Multicell RNA-Seq: Bridging the Gap between Bulk & Single Cell RNA-Seq

Jules Cazaubiel (BCB) & Nicholas Tourtillott (BCB)
Advisor: Dmitry Korkin (CS & BCB)

This report represents the work of one or more WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on the web without editorial or peer review.

Abstract

RNA sequencing has expanded nearly exponentially over the past decade and has allowed scientists an intimate glimpse into the expression patterns of cells and tissues. These techniques come in two main forms: bulk tissue RNA seq which captures a wide array of transcripts at a shallow depth across a tissue sample and single cell RNA seq which captures transcripts at a greater depth and tags them to individual cells. Here we present Multicell RNA-Seq, a pipeline of bioinformatics tools which together can be used to bridge the gap between bulk tissue and single cell techniques. Our approach allowed us to map specific transcript isoforms to single cell clusters and identify clustering levels beyond which transcript isoforms are no longer detectable within a single cell dataset.

Introduction

RNA sequencing has expanded nearly exponentially over the past decade as an area of biological research. It allows scientists to gain an intimate glimpse into expression patterns of cells or tissue at any given time and is leading the forefront of many different fields (Stark et al.). From immunological research to oncology and more we see RNA sequencing techniques being leveraged to explore how we understand the inner workings of cells and particularly their associated disease states. While RNA sequencing is an umbrella term for these experimental techniques, they do come in two main forms, bulk tissue sequencing and single cell sequencing, each of which have their own advantages and disadvantages (Stark et al.).

Compared to biological techniques of the past RNA sequencing allows for researchers to study hundreds of genes at once. Historically, scientific studies were often limited in their scope due to the ability to only observe the activity of a potential handful of genes at once, however modern techniques of RNA sequencing have blown those numbers out of the water. Comparing within the umbrella of RNA seq techniques you'll find that bulk seq and single cell seq lend themselves to different types of studies (Stark et al.). Bulk sequencing allows scientists to capture a wide array of transcripts across a tissue sample, often sacrificing sequencing depth for breadth (Thind et al.). However, this allows for the recognition of numerous transcript isoforms that single cell seq fails to capture. Comparatively, single cell seq allows for the capture of transcripts at a greater depth and even allows for scientists to associate transcripts with an individual cell (Haque et al.). Essentially, this allows for the investigation of individual cell states and expression patterns associated with a cell type, versus the general exploratory nature of bulk seq signatures which are not tied to a specific cell type.

We aim to introduce a new computational technique called Multicell RNA seq, which would allow scientists to bridge the gap between single cell and bulk tissue sequencing. Essentially, it would allow us to trace transcript isoforms captured in bulk tissue seq to a particular cell type within a single cell seq sample– thereby allowing for the sacrifice of some depth of single cell seq data, but bolstering it with the breadth provided by bulk tissue seq.

Additionally, this method relies entirely on preexisting bioinformatics tools, such as a transcript quantifier and a hierarchical clustering tool. Rather than create our own tool from scratch we found it simpler to amalgam existing tools together to produce our computationally transformed data given their acceptance and popularity in the bioinformatics field, however that has also proven difficult in finding how to fit the pieces together in such a way that this method proves functional. With that being said, multicell RNA seq provides scientists a novel method of applying the breadth of information captured in bulk seq techniques to the depth and cell specificity to single cell seq data and vice versa– enriching both datasets to further aid in all involved areas of biomedical research.

Background

Multicell RNA seq is a computational method that allows scientists to leverage previously published and widely used bioinformatics tools in a novel pipeline which can bridge the advantages of bulk tissue and single cell seq methods together, enriching the analysis of both datasets. Our pipeline relies on the infrastructure of popular bioinformatics tools and combines a transcript quantifier and a hierarchical clustering tool with the simple yet clever construction of a table. Through a unique balance of these tools we have established a novel computational pipeline which exploits the strengths of each RNA seq technique. As we progress through the necessary background information we will dive into each RNA seq technique in the context of this pipeline, alternative splicing in the context of our project, explain our rationale on tool selection, and finally provide a short overview of this paper's organization.

Firstly, RNA seq methods have revolutionized genetic research by expanding the breadth of genes studied at a particular time from a handful to hundreds within one experiment (Stark et al.). Rather than expanding upon the specific experimental methods used to capture and sequence RNA transcripts in single cell and bulk tissue experiments, we wish to delve into the unique forms of data provided by these techniques which we then leverage computationally for our pipeline.

As previously noted, RNA sequencing techniques allow scientists to capture and sequence the RNA transcripts of tissues and individual cells at a variety of depths and breadths (Stark et al.). These datasets are then leveraged by computational and bioinformatics tools for analysis given their sheer size and complexity, which would render this data simply unmanageable in their absence. Specifically, our methods leverage data from single cell RNA seq and bulk tissue seq experiments. Single cell RNA sequencing allows scientists to capture RNA transcripts at a great depth while also associating each transcript to a single cell within the sample (Haque et al.). These datasets allow researchers to delve into the expression patterns of individual cells for the exploration of specific cell types, subtypes, and states given the particular state of a sample whether diseased, healthy, cancerous, or otherwise. Essnetially, it gives scientists a snapshot of the gene expression of cells given a particular condition which has been leveraged in most reaches of biology to gain a more intimate glimpse at the genetic workings of disease and other biological processes (Haque et al.).

