Comparing Models of Subject-Clustered Single-Cell Data

Version 6.0-Description of Motivating Example

Lee Panter

Description of Motivating Example

Throughout the course of this paper, references are made to the 2018 manuscript entitled "The immune cell landscape in kidneys with lupus nephritis patients" [1]. In this manuscript Arazi, Rao, Berthier, et al. compared single-cell kidney tissue sample data from 45 Lupus Nephritis subjects vs. 25 population controls [1]. The kidney tissue samples were collected from ten clinical sites across the United States, were cryogenically frozen, and shipped to a central processing facility. At the central processing facility, the tissue samples were then thawed, and sorted into single-cell suspension across 384-well plates using FlowJo 10.0.7, 11-color flow cytomery [2]. Single-cell RNA sequencing was performed using modified CEL-Seq2 method [3] with $\sim$ 1 million paired-end reads per cell. The original experimental data may be accessed by visiting the Immport repository with accession code SDY997. Immport-SDY997: https://www.immport.org/shared/study/SDY997

Data Quality Control

The Seurat Guided Clustering Tutorial [4] was used to examine and perform quality control (QC) of the initial data.

This process quantifies the quality of each observation in two numerical measures (based upon two calculated variables, $nFeature$ and $PerctMT$, described below). Threshold values of these variables can then be chosen and used to filter calls not meeting the chosen criteria. The Seurat tutorial provides methods of automated calculation and filtering implemented by Arazi, Rao, Berthier, et al. in [1]. Identical variable calculations, with alternative threshold settings were independently implemented for this study.

The quality control variables are qualitatively defined as:

1. $nFeature$ is the number of unique genes detected to have a non-zero expression in each cell. This is used to identify cells with an abnormally low or high number of expressed genes. Low numbers may result from empty wells (zero content measurements) or broken-cells, while high numbers may result from observations of more than one cell.
2. $PerctMT$ is the percentage of reads that map to the mitochondrial genome. This is used to identify dead and/or broken cells since dead or dying cells will retain RNAs in mitochondria, but lose cytoplasmic RNA [5].

The pre-QC distribution of $PerctMT$ for each subject is displayed in (Figure 1) below:

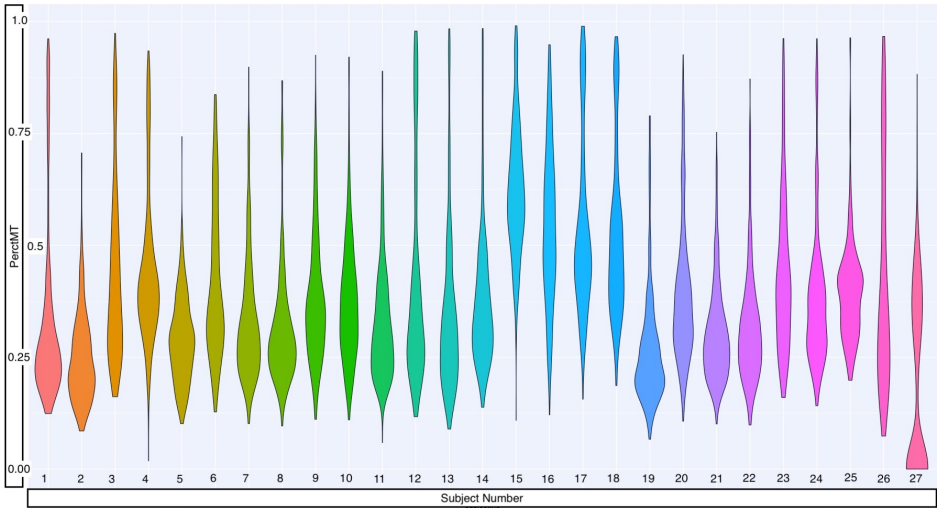

Figure 1: Pre-QC $PerctMT$ Distribution for each subject

The QC measures employed by Arazi, Rao, Berthier, et al. in [1] were:

1. $1,000 < nFeature < 5,000$
2. $PerctMT \leq 25\%$

All observations for which the calculated values of $nFeature$ and $PerctMt$ satisfied the inequalites in (1) and (2) above were kept, and the others were considered "low-quality" and removed. The resulting distribution of the $PerctMT$ variable is displayed in (Figure 2):

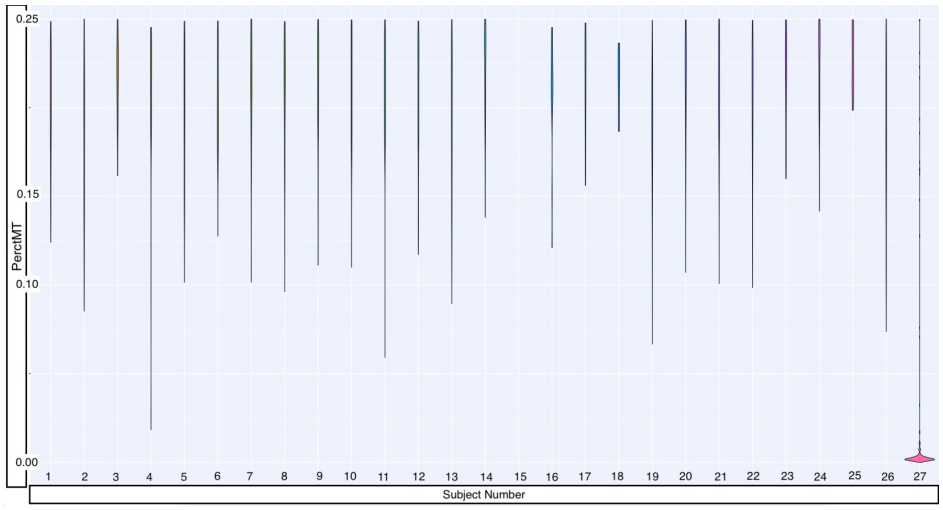

Figure 2: Post QC distribution of $PerctMT$ with thresholds implemented by Arazi, Rao, Berthier, et al

As 84% of cells were removed with the filters chosen by Arazi et al, we chose a more lenient threshold, removing observations with $PerctMT \leq 60\%$ to keep more cells. The additional subsetting measure of restricting the data to only B-cells was made in an effort to regularize (homogenize feature expression) the data sample. The resulting distribution of $PerctMT$ is displayed in (Figure 3) after filtering.

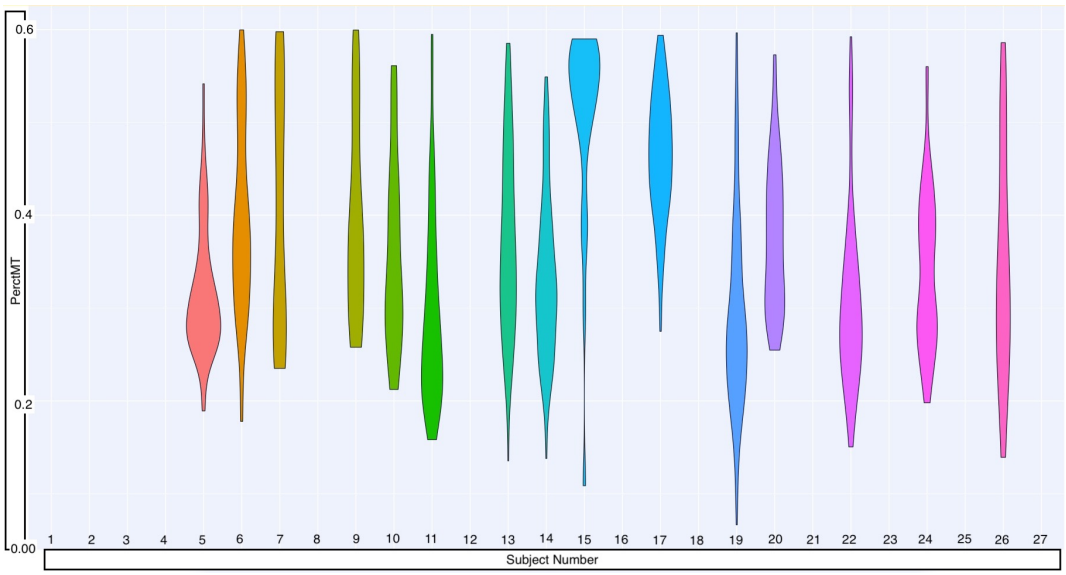

Figure 3: Post QC distribution of $PerctMT$ with thresholds implemented in this paper

