Supplemental materials of
"*Two-phase differential expression analysis for single cell RNA-seq*"

1 Lognormal-Poisson (LNP) versus Negative Binomial models

The negative binomial distribution (gamma-Poisson compound distribution) has been the most common choice in modeling the rate or concentration of expression for RNA-seq data. The reason is two fold: with Poisson distribution capturing the counting error given the expression level, a gamma-Poisson mixture model conveniently becomes the negative binomial model, with a closed form probability density and many existing methods associated with it. Second, Gamma distribution is a fairly flexible family of distributions.

This has worked reasonably well with bulk RNA-seq data. Since bulk RNA-seq measures counts from mean expression averaged over thousands to millions of cells, the dispersion parameters are very small for most genes. In such situations, as we pointed out in Wu *et al.* (2012), the gamma distribution and log-normal distribution are mostly exchangeable: both are right skewed, with variance proportional to the mean. In fact, the dispersion parameter ϕ in gamma corresponds to the σ^2 in the lognormal model, and both are the squared coefficient of variation (CV). When biological variation is low (small ϕ or σ^2), the expression in the log scale is either symmetrical under the LN model or close to symmetrical under the gamma model.

Table 1: Comparison of gamma and lognormal distributions

Probability Model	Gamma(α,β)	Lognormal(μ, σ^2)
	shape: α scale: β dispersion: $\phi = 1/\alpha$	location (for $\log X$) μ scale (for $\log X$) :σ
mean (E)	$\alpha\beta$	$\exp(\mu + \sigma^2/2)$
variance(V)	$\alpha\beta^2$	$\exp(2\mu + \sigma^2)(e^{\sigma^2} - 1)$
mean-variance relationship	$V = E^2/\alpha = E^2\phi$	$V = E^2(e^{\sigma^2} - 1) \approx E^2\sigma^2$ when $\sigma^2 \approx 0$
coefficient of variation $CV = SD/E = \sqrt{V/E^2}$	$\sqrt{\phi}$	$\sqrt{e^{\sigma^2} - 1} \approx \sigma$ when $\sigma^2 \approx 0$

When we deal with single cell RNA-seq data, the heterogeneity among cells is much greater than that from bulk samples. For example, we routinely observe counts with coefficient of variation greater than 1.

Figure S1 shows the CV as a function of the mean counts in two scRNA-seq datasets. The sample CV is lower for genes with higher mean counts. But even for genes that are almost always observed, the CV is often greater than 1, indicating large dispersion. To accommodate such high dispersion, a gamma model would be forced to take a shape that peaks at 0, no matter how high the mean expression is. This lack of flexibility in shape, when we need to handle large heterogeneity, makes the gamma model no longer suitable for scRNA-seq data. Lognormal distribution, on the other hand, can handle both small and large variations.

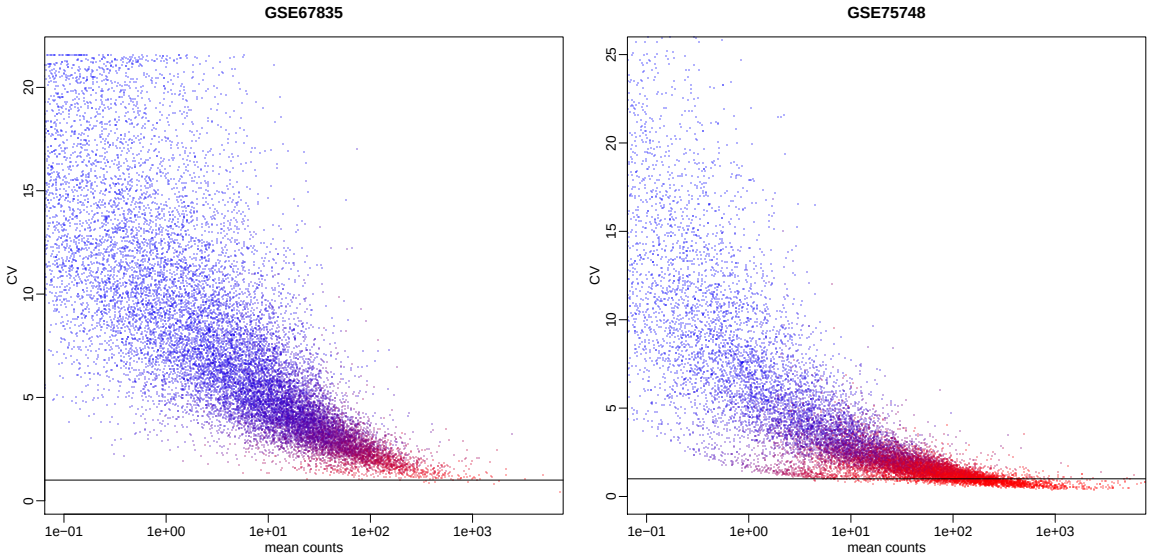

Figure S1: Coefficient of variation (CV) as a function of mean counts in single cell RNA-seq data. Color reflects the proportion of none zero counts for each gene across cells, with red for 1 and blue for 0.

2 Approximation of the lognormal-Poisson (LNP) distribution

A drawback of the LNP distribution is that it does not have closed-form distributional functions. We approximate the LNP density function as

$$\text{LNP}(y|\mu,\sigma) \approx \Phi\left(\frac{\log(y+0.5)-\mu}{\sigma}\right) - \Phi\left(\frac{\log[\max\{0,y-0.5\}]-\mu}{\sigma}\right) \tag{1}$$

where $\Phi(\cdot|\mu,\sigma)$ is the cumulative distribution function (CDF) of Gaussian distribution with mean μ and standard deviation σ. The idea behind this approximation is that the Poisson counting of a log-normal random variable still has a distribution very similar to log-normal, especially when the counts are relatively large. The approximation above discretizes the continuous log-normal distribution into LNP distribution by assigning the cumulative probability of each read count's ± 0.5 neighborhood to its point mass probability. Subsequently, the CDF of LNP can be approximated as:

$$\text{LNP}(Y \leq y|\mu,\sigma) \approx \Phi\left(\frac{\log(y+.5)-\mu}{\sigma}\right). \tag{2}$$

This provides very fast and accurate approximation. We illustrate this approximation in a few examples in Figure S2, by comparing this approximation with empirical CDF from Monte Carlo simulation with a large sample ($n = 5000$). We see that even for small μ the approximation is very accurate, and for large μ the approximation is almost identical to the truth.

