Análisis bioinformático de ensayos de RNA-Seq y expresión genética diferencial.

Análisis de datos de RNA Seq.

El análisis de datos de un experimento de RNA-Seq comienza con la inspección de los archivos crudos de las lecturas para eliminar datos de baja calidad, posteriormente se generarán los archivos de mapeo de secuencias para obtener archivos de localización de secuencias. Esta información es ordenada y filtrada para eliminar duplicados ópticos o artefactos experimentales y utilizada para obtener el conteo de frecuencias de expresión y obtener finalmente datos de expresión genética diferencial.

Dependencias.

Las siguientes versiones de los lenguajes **python**(2.7) y **R**(3.6) deben estar instaladas, además de los paquetes especificados en la siguiente lista.

1. **python 2.7, python 2.7-dev**
2. **python-numpy, python-matplotlib**
3. **python-HTSeq**
4. **tophat**
5. **bowtie**
6. **samtools**
7. **R** Versión 3.6 o superior.
8. **BiocManager**
9. **edgeR**
10. **limma**

Eliminación de secuencias de baja calidad con Trimommatic.

El filtrado de secuencias de baja calidad para encontrar el balance adecuado entre mayor calidad y menor pérdida de datos se realizará con la herramienta Trimommatic,

```
java -jar trimmomatic-0.36.jar PE -phred33 c01-R1.fastq
 c01-R2.fastq c01-trimmed-R1.fastq c01-unpaired-R1.fastq
 c01-trimmed-R2.fastq c01-unpaired-R2.fastq
SLIDINGWINDOW:5:28 MINLEN:35
```

Ejecutar el filtrado de secuencias para cada una de las condiciones experimentales.

Mapeo de secuencias con TopHat.

Mapeo de secuencias contra un genoma de referencia para obtener archivos con coordenadas de localización para cada secuencia.

```
tophat -r 200 -p 1 -a 10 --mate-std-dev 50
--no-coverage-search
--segment-length 17 -o tophat-c01 --bowtie1 hg19
c01-trimmed-R1.fastq c01-trimmed-R2.fastq
```

Ejecutar el mapeo de secuencias para cada una de las condiciones experimentales.

Inspección de los archivos de coordenadas y elminación de replicas redundantes con SamTools.

Los duplicados de PCR son artefactos que se producen en el paso previo a la secuenciacin y cuyo efecto principal es condicionar la profundidad de secuenciacin (DaCosta y Sorensen,2014), son causados por errores aleatorios en la fragmentación y amplificación por PCR donde dos lecturas proceden del mismo fragmento original de ADN provocando que el total de productos sea identico con al menos un fragmento de la librería. // Para eliminar esta información artefactual, debemos ordenar el archivo de mapeo por coordenada y después eliminar duplicados.

Inspección del resultado del mapeo de secuencias.

```
samtools view accepted_hits.bam | more
```

Ordenamiento del archivo de mapeo por coordenadas.

```
samtools sort accepted_hits.bam c01-srt
```

Eliminación de duplicados redundantes.

```
samtools rmdup c01-srt.bam c01-rmdupped.bam
```

Reordenamiento alfabético de las secuencias y conversión en archivo SAM.

```
samtools sort -n c01-rmdupped.bam c01

samtools view c01.bam > c01.sam
```

Conteo de frecuencias de expresión genética con HTSeq-count.

Mediante HTSeq-count se realiza una comparación entre las coordenadas de mapeo del archivo SAM contra las coordenadas de un archivo de anotación funcional del organismo estudiado.

```
htseq-count -s no -i gene_name -q tophat-c01/c01.sam
hg19.gtf > c01-counts.tsv
```

Construcción de la matriz de datos final.

En esta etapa se unen todos los archivos de conteos en una sola matriz de datos final.

```
join c01-counts.tsv c02-counts.tsv | join - t01-counts.tsv |
join - t02-counts.tsv > total_counts.tsv
```

Recursos

Jeffrey M. DaCosta, Michael D. Sorenson. Amplification Biases and Consistent Recovery of Loci in a Double-Digest RAD-seq Protocol https://journals.plos.org/plosone/article?id=10.1371/journal.pone.01067

Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.https://academic.oup.com/bioinformatics/article/26/1/139/

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.https://academic.oup.com/nar/article/43/7/e47/24

Trimmomatic: A flexible read trimming tool for Illumina NGS data. http://www.usadellab.org/cms/?page=trimmomatic

Samtools, a suite of programs for interacting with high-throughput sequencing data. http://www.htslib.org/doc/

TopHat is a fast splice junction mapper for RNA-Seq reads. https://www.nature.com/articles/nprot.2012.016

José Manuel S, http://github.com/J0MS/