Slide 3: Single-cell RNAseq

Whereas bulk RNA sequencing allows quantification of population-level gene expression of a tissue, the recent development of high-throughput, single-cell RNA sequencing (scRNAseq) enables quantification of gene expression in thousands of individual cells, allowing for the resolution of different cell states within a tissue. Dissecting cellular heterogeneity has implications in cell type discovery, cellular differentiation, stochastic expression analysis and lineage construction.

Slide 4: Experimental Workflow

There are many different protocols for scRNAseq, but the general workflow consists of the following steps:

1. Tissue sectioning, to acquire the sample of interest
2. Tissue dissociation, to separate the sample of interest from surrounding material and can be done mechanically, enzymatically or through microfluidics
3. Cell enrichment: an optional step to increase the relative abundance of the cell type of interest
4. Single cell isolation to isolate the cells in order to extract the RNA from a single nucleus and can be done with a variety of techniques
5. Cell lysis to obtain the genetic material from the interior of the cell
6. mRNA selection to isolate mRNA from the other contaminants
7. Library preparation to prepare for sequencing
8. Sequencing

Slide 5: Single Cell Isolation

Single cell isolation can be accomplished through some of the following steps:

- FACS (fluorescence-activated cell sorting): Antibodies bind to specific markers on cells, granting fluorescence. This suspension is then run through a laser that then sorts the cells based on their fluorophore (and thus, cell type).
- LCM (laser-capture microdissection): Cells are labeled with a marker and then physically separated using a laser. This tends to be hard on RNA.

Slide 6: Single Nucleus RNAseq (snRNAseq)

One limitation to single-cell sequencing is that it requires viable, intact cells to be separated for sequencing, which means it’s not well suited for use with frozen tissue. Additionally, the harsh tissue dissociation steps can disrupt sensitive cells.

A more robust alternative is snRNAseq, whereby whole tissue is homogenized and nuclei are isolated rather than cells. Studies have shown that the two methods have high concordance of expression with each other. The dataset we analyze is snRNAseq.

Slide 7: Bioinformatics

Single-cell sequencing data tends to have a lot more variability and noise than bulk RNA-seq, and therefore has some specific data processing steps to mitigate this.

- Unique molecular identifiers (UMIs) can be used (which can be added to protocols for more accurate expression quantification)
- We also have metadata for both datasets, including post-mortem inveral (PMI), sex, age and more factors as a comma-delimited file. to help normalize variability in depth.
- scRNAseq is prone to "drop-out" events, where lowly expressed genes may go undetected (due to sampling issues or etc.). These can be corrected by imputation with appropriate values statistically, or addressed using a hurdle model which fits a separate process for the artificial (drop-out) zeroes.
- Because of the high dimensionality and variability of the dataset, a dimensionality reduction and feature selection is often done to focus downstream analyses
- Finally, you need to identify and remove cells that have been compromised by the preparation protocol, and can therefore compromise your data. This can be carried out by features such as expression patterns of GO terms, RNA integrity numbers (RIN), high proportion of mitochondrial genes, and etc.

This leads to an important question, and the one we explore in our project: are some cell types more likely to be compromised than others?

Slide 9: Project Motivation and Hypothesis

Because scRNA and snRNAseq methods are not yet fully mature, they may demonstrate some uncharacterized bias. This bias would likely be technical due to library preparation protocols that impact cell survival based on morphological, physical or chemical features of different cell types. If that's the case, then these biases may be detected as differential gene expression between the bulk and single-cell data. Using both bulk tissue RNAseq and snRNAseq data from the same samples, we eliminate many confounding variables and can focus directly on the protocol-related variance. We hypothesize that cellular proportions will differ between single-cell and bulk RNAseq, and we will detect these differences by a differential expression of cell type marker genes from the bulk and single-cell data.

Slide 10: The Data

To test this hypothesis, we have access to data published in May 2019 in Science, which contains 41 post-mortem brain tissue samples from 32 patients. Two brain regions were sampled (anterior cingulate cortex and prefrontal cortex). Each sample has paired bulk and snRNAseq from the same tissue. This is one of very few studies that has this kind of data.