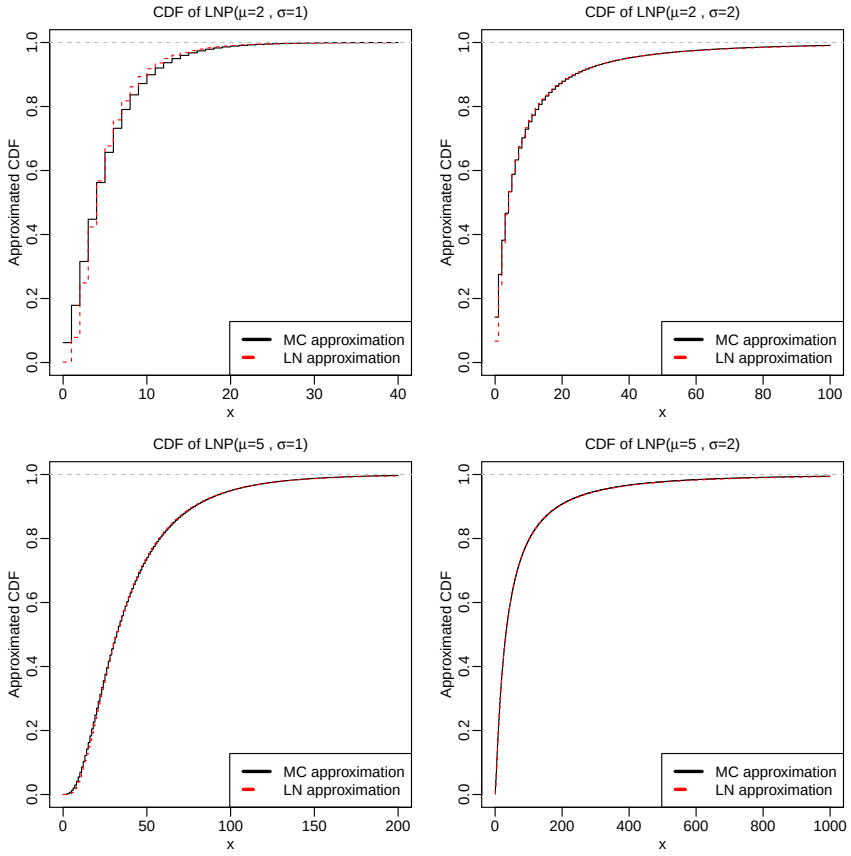

Figure S2: . Approximation of lognormal-Poisson distribution.

3 Additional DE comparison results from human brain dataset

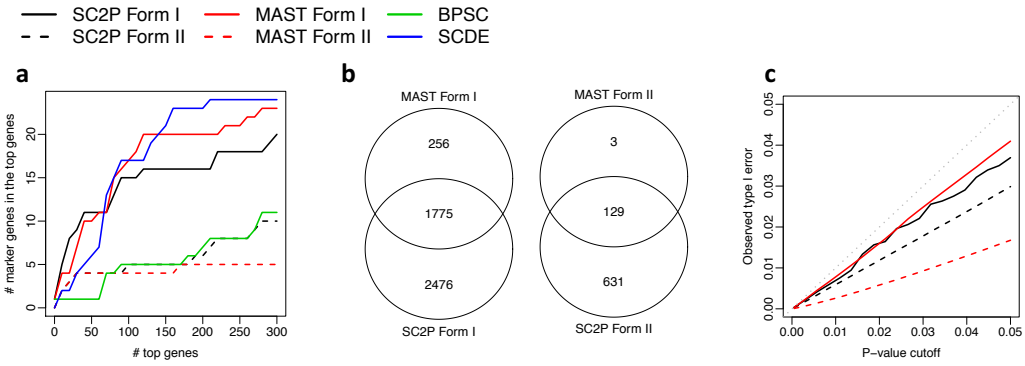

Figure S3: DE detection in human brain data, for neurons vs. oligodendrocytes comparison. The layout of the panels are the same as Figure 4 in the main text.

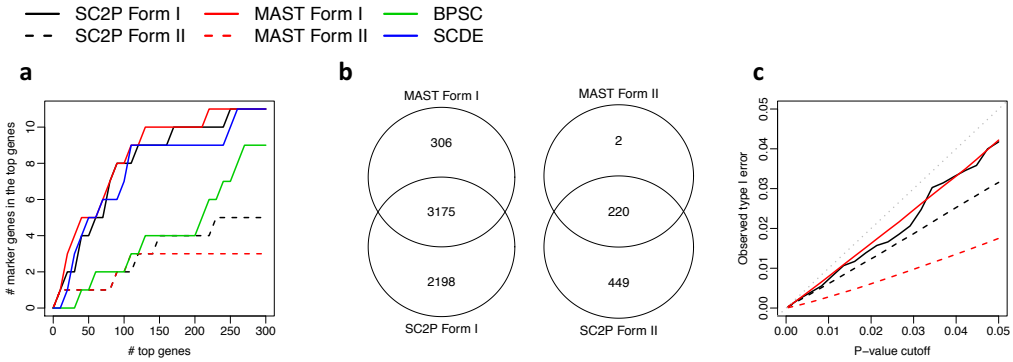

Figure S4: DE detection in human brain data, for astrocytes vs. neurons comparison. The layout of the panels are the same as Figure 4 in the main text.

