Proyecto: Etapa 1

- Sebastián Lemus Cadena
- Gustavo Gaitán

1. Motivación y Relevancia Biológica

El estudio del transcriptoma provee una gran cantidad de información con respecto a los fenómenos celulares que suceden bajo diferentes estados del medio en donde se encuentran. Esto significa que se pretende hallar cuál es el perfil de expresión génica bajo estas condiciones, con el objetivo de conocer las diferencias en cuanto a la respuesta celular (Finotello, 2015). De esta manera, se puede estudiar organismos a través de toda la biósfera y derivar diversas aplicaciones puntuales. Específicamente, los estudios de conteo de RNA, por medio de técnicas experimentales como RNA-seq, permiten conocer qué genes se están expresando en medidas diferenciales cuando se aplican diversos tratamientos al sistema celular en estudio, a comparación de una situación control de expresión basal sin dichos tratamientos (Wang, 2019). Como se ha indicado, el rango de aplicaciones de estos estudios es muy amplio. Por ejemplo, en el campo de la medicina, se suelen realizar estudios de expresión diferencial para comparar los perfiles de expresión de células sanas a comparación de células cancerígenas, con la finalidad de entender cómo funcionan estas últimas y buscar posibles mecanismos de tratamiento (Bao, 2008). Por otro lado, en estudios de resistencia antibiótica, se pueden realizar este tipo de procedimientos para conocer cómo responden diferentes especies y cepas bacterianas ante la presencia de diferentes sustancias antibióticas (Li, 2019). Esto permite indagar sobre mecanismos de resistencia antimicrobiana, lo cual permite diseñar estrategias para manejar este fenómeno. Consecuentemente, estos estudios proveen datos de gran tamaño, en función de la cantidad de genes a evaluar, en algunos casos, incluso el transcriptoma entero. Por este motivo, es fundamental que se desarrollen herramientas para analizar estos datos, de forma que se pueda obtener la mayor cantidad de información relevante y adquirir mecanismos para evaluar hipótesis biológicas. Teniendo presente la importancia de este problema, se pretende diseñar e implementar una herramienta que permita analizar datos provenientes de ensayos de conteo de transcritos y mejorar los tiempos de ejecución y la precisión y exactitud de los resultados. Este objetivo permitirá mejorar la eficiencia del análisis de los datos obtenidos en estos estudios.

2. Formalización Computacional

- ***Entradas:***

Una matriz M de tamaño $n\,x\,m$ donde n es el número de genes evaluados y m es el número de ensayos desarrollados. El conteo del $i-\text{ésimo}$ gen para el ensayo $j-\text{ésimo}$ se representa como $M_{i\,j} \in \mathbb{Z}^{+}$. Esta matriz debe ser resultado de cualquier metodología de conteo de transcritos génicos, obtenidos por técnicas como la de RNA-seq.

- **Pre-condiciones:**

Los datos en la matriz M deben ser crudos; no debieron haber sido transformados previamente por algún proceso de normalización.

- **Salidas:**

Una matriz E de tamaño $n\,x\,m\,x\,k$ donde n y m tienen la misma connotación de las dimensiones de la matriz M y, $1 \leq k \leq 2$ indica el tipo de resultado para el análisis de expresión diferencial: $E_{i\,j\,1} \in \mathbb{R}$ es el valor de expresión diferencial para el $i-\acute{e}simo$ gen en el ensayo $j-\acute{e}simo$ y, $E_{i\,j\,2} \in \mathbb{R}$ es el valor de significancia estadística.

- **Post-condiciones:**

E es de tamaño $n\,x\,m\,x\,k$ y $E_{i\,j\,1} \in \mathbb{R}$ (es decir, la matriz debe estar completa y sin valores faltantes).

3. **Alternativas (software) que solucionan el problema**

- DESeq 2**:** https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

- edgeR: https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

- DSS: http://www.bioconductor.org/packages/release/bioc/vignettes/DSS/inst/doc/DSS.html#references

4. **Datos a utilizar**

- Datos de experimentos de expresión diferencial https://www.ebi.ac.uk/gxa/release-notes.html

- Datos de caracterización molecular de Cáncer sintetizados en datos de genómica, epigenómica, transcriptómica y proteómica: https://portal.gdc.cancer.gov/repository?files_offset=20

- Datos de experimentos para interacción planta-patógeno cuyos resultados se asemejan a aquellos arrojados por un ensayo en RNA-seq:

https://github.com/ngaitan55/proyecto_tesis/tree/master/Datos_crudos/Proyecto%20tesis/paradavid/featurecounts

5. Artículos de investigación que abordan el problema

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, *40*(10), 4288-4297.

Wang, T., Li, B., Nelson, C.E. *et al.* Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20,** 40 (2019). https://doi.org/10.1186/s12859-019-2599-6

- **Referencias teóricas (Motivación y relevancia biológica)**

Bao, T., & Davidson, N. E. (2008). Gene expression profiling of breast cancer. *Advances in surgery*, *42*, 249–260. https://doi.org/10.1016/j.yasu.2008.03.002

Finotello, F. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis, *Briefings in Functional Genomics*, Volume 14, Issue 2, March 2015, Pages 130–142, https://doi.org/10.1093/bfgp/elu035

Li Z, Xu M, Wei H, Wang L, Deng M. RNA‑seq analyses of antibiotic resistance mechanisms in Serratia marcescens. *Mol Med Rep*. 2019;20(1):745-754. doi:10.3892/mmr.2019.10281

Wang, T., Li, B., Nelson, C.E. *et al.* Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20,** 40 (2019). https://doi.org/10.1186/s12859-019-2599-6