Práctica: Predicción de Rutas Metabólicas en Metagenomas

Sara Cuadros Orellana

INTRODUCCIÓN

En los datos *shotgun*, tenemos acceso a las secuencias de muchos genomas completos, los cuales están fragmentados. Para investigar las **funciones metabólicas** de la comunidad, necesitamos primero realizar el ensamblado estos genomas (*contigs*), y luego identificar los posibles genes y asociarlos con una función (anotación). Eso se suele hacer a través de BLAST contra diferentes bases de datos que contienen información funcional.

Tutorial # 1

En este ejercicio, vamos a analizar funcionalmente un metagenoma usando la herramienta HUMAnN (**H**uman Microbiome Project **U**nified **M**etabolic **An**alysis **N**etwork). HUMAnN analiza datos metagenómicos en el contexto de vías bioquímicas, aclarando la distribución de roles entre los miembros de la comunidad microbiana. Realiza la **tipificación metabólica** (***metabolic profiling***) de comunidades microbianas, estimando la presencia, ausencia y abundancia de familias de genes y vías metabólicas microbianas a partir de datos metagenómicos, lo que por veces se le denomina "**reconstrucción metabólica**".

HUMAnN está (normalmente) disponible en http://huttenhower.sph.harvard.edu/humann, y es una tubería de datos (pipeline) diseñado para uso en línea de comando (Unix/Linux). En estos momentos, la página está en mantención, entonces usaremos una implementación de HUMAnN en la plataforma Galaxy europea.

Galaxy es una plataforma de código abierto basada en la Web. Es posible descargarla y customizar sus herramientas localmente (incluir nuevas herramientas, modificar los parámetros *default* de las herramientas, habilitar la definición de todos los parámetros posibles en la interfaz gráfica, etc.). La conexión entre la ejecución de las herramientas, que ocurre en ambiente Linux (en un servidor remoto) y los comandos que se escriben en la interfaz gráfica de Galaxy se hace por programación simple (elaboración de programas de ayuda/utilidad "envoltorios" o "*wrappers*"). Hay repositorios de *wrappers* compartidos por la comunidad de bioinformáticos en todo el mundo. La capacidad de procesamiento dependerá del servidor remoto, así que Galaxy es una plataforma realmente muy útil cuando bien implementada.

Existen plataformas Galaxy en dominio público, por cortesía de algunas instituciones. Usaremos la versión publicada por el Grupo de Alemania (Friburgo), y estaremos limitados a los servidores y a la capacidad de procesamiento que ellos disponibilizan.

➢ Antes de empezar:

Ingresar a la plataforma Galaxy: https://usegalaxy.eu/

Registrarse e ingresar como usuario.

Reconocer la cantidad de espacio computacional que tiene disponible.

1. PREPARACIÓN DE LOS DATOS

1.1. Crear un nuevo historial y nombrarlo: **Metabolic_Profiling**

1.2. Obtener el archivo que vamos a analizar:

Ubicación del archivo:

Sample Hypersaline water prokaryotic community from SS33 pond from Santa Pola Saltern, Spain: https://www.ebi.ac.uk/metagenomics/samples/SRS478661

> Analysed associated runs
> > Analyses
> > > Download
> > > > Predicted CDS without annotation

Cómo obtenerlo:

a. Subirlo desde su computador
b. Subirlo usando protocolo FTP
c. Subirlo usando un enlace web

Usaremos la opción c. El archivo está compactado, y en este caso específico no será un problema. Si lo fuese, habría que descargar el archivo a su computador, descompactarlo y luego subirlo al Galaxy usando la opción a.

1.3. Confirmar que el archivo se encuentra en el historial, y visualizar su contenido.
SRR979792_FASTQ_CDS_unannotated.ffn

Antes de seguir con la reconstrucción metabólica, es importante estimar la composición taxonómica de la comunidad.

Es posible extraer información taxonómica y de la estructura de la comunidad a partir de los datos *shotgun*. Se pueden utilizar diferentes enfoques:

- Identificación y clasificación de OTUs

Dicho enfoque requiere un primer paso de clasificación de secuencias para extraer solo las secuencias 16S y 18S.

- Asignación de taxonomía en las secuencias completas utilizando bases de datos de genes marcadores.

En este tutorial, usaremos el segundo enfoque con la herramienta MetaPhlAn2 (Nature Methods 9: 811-814, 2012), que también está implementada en Galaxy Europe. Si desea usarla localmente, está disponible en https://bitbucket.org/biobakery/metaphlan2/overview

MetaPhlAn perfila la estructura y la composición de comunidades microbianas (bacterias, arqueas, eucariotas y virus) a partir de datos de secuenciación de escopeta metagenómica con resolución a nivel de especie. Es un enfoque que utiliza genes marcadores específicos de clado para asignar lecturas inequívocamente a clados microbianos con mayor precisión y> 50 veces más rápido que los enfoques actuales.

MetaPhlAn2 utiliza una base de datos de ~1M de genes marcadores específicos de clados únicos (no sólo los genes rRNA) identificados a partir de ~17.000 genomas de referencia. Asimismo, el usuario puede ocupar una base de datos (lamentablemente esta opción no fue habilitada en la plataforma pública de Galaxy Europe).

Metaphlan2 toma como entrada un archivo de secuencia en Fasta, FastQ, SAM.