Conversely, bulk tissue RNA sequencing techniques capture a wide breadth of RNA transcripts across a tissue sample (Thind et al.). This allows for scientists to even observe unique transcript isoforms within an experimental sample, however unlike single cell seq these techniques do not allow scientists to associate these transcripts with specific cells. The strength that bulk tissue seq allows scientists a wide breadth of sequencing which captures and observes numerous expression signatures and potential expression patterns, essentially providing a unique method of exploratory genetic analysis whose insights easily inform more targeted future experiments (Thind et al.).

The purpose of our method is truly to trace the alternative splicing of transcripts from one type of RNA seq data to the other. Basically, we wish to use our pipeline to define a level of cell clustering within single cell data which maintains a signal of alternative splicing when treating that data as if it were bulk RNA seq data, which adequately captures transcript isoforms. Our motivation for doing so is honestly quite simple. The alternative splicing of RNA transcripts has major implications for cell function and gene expression and while research within the RNA seq space has been exploring the implications of RNA expression for cell function such experiments have a great potential in understanding alternative splicing events (Zhao). While interest in alternative splicing events captured within RNA seq data has been increasing it largely remains an area of research with much untapped potential (Zhao). Thus, the method we present would give researchers a method into which they could explore the alternative splicing of transcripts within any given tissue or disease state for which they possessed the necessary datasets. This would allow for a unique perspective into the mechanics of cell function or disease and a closer more intimate look into how alternative splicing affects these processes– similar to how methods of RNA seq originally granted a more intimate look into how gene expression related to specific cell states.

Moving forward to computational tools utilized in our method we first start with a transcript quantifier. Transcript quantifiers take the raw reads from RNA seq experiments and perform a series of alignments and computations, allowing scientists a wide variety of information on the ubiquity of transcripts in the sample and allows for the annotation of those reads (Zhang et al.). These tools come in two varieties: those which rely on alignment to a reference file and those which use methods of pseudoalignment (Zhang et al.). We decided upon Kallisto for use in our pipeline because of its pseudo alignment based methods. Kallisto allows for transcript quantification of both forms of RNA seq data relevant to our method, while also not needing a reference genome to work with our data– as is the nature of pseudo alignment methods (Bray et al.). Our choice of Kallisto compared to other contemporary quantification methods came from its unique position in the bioinformatics world. The popularity of Kallisto means that its outputs are not only easily integrated into other tools published by the same lab, but that they are also easily used as inputs for a wide variety of tools within the bioinformatics space. This proves uniquely well suited for our task since we aimed to structure Multicell RNA seq as a novel combination of previously published tools to take advantage of their scientific acceptance to exploit and manipulate data in a novel manner. Additionally, we decided on

Kallisto as it does not require a reference file to generate transcript abundances from the input RNA seq data (Bray et al.). Meaning, our pipeline would then not rely on the existence of a high quality reference genome for a specific organism to be used on unique RNA seq datasets. This not only expands the use of our pipeline to less well characterized organisms, but frees it from the reliance of the existence of yet another high quality piece of data further expanding its use cases by reducing the resources necessary to run it.

The other backbone of our method is DUSC, an approach for cell type profiling within single cell data (Srinivasan et al.). Essentially, this method allows us to construct a cell type hierarchy of our single cell data, which we then utilize the various levels within that hierarchy to process as our pseudobulk data when compared to true bulk RNA seq data (Srinivasan et al.). The idea being that when comparing to bulk RNA seq transcript abundances we can find the level of cell type clustering within single cell data that still maintains a reasonable signal of alternative splicing of transcripts, whether that is at the level of cell super type, cell type, cell subtype, etc. DUSC allows us to construct this hierarchy as it first relies on a deep learning feature selection method, DAWN, which finds a number of latent features to allow for the accurate clustering of single cell data (Srinivasan et al.). Important to note, DUSC is also an unsupervised method which not only outperforms other unsupervised methods within this space, but whose accuracy actually approaches that of supervised methods (Srinivasan et al.). Similar to that of Kallisto, the use of an unsupervised method was important to our team because it would expand the use cases of our pipeline since it would not rely on the existence of some high quality reference data to operate.

Finally, as we progress through this paper we will discuss the construction and workflow of our pipeline, and its application to mouse sensory neuron datasets. Then we will explore further uses of multicell RNA seq, its limitations as a method, and our insights on potential next steps of our pipeline's use in research and future iterations of it.

Methods

To formulate the problem multicell RNA seq aims to address more specifically, we are establishing a pipeline of pre-existing bioinformatics tools to leverage single cell seq and bulk tissue seq datasets to then exploit the strengths of one dataset and apply them to the other, thereby bridging the gap between the two methodologies. With this method scientists could specifically trace the transcript isoforms captured by the breadth of bulk tissue seq to the specific cell types captured within the depth of single cell seq datasets. As we move forward we will first introduce our datasets of interest, then the workflow of our data processing pipeline, and finally the visualizations and analysis possible with Multicell RNA-Seq. However, for brevity, an overview of our pipeline can be seen below in Figure 1.

