UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Control de calidad y limpieza de secuencias NGS utilizando Galaxy. Manejo de archivos FASTQ.

Apellidos, nombre	Cardona Serrate, Fernando (fcardona@btc.upv.es)
Departamento	Departamento de Biotecnología. Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural.
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

En este artículo vamos a describir paso a paso cómo analizar los datos de secuenciación de segunda generación usando el servidor Galaxy. En concreto, estudiaremos cómo analizar los primeros archivos de secuencia (FASTQ) que obtenemos en secuenciación de nueva generación (NGS, de *Next Generation Sequencing*). Como ejemplo se utilizan archivos obtenidos mediante la plataforma Illumina (California, Estados Unidos) de segunda generación, pero es aplicable para cualquier archivo FASTQ obtenido de cualquiera de las plataformas de secuenciación NGS.

2 Objetivos

Una vez que el estudiante lea con detenimiento este documento, será capaz de:

- Comprender el contenido de los archivos FASTQ
- Obtener archivos FASTQ de los repositorios correspondientes
- Analizar la calidad de los archivos FASTQ
- Filtrar y recortar los archivos FASTQ por calidad

3 Introducción

La secuenciación de ADN ha sido revolucionaria en los campos de conocimiento de la biología molecular y la genética, ya que permite conocer la secuencia y estructura de los genes, así como las de otros elementos reguladores importantes en la regulación de la expresión génica. En el caso de humanos, es ampliamente conocido el proyecto genoma humano (1990-2001), que permitió conocer casi la totalidad de la secuencia del genoma humano. En este proyecto se utilizó el método de Sanger, conocido también primera generación de secuenciación [1].

Más recientemente (2008), las técnicas de secuenciación de segunda generación han permitido una secuenciación mucho más rápida de genomas humanos completos, en cuestión de días [2].

Por último (2010), los secuenciadores de tercera generación permiten obtener secuencias mucho más largas que los de segunda generación, facilitando el ensamblaje y análisis informático, y además no necesitan de un enriquecimiento previo de la muestra [3].

Todos los secuenciadores, a veces tras el paso inicial por un sistema de archivos propio, acaban dando las secuencias en formato FASTQ. Este tipo de archivos está basado en texto, y contienen, además de la secuencia de nucleótidos como texto (formato FASTA), las correspondientes puntuaciones de calidad de secuencia para cada nucleótido codificada en formato ASCII. Un archivo FASTQ tiene cuatro campos de información en diferentes líneas:

La línea 1 empieza con "@" seguido de un identificador de secuencia (nombre) y una descripción (opcional), similar al FASTA. La línea 2 es la propia secuencia. La 3ª comienza con "+" seguido (opcional) por el mismo identificador de secuencia (y cualquier descripción que se quiera añadir). La línea 4 muestra los valores de calidad de la secuencia de la línea 2

codificados en formato ASCII, y debe contener el mismo número de letras que esta secuencia.

Un archivo FASTQ que contenga una única secuencia tiene el siguiente aspecto:

@SEQ_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

El byte que representa la calidad va de 0x21 (calidad más baja; '!' en ASCII) a 0x7e (calidad más alta; '~' en ASCII). Estos son los caracteres de valor de calidad en orden creciente de izquierda a derecha (ASCII):

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Los archivos FASTQ originales de Sanger, al igual que las lecturas largas de NGS de 3ª generación, dividen las secuencias largas y las cadenas de calidad en varias líneas, como suele ocurrir también con los archivos FASTA.

El procesado de estas secuencias en formato FASTQ implica un paso de alineado o mapeo con el genoma de referencia del organismo correspondiente, de manera que pueda añadirse la información de coordenadas genómicas. Es decir, el análisis de la secuencia implica el paso a archivos SAM (*Sequence Alignment Map*) o BAM (*Binary Alignment Map*, SAM comprimido). Previamente a esta conversión, son necesarios una serie de procesos que incluyen el análisis de la calidad, así como eliminar las secuencias de baja calidad (filtrado o *filtering*) o los fragmentos de estas que no cumplan los requisitos de calidad establecidos (recorte o *trimming*) que no cumplan con los parámetros de calidad que hemos establecido.

