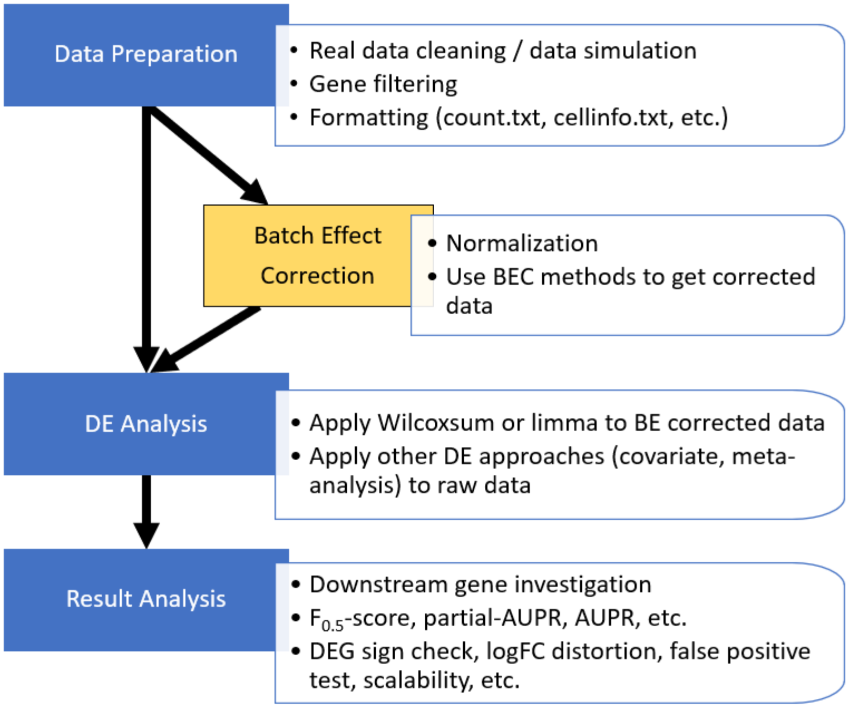
Benchmarking integration of single-cell differential expression

How-to-use-the-code



Data Preparation

- Real data cleaning / data simulation
- Gene filtering
- Formatting (count.txt, cellinfo.txt, etc.)

Batch Effect Correction

- Normalization
- Use BEC methods to get corrected data

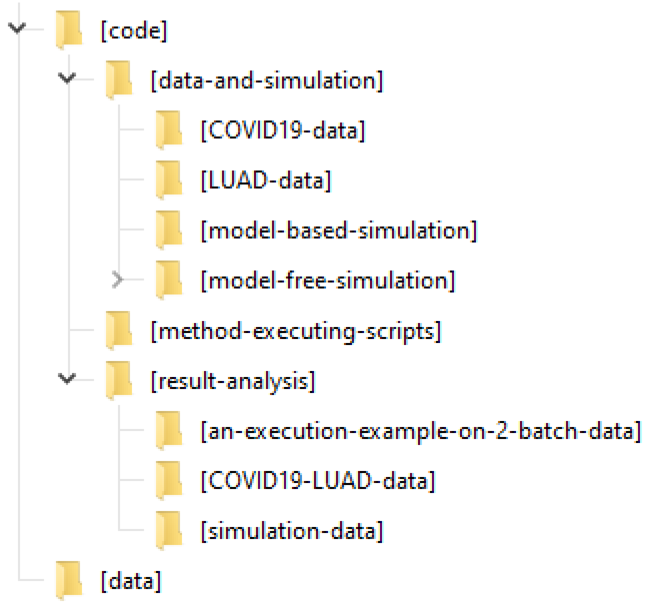
DE Analysis

- Apply Wilcoxsum or limma to BE corrected data
- Apply other DE approaches (covariate, meta-analysis) to raw data

Result Analysis

- Downstream gene investigation
- $F_{0.5}$-score, partial-AUPR, AUPR, etc.
- DEG sign check, logFC distortion, false positive test, scalability, etc.

