Comunidad de Desarrolladores de Software en Bioinformática

# Proyecto Colaborativo de Desarrollo de Software

**Nombre del proyecto:**
**Paquete para realizar un análisis datos de metilación de Illumina 450K**

**Responsable principal:**
**Dra. Yalbi Balderas**

**Otros responsables:**
**Candidato Dr. Miguel Negreros, Mtra. Queletzu Aspra, LCG. Semiramis Castro**

**Breve planteamiento del problema:**

La metilación del ADN se define como la adición de un grupo metilo en el carbono 5 (5mC) de algunas citosinas que se encuentran seguidas de una guanina (CpG), y se ha estudiado ampliamente por su relación con la regulación génica (Illingworth RS, et al., 2009). El dinucleótido se puede encontrar en acumulaciones a las cuales se les conoce como islas. Por mucho tiempo sólo se estudió la metilación de estas islas CpG ya que su presencia en la región promotora de los genes se ha asociado con la represión génica (Deaton AM et al., 2011), sin embargo, la metilación puede estar alejada de dichas islas (playas, arrecifes, cañones, etc) o en otras regiones génicas (cuerpo del gen, 5'UTR, 3'UTR, etc) generando resultados distintos (Jones PA, 2012). La plataforma de Illumina Infinium Human Methylation 450K utiliza sondas cuyo objetivo es cubrir la mayor cantidad de genes y categorías (respecto a las islas y las regiones génicas) posibles (Bibikova M, et al., 2011).

Para cada sitio CpG hay dos mediciones: una intensidad de metilación (M) y otra de ausencia de metilación (U). Estos valores de intensidades pueden ser usados para determinar la proporción de la metilación en cada locus CpG. Los niveles de metilación son comúnmente reportados como β-values ($\beta$ = M/(M+U+$\alpha$). Un sitio CpG puede estar metilado en el 100% de las hebras muestreadas, o en el 50% o en el 5%. Cuando se obtiene la distribución de los sitios se observa que los datos no siguen una distribución normal, sino que esta puede ser bimodal o de otro tipo. Normalmente se realiza una transformación de los β-values en M-values (M-value = log2(M/U) porque son más apropiados para las pruebas estadísticas según Du et al., 2010. Actualmente existen diversos paquetes en R para el análisis de metilación, y en particular existe un pipeline que utiliza diferentes paquetes (Maksimovic, 2017) y ha sido bastante funcional con algunas limitaciones, como el que se pueda utilizar una estadística que no asuma que las sondas siguen una distribución normal.

**Objetivo del proyecto:**

Para lograr un mejor entendimiento del comportamiento de los estados de metilación en las sondas del microarreglo de Illumina 450K u 850k proponemos como objetivo general:

1) Crear un paquete que incluya el pipeline de Maksimovic y conteste las siguientes preguntas:

- Evaluar la distribución de las sondas - cuáles sondas se distribuyen de manera bimodal y cuáles de manera normal u de otro tipo- Checar los paquetes citados (e.g. mclust), para probar bimodalidad al final en: https://en.wikipedia.org/wiki/Multimodal_distribution. Para tener una idea de por qué nos interesa realizar esto, revisar un ejemplo de aplicación en Wang, 2009. El pipeline de Maksimovic, incluye cómo realizar la gráfica de la distribución con todas las sondas, un material de apoyo también se puede revisar en: https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf
- Clasificar las sondas en los diferentes tipos existentes de distribuciones.
- Para cada grupo de sondas, identificar si varían de acuerdo a la posición del sitio CpG (*island, shore, open sea*, etc) – El pipeline de Maksimovic, incluye cómo relacionar esta información con los datos de metilación.
- Evaluar si las distribuciones cambian de acuerdo a las condiciones experimentales estudiadas, y cuáles son las sondas cuya distribución cambia.
- **Bonus extra:** En caso de que la distribución no sea normal, seleccionar una prueba estadística (https://stats.stackexchange.com/questions/58396/fitting-mixture-distributions-and-computing-goodness-of-fit) (e.g., Kolmogorov-Smirnof, Anderson-Darling test) que se adapte a este tipo de información para realizar el análisis de metilación diferencial (e.g., utilizar mixtools package, ADGofTest o el paquete correspondiente).

**Datos con los que se cuenta:**

Metodología (v.g. RNA-seq): **Metilación**
Plataforma (v.g. Illumina): **Illumina 450K**
Condiciones experimentales (v.g. tejido enfermo, tejido sano): **Fibroblastos obtenidos de pacientes con fibrosis pulmonar idiopática vs fibroblastos de personas sanas**
Réplicas (v.g. 3 réplicas por condición): **4 por condición**
Controles: **Normal lung fibroblast (4 réplicas)**
Información extra sobre los sets de datos: Se pueden encontrar en el siguiente link: **https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107226**

**Resultado ideal que debe generar el software:**

En los diferentes objetivos se espera que se generen gráficas, y en último objetivo si se llega a completar, una tabla con los resultados de las estadísticas obtenidas en el análisis de metilación diferencial similar al que se obtiene con el paquete limma (siguiendo el pipeline de Maksimovic).

**Referencias útiles:**

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R. High density DNA methylation array with single CpG site resolution. Genomics. 2011

Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011

Jones PA, Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012

Illingworth RS, Bird AP. CpG islands--'a rough guide'. FEBS Lett. 2009

Maksimovic, J., Phipson, B., & Oshlack, A. (2016). A cross-package Bioconductor workflow for analysing methylation array data. F1000Research, 5, 1281. http://doi.org/10.12688/f1000research.8839.3

Wang, J., Wen, S., Symmans, W. F., Pusztai, L., & Coombes, K. R. (2009). The Bimodality Index: A Criterion for Discovering and Ranking Bimodal Signatures from Cancer Gene Expression Profiling Data. Cancer Informatics, 7, 199–216.