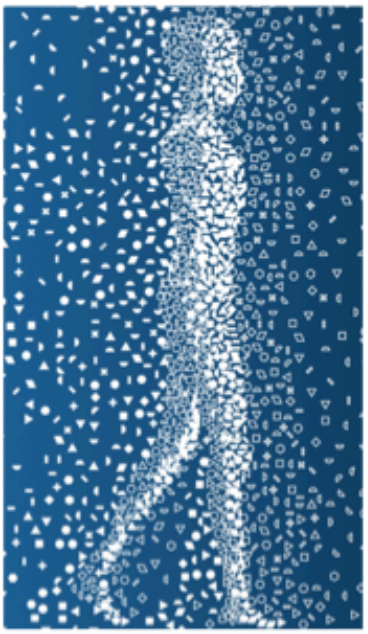
Sequence Trimming

Grupo de Bioinformática y Biología de Sistemas

Instituto de Genética - Universidad Nacional de Colombia



Andrés Pinzón Ph.D.
ampinzonv@unal.edu.co

INSTITUTO DE GENETICA

Acerca de este documento

Creado por: Andrés Pinzón, Grupo de Bioinformática y Biología de Sistemas, Instituto de Genética, Universidad Nacional de Colombia. email:ampinzonv@unal.edu.co

Fecha de creación: Octubre 18 de 2016.

Última actualización: Mayo 6 de 2024. Por: Andrés Pinzón.

Los datos

Las secuencias utilizadas en este tutorial fueron obtenidas del trabajo de Wu Fan y col. "*A new coronavirus associated with human respiratory disease in China*" (2020)[1]. Las secuencias fueron descargadas de la base de datos SRA del NCBI, usando la siguiente dirección:

https://sra-pub-src-1.s3.amazonaws.com/SRR10971381/WH_R1.fastq.gz.1

https://sra-pub-src-1.s3.amazonaws.com/SRR10971381/WH_R2.fastq.gz.1

Otras secuencias de SARSCOV-2 pueden ser consultadas en:

https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/

Esta es una secuenciación PE mediante Illumina MiniSeq, cada una de estas librerías (Forward y Reverse) consta de 28.282.964 lecturas, con longitud entre 31pb-151pb.

Trimming

En esta oportunidad vamos a utilizar el programa TRIMMOMATIC* para realizar la limpieza de las secuencias del trabajo de Wu Fan y col. (2020).
Por favor siga las indicaciones dadas a continuación y asegúrese de entender cada uno de los pasos que está llevando a cabo.

1. Crear un acceso directo (enlace simbólico a nuestras secuencias).

```
ln -s /datos/resources/examples/rawdata/WH_R1.fastq   WH_R1.fastq
ln -s /datos/resources/examples/rawdata/WH_R2.fastq   WH_R2.fastq
```

2.Realizar análisis de calidad

```
fastqc WH_R*
```

3.Realizar el trimming

Con base en los resultados obtenidos en el paso anterior y de acuerdo a su criterio realice un trimming de las dos secuencias. La siguiente línea de comandos es solamente un ejemplo **el cual usted debería modificar de acuerdo a su análisis y consultas realizadas**.
Recuerde que en trimmomatic los distintos pasos de limpieza suceden en el orden en que estos aparecen en la línea de comandos. Antes de continuar asegúrese de entender en qué consisten los siguientes argumentos de la línea de comandos (*phred33* es la codificación usada más comúnmente en la actualidad[2][3]).

---

[1] https://www.ncbi.nlm.nih.gov/pubmed/32015508
[2] Si le interesa determinar el tipo de codificación usada en sus archivos fastq, revise el siguiente enlace: https://wiki.bits.vib.be/index.php/Identify_the_Phred_scale_of_quality_scores_used_in_fastQ
[3] Más en relación a los scores Phred: https://drive5.com/usearch/manual/quality_score.html

TrimmomaticPE -phred33 WH_R1.fastq WH_R2.fastq R1_clean.fastq R1_unpaired.fastq R2_clean.fastq R2_unpaired.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:31

IMPORTANTE: Recuerde usar las opciones **nohup** MICOMANDO **&**

nohup TrimmomaticPE -phred33 WH_R1.fastq WH_R2.fastq R1_clean.fastq R1_unpaired.fastq R2_clean.fastq R2_unpaired.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:31 &

4.Realizar análisis de calidad

Sobre los resultados obtenidos realice un nuevo análisis de calidad de sus secuencias y compárelos con el análisis de calidad realizado en el paso 2.
Reporte sus resultados mediante una tabla, la cual debe contener **como mínimo**:

- Número de secuencias en datos crudos.
- Número de secuencias en datos procesados.
- Calidad promedio antes y después de la limpieza.
- Número de reads (porcentaje) perdidos.
- Tamaño promedio de reads antes y después de la limpieza.

*http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf
http://www.usadellab.org/cms/?page=trimmomatic

Grupo de Bioinformática y Biología de Sistemas

Instituto de Genética - Universidad Nacional de Colombia