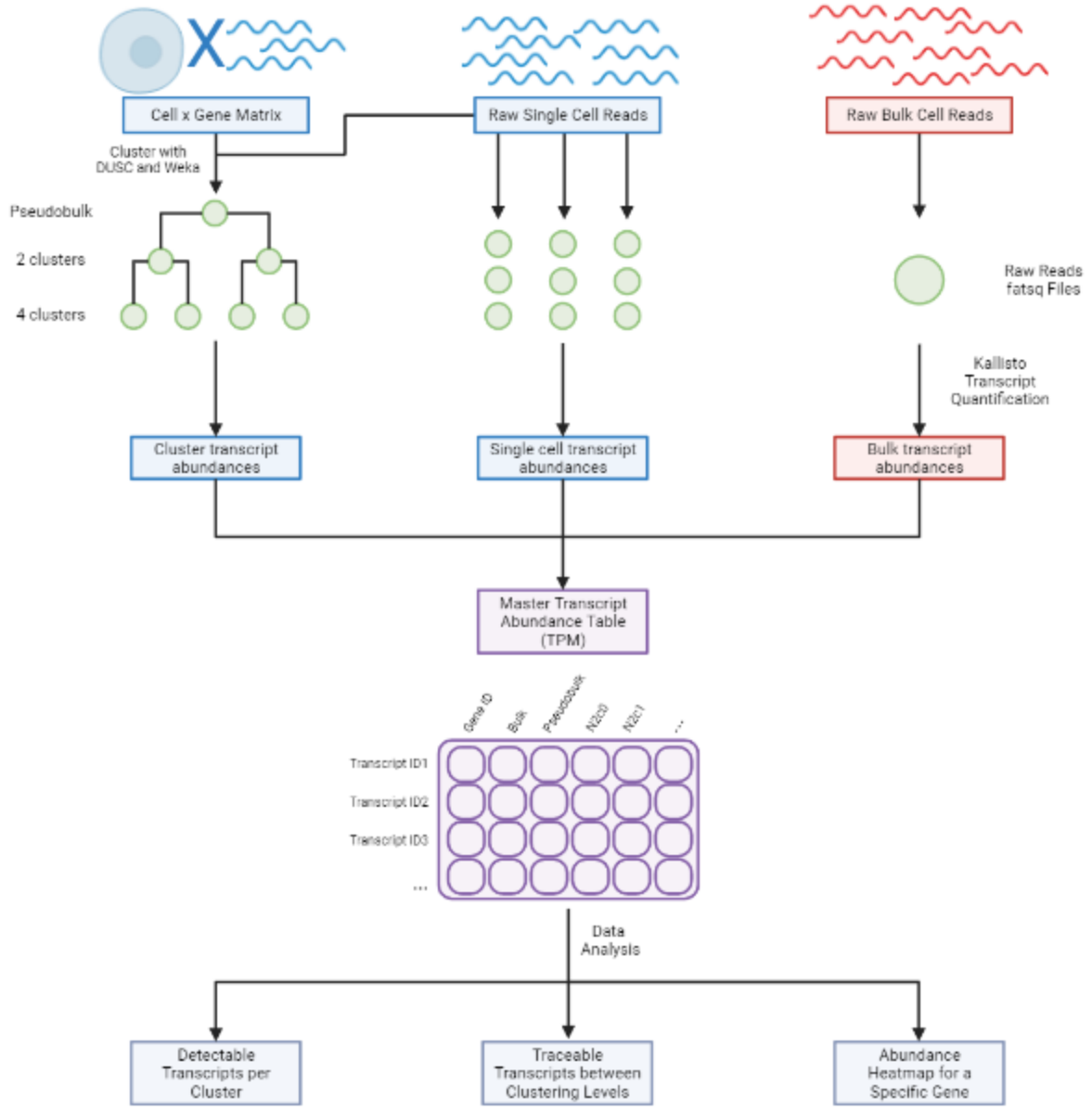

Figure 1. Overview of the Multicell RNA-Seq pipeline. A cell x gene matrix is obtained for the single cell dataset, along with the raw reads for both the single cell and bulk RNA-seq experiments. DUSC is then used to create a hierarchical clustering for the single cells, and clusters are created. Transcript abundances in TPM for all raw data files are obtained using Kallisto, and a master transcript abundance table is produced. This table can then be used for data analysis, to quantify and trace detectable transcripts or study variation of abundances of alternative splicing isoforms for specific genes.

Dataset Acquisition

In order to build and test our pipeline, we needed robust datasets. In an ideal situation, both the bulk and single cell datasets used for our study would have originated from the same

source, but we settled for two publicly available RNA-seq datasets from peer-reviewed publications. It is important to note that both RNA-seq datasets were drawn from the same tissue type, namely mouse sensory neurons, and that they came in the form of fastq files. We obtained raw RNA reads from two separate experiments: a single cell RNA study by Usoskin et al. from 2014 (with 863 cells sequenced) and a bulk RNA study by Zheng et al. in 2019. Obtaining raw RNA-seq reads was crucial to our process, as the merging of single cell RNA-seq data by our pipeline to amplify signals required reads, rather than abundances. Additionally, we also obtained a cell x gene matrix for the single cell dataset.

In order to download and store these large datasets on our server, we used the SRA accession codes provided in the original publications, along with the SRA Explorer tool. This allowed us to generate and run downloading scripts, leading to an easy and automated process.