4 DE comparison in T2D dataset

We compared cell populations from patients versus normal controls in the most abundant cell types in the pancreatic islet. Cell types are determined as described in Lawlor *et al.* (2017).

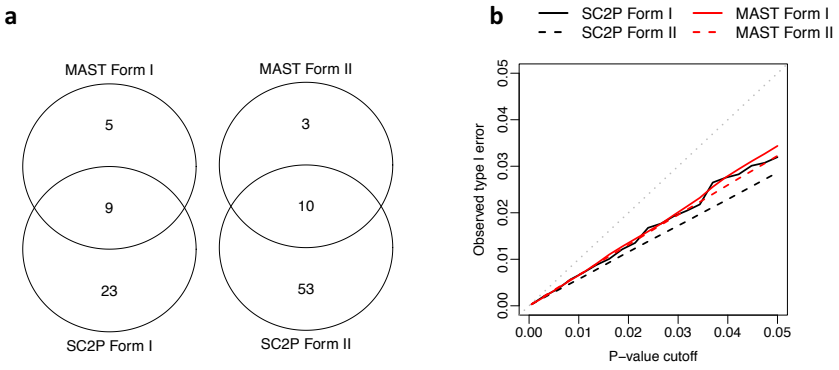

Figure S5: DE detection in T2D dataset comparing alpha cells from T2D patients versus normal controls. Genes with FDR<0.05 from the statistical tests are declared DE. (a) Overlaps of DE genes in both forms from MAST and SC2P; (b) Assessment of type I error control based from permutation.

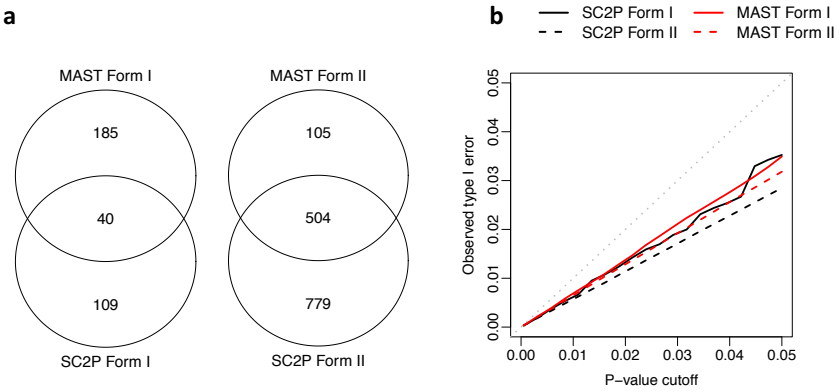

Figure S6: DE detection in T2D dataset comparing beta cells from T2D patients versus normal controls. Genes with FDR<0.05 from the statistical tests are declared DE. (a) Overlaps of DE genes in both forms from MAST and SC2P; (b) Assessment of type I error control based from permutation.

5 DE comparison in human brain dataset, comparing astrocytes and oligodendrocytes

Figures S7 and S8 are heatmaps of the log(1+counts) of genes declared as DE by MAST and SC2P.

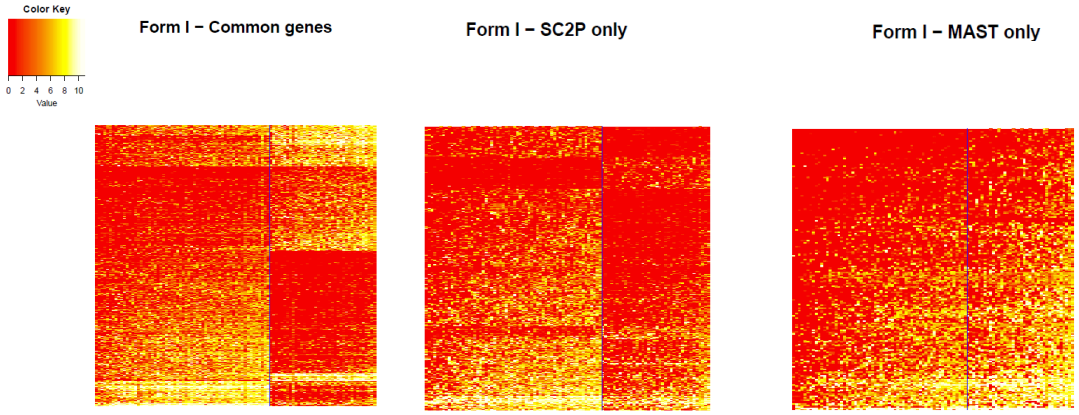

Figure S7: Form 1 DE (phase transition) reported by MAST and SC2P. There are 1695 genes identified by both, 493 genes by SC2P only and 216 by MAST only.

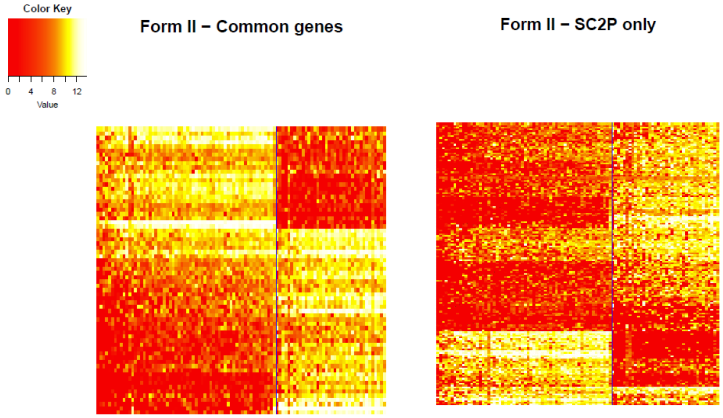

Figure S8: Form 2 DE (expression magnitude difference) genes comparing astrocytes and oligodendrocytes. There are 68 genes identified by both, 164 genes by SC2P only and none by MAST only.

