Práctica:

Análisis de Rutas Metabólicas en Meta-transcriptomas

Sara Cuadros Orellana

INTRODUCCIÓN

El microbioma humano juega un papel clave en la salud y la enfermedad. Gracias a la meta-transcriptómica comparativa, las funciones celulares que el microbioma desregula en la enfermedad ahora pueden explorarse computacionalmente. A diferencia de los enfoques centrados en genes, los métodos basados en vías proporcionan una visión sistémica de tales funciones; sin embargo, generalmente consideran cada camino de forma aislada y en su totalidad. Por lo tanto, pueden pasar por alto las diferencias clave que (i) abarcan múltiples rutas, (ii) contienen componentes desregulados bidireccionalmente, (iii) están confinados a una región de una determinada ruta. Para capturar estas propiedades, se necesitan métodos computacionales que van más allá del alcance de las rutas predefinidas.

¿Qué vías están contribuyendo potencialmente a la caries dental?

Para responder a esta pregunta, vamos ocupar un tutorial de Galaxy, desarrollado por Chao (Cico) Zhang, para realizar un análisis de red de datos de meta-transcriptómica utilizando Heinz en Galaxy para identificar la subred de puntuación óptima.

Usaremos conjuntos de datos usados por May et al. (2016) indicado a continuación, y reproduciremos algunos de los pasos computacionales de este estudio con datos y parámetros simplificados para acelerar el análisis a los fines de este tutorial.

> Ali May, Bernd W. Brandt, Mohammed El-Kebir, Gunnar W. Klau, Egija Zaura, Wim Crielaard, Jaap Heringa, Sanne Abeln, **metaModules identifies key functional subnetworks in microbiome-related disease**, Bioinformatics, Volume 32, Issue 11, 1 June 2016, Pages 1678–1685, https://doi.org/10.1093/bioinformatics/btv526

El estudio May et al. (2016) incluye muchos pasos que comienzan desde el pre-tratamiento de los conjuntos de datos RNAseq sin procesar. Las operaciones de procesamiento de los datos hasta la obtención de datos interpretados que usaremos aquí están más allá del alcance de este tutorial. En resumen, trabajaremos con datos que ya pasaron por *trimming* de adaptadores y de secuencias de baja calidad, por la remoción de ARNs no codificantes, la remoción de mRNAs humanos, y por el alineamiento (USEARCH) contra bases de datos de tRNAs y rRNAs para obtener un conjunto de datos de mRNA microbiano. Estas secuencias fueron mapeadas a una base de datos KEGG, y los recuentos de genes se estimaron utilizando un enfoque de "mejor resultado" contando el número de veces que un gen se asignó a una lectura como el objetivo de mejor coincidencia.

Aquí, hubiera sido posible usar el pipeline HUMAnN, que tiene en cuenta la calidad de alineación y las diferencias de longitud de genes (RPK), para estimar estos recuentos.

En este tutorial, comenzamos entonces con los datos de mRNA que ya contienen el recuento de KO (KEGG Orthology). Todos los datos necesarios para este tutorial están disponibles en Zenodo.

El conjunto de datos de caries dental (DC) que usaremos en este tutorial provino de un experimento que consistió en muestras de placa supra-gingival recolectadas de las superficies dentales de 36 individuos que tenían un perfil oral caries-positivo (CP, enfermedad) o caries-negativo (CN, salud). Cada una de las 36 muestras fue secuenciada, pre-procesada y transformada en recuentos de KO. Utilizaremos estos datos de conteo como punto de partida para realizar el análisis de red.

1. PREPARACIÓN DE LOS DATOS

1.1. Crear un nuevo historial y nombrarlo: Network_Analysis

1.2. Obtener los archivos CP y agruparlos:

Usar la opción Collection y Paste/Fetch data para otenerlos todos a la vez.

https://zenodo.org/record/1344105/files/2241_CP_DZ_PairTo_2242.txt
https://zenodo.org/record/1344105/files/2126_CP_MZ_PairTo_2125.txt
https://zenodo.org/record/1344105/files/2991_CP_DZ_PairTo_2992.txt
https://zenodo.org/record/1344105/files/2931_CP_DZ_PairTo_2930.txt
https://zenodo.org/record/1344105/files/2284_CP_DZ_PairTo_2283.txt
https://zenodo.org/record/1344105/files/2125_CP_MZ_PairTo_2126.txt
https://zenodo.org/record/1344105/files/4131_CP_DZ_PairTo_4132.txt
https://zenodo.org/record/1344105/files/2954_CP_DZ_PairTo_2955.txt
https://zenodo.org/record/1344105/files/2170_CP_MZ_PairTo_2169.txt
https://zenodo.org/record/1344105/files/2955_CP_DZ_PairTo_2954.txt
https://zenodo.org/record/1344105/files/2011_CP_DZ_PairTo_2012.txt
https://zenodo.org/record/1344105/files/2012_CP_DZ_PairTo_2011.txt
https://zenodo.org/record/1344105/files/2269_CP_DZ_PairTo_2270.txt
https://zenodo.org/record/1344105/files/3215_CP_MZ_PairTo_3214.txt
https://zenodo.org/record/1344105/files/2354_CP_DZ_PairTo_2355.txt
https://zenodo.org/record/1344105/files/3306_CP_DZ_PairTo_3307.txt
https://zenodo.org/record/1344105/files/2061_CP_DZ_PairTo_2062.txt
https://zenodo.org/record/1344105/files/2355_CP_DZ_PairTo_2354.txt
https://zenodo.org/record/1344105/files/2242_CP_DZ_PairTo_2241.txt