Generating Hierarchical Clustering

We started our process by generating a hierarchical clustering of the cells in the single cell dataset using the DUSC tool (Srinivasan et al.).. To do so, we processed the cell x gene matrix, removing unnecessary columns and transposing it, leaving only the numerical expression values in the format required by the tool. We then ran the feature learning part of DUSC, Dawn, on this cell x gene matrix, resulting in the production of latent features that could be utilized for hierarchical clustering. Then, these latent features were transferred to Weka to generate the hierarchical clustering assignments. We selected Weka's hierarchical clusterer with the filtered distance function and Ward linkage, as this combination of parameters resulted in the best separation of clusters. This resulted in the production of a stable hierarchy for our single cells, and cluster assignments for different levels of clustering were obtained for each cell. Overall, we selected a range of clustering levels: 2, 4, 16, 64, and 128 clusters per level respectively. Each cell's cluster assignment for each clustering level was recorded in a comprehensive csv file, for ease of use later on. These clusters followed a custom naming convention, with N# being the total number of clusters at that level, and c# being the individual cluster number. As an example, the N2c0 cluster belonged to the clustering level containing 2 clusters, and was the first cluster of this level.

Aggregation of Clusters

We then set out to merge our single cell data according to the cluster assignments obtained from DUSC. Due to the nature of our dataset, it was not uncommon to have duplicate files for individual cells. In order to match the cluster assignment to only one file per cell, we opted to use the first raw reads file matching each cell. While this may not be the best method to deal with duplicate files in single cell experiments, we opted to do so due to time constraints and ease of implementation. Future refinement of our pipeline may lead to changes in the handling of duplicates, such as merging them or going through a more rigorous selection process.

In order to implement the merging of single cell data to create raw data files for our clusters, we built a custom python script and ran it on our server. This script iterated through the

csv containing the single cells' cluster assignments, and merged them accordingly. This resulted in each cluster having its own fastq file, a collection of all the raw reads belonging to all cells in the cluster. We also created a pseudobulk file by merging the raw reads for all the cells in the dataset (equivalent to a clustering level with only one cluster).

Transcript Quantification

The next step in our pipeline was to produce abundance files. We used Kallisto to do so, a well known tool used to quantify transcript abundance from raw reads of both bulk and single cell RNA-seq experiments (Bray et al.). As a well established and versatile tool, Kallisto is widely used to generate transcript abundance files from RNA-seq data, and requires only raw data files along with a reference index file for the organism being considered. We obtained the Ensembl Mus musculus GTF file and used it to create the index required by Kallisto, before processing all raw data files including the previously created cluster files.

Aggregation of Results and Visualization

We then created a master transcript abundance table to carry out our analysis. This table was created using a custom python script and the transcript abundance files from Kallisto, with the metric being used for abundance being transcripts per million (TPM). This script iterated through kallisto abundance files and recorded their TPM column, using the custom cluster number as an identifier. With it, we were able to quickly and efficiently assess the differences in transcript abundances between clusters or clustering levels. A major part of our analysis was looking at the proportion of detectable transcripts (with TPM above zero), along with tracing detectable transcripts from the bulk dataset to different clustering levels. Custom python scripts were used to do so, recording a list of detectable transcripts for each cluster before making use of them to make comparisons across clusters and clustering levels.

Results

As a team we created Multicell RNA-Seq to be a pipeline which could investigate alternative splicing events captured by bulk experiments within a single cell dataset, thereby leveraging the strengths of one dataset to bolster the other. This is especially the case when considering through the tracing of alternatively spliced isoforms, we can now identify the cell type from which those expression signatures are coming from, which traditionally is information that is unavailable from a standalone bulk RNA seq dataset. The importance of our method also highlights the fact that existing gold standard bioinformatics tools may be repurposed or combined in novel ways to create a new scientific product. Not every method needs to reinvent the wheel or establish the next big algorithm to have a measurable impact on the scientific community.

Moving forward we would like to frame the structure of this section. Understandably, since Multicell RNA-Seq is a computational pipeline our results section will read unlike other traditional articles which are concerned with the analysis of a particular dataset. Rather than

display specific findings, we will outline a variety of applications for this method and comment on limitations faced by the current iteration of Multicell RNA-Seq.

Current Applications

Broadly, Multicell RNA-Seq allows us to detect and trace alternative splicing events from bulk RNA seq and single cell RNA seq datasets, assuming they come from the same tissue types. This analysis truly comes in a few main branches which can be applied in scientific investigations. These branches being: transcript isoform detection, transcript isoform tracing, and genetic and specific transcript isoform investigation. Generally, these branches are all steps in a workflow of exploratory data analysis and are by no means the limits of the applications offered by Multicell RNA-Seq, though they are likely the first steps in any application of this pipeline.

First off our method relies on setting a TPM cut off for transcript detection within a sample. Showcased below in Figure 2, this allows you to account for the number of transcripts which are reliably detectable at each sample level. This step in the exploratory workflow allows you to establish a baseline for your intersample comparisons and familiarize yourself with the distribution of the newly generated dataset. Generally, you should see trends as you do in the figure below where the bulk and pseudobulk levels of clustering should have the most detectable transcripts. From there you should see a steady decline of detectable transcripts between each new level of clustering, though it is normal to see decent variation of detectable transcripts within a clustering level. While this branch is admittedly poor in groundbreaking scientific insights it allows users to establish a ground truth for their dataset and see greater trends within the behavior of transcript isoforms at a cell type/subtype level in their data.

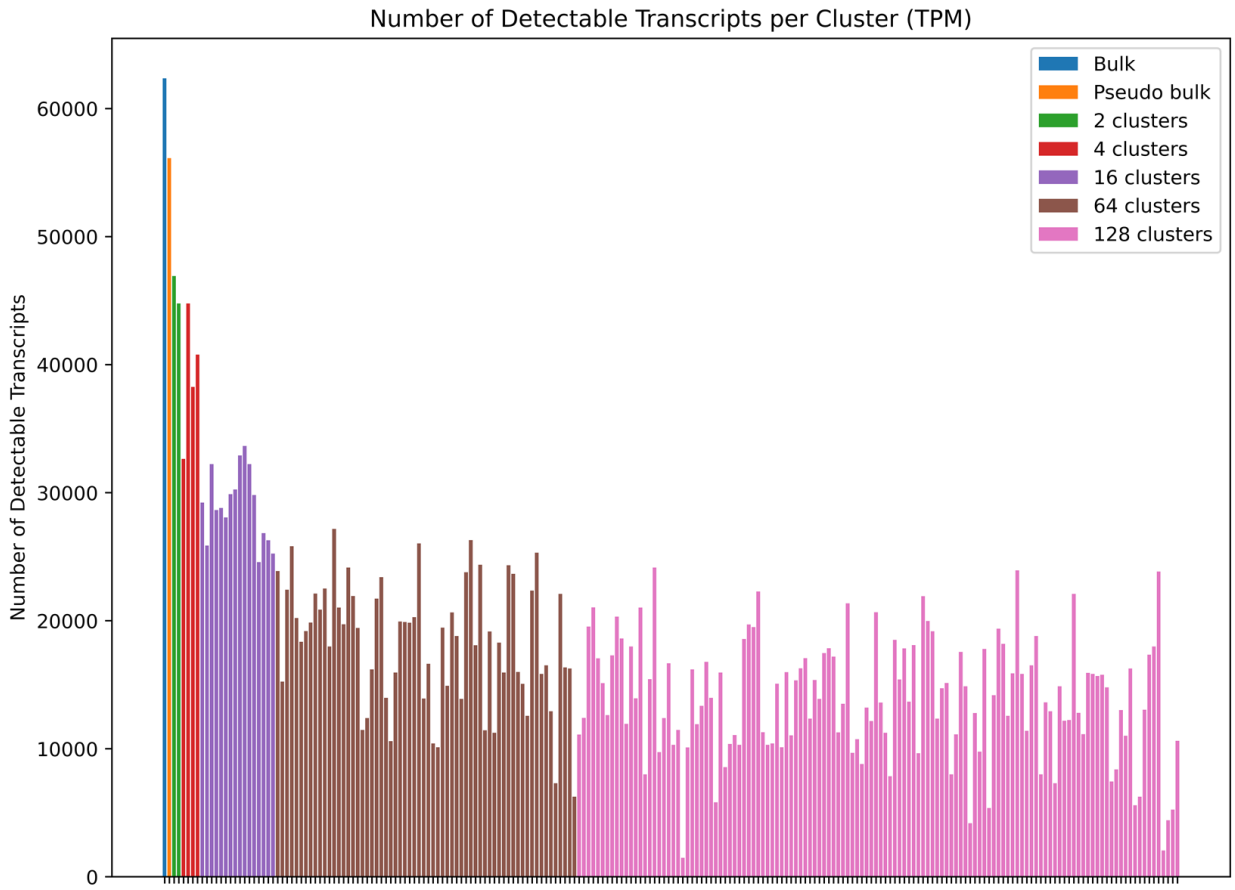

Figure 2. Bar chart showing the number of detectable transcripts (with TPM > 0) for each cluster. This allows investigators to account for the number of transcripts which are reliably detectable at each sample level. Most datasets should see the most detectable transcripts at the bulk and pseudobulk levels of clustering and from there you should see a steady decline of detectable transcripts between each new level of clustering, though it is normal to see decent variation of detectable transcripts within a clustering level. Such a visualization allows investigators to establish a ground truth for detectable transcript distributions at cell clustering levels within their dataset.

Moving forward in the workflow, the next step would be to identify traceable transcripts within your sample. Essentially, a transcript is traceable from your bulk dataset to a specific level of clustering within your single cell dataset if it is detectable at the bulk level and every other level of clustering up to that specific level. This is the truly novel application of Multicell RNA-Seq. What you will notice in using this analysis in your own datasets is that at some level of single cell clustering there will be a sheer drop off of traceable transcripts. This is likely due to the sparse nature of single cell data, because for a transcript to be traceable to a particular level the combined signal of expression from all cells in that cluster must be loud enough to be detected. However, given the sparsity of single cell data at lower levels of clustering the signal

simply will not be strong enough to be detected. Such a behavior can be seen below in Figure 3, where a massive drop off of traceable transcripts takes place between the 16 and 64 clustering levels. The point of such an analysis then is to find the sweet spot between signal and noise for transcript tracing, where we can trace as many transcripts from bulk datasets to clusters of single cell without succumbing to the noise of housekeeping genes or being lost to the granularity and sparsity of single cell data. After which investigators would be able to delve into the unique transcript isoform behaviors displayed in particular cell clusterings.