6 Robustness

We compared the p-values obtained from the full dataset (100 cells) to the p-values from reduced dataset with obtained by randomly removing 5 cells in the neurons vs. oligodendrocytes comparison. In the main text we presented one set of results from MAST and SC2P. Here we include 10 runs of the same comparison, shown in Figure S9.

we perform additional analyses by removing 10%, 20%, and 50% cells. Each analysis is run 10 times. We compute the Pearson's correlation of p-values from MAST and SC2P, and present the distributions of the correlation coefficient as the boxplot in Figure S10. In all scenarios, SC2P has much higher correlations than MAST, indicating better robustness.

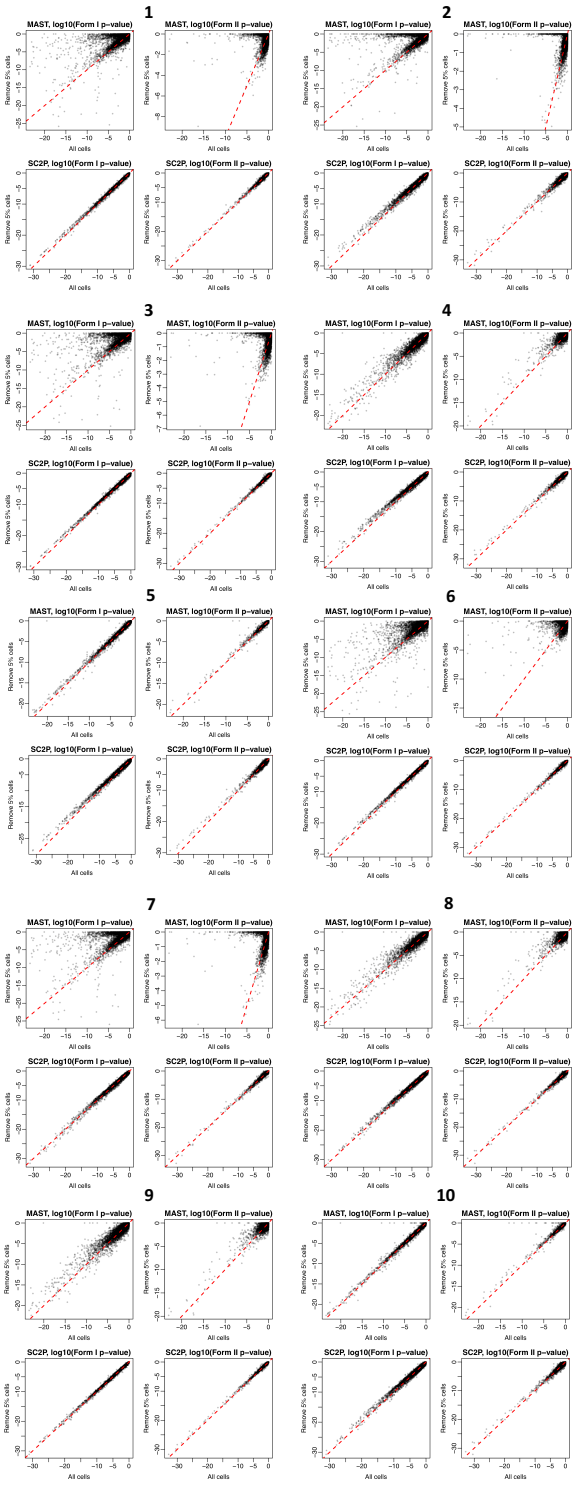

Figure S9: Robustness of DE detection. Figures show comparison of p-values from testing DE using all cells in the dataset or a subset with 5 cells randomly removed, in the human brain data (neurons vs. oligodendrocytes). Form 1 (phase transition) and Form 2 (magnitude difference in phase II) DE are compared separately.

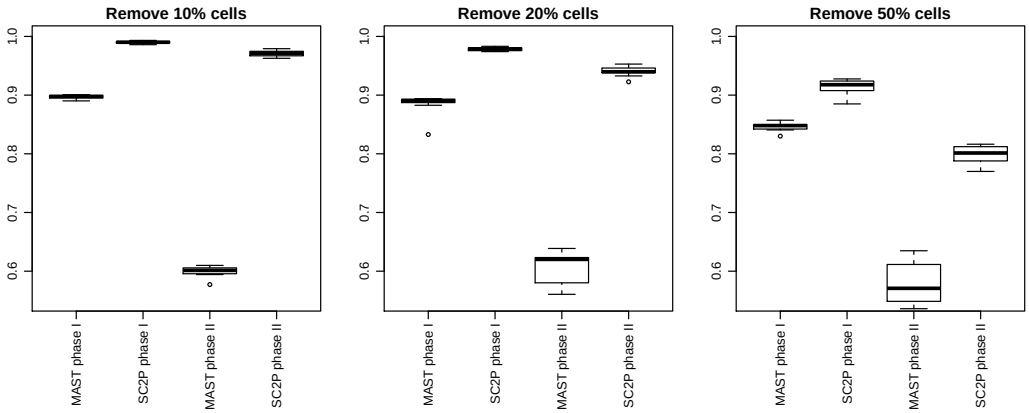

Figure S10: Correlation of p-values before and after removing different numbers of cells.

7 Proportion of genes in Phase II

The proportion of genes detected in Phase II varies substantially among cells, and appears to differ between cell types, as shown in Figure S11.