1. GitHub folder structure

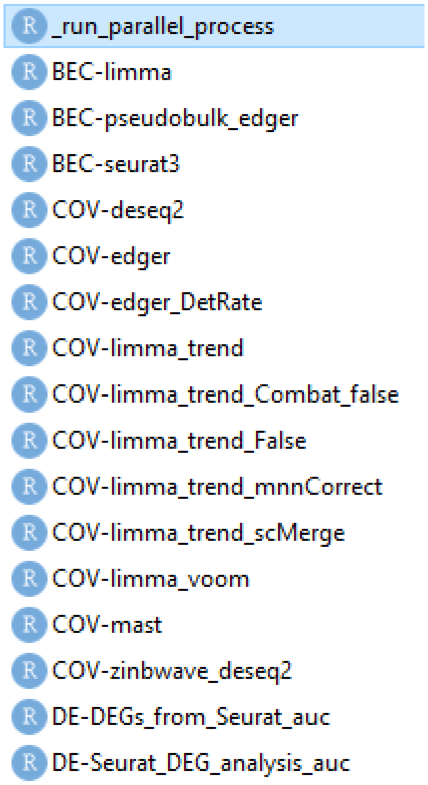


```
[code]
    [data-and-simulation]
        [COVID19-data]
        [LUAD-data]
        [model-based-simulation]
        [model-free-simulation]
    [method-executing-scripts]
    [result-analysis]
        [an-execution-example-on-2-batch-data]
        [COVID19-LUAD-data]
        [simulation-data]
[data]
```

- **‘code’** contains the core analysis Python & R scripts for this study
  - **‘data-and-simulation’** contains sample data and scripts for data preparation step
    - **‘COVID19-data’** gives scripts for data preparation using COVID-19 data
    - **‘LUAD-data’** gives scripts for data preparation using LUAD data
    - **‘model-based-simulation’** gives scripts for simulating data using MCA and Pancreas data
    - **‘model-free-simulation’** gives scripts for simulating data using Splatter
  - **‘method-executing-scripts’** contains implementation for each considered method
  - **‘result-analysis’** contains scripts for analyzing

- **‘data'** contains figures and tables of the experimental results for illustration

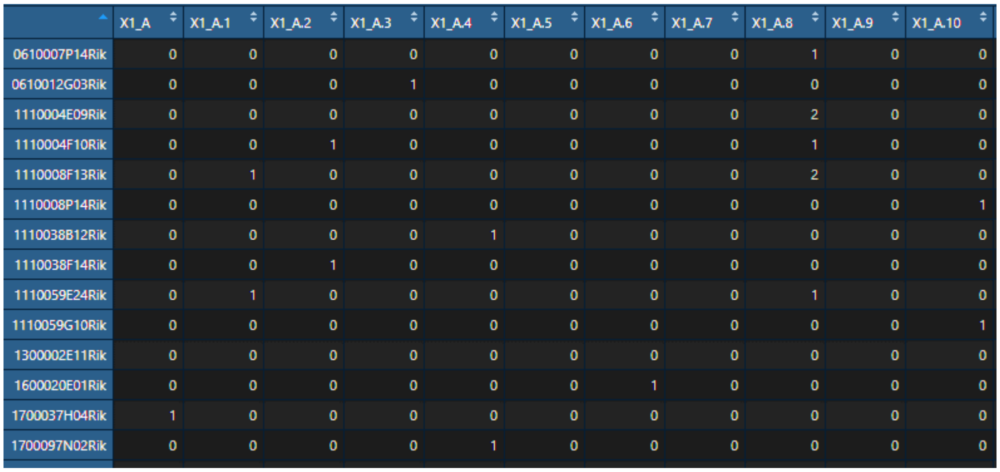
2. Analysis code input-output



_run_parallel_process
BEC-limma
BEC-pseudobulk_edger
BEC-seurat3
COV-deseq2
COV-edger
COV-edger_DetRate
COV-limma_trend
COV-limma_trend_Combat_false
COV-limma_trend_False
COV-limma_trend_mnnCorrect
COV-limma_trend_scMerge
COV-limma_voom
COV-mast
COV-zinbwave_deseq2
DE-DEGs_from_Seurat_auc
DE-Seurat_DEG_analysis_auc

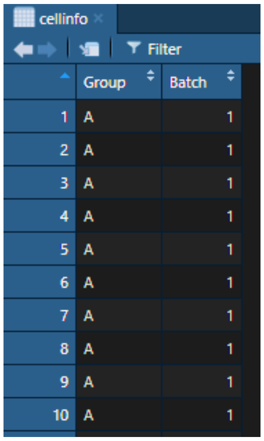
- **'_run_parallel_process':** a list of all testing methods to run multiple scripts at the same time.
- **'<method-category>_<method-name>':** script to specifically test a particular method
    - **< method-category >:** includes 'BEC', 'COV', 'META', 'DE' indicating the characteristic of a method
    - **< method-name >:** indicate the specific method

- **input:**
    - a count matrix (genes × cells)



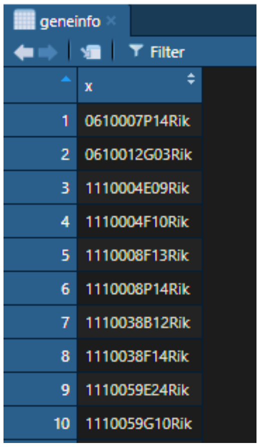
	X1_A	X1_A.1	X1_A.2	X1_A.3	X1_A.4	X1_A.5	X1_A.6	X1_A.7	X1_A.8	X1_A.9	X1_A.10
0610007P14Rik	0	0	0	0	0	0	0	0	1	0	0
0610012G03Rik	0	0	0	1	0	0	0	0	0	0	0
1110004E09Rik	0	0	0	0	0	0	0	0	2	0	0
1110004F10Rik	0	0	1	0	0	0	0	0	1	0	0
1110008F13Rik	0	1	0	0	0	0	0	0	2	0	0
1110008P14Rik	0	0	0	0	0	0	0	0	0	0	1
1110038B12Rik	0	0	0	0	1	0	0	0	0	0	0
1110038F14Rik	0	0	1	0	0	0	0	0	0	0	0
1110059E24Rik	0	1	0	0	0	0	0	0	1	0	0
1110059G10Rik	0	0	0	0	0	0	0	0	0	0	1
1300002E11Rik	0	0	0	0	0	0	0	0	0	0	0
1600020E01Rik	0	0	0	0	0	0	1	0	0	0	0
1700037H04Rik	1	0	0	0	0	0	0	0	0	0	0
1700097N02Rik	0	0	0	0	1	0	0	0	0	0	0

- a data frame of cell descriptions (group, batch, … information)



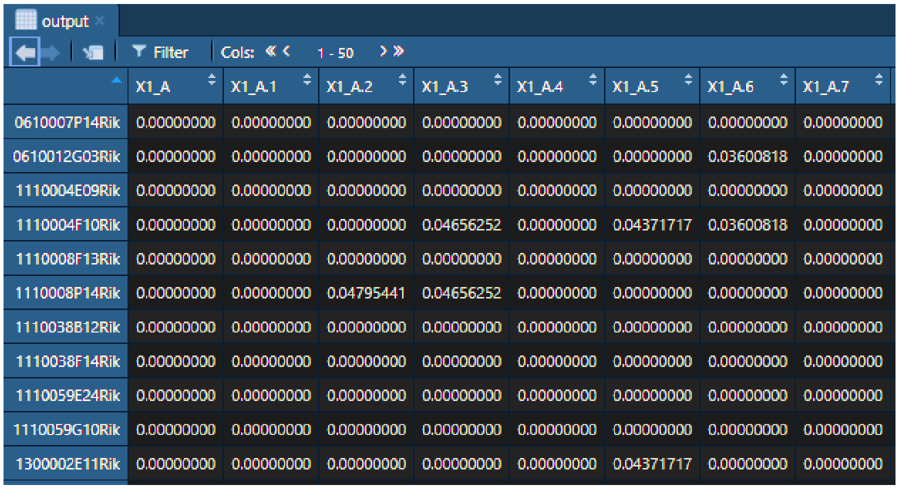
	Group	Batch
1	A	1
2	A	1
3	A	1
4	A	1
5	A	1
6	A	1
7	A	1
8	A	1
9	A	1
10	A	1

- a data frame of gene descriptions (id, name, code, …information)



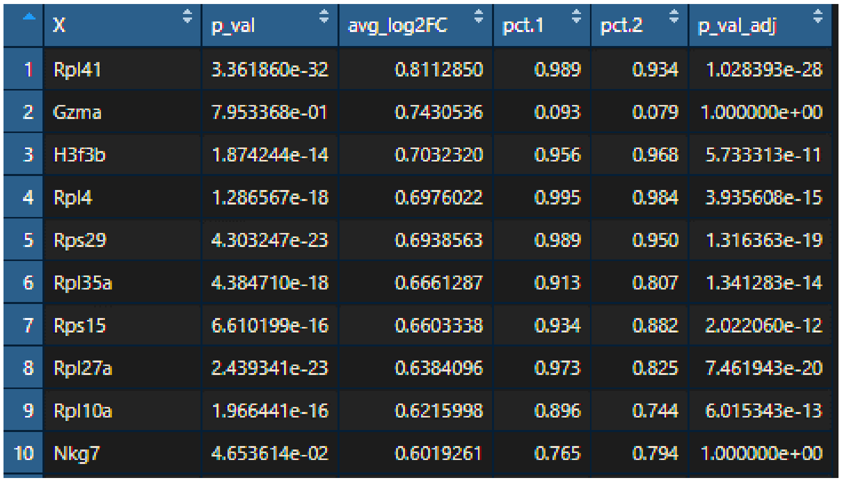
	x
1	0610007P14Rik
2	0610012G03Rik
3	1110004E09Rik
4	1110004F10Rik
5	1110008F13Rik
6	1110008P14Rik
7	1110038B12Rik
8	1110038F14Rik
9	1110059E24Rik
10	1110059G10Rik

