Universidad de los Andes

BCOM4102-Ecologia Microbiana

Programa de Maestría en Biología Computacional

Taller Semana 10: Metatranscriptómica

Abril 15 de 2021

**Taller Semana 10**

Carlos Andrés Díaz - **código: 202010343**

David León - **código: 201615216**

Cesar Patiño - **código: 201924259**

**Inicie una sesión en el cluster (magnus.uniandes.edu.co), recuerde hacerlo vía ssh (desde Terminal o MobaXterm), su login es bcom4102 y el password es 202110bcom4102.**

**Recuerde ejecutar siempre sus comandos desde un nodo interactivo.**

**Objetivos: Después de ejecutar este tutorial, usted debería poder:**

**1. Describir los pasos generales involucrados en el ensamblaje de novo de metatranscriptomas a partir de secuencias de lectura corta.**

**2. Conocer los pasos a seguir para analizar el metatranscriptoma de una comunidad microbiana.**

**3. Evaluar la calidad de un ensamblaje de metatranscriptoma.**

**4. Interpretar la métrica más común de expresión relativa de los genes en una comunidad microbiana.**

**Material orientación: (máx. 15 min)**

**1. Inicialmente, le invitamos a ver dos videos acerca de la herramienta más usada para ensamblaje de novo de datos de transcripción:** [**Trinity**](https://github.com/trinityrnaseq/trinityrnaseq/wiki)**, creados por su desarrollador principal, Brian Haas.** [**Los videos los encuentra en este enlace**](https://www.broadinstitute.org/broade/trinity-screencast)**, y son particularmente:**

**1. Introduction to De Novo RNA-Seq Assembly using Trinity.**

**2. Trinity – How it works.**

**2. Debido a la complejidad y el gran número de pasos necesarios en el análisis de un metatranscriptoma, en esta actividad la idea es seguir algunos de los pasos referenciados en** [**este taller**](https://parkinsonlab.github.io/Metatranscriptome-Workshop/)**. En él, se describe el análisis detallado de un metatranscriptoma de colon de ratón, con datos generados con Illumina SE 150bp.**

**Nota importante: Todos los comandos del tutorial ya fueron ejecutados para minimizar el tiempo de ejecución y enfocarse en familiarizarse con el proceso de análisis e interpretación de resultados.**

**Datos: Localización de los datos a trabajar:**

**/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Taller\_Semana10**

**Datos Nuestros:**

**/hpcfs/home/ciencias/biologia/cursos/bcom4102/patino\_leon\_diaz/Taller9**

**Indicaciones**

I. **En la primera fase de análisis, los datos son limpiados y filtrados por calidad. También han pasado por un proceso de de-replicación con** [**CD-HIT**](https://github.com/weizhongli/cdhit) **con el fin de reducir la cantidad de tiempo de cómputo requerido para la identificación y filtrado de lecturas de RNA ribosomal.**

**Para entregar**:

**¿Cuántas lecturas únicas hay en el archivo de-duplicado? ¿A qué porcentaje de lecturas iniciales corresponde?**

Para determinar el número de lecturas únicas se utiliza la siguiente línea de comando:

**grep -o -i Cluster mouse1\_unique.fastq.clstr | wc -l**

**grep -o -i @ mouse1.fastq | wc -l**

Existen 65782 lecturas únicas, las cuales corresponden al 65,78% de las lecturas encontradas del archivo base mouse1.fastq o el 62,1212% al compararlo con el valor post triming encontrado en mouse1\_trim.fastq

¿**Cuál es el proceso a seguir para identificar y filtrar lecturas de fuentes de contaminación como vectores, adaptadores, cebadores, y RNA proveniente del hospedero?**

Para identificar y filtrar lecturas se empieza usando el alineador de Burrows Wheeler (BWA) y la herramienta BLAT (la cual realiza un análisis similar a un BLAST) con el fin de comparar con respecto a un conjunto de bases de datos que facilita la identificación de las secuencias al compararlas con hospedero. Posteriormente, se genera un índice para las secuencias generadas por BWA y BLAT con el fin de realizar alineamientos para las lecturas y filtrar aquellas que se encuentren en la base de datos de vectores a través del uso de la herramienta Samstools.

**¿Cuántas lecturas alinearon BWA y BLAT contra la base de datos de las secuencia provenientes del hospedero?**

Para determinar las lecturas alineadas se utiliza la siguiente línea de comando:

**vsearch --fastq\_filter mouse1\_univec\_bwa.fastq --fastaout mouse1\_univec\_bwa.fasta**

**vsearch --fastq\_filter mouse1\_univec\_blat.fastq --fastaout mouse1\_univec\_blat.fasta**

Se alinearon un total de 65598 lecturas de las cuales 0 fueron truncadas y 0 fueron descartadas.

**¿Cuál es la ventaja de usar BLAT adicionalmente a BWA?**

La principal ventaja de usar BLAT es que se encuentra la posibilidad de realizar alineamientos adicionales para las lecturas, al compararlas con la base de datos de fuentes de contaminación de vectores.

**¿Para qué se utiliza la herramienta** [**Infernal**](http://eddylab.org/infernal/)**? ¿Cuál es la importancia de este paso en el flujo de análisis?**

La herramienta infernal se usa para hacer un “screening” de las muestras con el fin de evitar largos tiempos de procesamiento en el proceso de ensamblaje y anotación. Este paso es importante ya que esta herramienta presenta una sensibilidad más alta con respecto a los modelos de covarianza usuales.