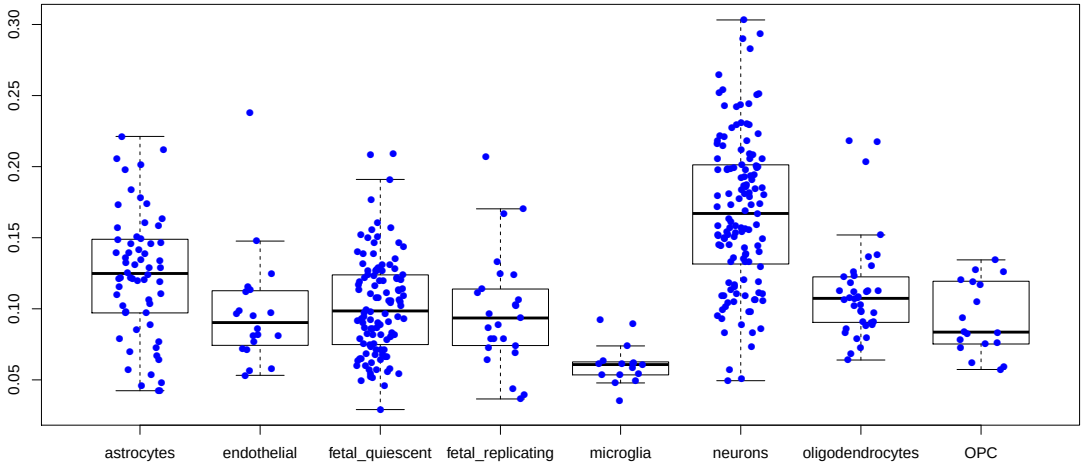

Figure S11: Estimated proportion of genes in Phase II in different cell types from the Human Brain dataset.

Many researchers use a simple statistic to estimate a global detection rate of gene expression in a cell, such as the proportion of genes with non-zero count. Finak *et al.* (2015) refers to this as the "cellular detection rate". We find that this proportion can be quite misleading. Figure S12 shows that both the proportion of genes in Phase II expression, as estimated by SC2P, is highly correlated with the proportion of genes with non-zero counts in most cells, as expected. However, It is worth noting that there is considerable variation, and in some cells this discrepancy is substantial. Also, though there appears to be a linear relationship, the slope is not 1.

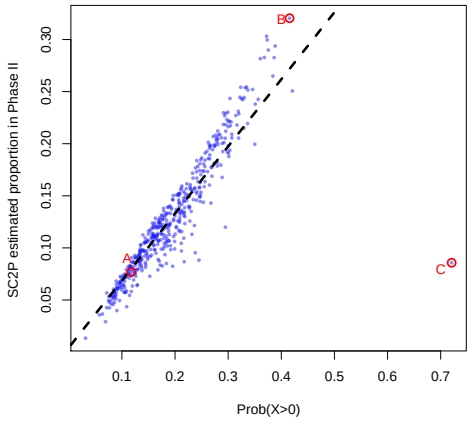

Figure S12: Estimated proportion of genes in Phase II versus genes with non zero count. The three cells in Figure 1 in the main text is highlighted.

8 Simulation

We perform simulation studies to compare the performances of DE detection from MAST and SC2P. The simulated data are generated based on the human brain data so that they mimic the real data characteristics. To be specific, we take the neuron cells (131 cells) from the human brain data, and estimate the statistical model parameters including

- π_{gi}: gene-specific probability for being in phase II.
- μ_g and σ_g^2: gene-specific parameters for the log-normal distribution.
- p_i and λ_i: cell-specific parameters for the zero-inflated Poisson distribution.
- S_i: size factors.

The counts Y_{gi} are simulated based on these parameters according to the data generative model presented in Section 2.1. To simulate counts, we keep the gene-specific parameters (π_{gi}, μ_g and σ_g^2) intact. The cell specific parameters (p_i, λ_i and S_i) are randomly drew with replacement from the estimates from the real data, given number of cells. In the simulation, we first generate the counts in the first treatment group. We then randomly picked 500 DE genes in form I and form II. For form I DE genes, the differences in probability for being in phase II between two groups are randomly generated from uniform [0.1, 0.3]. For form II DE genes, the log fold change in μ_g between two groups are randomly sampled from a mixture of normal distribution $0.5N(-1,1)+0.5N(1,1)$. Based on these, we compute the gene-specific parameters in the second treatment group for the DE genes, and then generate the counts.

We run MAST and SC2P on the simulated counts, and compare their performances in three areas:

1. Detection accuracy, as measured by the number of true DE genes among the top ranked DE genes.
2. Sensitivity: number of DE genes detected under certain FDR threshold.
3. Inference: comparison of observed and nominal FDR.

We run the simulation under different numbers of cells (50, 100, and 200). Under each setting, simulation was run for 20 times, and average values are obtained. Figure S13 summarizes the results when there are 50 cells in each treatment group. Figures in the first row compare the DE detection accuracies in both forms of DE. The results are essentially the same from MAST and SC2P. The second row compares the sensitivities of DE detection, and SC2P significantly outperforms MAST as it detects more DE genes under all FDR thresholds, in both forms. The third row compares the FDR estimation. SC2P has very accurate FDR, whereas MAST is overly conservative. This explains the better sensitivities showed in the figures in second row: the higher sensitivity from SC2P is due to the better FDR estimation. Figures S14 and S15 shows the results from having 100 and 200 cells in each treatment group, and the conclusions are essentially the same.

Overall, these results are consistent with the real data results: SC2P and MAST provide comparable gene ranks, but SC2P is more sensitive due to better statistical inference. The better sensitivity, combined with better robustness and computational efficiency, make SC2P a more desired method for scRNA-seq DE detection.