1.4. Encontrar la herramienta MetaPhlAn2 en el menú de herramientas a la izquierda.

MetaPhlAn2 to profile the composition of microbial communities (Galaxy Version 2.6.0.0)

Realizar la asignación taxonómica de las lecturas del archivo SRR979792_FASTQ_CDS_unannotated.ffn

1.5. Input file: SRR979792_FASTQ_CDS_unannotated.ffn

"Database with clade-specific marker genes": locally cached

"Cached database with clade-specific marker genes": MetaPhlAn2 clade-specific marker genes.

Se generan 3 archivos:

- Un archivo tabular con la estructura de la comunidad.
  Cada línea contiene un taxón y su abundancia relativa encontrada para nuestra muestra. El archivo comienza con taxones a nivel de reino (kingdom: k_) y va a taxones más precisos.
- Un archivo BIOM con la misma información que el archivo anterior, pero en formato BIOM. Ese formato de archivo puede ser utilizado por herramientas que requieren información de la estructura de la comunidad en formato BIOM, tales como MOTHUR, QIIME, PICRUSt, MG-RAST, MEGAN y otras.

- Un archivo SAM (Sequence Alignment Map) con los resultados del mapeo de las secuencias en la base de datos de referencia.

2. IDENTIFICACIÓN DE FUNCIONES METABÓLICAS:

2.1. Encontrar la herramienta HUMAnN2 en el menú de herramientas a la izquierda.

HUMAnN2 to profile presence/absence and abundance of microbial pathways and gene families (Galaxy Version 0.11.1.0)

En este tutorial, podemos usar las bases de datos implementadas en Galaxy Europe, que actualmente se limitan a la base de datos UniRef, o podemos subir una base de datos propia. Una opción es usar la base de datos KEGG reducida (*last free version* - v56) que se puede obtener por este enlace: https://www.dropbox.com/s/awa46dar435nn0w/kegg.reduced.fasta.tar.bz2?dl=1

Sin embargo, debido a su tamaño, a la limitación de recursos computacionales que se nos hacen disponibles en los servidores de Galaxy, y al poco tiempo que tenemos para esta clase, podemos usar como demostración (por mientras se sube la base de datos Kegg) la comparación con la base de datos UniRef50, ya disponible ("locally cached").

La búsqueda de similitud se puede hacer usando varias herramientas como BLAST, USEARCH, RapSearch, etc. En este tutorial usaremos una herramienta de búsqueda de similitud relativamente nueva, que es mucho más rápida y tan sensible como Blast, y funciona muy bien para grandes conjuntos de datos metagenómicos. Se llama DIAMOND (Nature Methods, 12: 59, 2015).

2.2. Realizar la clasificación de las funciones metabólicas del metagenoma.

"Input sequence file": SRR979792_FASTQ_CDS_unannotated.ffn

"Use of a custom taxonomic profile": Yes

"Taxonomic profile file": archivo de salida de MetaPhlAn2 (Community profile)

"Nucleotide database": Locally cached

"Nucleotide database": Full

"Protein database": (1) Locally cached

"Protein database": (1) Full UniRef

"Search for uniref50 or uniref90 gene families?": uniref90

"Database to use for pathway computations": MetaCyc

Si desean, posteriormente pueden importar la base de datos kegg y repetir el análisis cambiando algunos parámetros:

"Protein database": From history

"Protein database": kegg.reduced.fasta

"Advanced Options": **Remove stratification from output? Yes**

HUMAnN2 genera 3 archivos:

- Un archivo con la abundancia de familias de genes.
  Las famílias génicas usan el código UniProt. Tome un ejemplo y búscalo en https://www.uniprot.org/

  La abundancia de la familia de genes se informa en unidades RPK (lecturas por kilobase) para normalizar la información por la longitud del gen. Refleja el número relativo de copia del gen en la comunidad.
  El valor "UNMAPPED" es el número total de lecturas que permanecen sin asignar después de los dos pasos de alineación (nucleótido y búsqueda traducida). Debido a que otras características de los genes en la tabla se cuantifican en unidades RPK, "UNMAPPED" se puede interpretar como un único gen desconocido de 1 kilobase de longitud que recluta todas las lecturas que no pudieron mapearse a secuencias conocidas.
  Es posible re-normalizar estos datos para convertirlos a Abundancia Relativa o Copias por Millón. (Etapa 3, que sigue)

- Un archivo con la cobertura de vías metabólicas.
  La cobertura de vías proporciona una descripción alternativa de la presencia (1) o ausencia (0) de vías en una comunidad, independientemente de su abundancia cuantitativa.

- Un archivo con la abundancia de vías metabólicas.

3. NORMALIZACIÓN DE LAS ABUNDANCIAS DE FAMÍLIAS DE GENES

3.1. Encontrar la herramienta RENORMALIZE a HUMAnN2 generated table.

Renormalize a HUMAnN2 generated table (Galaxy Version 0.11.1.0)

a. Usar la herramienta Renormalize a HUMAnN2 generated table:

"Gene/pathway table": **elegir "multiple datasets" y marcar lost res archivos de output de HUMAnN a la vez.**

"Normalization scheme": **Relative abundance**

"Normalization level": Normalization of all levels by community total

Repetir el análisis cambiando lo siguiente:

"Normalization scheme": Copies per million

Podemos resumir los análisis con datos metagenómicos *shotgun*: