Anásisis de Datos de ultrasecuenciación

https://github.com/IagoLast/ADO_PEC_2

Iago Lastra Rodríguez

Resumen

En este trabajo se realiza un análisis de expresión génica comparando diferentes niveles de expresión de muestras pertenecientes a un análisis del tiroides en donde se compara tres tipos de infiltración medido en un total de 292 muestras pertenecientes a tres grupos, Not infiltrated tissues (NIT) Small focal infiltrates (SFI) y Extensive lymphoid infiltrates (ELI).

Índice

1. Introducción y objetivos **2**

2. Materiales y Métodos **3**

2.1. Muestreo aleatorio simple de los datos 3
2.2. Análisis inicial. 3
2.3. Filtrado y normalización 3
2.4. Heatmap 4
2.5. Análisis de expresión diferencial 4
2.6. Resultados 4

3. Discusión **5**

4. Apéndice **5**

1. Introducción y objetivos

Las nuevas técnicas de ultra secuenciación, también conocidas cómo next-generation sequencing (NGS) han revolucionado la práctica clínica.

Las técnicas de NGS permiten obtener de forma barata y rápida la secuencias de ADN completas a partir de muestras de organismos vivos. En el caso de los humanos, aproximadamente el 1-2 % del ADN se utiliza para generar proteínas (exomas) y hay estudios que sugieren que la causa de muchas enfermedades genéticas afecta específicamente a los exones.

Por todo esto en este trabajo se analizarán muestras de ultrasecuenciación procedentes de un estudio que utiliza NGS presente en la web de GTEx pertenecientes a un análisis del tiroides en donde se compara tres tipos de infiltración medido en un total de 292 muestras pertenecientes a tres grupos:

- Not infiltrated tissues (NIT): 236 samples
- Small focal infiltrates (SFI): 42 samples
- Extensive lymphoid infiltrates (ELI): 14 samples.

Por motivos didácticos y de eficiencia computacional se ha reducido el número de muestras a 10 por grupo escogidas mediante un muestreo aleatorio simple sin repetición. Este trabajo solamente contiene la comparación entre 2 de los 3 grupos.

A partir de estas muestras previamente clasificadas se intentará listar que genes presentan diferencias estadísticamente significativas en sus niveles de expresión.

2. Materiales y Métodos

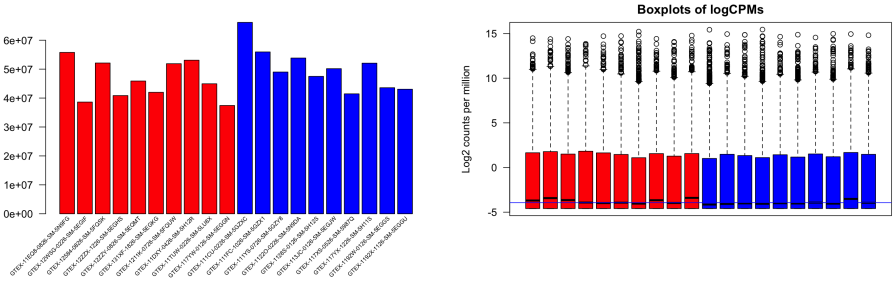
A continuación se resume el flujo de trabajo seguido en el análisis, este trabajo ha sido realizado utilizando R y R-Studio y todo el código y datos está disponible en un repositorio de GitHub de forma que los resultados sean fácilmente reproducibles y cualquier paso puede verse en detalle.

2.1. Muestreo aleatorio simple de los datos

Para crear el conjunto incial de datos se realizó un muestreo aleatorio simple eligiendo 10 muestras correspondientes a cada grupo (NIT) (SFI) (ELI).

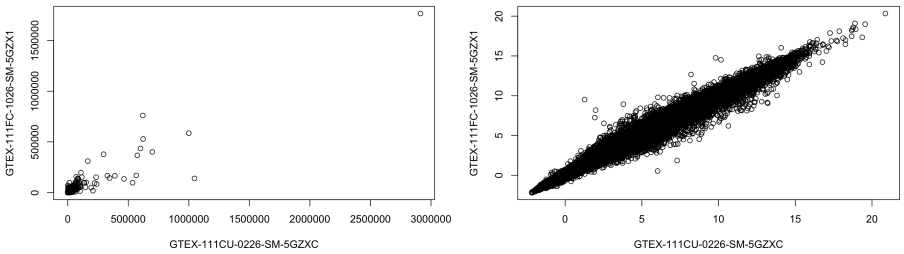
2.2. Análisis inicial.