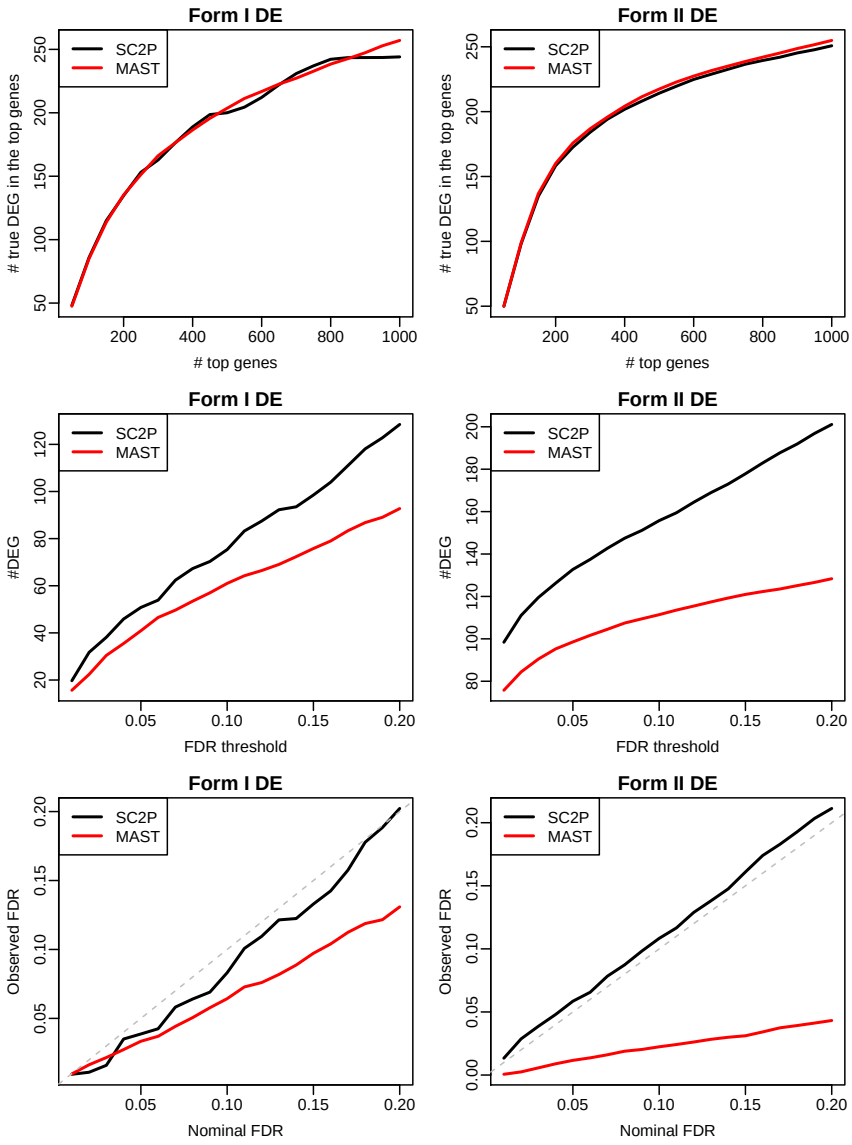

Figure S13: Simulation results when there are 50 cells in each treatment group.

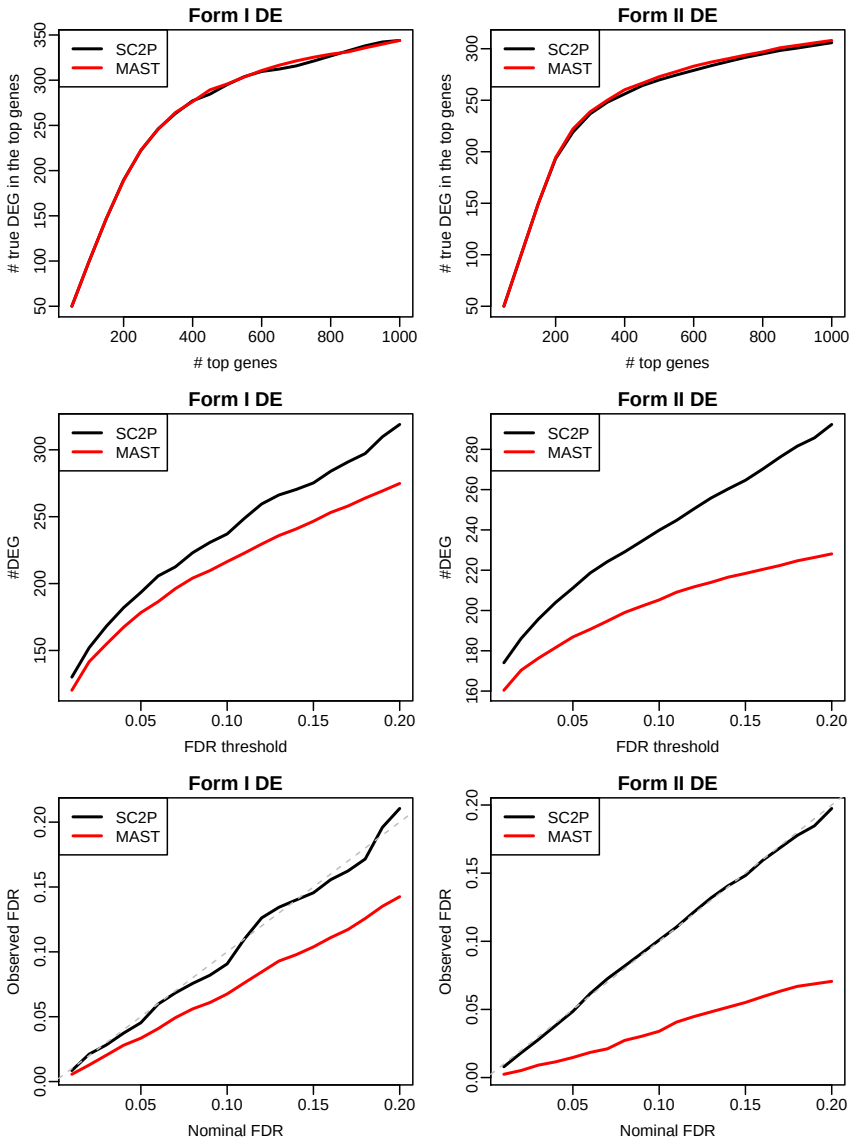

Figure S14: Simulation results when there are 100 cells in each treatment group.