El servidor Galaxy [4] es un sistema gratuito y de código abierto para el análisis de datos, la creación de flujos de trabajo, la formación y la educación, la publicación de herramientas, la gestión de infraestructuras, entre otros, que facilita el análisis de secuencias NGS sin necesidad de tener conocimientos avanzados de bioinformática. Esta herramienta aglutina distintas herramientas bioinformáticas, de análisis NGS y de otros tipos (por ejemplo, estructura de proteínas), que procesan los datos en su propia nube y están siempre accesibles. De esta forma, es una herramienta muy adecuada para este tipo de procesos en usuarios no avanzados con recursos de hardware limitados y/o pocos conocimientos informáticos. Es posible darse de alta en el sistema de forma gratuita con una cuenta académica en https://usegalaxy.org/.

4 Desarrollo

A la hora de iniciar el análisis de unos resultados de NGS, la primera tarea a realizar es cargar los archivos de secuencia en Galaxy. Existen varias formas de hacerlo, aquí se ejemplifica una de ellas utilizando 4 secuencias de la plataforma *Illumina* de un mismo trabajo disponibles en el repositorio. A continuación, se describe el proceso de forma detallada.

Cargar las secuencias en Galaxy en la herramienta *Faster download and extract reads in FASTQ from NCBI SRA*. Pueden cargarse todas a la vez separadas por comas: SRR8241113, SRR8241114, SRR8241115, SRR8241116 según se indica en la **imagen 1**:

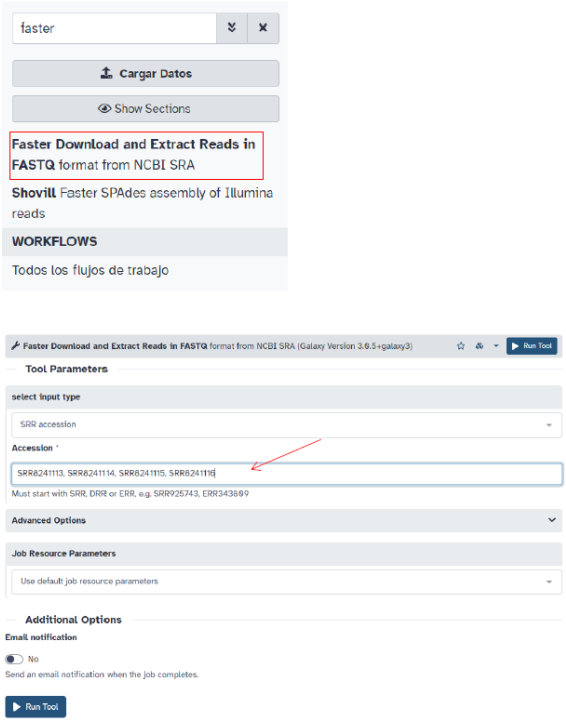

Imagen 1. Carga de archivos de secuencias FASTQ en Galaxy

4.1 Análisis de la calidad de las secuencias con FASTQC

Normalmente conviene analizar la calidad de las secuencias, lo que se puede hacer utilizando la herramienta FastQC y configurando los parámetros del programa como se indica en la imagen 2. Se pueden ir subiendo las secuencias una por una y cuando estén todas darle a “Run”. En el desplegable saldrán todas las secuencias que hayan subidas para analizar. Pueden analizarse todas a la vez arrastrando todos los archivos a analizar (MultiQC).

Los resultados pueden descargarse desde la historia de Galaxy y visualizar el análisis de calidad en modo local (archivo descargado, Imagen 3 flecha roja), o bien verlos en la propia herramienta (imagen 3, icono de gráficos).

Algunos ejemplos de resultados y sus interpretaciones pueden verse en la imagen 4 (FASTQC report).

