Description of Motivating Example

Throughout this paper references are made to the 2018 article entitled "The immune cell landscape in kidneys with lupus nephritis patients", in which Arazi, Rao, Berthier, et al. compare 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, cryogenically frozen, then 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 cytometry [2]. Single-cell RNA sequencing was performed using a 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

I use the Seurat Guided Clustering Tutorial [4] to perform quality control (QC) of the initial data. This process quantifies the quality of each single-cell observation in two numerical measures (based upon two calculated variables, **nFeature** and **PerctMT**). Threshold values of these variables are chosen and used to filter cells (observations) 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 are independently implemented for this study.

The quality control variables are conceptually 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 (partial) 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 as 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 X!X**) below:

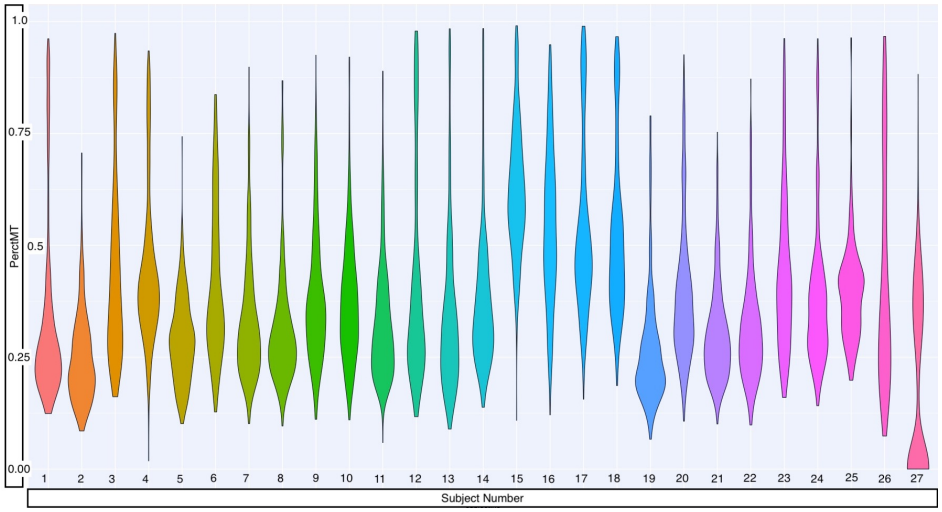

Figure X!X: Pre-QC **PerctMT** Distribution for each subject

The QC measure thresholds employed by Arazi, Rao, Berthier, et al. in [1] are:

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

All observations for which the calculated values of **nFeature** and **PerctMT** satisfied the inequalities 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 X!X**):

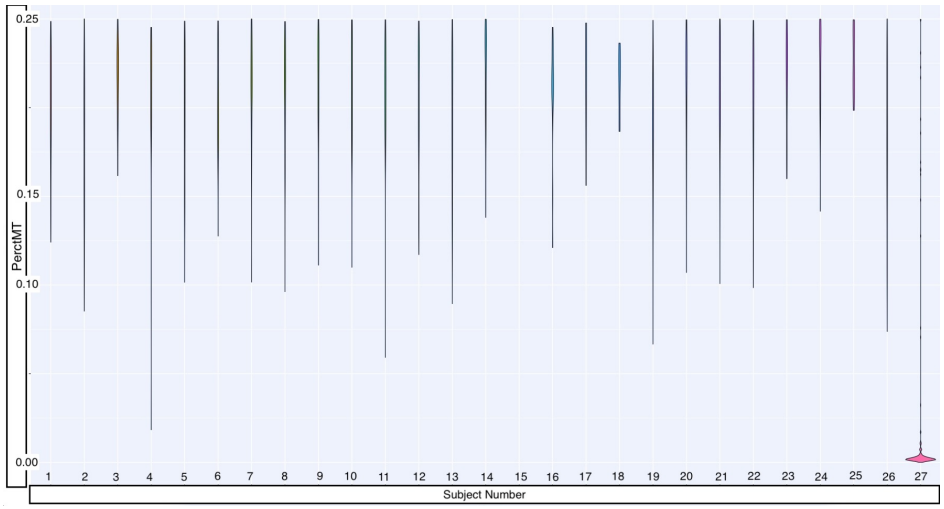

Figure X!X: Post QC distribution of **PerctMT** with thresholds implemented by Arazi, Rao, Berthier, et al

As 84% of cells as removed with the filters chosen by Arazi et al, I choose a more lenient threshold, removing observations with **PerctMT** $\leq 60\%$, in an effort to keep more cells.

An additional restriction of the data to only B-cells is made in an effort to regularize the data sample (i.e. homogenize feature expression). The resulting distribution of **PerctMT** is displayed in (**Figure X!X**) after filtering.