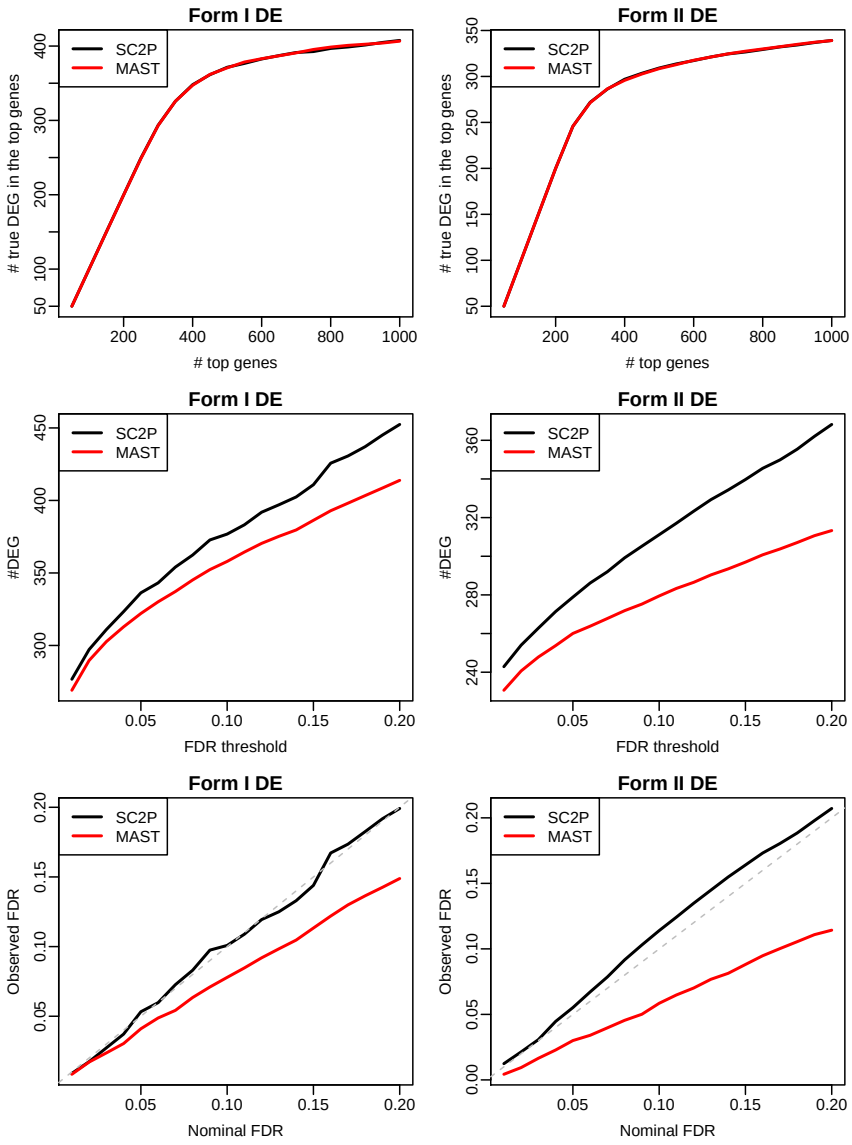

Figure S15: Simulation results when there are 200 cells in each treatment group.

9 Comparison with DESeq2

We run DESeq2 on the the human brain and T2D data sets, and compared its performance with other methods. Figure S16 shows the results from the human brain data astrocytes vs. oligodendrocytes comparison. First, in terms of recovering know marker genes (Figure S16A), DESeq2 performs in between the group of top performers (SC2P, MAST and SCDE) and the less sensitive BPSC. At various nominal FDR levels ranging from 1% to 20%, DESeq2 made more discoveries (Figure S16B), but at a cost of inflated type I error, especially at the lower end of type I error rate (Figure S16 C). The Venn diagram in S16 D shows that DESeq2 identified many of the genes reported by SC2P and MAST, but also has a large number of unique discoveries. Given the inflated type I error, these are likely false discoveries.

Figures for other comparisons from the human brain and T2D data are shown in Figures S17–S20. The general conclusions from these results are essentially the same.

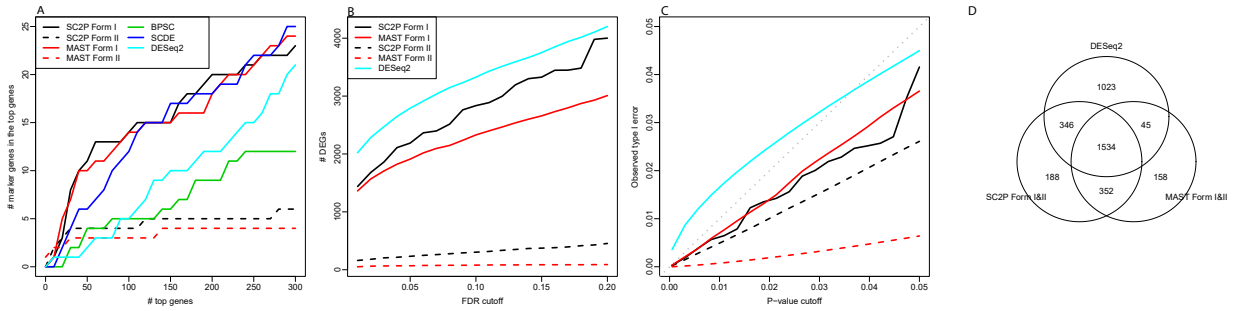

Figure S16: Comparing DESeq2 with other methods for DE detection in human brain data, astrocytes vs. oligodendrocytes comparison.