The distribution of observations for each subject before and after the quality control thresholds are imposed is also show numerically in Table 1:

Subject Number	1	2	3	4	5	6	7	8	9
Number of Observations Before QC	375	375	364	381	340	383	383	356	372
Number of Observations After QC	0	0	0	0	58	86	32	0	31

Subject Number	10	11	12	13	14	15	16	17	18	19
Number of Observations Before QC	327	311	379	375	345	371	381	381	377	380
Number of Observations After QC	21	107	0	107	100	25	0	122	0	127

Subject Number	20	21	22	23	24	25	26	27
Number of Observations Before QC	381	380	333	333	239	218	378	342
Number of Observations After QC	75	0	87	0	79	0	53	0

Table 1: Observation counts per-subject before and after Quality Control threshold filter restrictions

The process of eliminating observations through quality control threshold measures is compa-

rable to outlier detection and removal. Values defining the quality of an observations are determined by the context of the data being studied, as well as the distribution of values within the data. An observation should only be considered abnormal, poor-quality, uninformative, or unrealistic if it can be characterized as such in the context of its observational setting and compared to the data observed.

The pre-defined thresholds implemented by Arazi, Rao, Berthier, et al outline the expected observational circumstances surrounding the Lupus Nephritis data. However, these limits set unrealistic boundaries in the context of the data provided, and therefore were not reasonable for classifying poor-quality observations.

With this in mind, we also note that quality-control is dissimilar to outlier-detection and removal because the thresholds used define the sample of interest. In this way, an experimenter would conduct quality-control as a sub-sampling method, and would perform outlier detection and removal on the sub-sample.

This subtle, but important difference allows for the *Population of Interest* to be represented by the sample *after QC fiter have been implemented*. This allows us to reduce the data set distribution to subjects with positive observational counts, as they are part of the *Sample of Interest*. This distribution is displayed in Table 2:

Subject Group Number	5	6	7	9	10	11	13	14
Number of Observations	58	86	32	31	21	107	107	100

Subject Group Number	15	17	19	20	22	24	26
Number of Observations	25	122	127	75	87	79	53

Table 2: Observation count per-subject, subjects with positive counts

Table 3 displays the descriptive statistics for the number of observations per-subject.

MIN	1st Q	Median	Mean	3rd Q	MAX
21	42.5	79	74.0	103.5	127

Table 3: observation count per-subject descriptive statistics

Variable Selection and Summaries

We chose two pairs of variables from the 38,354 genetic markers in the Lupus Data to compare across the three methods. The variables we chose have higher values of correlation than arbitrary variable pairings as indicated by a high Pearson Correlation Coefficient (top 10% of all possible pairings), and have previously been associated with human diseases or conditions (e.g. cancer treatment research in the case of MALAT1 [6], or observed limb malformations in the case of FBLN1 [7]). An attempt was also made to assign predictor-pairings of interest. The CD19 marker (paired with MALAT1) is a transmembrane protein, encoded by the CD19 gene. Since the FlowJo cytometry measurements contain CD19 protein readings, the relationship between the "CD19 quantification" used as a predictor predictor and the outcome of interest can be modeled using proteomic or transcriptomics data. CD34, the predictor which we link with FBLN1 is also a transmembrane protein encoded by a gene, and similarly interesting.

Without undergoing the process of expression normalization, single-cell RNA sequencing data is represented as non-negative integer count data. Higher counts correspond to higher detection frequencies and (without compensating for expected expression frequency) these detection frequencies can be interpreted as a quantification of the magnitude of expression for a transcriptomic marker.