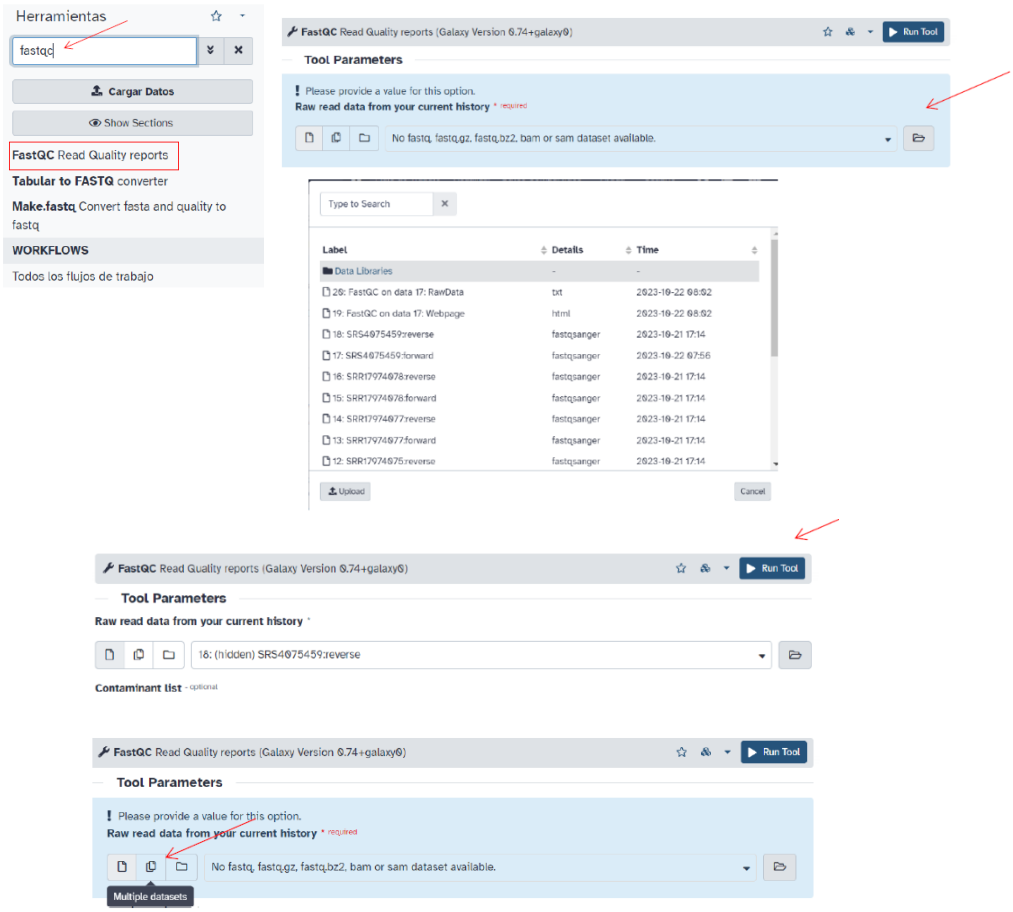

Imagen 2. Análisis de la calidad de las secuencias utilizando FASTQC en Galaxy

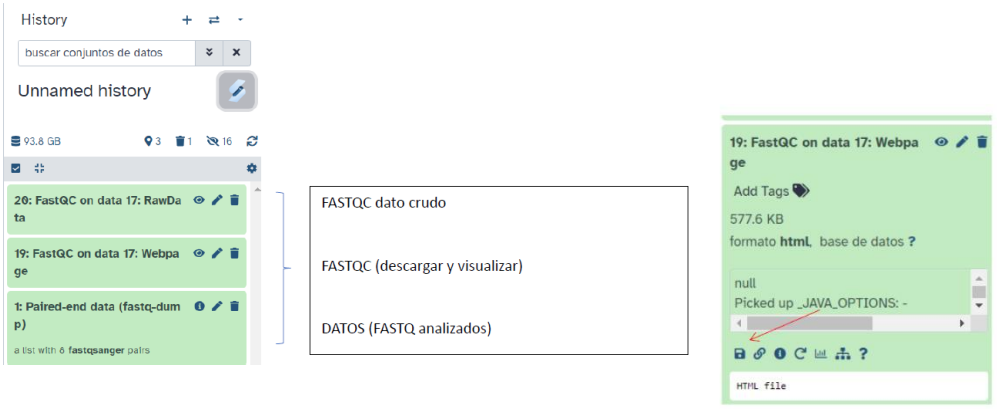

Imagen 3. Descarga de resultados obtenidos en FASTQC en Galaxy

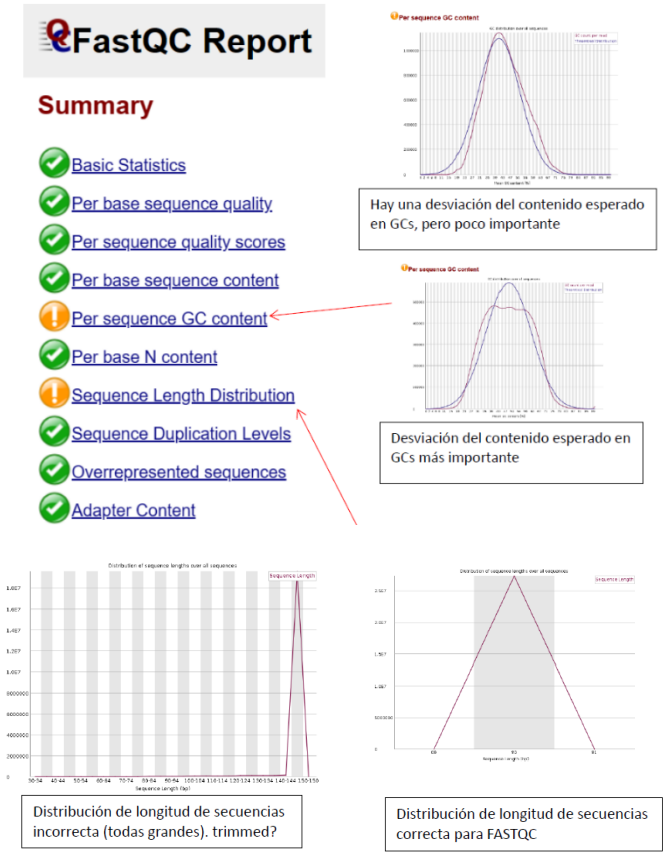

Imagen 4. FASTQC report. Ejemplos e interpretaciones de resultados obtenidos en FASTQC en Galaxy.

4.2 Procesado de las secuencias. Recorte y filtrado.

Tras analizar la calidad de las secuencias a analizar, frecuentemente es necesario hacer un procesado de estas, que puede consistir en eliminar las secuencias de baja calidad (filtrado o *filtering*) o los fragmentos de estas que no cumplan los requisitos de calidad establecidos (recorte o *trimming*) que no cumplan con los parámetros de calidad que hemos establecido. Algunas veces pueden ser necesarios ambos procesos, uno, o ninguno, dependiendo de los resultados obtenidos en FASTQC.

4.2.1 Procesado de las secuencias. Recorte.

El ***trimming*** puede hacerse en Galaxy, por ejemplo, utilizando la herramienta *Trimmomatic*. Un ejemplo de los parámetros a utilizar puede ser "corte al final por calidad" (trailing) y "Límite calidad 30" (imagen 5).

4.2.2 Procesado de las secuencias. Filtrado.

El ***filtering*** puede hacerse en Galaxy, por ejemplo, utilizando la herramienta *Trimmomatic*. Un ejemplo de los parámetros a utilizar puede ser "corte al final por calidad" (trailing) y "Límite calidad 30" (imagen 6).