Una vez elegidas las muestras se realizó un análisis visual sobre los datos en crudo y log2 de CPM para buscar posibles anomalías. El color rojo representa ELI y el azul NIT. Como se puede observar todas las muestras presentan una distribución similar por lo que no hay señales para desconfiar de la calidad de los datos.

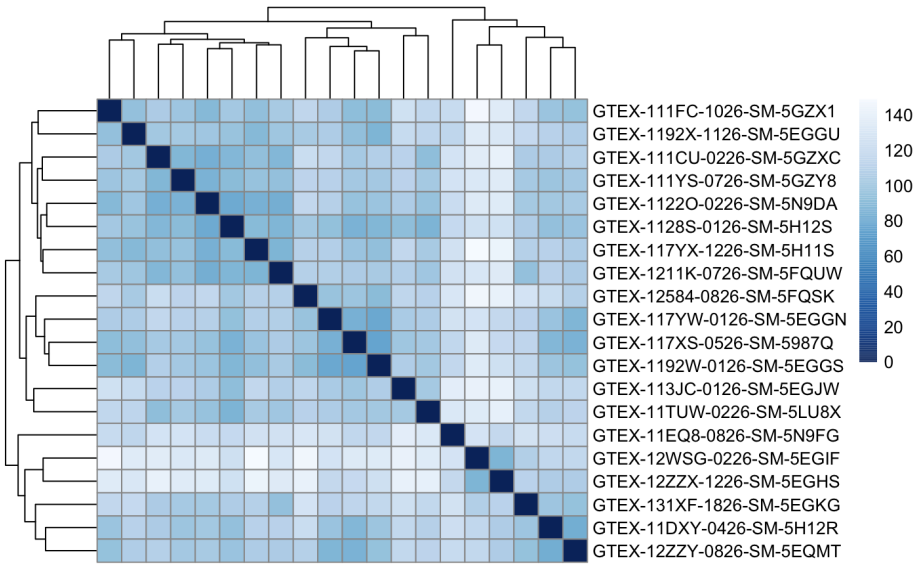


2.3. Filtrado y normalización

De los 56202 genes iniciales, se realiza un filtrado sencillo descartando aquellos que no se expresan quedando como resultado 40742 muestras. Sobre estas muestras se aplica el algoritmo rlog dado que el tamaño de la muestra es pequeño. A continuación se presenta un ejemplo del efecto de la transformación en una muestras. A la izquierda los datos en crudo y a la derecha los normalizados mediante rlog.



2.4. Heatmap



2.5. Análisis de expresión diferencial

De todas las opciones disponibles, por su simplicidad y cantidad de documentación disponible se opta por utiliza el paquete Bioconductor con las funciones `DESeqDataSetFromMatrix` y DESeq para crear una lista de genes cuya expresión es diferente entre los grupos.

De entre todos los genes obtenidos seleccionamos los más relevantes de acuerdo a su p-valor y su foldChange y añadimos SYMBOL derivada de los ENSEMBL IDs para facilitar la interpretación biológica de los resultados.

2.6. Resultados

La siguiente tabla muestra los genes que han diferenciado sus niveles de expresión entre grupos de una forma estadísticamete significativa:

log2FoldChange	lfcSE	stat	pvalue	padj	symbol
2.412671	0.5297461	4.554392	0.0000053	0.0009312	TNFRSF18
3.936202	0.6254408	6.293484	0.0000000	0.0000002	AMPD1
3.783565	0.7912021	4.782046	0.0000017	0.0003530	TTC24
5.708194	0.6624761	8.616453	0.0000000	0.0000000	FCRL5
4.869907	0.9200411	5.293141	0.0000001	0.0000313	FCRL2
3.881095	0.6107648	6.354484	0.0000000	0.0000001	SLAMF7

3. Discusión

- Por falta de tiempo no he podido analizar los otros grupos pero el proceso es exáctamente el mismo cambiando las variables al principio del script.
- Sería interesante contrastar los resultados con un biólogo para realizar una interpretación biológica completa.
- Existen muchas aproximaciones a este problema quizá estaría bien mezclar varias y buscar genes identificados por la mayoría.
  - El creador de Dseq propone un paso final de normalización que no se ha realizado en este estudio.
  - La universidad de Cambridge propone una aproximación basada en modelos lineales que parece más sencilla de comprender.

4. Apéndice

- Código en github
- Todas las referencias del texto son link al artículo original. (Hypercitation)