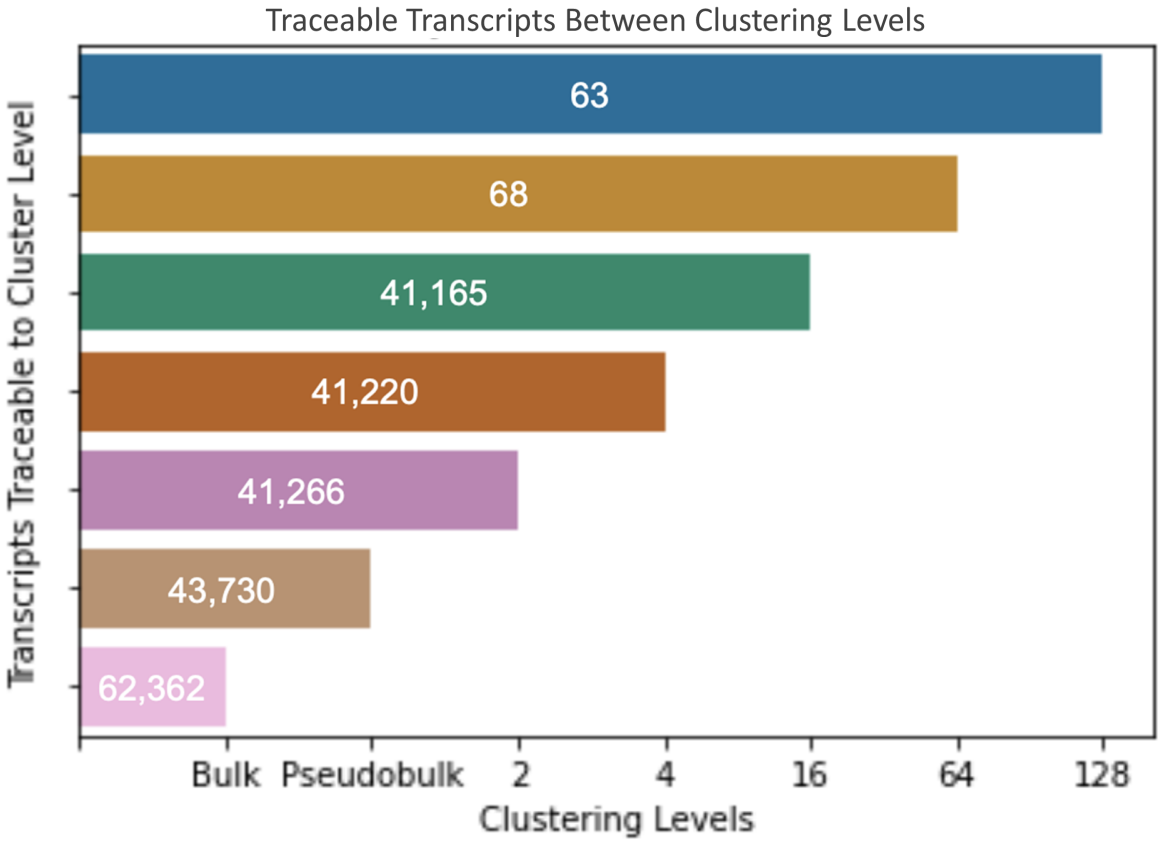

Figure 3. Bar plot showing the number of traceable transcripts between clustering levels. This plot depicts the number of traceable transcripts at a particular level of clustering. A transcript is traceable from a bulk dataset to a specific level of clustering within a single cell dataset if it is detectable at the bulk level and every other level of clustering up to that specific level. As can be seen above, at a particular level of cell clustering there will be a dramatic drop off of traceable transcripts. This behavior is likely due to a loss of signal for transcripts at lower levels of single cell clustering due to the sparsity of such datasets.

In a similar vein, our method is uniquely suited for the investigation of individual genes and more specifically the different isoforms of the transcripts for said gene which occur due to alternative splicing events. Likely the most common and first step in these types of investigations

will be through the comparison of expression patterns of each transcript isoform, an example of which can be seen in Figure 4. The gene of interest in this figure is the eukaryotic translation elongation factor 2 (Mus musculus, ensembl genome browser 106) and the multiple transcript isoforms of the gene appear to be highly expressed somewhat consistently at all levels of clustering of the dataset. However, of note such a visualization is vital in seeing differences in isoform expression as is the case of the second isoform from the other two which appears to be uniquely expressed in a few cell clusters compared to its other two relatives. Such a behavior could denote unique expression patterns in specific cell types and subtypes and could be the beginnings of an investigation into the expression patterns and biological implications of alternative splicing events in this gene of question.

