Revision

Reviewer 1

Question about number of clusters at each condition

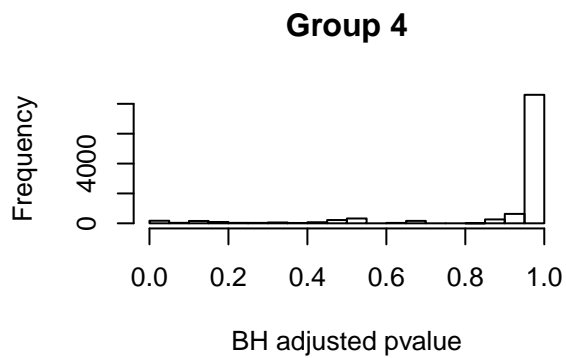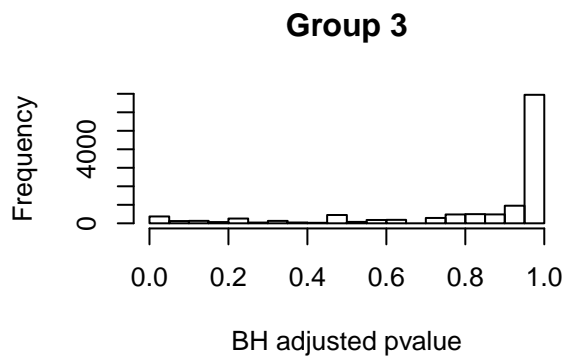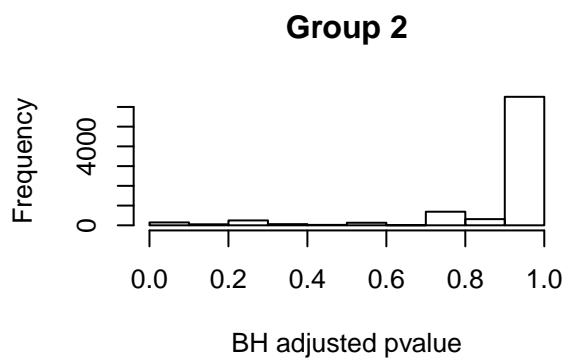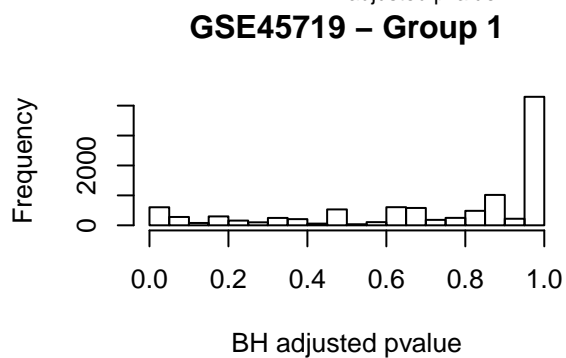(1) One key assumption they made is that the number of cell clusters in both conditions is the same (K). This does not allow the scenarios where different conditions may have different cell subtypes (e.g., normal versus controls). Of course, if the cell types are different, then all genes should have different distributions. Should not they first decide whether this is the case?

It is not an assumption for number of clusters to be same across two conditions. Our method pooled cells from both conditions first and identified a global clustering for all the cells, the number of clusters (K) is for the global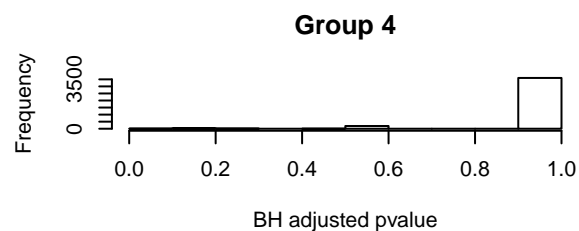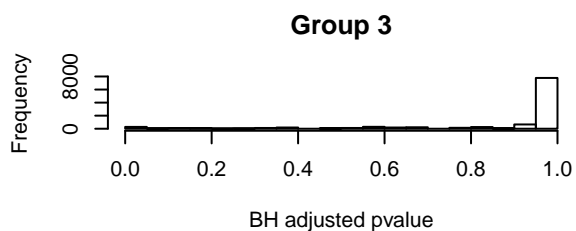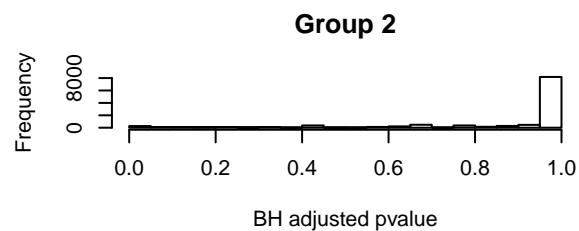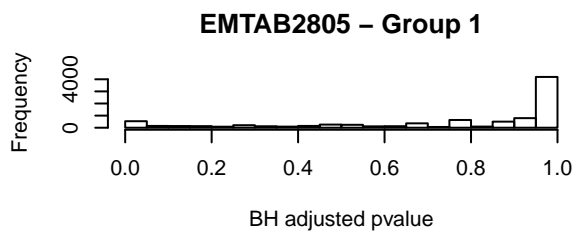 clustering. These two conditions can have different number of clusters given that one condition can have zero proportion of cells from a cluster. For example, $K = 3$, there are 3 clusters (labelled as A, B, C) globally. Condition 1 having 3 types, 60% from A, 30% from B and 10% from C, while condition 2 only have 2 types 50% from B and 50% from C. It is not necessary that different cell types will induce different distributions over the whole genome. Actually the major proportion of genes should still have equivalent distribution. Which is key observation we proposed in the paper and we propose an empirical Bayesian framework with a specifically constructed prior to handle the scenario when both means and proportions changed.