**¿Cuántas secuencias de RNA ribosomal se identificaron? ¿Cuántas lecturas quedan posterior a este filtrado?**

**grep -o -i @ mouse1\_unique\_rRNA.fastq | wc -l**

A partir del archivo que contiene las secuencias de rRNA, se identificaron 1422 secuencias y la cantidad de lecturas que quedan posterior al filtrado son 1134 secuencias.

**Ensamblaje de lecturas: Se ha demostrado que el ensamblaje de lecturas en contigs más grandes aumenta significativamente la capacidad para anotarlas en genes conocidos a través de búsquedas en bases de datos.**

**En esta sección, se seguirán las indicaciones del Paso 8 del tutorial enlazado, donde se usa el algoritmo de ensamblaje de SPAdes (utilizado la semana pasada), a través del modo específico para ensamblade de transcritos. Nota: utilice el archivo “mouse1\_trim.fastq” como entrada para el ensamblaje (no se ejecutan los demás pasos de la sección).**

**Módulos a cargar:**

**spades/3.14.0**

**quast/5.0.2**

Para entregar:

**¿Cuáles fueron los tamaños de *k*-mero utilizados por defecto?**

Los tamaños de kmeros utilizados por defecto son 53 y 79

**SPAdes ensambla las lecturas en contigs en un archivo “mouse1\_spades/transcripts.fasta”. Use QUAST, como en el taller de la semana pasada y explique según las métricas obtenidas cómo el número y tamaño de contigs generados se ajustan al contexto de un metatranscriptoma (recuerde que se trata de una comunidad simplificada por razones prácticas).**

**spades.py --rna -s mouse1\_trim.fastq -o mouse1\_spades**

**mv mouse1\_spades/transcripts.fasta mouse1\_contigs.fasta**

**Inspeccione el archivo pre-computado “mouse1\_proteins.fasta”. ¿Cuál es el número de proteínas putativas predichas? ¿Qué significa este número con respecto al total de genes putativos predichos?**

grep -c ">" mouse1\_proteins.fasta

grep -c ">gi" mouse1\_genes.fasta

El número de proteínas putativas predichas es 21699. Esto significa que existe un gran número de proteínas “hipotéticas” de las cuales no estamos 100% seguros de su presencia; sin embargo, fueron representadas en el archivo. Esto, con respecto al número de genes predichos equivalente a 1234.

**Generar valores de expresión normalizados asociados con cada gen: Una vez se han eliminado las bases y/o lecturas de baja calidad, los vectores, los adaptadores, las secuencias del hospedero y las secuencias de RNA ribosomal; además de contar con las lecturas anotadas tanto a nivel taxonómico como funcional, se procede a observar la expresión relativa de cada uno de los genes en el microbioma.**

**Para entregar:**

Visualice el archivo pre-computado “mouse1\_RPKM.txt” en Excel ó R.

**¿Qué filos parecen más activos?**

Los filos que aparecen más activos son Oscillospiraceae, seguido por Bacteroidetes, Alphaproteobacteria y Archaea

**Para el top 10 de los genes con mayor expresión, use las anotaciones de identificador de gen en la columna 1 e identifique las funciones potenciales de estos genes mediante búsquedas en bases de datos en línea. ¿A qué corresponden?**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **GeneID** | **Función**\* | **Especie**\* | **Longitud** | **Número de reads** | **Número E.C**\* | **RPKM** |
| WP\_004054720.1 | ATP-dependent zinc metalloprotease FtsH | *Eubacterium plexicaudatum* | 1848 | 2 | 3.6.4.6 | 999723876265 |
| WP\_051142467.1 | 1-deoxy-D-xylulose-5-phosphate synthase | *Anaerotruncus sp.* | 1848 | 2 | 2.2.1.7 | 999723876265 |
| CUP83737.1 | Amylosucrase | *Ruminococcus torques* | 1848 | 2 | 2.4.1.4 | 999723876265 |
|  | hydroxypyruvate reductase | *Parabacteroides distasonis* | 924 | 1 | 1.1.1.81 | 999723876265 |
| WP\_003527199.1 | [molecular chaperone DnaK](https://www.ncbi.nlm.nih.gov/protein/WP_003527199.1/) | *Clostridium*  *leptum* | 1848 | 2 | 3.6.1.3 | 999723876265 |
| WP\_023347813.1 | [molecular chaperone DnaK](https://www.ncbi.nlm.nih.gov/protein/WP_003527199.1/) | *Firmicutes bacterium* | 1848 | 2 | 3.6.4.10 | 999723876265 |
| WP\_013613613.1 | [membrane protein insertase YidC](https://www.ncbi.nlm.nih.gov/protein/WP_013613613.1/) | *Odoribacter splanchnicus* | 1848 | 2 | - | 999723876265 |
| AHF24012.1 | amylosucrase | *uncultured bacterium* | 1848 | 2 | 2.4.1.4 | 999723876265 |
| SCJ07986.1 | ATP-dependent zinc metalloprotease FtsH | *uncultured Clostridium sp* | 1848 | 2 | 3.6.4.6 | 999723876265 |
| SEG00983.1 | amylosucrase | *Eubacterium ruminantium* | 1848 | 2 | 2.4.1.4 | 999723876265 |

\*Los datos fueron obtenidos del NCBI y la base de datos KEGG

**Referencias:**

* Ruiz, E., Ramírez, C. A., Nocua, P., Requena, J. M., & Puerta, C. J. (2018). Identificación de proteínas reguladoras de la expresión génica en tripanosomátidos. *Rev. Acad. Colomb. Cienc. Ex. Fis. Nat.*, *42*(165), 306-318. https://doi.org/10.18257/raccefyn.671