Al finalizar, presionar “Build” para construir una lista, y llamarla CP.

1.3. Obtener los archivos CN y agruparlos:

Usar la opción Collection y Paste/Fetch data para otenerlos todos a la vez.

https://zenodo.org/record/1344105/files/2310_CN_DZ_PairTo_2309.txt
https://zenodo.org/record/1344105/files/2062_CN_DZ_PairTo_2061.txt

https://zenodo.org/record/1344105/files/2191_CN_MZ_PairTo_2192.txt
https://zenodo.org/record/1344105/files/2052_CN_MZ_PairTo_2051.txt
https://zenodo.org/record/1344105/files/2051_CN_MZ_PairTo_2052.txt
https://zenodo.org/record/1344105/files/2192_CN_MZ_PairTo_2191.txt
https://zenodo.org/record/1344105/files/2234_CN_DZ_PairTo_2233.txt
https://zenodo.org/record/1344105/files/2233_CN_DZ_PairTo_2234.txt
https://zenodo.org/record/1344105/files/2270_CN_DZ_PairTo_2269.txt
https://zenodo.org/record/1344105/files/2225_CN_MZ_PairTo_2226.txt
https://zenodo.org/record/1344105/files/4132_CN_DZ_PairTo_4131.txt
https://zenodo.org/record/1344105/files/2309_CN_DZ_PairTo_2310.txt
https://zenodo.org/record/1344105/files/2992_CN_DZ_PairTo_2991.txt
https://zenodo.org/record/1344105/files/3214_CN_MZ_PairTo_3215.txt
https://zenodo.org/record/1344105/files/2169_CN_MZ_PairTo_2170.txt
https://zenodo.org/record/1344105/files/2930_CN_DZ_PairTo_2931.txt
https://zenodo.org/record/1344105/files/3307_CN_DZ_PairTo_3306.txt

Al finalizar, presionar “Build” para construir una lista, y llamarla **CN**.

El análisis de expresión diferencial (DEA) significa tomar los datos del recuento de lecturas normalizadas y realizar análisis estadísticos para descubrir cambios cuantitativos en los niveles de expresión entre grupos experimentales. Por ejemplo, utilizamos pruebas estadísticas para decidir si, para un gen determinado, una diferencia observada en los recuentos de lectura es significativa, es decir, si es mayor de lo que se esperaría solo debido a la variación aleatoria natural.

Usaremos DEA para identificar los genes candidatos relevantes para la caries dental. Lo haremos usando una herramienta llamada DESeq

Posteriormente, usaremos Heinz para inferir las vías relacionadas.

2. ANÁLISIS DE EXPRESIÓN DIFERENCIAL

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

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.11.40.6)

“Specify a factor name”: dental_caries (under 1: Factor)

“Specify a factor level”: CP (under 1: Factor level)

param-files “Counts file(s)”: CP (collection)

“Specify a factor level”: CN (under 2: Factor level)

param-files “Counts file(s)”: CN (collection)

"Files have header": No

"Visualising the analysis results": No

El resultado es un archivo tabular con siete columnas.

Recortaremos apenas los valores de la columna 6 (p-values) para ver cómo se distribuyen y cómo se ajustan a un modelo mixto (BUM).

2.2. Recortar los valores de p-value (columna 6) usando "**Cut**".
Cut columns from a table (Galaxy Version 1.0.2)

"Cut columns": c6

"Delimited by": TAB

"From": archivo de salida de DESeq2

2.3. Ajustar un modelo BUM usando "**Fit the BUM model**".

Fit a BUM model with p-values (Galaxy Version 1.0)

"Input file": archivo de salida de la etapa de corte (c6)

Después de obtener los parámetros del modelo BUM del último paso, usaremos Heinz para identificar la subred óptima. Entonces podríamos buscar las vías clave relevantes para la subred.

