Curso de Iniciación a la Secuenciación Masiva

BU-ISCIII

Práctica 1 día 9: Ensamblado, anotación, tipificación e identificación de genes de resistencia

Descripción

Este tutorial está parcialmente basado en el material suplementario al artículo [1]. Primero crearemos un ensamblaje del genoma E. coli O14:H4, comprobaremos su calidad. Adicionalmente haremos un MLST e identificaremos los genes de resistencia.

Se utilizarán los datos de secuenciación de Escherichia coli O104:H4, responsable del brote letal del síndrome urémico hemolítico (HUS) en Alemania en 2011 [2, 3]. La cepa que ocasión el brote pertenece al linaje E. coli enteroagregativa (EAEC), que ha adquirido una toxina Shiga [4], normalmente asociada con E. coli enterohemorragica (EHEC) y con múltiples genes de resistencia.

Ejercicio 1

```
# Primero vemos a la home y de ahí a la carpeta del curso
cd
cd introduction_to_bioinformatics_handson/
pwd
# Output: /home/alumno/introduction_to_bioinformatics_handson

# Copiamos los datos de la práctica
cp -r
/mnt/ngs_course_shared/introduction_to_bioinformatics_handson/05_handson_as
sembly/ .

# Cuando acabe nos movemos a la carpeta de los datos
cd 05_handson_assembly
```

El objetivo de este ejercicio es poner en práctica los conocimientos adquiridos en este curso sobre el preprocesamiento de datos NGS para comprobar la calidad de las lecturas sobre las que vamos a trabajar. Para ello vamos a utilizar la herramienta FastQC. Puedes obtener más información sobre la aplicación en http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Desde la línea de comandos ejecutamos:

```
#Inicializamos el environment de conda
conda activate ngs_course

# Realizamos el análisis de calidad
fastqc -t 2 RAW/SRR292770_* --outdir RESULTS/fastqc
```

Vemos todos los reports abriendo los archivos .html que están dentro de la carpeta `RESULTS/fastqc/` (SRR292770_1_fastqc.html y SRR292770_2_fastqc.html)

Comprueba como las secuencias superan la mayoría de los test.

Ejercicio 2

El objetivo de este ejercicio es realizar un ensamblaje de las lecturas mediante el programa Unicycler.

```
#Ejecutamos unicycler
cd RESULTS
unicycler --threads 2 -1 ../RAW/SRR292770_1.fastq.gz -2
../RAW/SRR292770_2.fastq.gz -o unicycler_output

#Copiamos el fichero de contigs a nuestro directorio de trabajo
cp unicycler_output/assembly.fasta SRR292770_unordered.fasta
```

Al ejecutar este comando se generaran una serie de ficheros, incluido el fichero que contiene los contigs (assembly.fasta). Copiamos el archivo assembly.fasta y lo renombramos.

Ejercicio 3

Una vez que tenemos los contigs con las lecturas ensambladas es útil ordenarlos frente a un genoma de referencia. Una manera sencilla de conseguirlo es utilizando la opción Move Contigs de Mauve (http://asap.ahabs.wisc.edu/mauve).

Utilizaremos los contigs obtenidos anteriormente y la referencia E. coli Ec55989 (NCBI accession NC 011748) que corresponde con una cepa muy cercana con un genoma completo disponible, localizado en la carpeta `/home/alumno/introduction_to_bioinformatics_handson/05_handson_assembly/REFERENCES`.

Se abre la aplicación desde la terminal:

```
#Inicializamos el environment de conda
conda activate mauve

#Ejecuta Mauve
mauve
```

En el menú elegimos la opción Tools -> Move Contigs.

Aparece una ventana de diálogo etiquetada Choose location to keep output files and folders. Navega hasta la carpeta donde estén la referencia y los contigs y crea una nueva carpeta denominada Mauve-Output.

Se muestra un mensaje acerca del proceso iterativo de reordenamiento de los contigs. Pulsa OK para continuar.

Aparece una caja de dialogo etiquetada como Align and Reorder Contigs. Pulsa el boton Add Se-quence... y elige el genoma de referencia frente al que quieres alinear que en este caso es NC_011748.fna `(/home/alumno/introduction_to_bioinformatics_handson/05_handson_assembly/REFERENCES/NC_011748.fna)`.

Pulsa el boton Add Sequence... nuevamente y elige el fichero con los contigs SRR292770_unordered.fasta

Pulsa Start para comenzar la reordenación. Este proceso puede llevar un tiempo. La reordenación se lleva a cabo mediante una serie de iteraciones. Por cada una de ellas se genera una ventana Mauve unknown - alignmentX, donde X es el número de iteración.

El conjunto final de contigs ordenado y orientado se encuentra en el fichero fasta de la última iteración (alignment10). Para localizarlo mira dentro del directorio MauveOutput creado anteriormente. Por cada iteración debe haber una carpeta alignmentX. Copia el fichero SRR292770_unordered.fasta del directorio que tenga el valor de X más alto a tu directorio de trabajo; nombrándolo como SRR292770.fasta. Una vez que hayas copiado el fichero puedes eliminar los directorios alignmentX.

Ejercicio 4

objetivo de este ejercicio es visualizar los contigs ordenados utilizando Mauve. Generaremos un alineamiento múltiple de los contigs ordenados del genoma del brote O104:H4, la referencia Ec55989 y un ensamblaje adicional creado utilizando más lecturas que nuestro draft.

El ensamblaje alternativo de la cepa TY-2482 (NCBI accession AFVR01) está disponible para descargar en http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AFVR01. Deberíamos descargarnos dicho ensamblaje y ordenarlo a la referencia Ec55989 siguiendo el método detallado en el ejercicio anterior. Como este paso consume mucho tiempo; con los restantes ficheros de la practica encontrareis el fichero AFVR01.1_sorted.fasta que puede ser utilizado directamente `(/home/alumno/ngs_course_exercise/05_handson_assembly/RAW/AFVR01.1_sorted.fasta)`.

- Ejecuta Mauve
- Elige la opcion de menu File -> Align with progressive Mauve...
- Aparece una caja de dialogo etiquetada como Sequences to align. Pulsa el boton Add Sequence... y elige el fichero de contigs SRR292770.fasta
- Pulsa Add Sequence... nuevamente y elige el fichero AFVR01.1_sorted.fasta con el ensamblaje alternativo. Pulsa Add Sequence... y añade el fichero NC_011748.fna `(/home/alumno/introduction_to_bioinformatics_handson/05_handson_assembly/REFERENCES/NC_011748.fna)` con el genoma de referencia Ec55989. Especifica Multiple como nombre para el fichero de salida y pulsa Save.
- Pulsa Align... para ejecutar el alineamiento. Este proceso puede llevar un tiempo.
- Cuando el alineamiento finaliza puedes simplificar la imagen resultante si demarcas la opción View -> Style -> LCB connecting lines.
- Los bloques de colores indican las regiones de secuencia con homología en otros genomas. Las líneas rojas indican los bordes de los contig.

- Comprueba que el orden de los contigs de nuestro ensamblaje con Unicycler y el ensamblaje alternativo es similar. Ambos ensamblajes contienen contig que no mapan al genoma de referencia.
- Puedes almacenar una imagen de lo que estas visualizando con Tools -> Export -> Export image...

Ejercicio 5

El objetivo de este ejercicio es analizar la calidad de un ensamblado. Para ello vamos a utilizar la herramienta QUAST (http://bioinf.spbau.ru/quast).

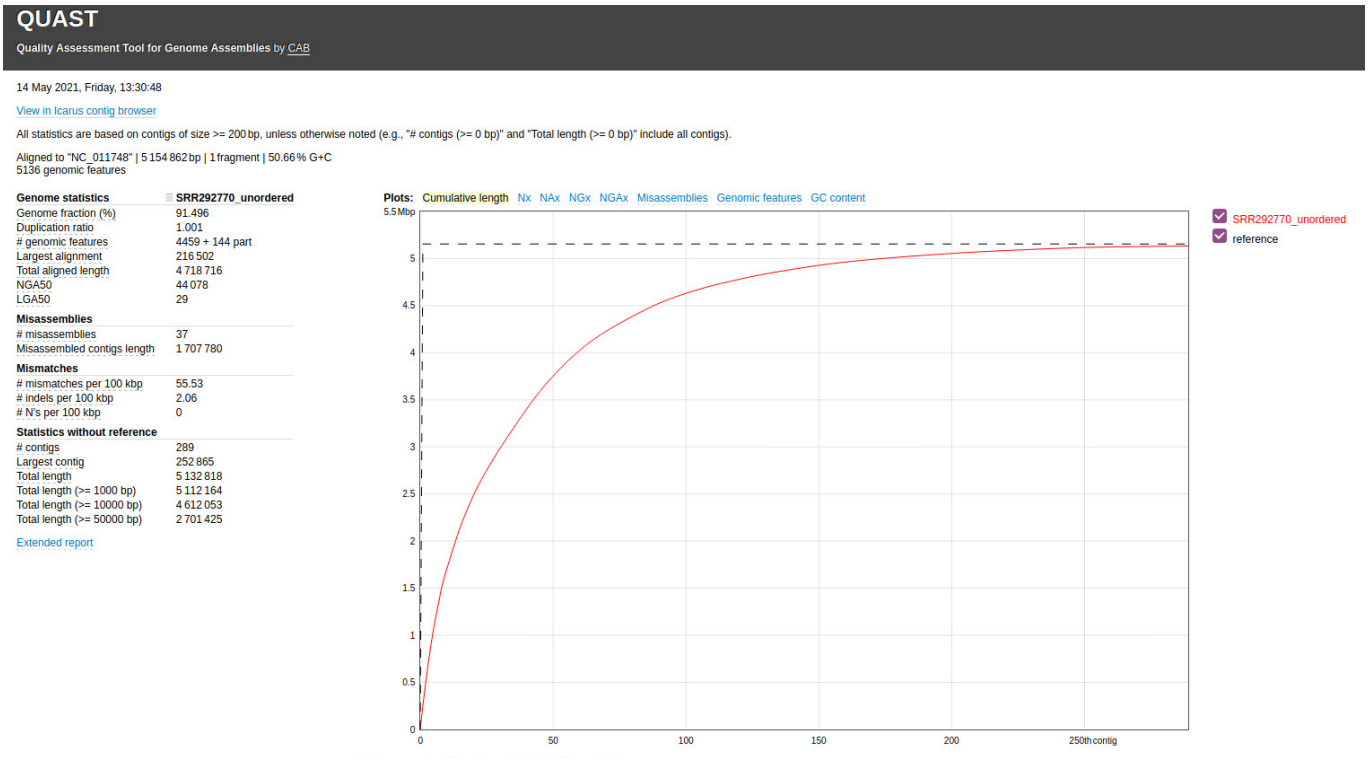
```
# Comprobamos donde estamos
pwd
# Output: /home/alumno/introduction_to_bioinformatics_handson/

#Cambiamos el environment de conda
conda deactivate
conda activate ngs_course
#Ejecutamos QUAST
quast.py SRR292770_unordered.fasta -o quast \
-R ../REFERENCES/NC_011748.fna -G ../REFERENCES/NC_011748.gff \
-m 200 -t 2
```