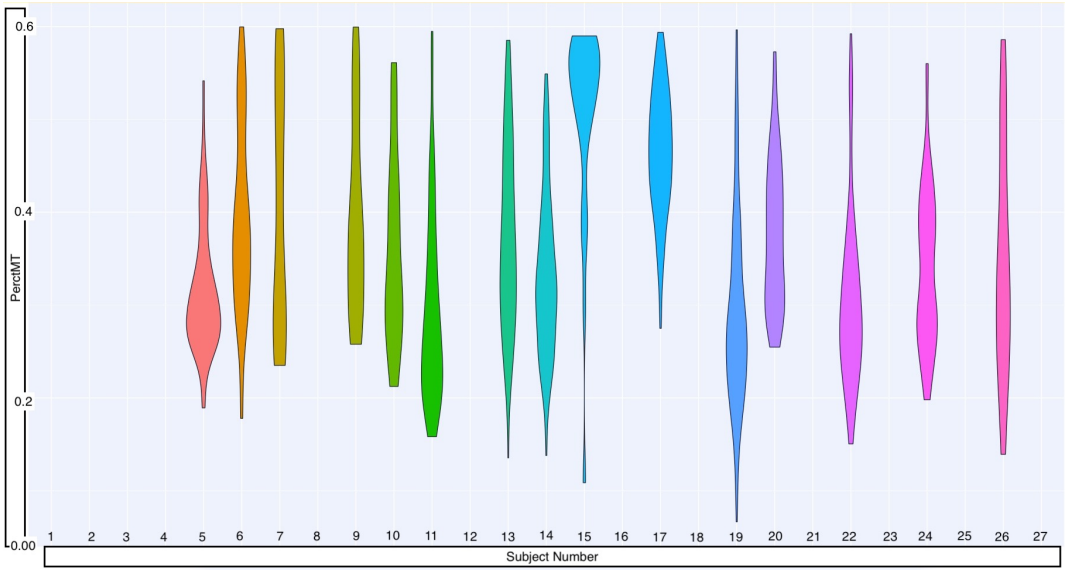

Figure X!X: Post QC distribution of **PerctMT** with thresholds implemented in this paper

The distribution of observations for each subject Pre/Post QC (with updated

PerctMT $\leq 60\%$ threshold value) is shown numerically in **Table X!X**:

Subject Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
NO of Obs Before QC	375	375	364	381	340	383	383	356	372
NO of Obs After QC	0	0	0	0	58	86	32	0	31
Subject Number	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
NO of Obs Before QC	327	311	379	375	345	371	381	381	377
NO of Obs After QC	21	107	0	107	100	25	0	122	0
Subject Number	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
NO of Obs Before QC	380	381	380	333	333	239	218	378	342
NO of Obs After QC	127	75	0	87	0	79	0	53	0

Table X!X: Observation counts per-subject Pre/Post QC with updated **PerctMT** $\leq 60\%$ threshold value.

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

Table X!X: observation count per-subject summary statistics Post QC with updated **PerctMT** $\leq 60\%$ threshold value

Variable Selection and Summaries

I select two pairs of variables from the 38,354 genetic markers in the Lupus Data to compare across the five modeling methods. The variables I choose have higher values of correlation than arbitrary variable pairings as indicated by a high Pearson Correlation Coefficient (both selected pairings are within the top 10% of highest Pearson Correlation Coefficients 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]-used as the first outcome, or observed limb malformations in the case of FBLN1 [7]-used as the second outcome). I also

attempt to assign predictor-pairings of interest. The CD19 marker (the predictor paired with MALAT1) is a transmembrane protein encoded by the CD19 gene. The FlowJo cytometry measurements contain CD19 protein readings, so the relationship between CD19 as a predictor and the outcome of interest (MALAT1) 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 values. Higher counts correspond to higher detection frequencies and these detection frequencies can be interpreted as a quantification of the magnitude of expression for a transcriptomic marker (e.g CD19, CD34, MALAT1, FBLN1).

I provide numerical summaries of the four selected variables in Appendix Tables (A1) - (A4). Each describes selected variable summary statistics (minimum, maximum, average, and median) for the positive observational count subjects in **Table X!X**.

Measurements of scRNA-seq data are specific to precise transcriptomic targets. This means that single-cell expression profiles (a single observation) can be limited to a small transcriptomic scope. So while the agglomerated scope of gene expression across a sample is the same as (or broader than) 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. Together, these factors contribute to right-skewed variable distributions (**Figure X!X**)