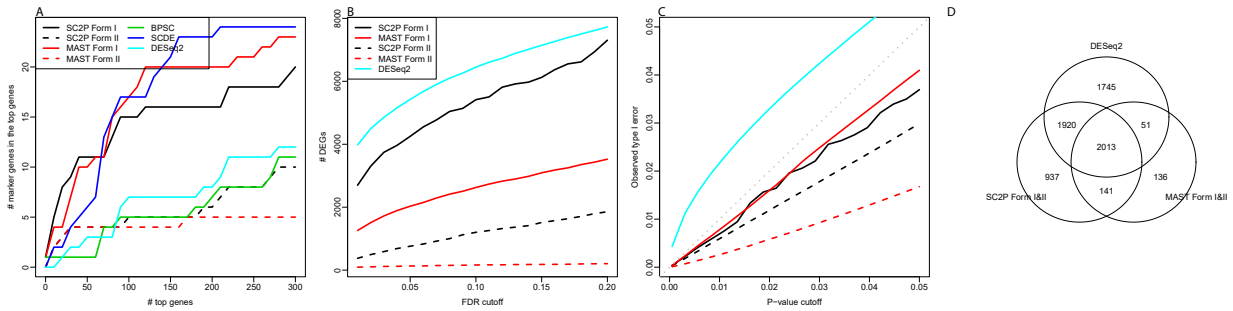

Figure S17: Comparing DESeq2 with other methods for DE detection in human brain data, neurons vs. oligodendrocytes comparison.

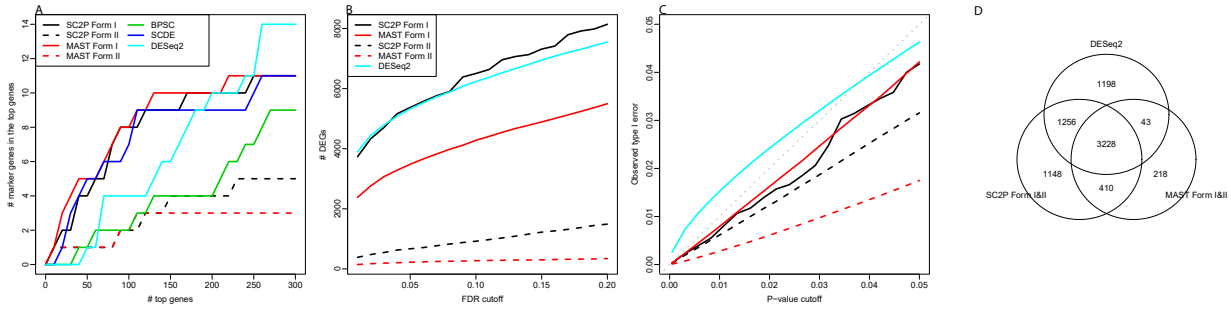

Figure S18: Comparing DESeq2 with other methods for DE detection in human brain data, astrocytes vs. neurons comparison.

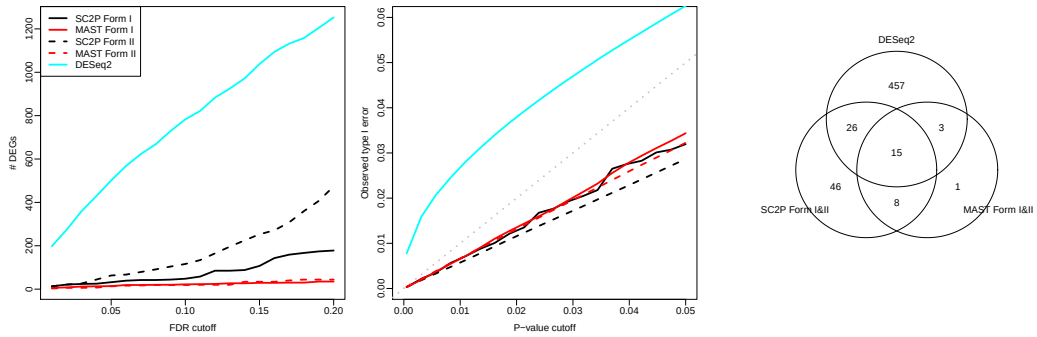

Figure S19: Comparing DESeq2 with other methods for DE detection in T2D data, comparing alpha cells between T2D patients and normal control.

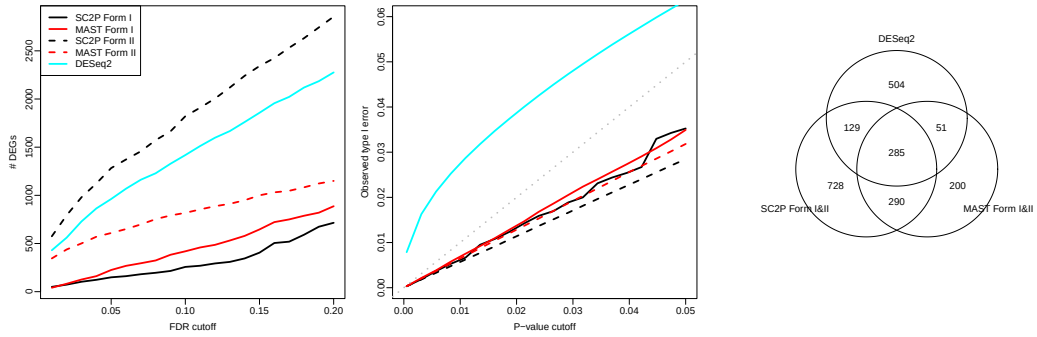

Figure S20: Comparing DESeq2 with other methods for DE detection in T2D data, comparing beta cells between T2D patients and normal control.

References

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., *et al.* (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, **16**(1), 278.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome research*, **27**(2), 208–222.

Wu, H., Wang, C., and Wu, Z. (2012). A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, **14**(2), 232–243.