The variables that we study here are summarized in Appendix Table (A1) - (A4). Each describes selected variable summary statistics (minimum, maximum, average, and median) for the subset samples specific to the subject identifiers used in Table (2).

Measurements of scRNA-seq data can be highly specific to very precise transcriptomic targets (expression profiles can be limited to very small transcriptome scope), so while the agglomerated scope of gene expression across a sample is the same as a traditional bulk experiment, individual observations have a biologically inflated zero-component. There are also *technical* zero-inflation components that are associated with protocol variations, and measurement error.

This is evident in the case of the FBLN1 ~ CD34 pairing, where we see that expression values for several subjects exhibit:

$$\min_j(FBLN1_{ij}) = \min_j(CD34_{ij}) = 0 = \max_j(CD34_{ij}) = \max_j(FBLN1_{ij})$$

where

$$i \in \{5, 6, 7, \ldots, 26\}$$

$$j \in \{1, \ldots, n_i\}$$

Which implies that:

$$(FBLN1_{ij}) = (CD34_{ij}) = 0 = (CD34_{ij}) = (FBLN1_{ij}) \quad \forall i, j$$

We expect the additional presence of zeros to be attributable to both biological and technical sources. Together, these factors contribute to heavily right-skewed variable distributions (Figure 4)

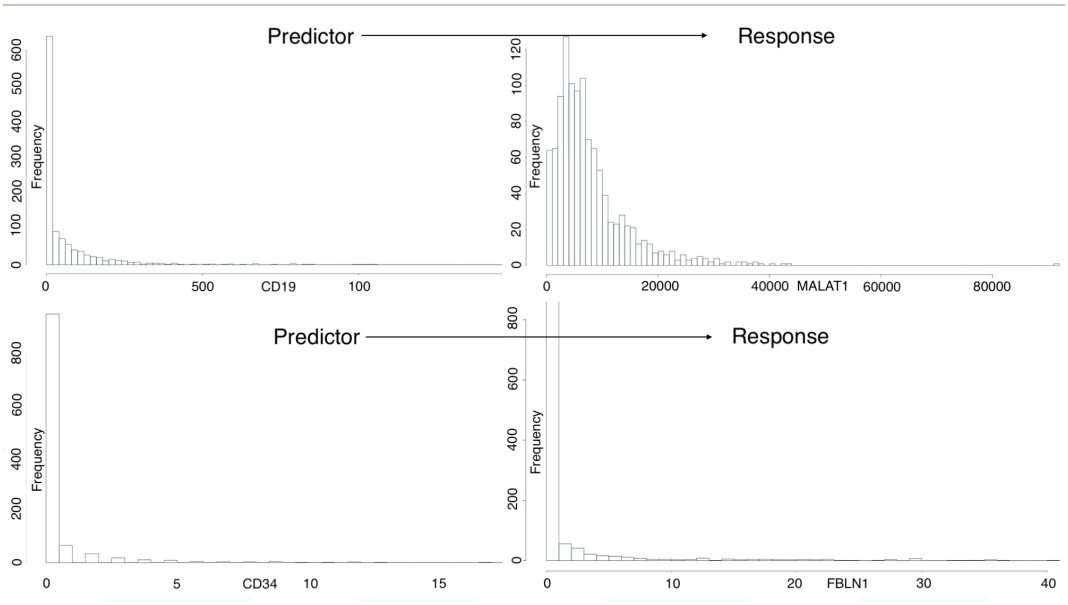

Figure 4: Predictor-Response pairing variable distributions

The MALAT1 variable had a large minimum outcome compared to the other variables. All measurements of this variable are positive in their raw state, so we translate the raw observations negatively by the minimum (67) value. This gives a minimum expression value of zero, which coincides with our intuition as well as the other variables under investigation. It should be noted that this process would be incorperated into the model-fitting procedure automatically through the intercept term.