Slide 11: The Data

For the snRNAseq data, single cells were isolated and prepared by the droplet-based DroNucSeq protocol, which isolated around 3000 viable nuclei/sample. These were then sequenced on Illumina NovaSeq6000 to an average depth of 70k reads per nucleus, and expression was quantified by UMIs using 10x Genomics CellRanger. We received data after alignment to the genome as a UMI normalized count matrix, after filtration steps taken by the authors (to remove low quality reads).

Bulk RNAseq was also done on an Illumina NovaSeq6000 to an average depth of 100x. We received this data as a filtered raw count matrix.

We also have metadata for both datasets, including post-mortem inveral (PMI), sex, age, brain region and more factors as a comma-delimited file.

Slide 12: Statistical and Computational Methodology

To do this analysis, first we “bulkized” the snRNAseq data by summing UMI counts of each gene for all cells coming from the same sample. This is done to convert the cell-level data to sample-level, allowing it to be compared to bulk RNAseq data. To ensure the bulkized data was comparable to the bulk data, we assessed correlation between the two, expecting high levels of concordance.

Then, we performed differential expression analysis between bulk and bulkized data for each sample, by using limma with “sample of origin” as covariate in the linear model. PCA indicated that the brain region accounted for a significant proportion of variation, so the brain region was used as a second covariate. We suspected that differential expression analysis alone wasn’t perfectly suited to this kind of analysis since we may want to treat the samples as independent variables, so we also looked for interesting genes by inspecting those that diverged significantly from linear regression on the data.

With “interesting” genes identified, we performed enrichment analysis on differentially expressed genes, mainly by looking for enrichment of cell types, as indicated by previously published cell type markers. This is done to find any patterns within the differentially expressed genes that could indicate proportional differences in cell populations.

Slide 14: QC: Expression Data Correlates as Expected

We tried multiple ways to bulkize the single cells, and found that the best way was to sum UMI counts, rather than original raw counts, and without using voom normalization, which we thought would help as it better models the variance for genes with lower counts. This method gave maximal mean r^2 of 0.8 (range of 0.76 to 0.84). This indicated that the data was highly significantly correlated at the gene level and appropriate for further DE analysis.

Slide 15: QC: Heatmaps

We then looked at sample-sample correlations on both datasets, finding that correlations and clusters did change between the two but not dramatically, and both datasets did not cluster obviously with any covariate. However, in the bulkized dataset there is slight clustering based on batch and brain region, as well as sex.

Slide 16: QC: PCA

Similarly, we did PCA on both datasets. In bulk, PC1 and PC2 contributed to 19% and 11% of the variance, but were not found to separate any known covariates. However in the bulkized dataset, PC1 and PC2 contributed to 45% and 11% of the variance, and PC2 was found to modestly separate by brain region, where anterior cingulate cortex (ACC) samples cluster to some extent. For this reason, brain region was selected as a covariate in our linear model.

Additionally, looking at the distribution of expression of all genes shows that bulk data are more skewed than bulkized.

Slide 17: Gene Information and Cell Type Markers

As a crucial part of this project is in mapping genes to cell types, marker genes (from mouse) for 18 prefrontal cortex cell types were obtained from NeuroExpresso and converted to human homologs as HGNC gene symbols using homologene. These were then used to obtain length, GC %, and GO terms using BiomaRt.

Slide 18: Differential Expression Analysis

Interestingly, after conducting DE analysis with limma we found that the large majority of the genes detected were also differentially expressed. Since this is potentially driven by an inherent bias due to the difference in the kind of data, we took the genes with the fold changes that were at least 25% of the maximal with the reasoning that this is where interesting genes would be found. We also confirmed that differential expression was not driven in one direction (as would happen if all the data of one dataset was consistently higher than the other) by normalizing their means and confirming that the proportion of up- to down-regulated genes was roughly equal.

