­­

**Check the data features from the simulation**

Due to the complex of the real world, I plan to first check the reasons for un-even distributed permutated pvalues from the simulation. Here’s the result.

**Part I: For jsd\_zinb inflation of type I error…**

It seems like if the variance of the distance are larger, then the p-values might be inflated

jsd\_zinb Case 1

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="disp";  r\_mean=1.2;  r\_var=1.2;r\_disp=1.1;  r\_mult=0.4;file\_tag=2;  n\_ind=10;n\_cell=60;  dist\_method="jsd";  fit\_method="zinb"; | quantile(50%, 75% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.12.45 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.06.03 AM.png |

jsd\_zinb Control 1

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="disp";  r\_mean=1.2;  r\_var=1.2;r\_disp=1.1;  r\_mult=0.4;file\_tag=2;  n\_ind=40;n\_cell=100;  dist\_method="jsd";  fit\_method="zinb"; | quantile(50%, 75% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.23.13 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.23.46 AM.png |

For jsd\_zinb inflation of type I error…

jsd\_zinb Case 2

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="mean";  r\_mean=1.1;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=20;n\_cell=80;  dist\_method="jsd";  fit\_method="zinb"; | quantile(50%, 75% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.34.15 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.58.30 AM.png |

jsd\_zinb Control 2

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="mean";  r\_mean=1.1;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=20;n\_cell=60;  dist\_method="jsd";  fit\_method="zinb"; | quantile(50%, 75% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.42.02 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.59.57 AM.png |

jsd\_zinb Control 2.2

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="mean";  r\_mean=1.1;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=20;n\_cell=100;  dist\_method="jsd";  fit\_method="zinb"; | quantile(50%, 75% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 12.46.51 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.01.45 AM.png |

Based on the case2 vs control 2 senario, we extract the cells between n\_cell=60 to n\_cell=80;

Then we extract the count from our simulation and plots the count distribution.

> table(sim\_matrix[1:50,])

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 24 26 32 35

12310 4022 1694 801 399 247 157 98 82 59 41 22 20 16 9 5 4 3 3 2 1 1 1 1 1 1

> table(sim\_matrix[1501:2050,])/10

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

15660.7 4497.2 1223.2 358.0 119.2 53.5 26.4 17.5 12.0 8.6 6.7 3.9 3.6 2.4 1.8 1.2 0.8 1.3 0.8

19 20 21 22 25 26 29

0.3 0.2 0.3 0.1 0.1 0.1 0.1

Thus it gives the hint that there are a small amount of cell have extra high expressions

Then I focus on the parameters(logmean, overdisp, dropout) of the “gap 20 cells”

From the 100 sig genes vs middle genes.

Pval of senario:

perm\_label=1;param\_tag="mean";

r\_mean=1.1;

r\_var=1.2;r\_disp=1.2;

r\_mult=0.4;file\_tag=1;

n\_ind=20;n\_cell=60-80;

dist\_method="jsd";

fit\_method="zinb";

|  |  |  |  |
| --- | --- | --- | --- |
| mean | overdisp | dropout | Dist\_array |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.51.27 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.51.36 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.51.44 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.51.58 AM.png |

It seems that the situation large distance is related to the mean parameters.

Then, we calculate the correlation between maxium/median distance of a dist matrix and the maxium of mean parameter for the corresponding cells who seems to introduce significant p-values. We found the correlation is super high.

Pval of senario:

perm\_label=1;param\_tag="mean";

r\_mean=1.1;

r\_var=1.2;r\_disp=1.2;

r\_mult=0.4;file\_tag=1;

n\_ind=20;n\_cell=60-80;

dist\_method="jsd";

fit\_method="zinb";

|  |  |
| --- | --- |
| maxium distance of a dist matrix and the maxium of mean parameter | median distance of a dist matrix and the maxium of mean parameter |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.58.45 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 1.59.37 AM.png |

Corresponding controls

Pval of senario:

perm\_label=1;param\_tag="mean";

r\_mean=1.1;

r\_var=1.2;r\_disp=1.2;

r\_mult=0.4;file\_tag=1;

n\_ind=20;n\_cell=60;

dist\_method="jsd";

fit\_method="zinb";

|  |  |  |  |
| --- | --- | --- | --- |
| mean | overdisp | dropout | Dist\_array |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 2.06.56 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 2.07.09 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 2.07.21 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 2.08.14 AM.png |

Corresponding controls

Pval of senario:

perm\_label=1;param\_tag="mean";

r\_mean=1.1;

r\_var=1.2;r\_disp=1.2;

r\_mult=0.4;file\_tag=1;

n\_ind=20;n\_cell=60;

dist\_method="jsd";

fit\_method="zinb";

|  |  |
| --- | --- |
| maxium distance of a dist matrix and the maxium of mean parameter | median distance of a dist matrix and the maxium of mean parameter |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 2.12.12 AM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-02 at 2.14.37 AM.png |

Thus we can see, although there are still relatively strong correlation between distance matrix distribution vs the simulation mean. There extreme values of mean introduce the variance ditances and this effect is correlated to the uneven permutated pvalues from permutation based on the JSD zinb method

**Part II: For jsd\_empirical inflation of type I error…**

It seems like the variance of overdispersion are associated with type I error inflation

jsd\_empirical Case 1

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="var";  r\_mean=1.2;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=40;n\_cell=100;  dist\_method="jsd";  fit\_method="empirical"; | quantile(min, 50% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.09.14 PM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.10.18 PM.png |

jsd\_empirical Control 1

|  |  |
| --- | --- |
| Pval of senario:  perm\_label=1;param\_tag="var";  r\_mean=1.2;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=40;n\_cell=80;  dist\_method="jsd";  fit\_method="empirical"; | quantile(min, 50% and max) of dist\_array of genes with first 500 sig pval and 1001-1500 sig pval |
| Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.21.47 PM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.22.44 PM.png |

|  |  |  |
| --- | --- | --- |
| a dist matrix and the overdisp  parameter | **Case**  Pval of senario:  perm\_label=1;param\_tag="var";  r\_mean=1.2;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=40;n\_cell=100;  dist\_method="jsd";  fit\_method="empirical"; | **Control**  Pval of senario:  perm\_label=1;param\_tag="var";  r\_mean=1.2;  r\_var=1.2;r\_disp=1.2;  r\_mult=0.4;file\_tag=1;  n\_ind=40;n\_cell=80;  dist\_method="jsd";  fit\_method="empirical"; |
| Max | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.13.57 PM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.24.24 PM.png |
| median | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.14.07 PM.png | Macintosh HD:Users:grasscliff:Desktop:Screen Shot 2020-02-03 at 3.24.35 PM.png |

Thus it seems like some genes with large variance of overdispersion are associated with the inflated type I error…

**Part III: Conclusions…**

According to current situation, we can

1. Refer DESeq2 an use cook distance to move further points closer for ZINB method.

2. For empirical method, currently no idea…