The modeling methodologies we employ motivates a log-transformation in an attempt to achieve approximate normality, especially for the outcome variable's distribution. We perform the "log plus +1" transformation on all variables:

$$X \mapsto \log(X + 1)$$

The resulting distributions are shown in Figure (5):

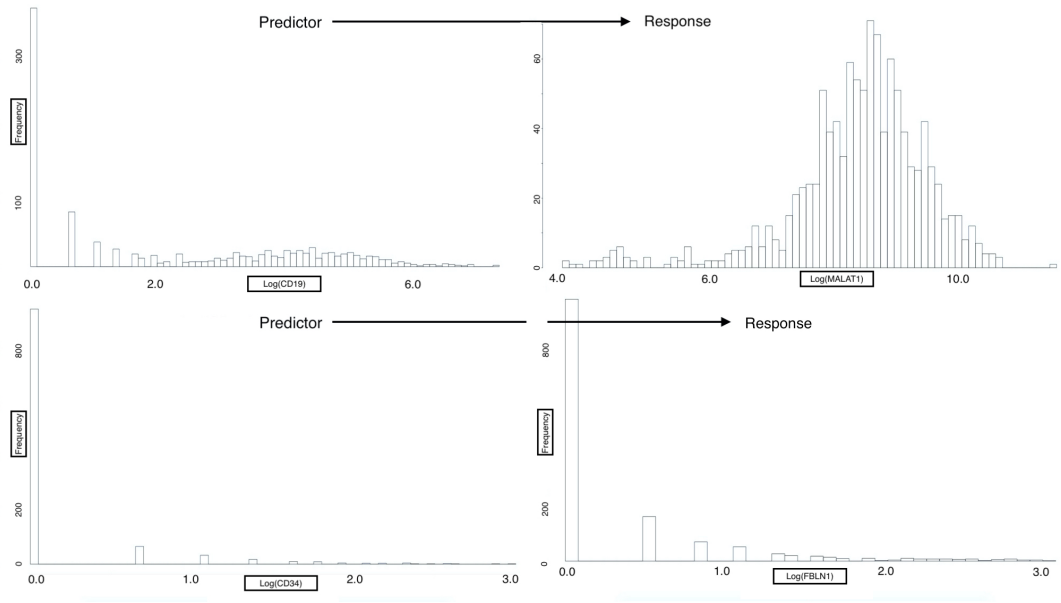

Figure 5: Predictor-Response variable pairings, post-transformation distributions

We see that the log-transformed response MALAT1 is approximately normal distribution. Conversely, the log-transformed response FBLN1 is not inherently better than the untransformed response. We can clearly see the heavy influence of zero-inflation in these variables as is apparent from the dominance of the "zero-bins" in Figure (5).

Regardless, we model each outcome under the assumption that: compensating for observational correlation will sufficiently account for non-normality of the responses. This may not generally the case, and additional transformations or modeling methodologies may be needed to improve model error distributions. However, for the purpose of comparing the previously mentioned models on subject-correlated single-cell data, we will proceed with this assumption and verify ridual homoscedasticity, normality and independence using fitted vs residual plots and quantile-quantile plots.

1. Arazi A, Rao DA, Berthier CC, et al. (2018) The immune cell landscape in kidneys of lupus nephritis patients. *bioRxiv* 363051.

2. FlowJo X V10. 0.7 r2 flowjo. *LLC https://www flowjo com.*

3. Hashimshony T, Senderovich N, Avital G, et al. (2016) CEL-seq2: Sensitive highly-

multiplexed single-cell rna-seq. *Genome biology* 17: 77.

4. Satija R, others (2018) Seurat: Guided clustering tutorial. *Satija Lab http://satijalab org/seurat/pbmc3k_tutorial html*.

5. Bacher R, Kendziorski C (2016) Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* 17: 63.

6. Gutschner T, Hämmerle M, Diederichs S (2013) MALAT1—a paradigm for long noncoding rna function in cancer. *Journal of molecular medicine* 91: 791–801.

7. Debeer P, Schoenmakers E, Twal W, et al. (2002) The fibulin-1 gene (fbln1) is disrupted in at (12; 22) associated with a complex type of synpolydactyly. *Journal of medical genetics* 39: 98–104.