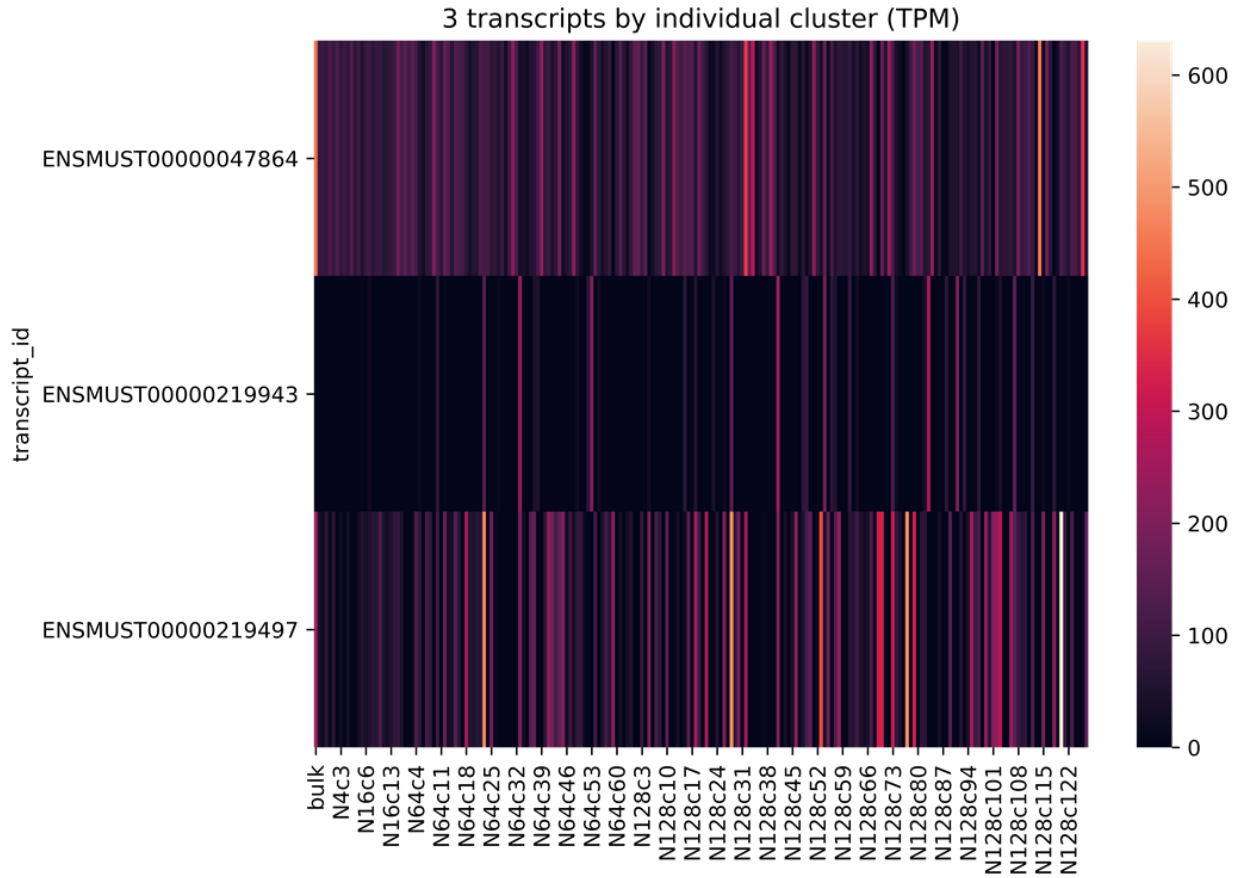

Figure 4. Heatmap showing the abundance in TPM of three alternative splicing isoforms of the Eef-2 (eukaryotic translation elongation factor 2) gene in all clusters. The lighter the band, the more this particular transcript is expressed in that cluster. This allows for a quick visualization of the differences in expression levels for isoforms of the same gene, and allows mapping of specific expression patterns to singular clusters. As an example, we can see that the second transcript is only expressed in a select few clusters, hinting at different expression patterns in cellular types or subtypes.

Of note, these applications described above are merely the beginnings of possible investigations of transcriptomics data using Multicell RNA-Seq. They truly all fall in the realm of exploratory data analysis and the true strengths of this method can only be leveraged through the targeted application of analysis. With that in mind we suggest that investigations into alternative splicing events and isoform expression are in the best setting to leverage the analysis that Multicell RNA-Seq makes possible and that such analyses could likely benefit from being combined with traditional methods of single cell or bulk RNA seq analysis.

Current Limitations

The most glaring limitation we face comes in the form of the metric we use to quantify RNA abundance in our samples. Currently we use transcripts per million (TPM) which scales the counts for transcripts within a sample according to their proportion within that individual sample. However, as noted by current research TPM is a poor metric for intersample comparison of RNA seq data (Zhao, 2021). The desired metric is normalized counts, a metric which (as the name suggests) normalizes all RNA transcript counts across samples– allowing for a more reliable comparison to be made between different Kallisto sample runs.

To achieve such a metric we suggest a method provided by DeSeq2 (Love, 2014). Not only does this method easily integrate with Kallisto, the output normalized counts can easily be extracted from this R package and input directly into our table as described previously in the methods. Correcting this RNA abundance metric will add another step to Multicell RNA-Seq, however it would keep in line with the original intent of cleverly using gold standard bioinformatics tools to leverage and analyze data in a new way. It would also shore up the insights drawn from any analysis performed with Multicell RNA-seq by providing a proper ground truth to draw intersample relationships from.