Slide 19: Linear Regression Outliers

Parallel to this investigation, we looked for interesting genes using linear regression per sample. This was mainly a branching exploratory analysis that should yield similar findings to the differential expression analysis. To find interesting genes, we used thresholds which are the first and third quartiles of all residuals taken together.

Here, a positive residual indicates that the gene was expressed at a higher level in bulk data as compared to bulkized (and vice versa for a negative residual). If it is more than one quartile from the mean, it is considered an outlier. To make sure we capture consistently interesting genes, we take only those that are outliers in at least two thirds of all samples. The distribution of residuals for all genes was uninteresting, with an equal proportion higher in bulk as lower in bulk.

Slide 20: Linear Regression Outliers

However, if we only plot **marker genes** for the present cell types, their distribution diverges significantly from the background. Since these distributions are clearly not normal, a one-sided Mann-Whitney test was used to assess the deviation of the median (formally testing that the location shift is positive), and significant p-values (on the order of 10^{-20}) were obtained for all cell types except two pyramidal subtypes, demonstrating a clear decrease in expression of most cell types in the bulkized single cell data as compared to the bulk data.

Slide 21: Overrepresentation of Cell Type Markers (DE)

These findings are replicated in the differential expression analysis. This table summarizes with a color scale, for each cell type, what fraction of its marker genes are up or downregulated in bulkized compared to bulk data (and raw counts in each cell). The cell types are sorted with the ones with the highest fraction of downregulated markers on the left. The darker the blue color, the higher this fraction is. Since our set of marker genes specify more specific cell subtypes than our original data does, some “cell counts” are excluded.

Slide 22: Overrepresentation of Cell Type Markers (Linear Regression)

Similar to the previous slide, this table shows that the marker genes are expressed at lower levels in bulkized data. There is some overlap with the previous table looking at the top cell types whose marker genes are downregulated and the number of more lowly expressed cell types in bulkized data strongly correlates in these two analyses. Together, this shows an overall signature of decreased gene expression in bulkized single cell data, as compared to bulk data, with most pyramidal subtypes at the most extreme end of this bias.

Slide 24: General Decrease of Expression in All Cell Types

Interestingly, most cell types and subtypes we have marker genes for demonstrate a strong pattern of decreased expression in bulkized single cell data. This finding is unexpected. As this data is proportional and all cells in the single nucleus RNAseq are presumed to be broadly brain-related, it isn’t possible for all cell types to be downregulated. One possibility is that the single nucleus RNAseq data has some contamination with non-neuronal cell types, which would help explain the lack of stark down-regulation of non-neuronal markers.

Slide 25: Pyramidal Cells Have Most Significant Bias

Our analysis revealed pyramidal cells of many subtypes to have the highest biases in this dataset. This may be explained by the chemical and physical properties of this cell type, as reported in literature. One possible explanation may lie in the abundance of pyramidal cell-specific membrane channels, altering their membrane properties and affecting their response to lysis buffer used in snRNAseq lysis protocols. Another possibility is that the intricate branching structure of pyramidal cell dendrites may form a globular structure in suspension, thus shielding them from lysis (and thus, preventing the release of cellular mRNA). In all, this indicates some cell type bias in snRNAseq, in favor of our hypothesis, even with the uncertainties mentioned in the previous slide.

Slide 26: Weaknesses

While initial results seem to indicate some bias, this analysis is not without its flaws. One significant limitation is that we cannot confidently state that this bias is due to cellular composition differences resulting from differences in library preparation. While there seems to be a significant correlation, it doesn't indicate causation. Our confidence is also limited by the fact that we only have one dataset of this type to analyze, as our findings may not be generalizable. Finally, and importantly, the data may just be somewhat unsuitable for this task, indicated by the large proportion of differential expression hits (and similarly, outliers). Thresholds were applied somewhat *ad hoc* (on logFC and quantile cutoffs) to reduce the size of our hit list, and results may be sensitive to these cutoffs. Another subtle weakness is that the markers we use to report cell types originate from studies in mice, and they may not be truly unique cell type markers in human. Further studies may aim to characterize this bias directly by directly counting cells after staining with some cell type markers.