos
- Plásmido: Plásmido contra el que se ha alineado el genoma de entrada.
- Identidad: Porcentaje de identidad en el alineamiento entre el plásmido
  que mejor coincide en la base de datos y la secuencia correspondiente en
  el genoma de entrada (también llamado par de segmentos de alta puntuación
  (HSP)). Una alineación perfecta es del 100%, pero también debe cubrir toda
  la longitud del plásmido en la base de datos (compare los ejemplos 1 y 3).
- Longitud de la consulta/plantilla: La longitud de consulta es la longitud
  del plásmido que mejor coincide en la base de datos, mientras que la
  longitud HSP es la longitud de la alineación entre el plásmido que mejor
  coincide y la secuencia correspondiente en el genoma (también denominada
  par de segmentos de alta puntuación (HSP)).
- Contig: Nombre del contig en el que se encuentra el plásmido.
- Posición en el contig: Posición inicial del gen encontrado en el contig.
- Nota: Notas sobre el plásmido
- Número de acceso: Número de acceso al Genbank de referencia según el NCBI
  para el plásmido en la base de datos.
```

`plasmid.fasta`: Un archivo fasta que contiene las mejores secuencias coincidentes del genoma de consulta `hit_in_genome.fasta`: Un archivo fasta que contiene los genes plasmídicos que mejor coinciden con los de la base de datos

Responda las siguientes preguntas

- ¿Cuántas secuencias de plásmidos se encontraron?
- ¿Dónde se encuentran?
- ¿Están todas estas secuencias asociadas a *Staphylococcus aureus*?
- ¿Qué podemos concluir sobre contig00019?

Integrones

Los integrones son mecanismos genéticos que permiten a las bacterias adaptarse y evolucionar rápidamente mediante el almacenamiento y la expresión de nuevos genes. Un integrón está compuesto mínimamente por

- un gen que codifica para una recombinasa de sitio específico (intI)
- un sitio de recombinación proximal (attI), que es reconocido por la integrasa y en el que pueden insertarse casetes de genes
- un promotor (Pc) que dirige la transcripción de los genes codificados en casetes.

Para detectar los integrones, utilizaremos `IntegronFinder` (Néron et al. 2022). Esta herramienta

1. Anota el CDS con Prodigal
2. Detecta de forma independiente
    - integrón integrasa utilizando la intersección de dos perfiles HMM: uno específico de tirosina-recombinasa (PF00589) y otro específico del integrón integrasa, cerca del dominio patch III de tirosina recombinasas
    - attC con un modelo de covarianza (CM), que modela la estructura secundaria además de las pocas posiciones de secuencia conservadas.
3. Integra los resultados para distinguir 3 tipos de elementos
    - Integrón completo: Integrón con integrasa de integrón cerca de sitio(s) attC
    - Elemento In0: Integrón integrasa solamente, sin ningún sitio attC cercano
    - Elemento CALIN: Grupo de sitios attC Sin integrasa cercana

Ejecutar En el panel de herramientas busque `PlasmidFinder` y ajuste los siguientes parametros: - "Replicon file": `DRR187559_contigs.fasta` - "Thorough local detection": `Yes` - "Search also for promoter and attI sites?": `Yes` - "Remove log file": `Yes`

`IntegronFinder` genera 2 salidas:

1. Un resumen con para cada secuencia en la entrada el número de elementos CALIN identificados, elementos In0, e integrones completos.
2. Un archivo de anotación de integrones en forma de tabla.

Responda las siguientes preguntas

- ¿Cuántos elementos de integrón se han encontrado?

Elementos IS (secuencia de inserción)

El elemento de secuencia de inserción (IS) es una secuencia corta de ADN que actúa como elemento transponible simple. Los IS son los elementos transponibles autónomos más pequeños pero más abundantes en los genomas bacterianos. Sólo codifican proteínas implicadas en la actividad de transposición. Desempeñan, pues, un papel clave en la organización y evolución del genoma bacteriano.

Para detectar elementos IS, utilizaremos `ISEScan` (Xie y Tang 2017). ISEScan es un software altamente sensible basado en modelos de Markov ocultos construidos a partir de elementos IS curados manualmente.

1. En el panel de herramientas busque `ISEScan` y ajuste los siguientes parametros:
    - "Genome fasta input": `DRR187559_contigs.fasta`

Responda las siguientes preguntas

- ¿Cuántos elementos IS se han detectado?
- ¿Dónde se encuentran?
- ¿Cuáles son las distintas familias de especies invasoras?

Bibliografia

Seemann, T., 2014 Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069. 10.1093/bioinformatics/btu153

Xie, Z., and H. Tang, 2017 ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics 33: 3340–3347. 10.1093/bioinformatics/btx433

Hikichi, M., M. Nagao, K. Murase, C. Aikawa, T. Nozawa et al., 2019 Complete Genome Sequences of Eight Methicillin-Resistant Staphylococcus aureus Strains Isolated from Patients in Japan (I. L. G. Newton, Ed.). Microbiology Resource Announcements 8: 10.1128/mra.01212-19

Carattoli, A., and H. Hasman, 2020 PlasmidFinder and in silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). Horizontal gene transfer: methods and protocols 285–294. 10.1007/978-1-4939-9877-7_20

Schwengers, O., L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom et al., 2021 Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial genomics 7: 000685. 10.1099/mgen.0.000685

Néron, B., E. Littner, M. Haudiquet, A. Perrin, J. Cury et al., 2022 IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in Klebsiella. Microorganisms 10: 700. 10.3390/microorganisms10040700

Diesh, C., G. J. Stevens, P. Xie, T. De Jesus Martinez, E. A. Hershberg et al., 2023 JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome biology 24: 1–21. 10.1186/s13059-023-02914-z