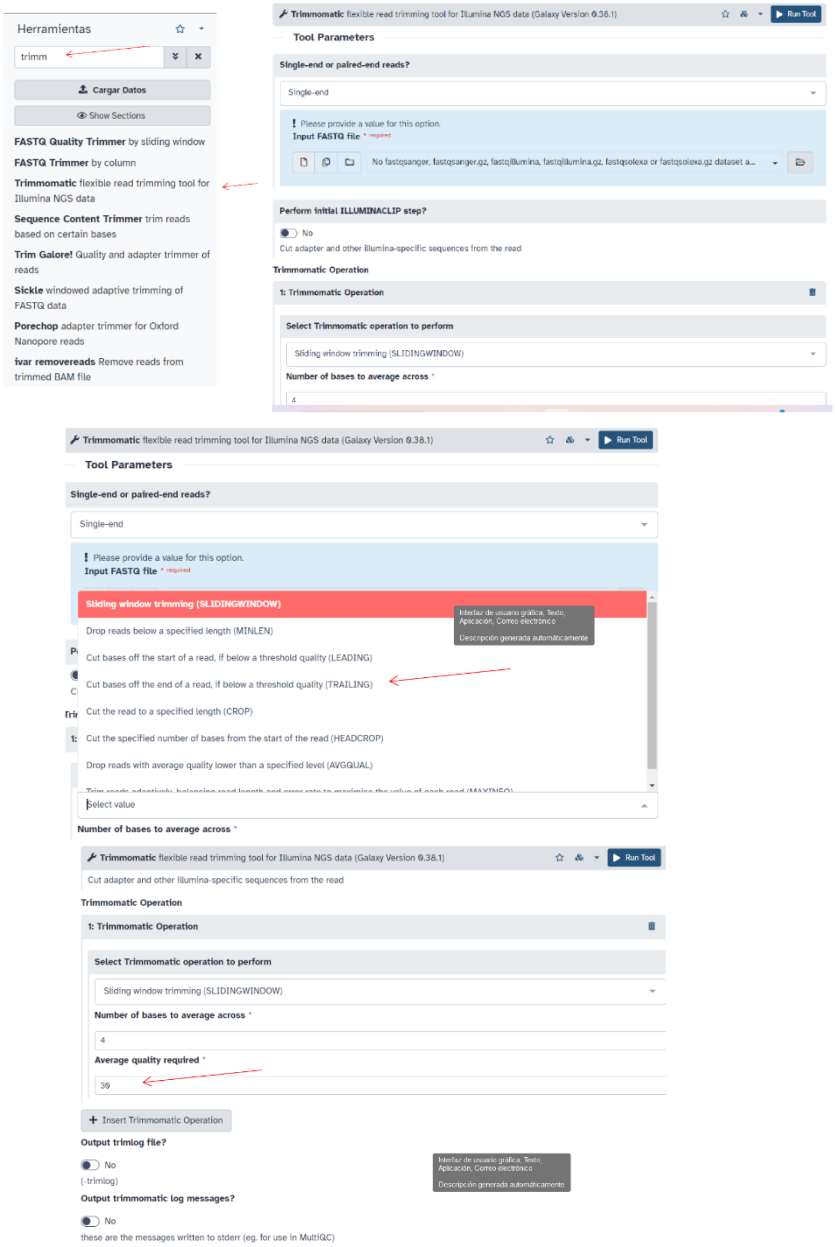

Imagen 5. Trimming utilizando Trimmomatic en Galaxy.

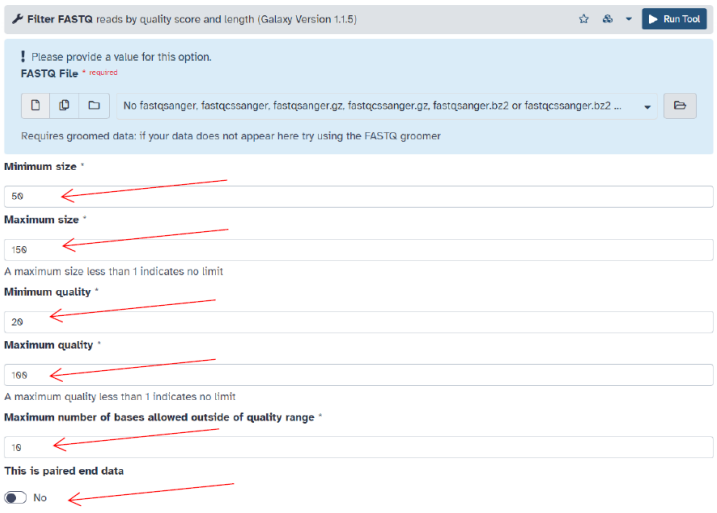

Imagen 6. Filtering utilizando Filter FASTQ en Galaxy.

Después del procesado es necesario volver a analizar las secuencias con FASTQC para ver que el procesado ha mejorado la calidad de las secuencias como pretendíamos. Si se tienen unos parámetros de calidad correctos de acuerdo con los resultados obtenidos en FASTQC, ya se pueden alinear las secuencias con el genoma de referencia para obtener los archivos SAM (*Sequence Alignment Map*) o BAM (*Binary Alignment Map*, SAM comprimido).

5 Cierre

A lo largo de este objeto de aprendizaje hemos visto qué formato tienen los archivos de secuencia FASTQ, además de cómo se puede analizar y mejorar su calidad, por filtrado o recorte, para su mapeo posterior, todo ello utilizando la herramienta Galaxy y sin necesidad de conocimientos bioinformáticos avanzados. Tras analizar inicialmente la calidad, si esta no es adecuada para continuar con el análisis, puede mejorarse mediante *filtering* y/o *trimming*. Posteriormente es necesario volver a analizar la calidad de los archivos para continuar con los análisis posteriores. En el siguiente gráfico se puede ver un diagrama del proceso a seguir.

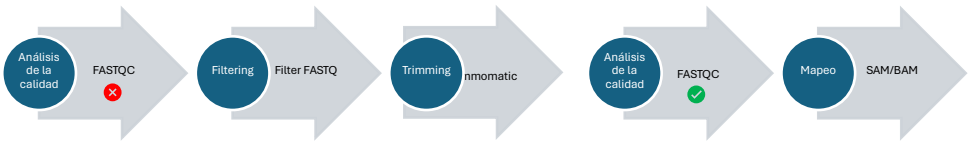

Gráfico 1. Diagrama de flujo para el análisis de archivos FASTQ en Galaxy.

6 Bibliografía

1- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* **409**, 860–921 (2001). https://doi.org/10.1038/35057062

2- Wadman, M. James Watson's genome sequenced at high speed. *Nature.* 452, 788 (2008). https://doi.org/10.1038/452788b

3- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*. **7(6)**, 461-465 (2010). https://doi.org/10.1038%2Fnmeth.1459

4- The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*. gkae410 (2024). https://doi.org/10.1093/nar/gkae410