Questions about fitting data with mixture of NB

(2) Single cell RNA-seq data often have lots of zeros. I was wondering how well the mixture of multinomial distributions really fit the data. Some plots that show the model fits would be useful.

We are using mixture of negative binomial to model the counts instead of multinomial distributions. Mixture of NB are flexible to approximate a lot of distributions. Specifically, NB can approximate constant 0 by arbitrary accuracy. As we know the density of $P(X = 0) = 1 - p$ and $P(X > 0) = p$ given $X \sim \text{NB}(1, p)$. Further, scRNA-seq data tends to be overdispersed so that many people are using NB to model it. (cite DESeq2, cite Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression). Also, the empirical data (EMTAB2805, GSE45719) showed that more than 80% of genes their nonzero counts are overdispersed. In addition, given our estimated group labels, we test whether there is a strong evidence to reject fitting NB model at each group. Given the cells from same group, we excluded those genes with more than 70% of 0s, as we know NB can approximate 0 well. For the remaining genes, we do the goodness of fit tests following the procedure proposed in the paper "Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation(Guosheng Yin,Yanyuan Ma)". We found very few genes would reject the negative binomial model

EMTAB2805 − Group 1

Frequency

4000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 2

Frequency

8000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 3

Frequency

8000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 4

Frequency

3500

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 5

Frequency

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

GSE45719 − Group 1

Frequency

2000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 2

Frequency

4000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 3

Frequency

4000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Group 4

Frequency

4000

0

0.0   0.2   0.4   0.6   0.8   1.0

BH adjusted pvalue

Questions not sure

(3). For the transcriptional bursting data analysis,
some comparisons with other methods such as MAST, DESEQ2, scDD would be useful, as they did for
synthetic data sets.

Reviewer 2

Questions not sure

1.How does the model behave when both the proportions and the means are different? The calculation p7
(before the key issue) should be detailed. What hypotheses on $f_{g,k}$ allow this result ? The distribution can
remain the same even both proportions and means changed. We assume the only parameter differs $f_{g,k}$ for k
is the mean.

The distribution can remain the same even both proportions and means changed. We assume the only
parameter differs $f_{g,k}$ for k is the mean.

2.It is assumed that the shape parameter is constant and independent to the population classes. Is it a
realistic hypothesis ? What is it impact on the results ? Is it possible to relax this hypothesis to consider a
shape parameter depending on k ? A discussion on this hypothesis should be added in the discussion.

For the fixed shape parameter of NB among groups, we performed tests on each gene. The testing procedure
is following: the null hypothesis is for different groups having the same shape parameter. Under the null we
pool the data and find the MLE and obtain it's condifence interval by fisher information. Under alternative,
we use method of moments to obtain point estimator of shape parameter separately at each group. If there is
at least one group's point estimator being outside of the confidence interval, we flagged the corresponding
gene. We run the test on the GSE75748. We found 3305 out of 19097 genes (17.3%) having been identified as
genes with hetergenous shape parameters. So for the majority of genes, it is ok to assume a shared shape
parameter across groups. Further, our aim is to infer the distribution pattern of mixing components. If for
some genes and some groups both mean and shape parameters are different. It may still be ok only testing
the change pattern of means. Among those 3305 genes, 1327 (40%) of them also demonstrated different
means between groups. We also run the same procedure on GSE45719. We found 23% genes having been
flagged and 45% among them are also having differential means between groups. I would say overall the
proportion of genes may be affect by the same shape parameter assumption is low

Question about sensitivity to clustering

3.For the clustering, the authors should better justify their choice. The results seem to be strongly related to
the clustering method.

We have tried different clustering method, sc3 on 3 datasets, (EBTAB2805, GSE45719, GSE79102). the
correlation coefficients of the local fdr are 0.92,0.94,0.96. Our method is not very sensitive to the clustering
methods.

Questions for simulations

Finally for the simulation part, the data generation should be explained and it would be nice to have criteria
evaluating the difficulty of the simulated datasets. For example $K = 12$ seems very difficult, and $K = 7$ also
(On Figure S7, some ROC curves are closed to the first bisector. Moreover on this plot, configurations are
not precised).

More detailed information of simulation has been placed in the supplementary material, those pca plots
showed that groups are collapsed together when there are more groups, which make all the methods difficult

to detect DD genes. In addition, we have fixed number of cells(400) so more number of groups will decrease the number of cells per group which makes it harder to detect the change between groups also mixing of more groups weaken the overall signal between conditions.

Questions for simulations

The results on Figures 4 and 5 are questionable. ScDDboost has the best TPR and a FDR closed to 0, whereas this latter should be controled at 5%. DESeq2 seems to better control the trade-off. Can the authors comment this remark ? In general, I am very surprised by the very small number of replicate datasets per scenario. Is it possible to increase it and to use boxplots and ROC curves to summary the results instead of one figure for the TPR and one figure for the FDR ?

We have done more replicates for each simulation setting, now we have 10 replicates under each settings. And we have similar results as before. (add plots later)
Our method is not a trade-off for more power by using FDR.