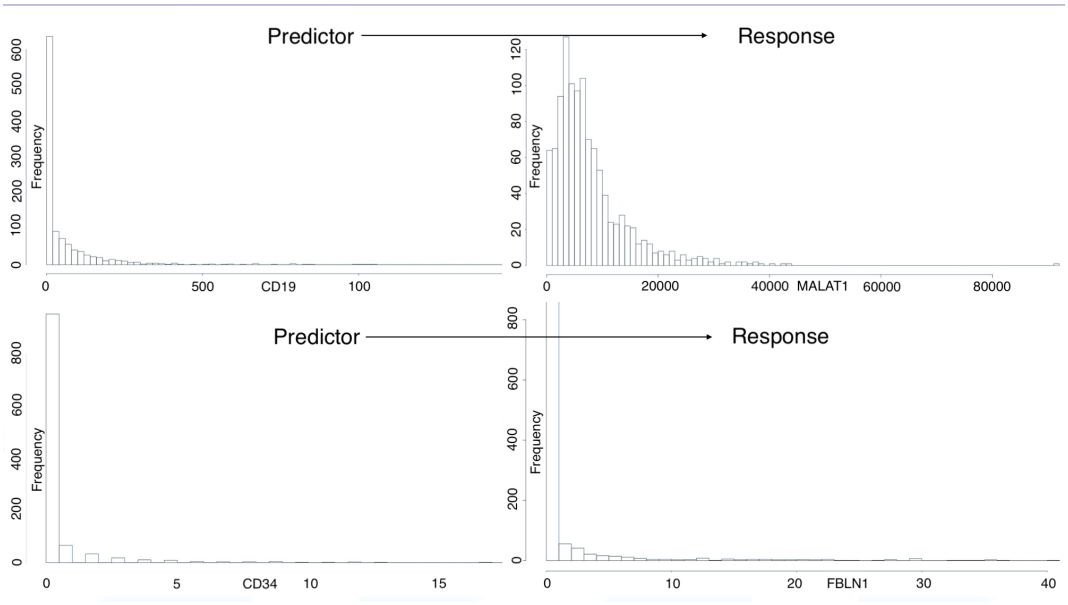

Figure X!X: Predictor-Response pairing variable distributions

The MALAT1 variable has a large minimum outcome compared to the other variables, so I translate all the values of this variable *negatively* by the minimum value.

$$min(\text{MALAT1}) = 67$$

This gives a minimum expression value of zero, which coincides with intuition as well as the minimum value of the other variables under investigation.

The modeling methodologies I employ motivate a log-transformation in an attempt to achieve approximate normality, especially for the outcome variable's distribution. I perform the "log plus +1" transformation on all variables (predictor and outcome):

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

The resulting distributions are shown in (**Figure X!X**)

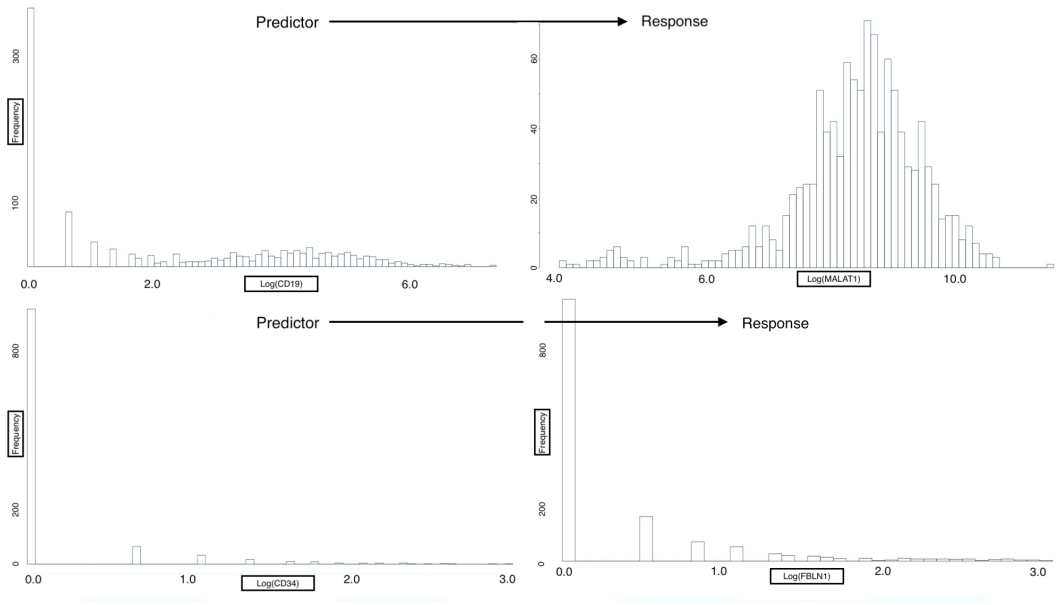

Figure X!X: Predictor-Response variable pairings, post-transformation distributions

The log-transformed response of MALAT1 is approximately normally distributed; however, the log-transformed response FBLN1 is not inherently better than the un-transformed response.

Regardless, each outcome is modeled under the assumption that: compensating for observational correlation will sufficiently account for non-normality of the responses. It may be the case that additional transformations and/or alternative modeling techniques may be needed to improve model error distributions. However, for the purpose of comparing the previously mentioned models on subject-correlated single-cell data, I proceed with this assumption and I verify residual homoscedasticity, normality and independence using fitted vs residual plots and quantile-quantile plots.

References

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.