Heinz es un algoritmo para buscar una subred óptima desde una red más grande.

A través de los pasos anteriores, tenemos una lista de identidades, que es una lista de ID de genes con valores p, que forman los nodos de 'la red más grande', las relaciones entre los nodos, es decir, los bordes, deben ser obtenidas de una red de fondo, que representa una vía de acceso a bases de datos, como Reactome y STRING.

En este tutorial, usaremos una red de fondo **pequeña**, para fines de demostración. La red de fondo se representa como bordes en un archivo TXT donde cada línea denota un borde (*edge*).

2.4. Obtener el archivo "edge file" de demostración.
https://zenodo.org/record/1344105/files/edge.txt

3. IDENTIFIQUE LAS VÍAS CLAVE CON HEINZ

Como primer paso, necesitamos calcular una puntuación de Heinz (Heinz score) para cada nodo, utilizando los parámetros del modelo BUM que obtuvimos.

También necesitamos especificar un valor FDR (*False Discovery Rate*) como entrada. FDR es un método para conceptualizar la tasa de errores de tipo I en pruebas de hipótesis nulas al realizar comparaciones múltiples. Para diferentes conjuntos de datos y problemas, probablemente necesitemos recoger un valor FDR por separado. Aquí configuraremos FDR a 0.11.

La herramienta "Calculate a Heinz score" nos pedirá un archivo de nodos con p-values. Obtendremos esa información nuevamente del archivo tabular generado por DESeq, recortando las columnas 1 y 6 (KO IDs y p-values).

3.1. Recortar las columnas 1 y 6 usando "**Cut**".
Cut columns from a table (Galaxy Version 1.0.2)

"Cut columns": c1, c6

"Delimited by": TAB

"From": archivo de salida de DESeq2

3.2. Calcular el Heinz Score.
Calculate a Heinz score for each node (Galaxy Version 1.0)

"A node file with p-values": archivo de salida de la etapa anterior (c1,c6)

"FDR value": 0.11

"Choose your input type for BUM parameters": archivo de salida de "BUM model"

"Output file of BUM model as input: lambda on the first line and alpha, the second": archivo de salida de "Fit a BUM model"

4. EJECUTE HEINZ: IDENTIFIQUE LA SUBRED ÓPTIMA

Después de obtener los puntajes de Heinz, ejecutaremos el programa Heinz para encontrar la subred óptima de la red de fondo que mencionamos anteriormente.

4.1. Identificar la subred de puntuación óptima.
Identify optimal scoring subnetwork using Heinz (Galaxy Version 1.0)

"File containing Heinz scores": archivo de salida de "Calculate a Heinz score"

"Edge file": archivo **edge** que importamos de Zenodo

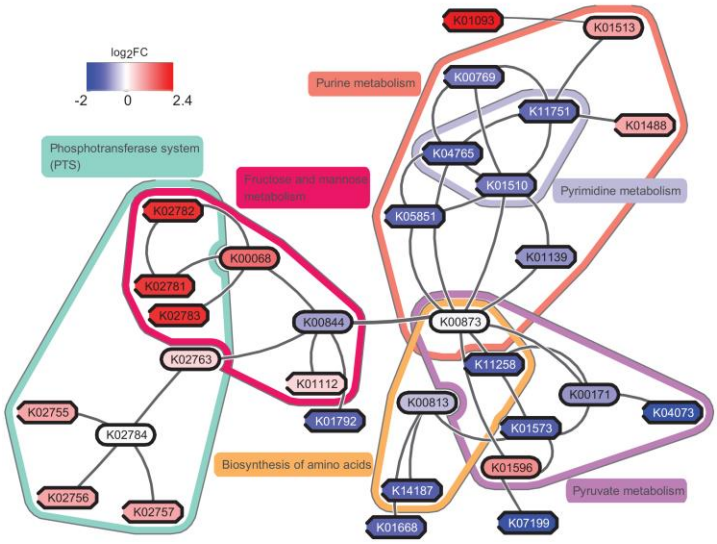
El resultado no es muy comprensible, así que necesitamos convertirlo en un gráfico.

4.2. Visualizar la subred.

Visualize the optimal scoring subnetwork (Galaxy Version 0.1.0)

La forma redonda corresponde al puntaje positivo de Heinz; la forma cuadrada corresponde al puntaje negativo de Heinz.

Además de la herramienta implementada en Galaxy, se puede usar el complemento eXamine de Cytoscape para una visualización más rica:



Para responder a la pregunta "¿qué vías están contribuyendo potencialmente a la caries dental?", Necesitamos encontrar las vías clave relevantes para la subred.

Para lograr eso, probablemente necesitemos combinar información de otros servicios de bioinformática y literatura.

Un resumen gráfico del análisis realizado: