General Assembly Studies Sixth Post

For my capstone project, I am working with genetic data taken from the blood cells of four healthy human donors. One of the primary goals of the Kaggle competition this project is based on is to build a model that predicts the presence of 136 proteins on the surface of a cell given the distribution of genes being expressed in that same cell, the transcriptome. The manner the data is acquired is quite noteworthy and provides valuable background for this project. Access to multimodal data like this (having transcriptome and protein presence data with resolution down to the cellular level) is a recent development. Measuring the transcriptome for thousands of cells simultaneously is new, seemingly formally introduced in 2015 in a paper by Macosko – McCarroll. I will describe that process first via my summary of Macosko – McCarroll (2015) and then expand to how 10X genomics simultaneously gathers surface protein data at a cellular resolution.

The first step is the creation of unique DNA barcoded microparticles (beads that are similar to the size of an individual cell) covered in special primers, such that, if this bead encounters free-floating RNA (in a solution), the RNA will connect to these primers. These primers are themselves RNA and contain two crucial sections. The first is a section of RNA unique to a given microparticle, called a barcode. The other RNA section is unique to each primer on the microparticle, called a unique molecular identifier (UMI).

Each microparticle is placed in a micro-droplet containing one cell using microfluidic devices. This cell is then lysed; its contents are released. Then the RNA within the cell is free to join to a primer on the microparticle. (These RNA sequences are associated with genes transcribed from DNA to RNA and released from the nucleus to produce proteins in the cytoplasm.) The droplets are combined, and copies of these RNA molecules (RNA encoding genes) and their associated primer are made using PCR. This tool amplifies the number of DNA samples (or RNA samples with a few extra minor steps) in a solution as an initial step before sequencing them.

The sequencing process produces a list of sequences for all the RNA encountered with the two critical components in the primer, the barcode and the UMI. Using the barcode, all the RNA snippets can be categorized by their cell of origin, and UMI can categorize everything to their associated primer; this allows for a count of the number of the given RNA snippet in the cell. (Remember we had to do an amplification step before sequencing to get accurate counts, we use the UMI to prevent overcounting an RNA snippet.) So, this process results in a table with rows as the cell and columns as the RNA fragment (RNA encoded gene) found, with the element as the amount encountered.

10X Genomics uses this process and adds one more complication. They supply antibodies created to target specific proteins embedded in the cell membrane’s outer surface. These antibodies also have an RNA tail that keys for this antibody type. Back in the microdroplet, when the cell and the micro-particle studded with RNA primers first encounter each other, these antibodies are also added. The antibodies bind to the proteins they have been keyed to and release their RNA tag, which also attaches to a primer on the microparticle. This microparticle now has RNA associated with the transcriptome and RNA keys related to the antibodies that were able to bind with the cell, indicating the presence of the protein they were keyed to. When the sequencing process from before is completed, there is a table describing the transcriptome of each cell and a table recording the amount of each of the protein types an antibody was keyed to on that same cell.

This Kaggle competition aims to build a model that maps a cell’s transcriptome to a prediction for the surface protein of each type recorded for that cell. I am excited to explore this data set and work to build these models.

Works Cited:

1. Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A. and McCarroll, S., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202-1214.
2. Kaggle.com. 2022. *Open Problems - Multimodal Single-Cell Integration | Kaggle*. [online] Available at: <https://www.kaggle.com/competitions/open-problems-multimodal/data> [Accessed 11 October 2022].
3. 2021. *Single Cell Gene Expression with Feature Barcode technology*. [ebook] 10x Genomics, Inc. Available at: <https://pages.10xgenomics.com/rs/446-PBO-704/images/10x\_LIT093\_Product-Sheet\_Explore-Cellular-Diversity\_Letter\_digital.pdf> [Accessed 11 October 2022].