Especificando con el parámetro -o el directorio donde se almacenarán los resultados. Con las opciones -R y -G le especificamos un genoma de referencia muy cercano a la cepa secuenciada y el modelo de genes correspondiente. De este modo, Quast es capaz de calcular algunas métricas adicionales.

Una vez Quast haya finalizado la ejecución podemos visualizar los resultados abriendo el fichero report.html.

Situándote encima de las etiquetas se muestra una breve descripción sobre la métrica correspondiente.

Trata de comprender cada una de las métricas y gráficos generados; y como te ayudan a determinar la calidad del ensamblado.



Ejercicio 6

El objetivo de este ejercicio es realizar tipado (typing) MLST e identificación de genes de resistencia utilizando datos de secuenciación masiva. Para ello se utiliza la herramienta SRST2 (http://katholt.github.io/srst2).

Para hacer el MLST debemos descargarnos la BD correspondiente y luego cotejar las lecturas frente a esa base de datos:

```
# Comprobamos que nos estamos en la carpeta RESULTS de 05_handson_assembly
o nos movemos a ella
pwd
#
/home/alumno/introduction_to_bioinformatics_handson/05_handson_assembly/RES
ULTS/

# sino estamos ahí ejecutar:
cd
/home/alumno/introduction_to_bioinformatics_handson/05_handson_assembly/RES
ULTS/

# Creamos un directorio para MLST
mkdir ecoli_mlst

#Nos movemos a ese directorio
cd ecoli_mlst

#Inicializamos el environment de conda
conda deactivate
conda activate srst2
```

```
#Bajamos la BD
getmlst.py --species "Escherichia coli#1"
#Saldrá un warning, pero no es un error, solo un aviso

#MLST
srst2 --input_pe ../../RAW/SRR292770_1.fastq.gz
../../RAW/SRR292770_2.fastq.gz \
--output ./ --log --mlst_db Escherichia_coli#1.fasta \
--mlst_definitions ../../REFERENCES/ecoli.txt --mlst_delimiter '_'

#Visualizamos los resultados
less __mlst__Escherichia_coli#1__results.txt
```

SRST2 ya incorpora en su instalación una BD de genes de resistencia y, por tanto, no es necesario descargarse una:

```
# Volvemos al directorio anterior
cd ..

#Creamos un directorio para la resistencia
mkdir ecoli_resist

#Nos movemos a ese directorio
cd ecoli_resist

#Identificacion de genes de resistencia
srst2 --input_pe ../../RAW/SRR292770_1.fastq.gz
../../RAW/SRR292770_2.fastq.gz \
--output ./ --log --gene_db ../../REFERENCES/ARGannot.r1.fasta

#Visualizamos los resultados
less __genes__ARGannot.r1__results.txt
```

Referencias

[1] David J. Edwards, Kathryn E. Holt: Beginner's guide to comparative bacterial genome analysis using next- generation sequence data. BMC Microbial Informatics, 2013.

[2] Buchholz U, Bernard H, Werber D, Bohmer MM, Remschmidt C, Wilking H, Delere Y, an der Heiden M, Adlhoch C, Dreesman J, et al: German outbreak of Escherichia coli O104:H4 associated with sprouts. N Engl J Med 2011, 365:1763{1770.

[3] Bielaszewska M, Mellmann A, Zhang W, Kock R, Fruth A, Bauwens A, Peters G, Karch H: Charac-terisation of the Escherichia coli strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect Dis 2011, 11:671{676.

[4] Frank C,Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A, et al: Epidemic pro le of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. N Engl J Med 2011, 365:1771{1780.