In another direction, Multicell RNA-Seq is limited by the availability of similar datasets. Ideally, this method would be employed on bulk and single cell datasets which were derived from the same tissue sample or from the same tissue type of the same host. The more similar the datasets the more sound the conclusions and insights drawn from the expression patterns will be. However, such an ideal scenario is unique to say the least. Though in that vein as both forms of RNA Seq experiments become more abundant and cheaper in cost the abundance of said datasets will increase. While it will still be rare to see the exact same sample donors, it will hopefully become more common to have both sets of datasets for organism tissues so that Multicell RNA-Seq can be employed in those cases.

Conclusions

With the aforementioned limitations in mind, we would finally like to discuss potential areas of research which could benefit from the application of Multicell RNA-Seq and the future directions of this pipeline. However, before we move on to future applications we would like to note one more takeaway from this project. New insights and forms of analysis can be achieved

through the clever use of published gold standard bioinformatics tools. In some areas of the scientific community there may be pressure to create the next tool or write the next breakthrough algorithm, but to achieve new insights from data we need only get clever not invent another wheel.

Multicell RNA-Seq as a method would be best applied in areas of research interested in the biological implications of alternatively spliced isoforms on cell types and disease states. That is the true strength of our method– the ability to trace transcript isoforms from the breadth of bulk RNA seq to the granularity of single cell RNA seq. To some extent we can not predict which areas of research in particular will benefit the most from the application of our pipeline, however, we can make a few educated guesses based on our current understanding of alternatively spliced isoform behavior. With that in mind, we would think that Multicell RNA-Seq would be the most beneficial in the investigation of developmental biology as well as any number of disease states since alternatively spliced isoforms remain a largely untapped area of biology (Zhao).

Finally, we would like to address the next steps for Multicell RNA-Seq as a project on its own. While we did address some limitations of our pipeline previously, we would be remiss to repeat this one in particular. Future researchers should first take our method and adapt it to utilize normalized counts rather than TPM as the metric used for intersample comparisons. Updating this comparative metric will only add more weight to the biological insights taken away from the analyses made possible by our pipeline. Moreover, we would like to suggest the addition of cell type annotation to the single cell dataset used. As it stands we are able to trace alternatively spliced isoforms to specific cell clusters, though to easily understand the data we are looking at in a biological sense we would need to know what cell types are in the cluster that we are looking at which a method of cell type annotation would provide. In a similar vein, adding a method of transcript or gene function annotation would add more easily accessible biological insight to our pipeline and doing so would only further empower investigators in their research. So in short, we suggest that future directions of this project take the key driver of this pipeline even further– to empower investigators to delve into the implications of alternatively spliced isoforms on biological processes.

Github Access:
https://github.com/Jcazaubiel/Multicell-RNA-seq

Bibliography

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Froussios, K., Mourão, K., Simpson, G., Barton, G., & Schurch, N. (2019). Relative abundance of transcripts (rats): Identifying differential isoform abundance from RNA-seq. *F1000Research*, *8*, 213. https://doi.org/10.12688/f1000research.17916.1

Hamilton, M. J., Girke, T., & Martinez, E. (2018). Global isoform-specific transcript alterations and deregulated networks in clear cell renal cell carcinoma. *Oncotarget*, *9*(34), 23670–23680. https://doi.org/10.18632/oncotarget.25330

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for Biomedical Research and Clinical Applications. *Genome Medicine*, *9*(1). https://doi.org/10.1186/s13073-017-0467-4

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). https://doi.org/10.1186/s13059-014-0550-8

Patrick, R., Humphreys, D. T., Janbandhu, V., Oshlack, A., Ho, J. W. K., Harvey, R. P., & Lo, K. K. (2020). Sierra: Discovery of differential transcript usage from Polya-captured single-cell RNA-seq data. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02071-7

Srinivasan, S., Johnson, N. T., & Korkin, D. (2019). A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. https://doi.org/10.1101/511626

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Summary - mus_musculus - ensembl genome browser 106. (n.d.). Retrieved April 26, 2022, from https://uswest.ensembl.org/Mus_musculus/Gene/Summary?db=core%3Bg

Tekath, T., & Dugas, M. (2021). Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle. *Bioinformatics*, *37*(21), 3781–3787. https://doi.org/10.1093/bioinformatics/btab629

Thind, A. S., Monga, I., Thakur, P. K., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M., & Ashford, B. (2021). Demystifying emerging bulk RNA-seq applications: The application and utility of Bioinformatic Methodology. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab259

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P. V., Linnarsson, S., & Ernfors, P. (2014). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience*, *18*(1). https://doi.org/10.1038/nn.3881

Zhang, C., Zhang, B., Lin, L.-L., & Zhao, S. (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-4002-1

Zhao, S. (2019). Alternative splicing, RNA-seq and drug discovery. *Drug Discovery Today*, *24*(6), 1258–1267. https://doi.org/10.1016/j.drudis.2019.03.030

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *Journal of Translational Medicine*, *19*(1). https://doi.org/10.1186/s12967-021-02936-w

Zheng, Y., Liu, P., Bai, L., Trimmer, J. S., Bean, B. P., & Ginty, D. D. (2019). Deep sequencing of somatosensory neurons reveals molecular determinants of intrinsic physiological properties. *Neuron*, *103*(4). https://doi.org/10.1016/j.neuron.2019.05.039