- **output:**
    - batch effect correction methods: a matrix of corrected values (genes × cells)



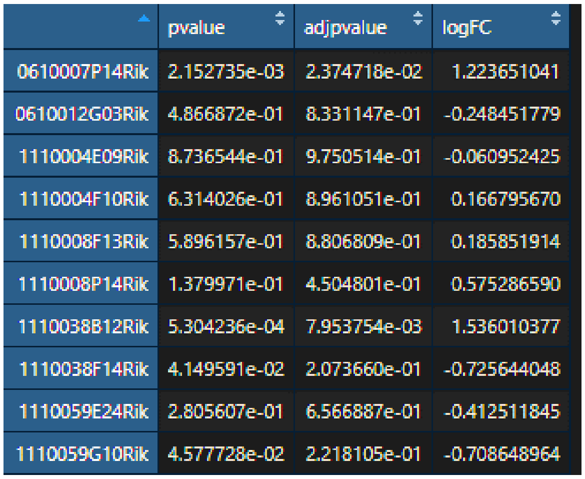
	X1_A	X1_A.1	X1_A.2	X1_A.3	X1_A.4	X1_A.5	X1_A.6	X1_A.7
0610007P14Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
0610012G03Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.03600818	0.00000000
1110004E09Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
1110004F10Rik	0.00000000	0.00000000	0.00000000	0.04656252	0.00000000	0.04371717	0.03600818	0.00000000
1110008F13Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
1110008P14Rik	0.00000000	0.00000000	0.04795441	0.04656252	0.00000000	0.00000000	0.00000000	0.00000000
1110038B12Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
1110038F14Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
1110059E24Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
1110059G10Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
1300002E11Rik	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.04371717	0.00000000	0.00000000

- Wilcoxson rank sum test: a data frame of gene ranking analysis



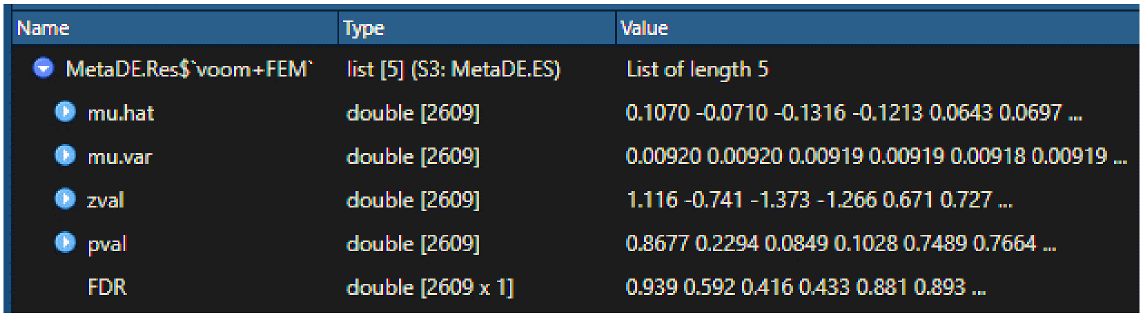
	X	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
1	Rpl41	3.361860e-32	0.8112850	0.989	0.934	1.028393e-28
2	Gzma	7.953368e-01	0.7430536	0.093	0.079	1.000000e+00
3	H3f3b	1.874244e-14	0.7032320	0.956	0.968	5.733313e-11
4	Rpl4	1.286567e-18	0.6976022	0.995	0.984	3.935608e-15
5	Rps29	4.303247e-23	0.6938563	0.989	0.950	1.316363e-19
6	Rpl35a	4.384710e-18	0.6661287	0.913	0.807	1.341283e-14
7	Rps15	6.610199e-16	0.6603338	0.934	0.882	2.022060e-12
8	Rpl27a	2.439341e-23	0.6384096	0.973	0.825	7.461943e-20
9	Rpl10a	1.966441e-16	0.6215998	0.896	0.744	6.015343e-13
10	Nkg7	4.653614e-02	0.6019261	0.765	0.794	1.000000e+00

- parametric and integration methods: a data frame of gene ranking analysis



	pvalue	adjpvalue	logFC
0610007P14Rik	2.152735e-03	2.374718e-02	1.223651041
0610012G03Rik	4.866872e-01	8.331147e-01	-0.248451779
1110004E09Rik	8.736544e-01	9.750514e-01	-0.060952425
1110004F10Rik	6.314026e-01	8.961051e-01	0.166795670
1110008F13Rik	5.896157e-01	8.806809e-01	0.185851914
1110008P14Rik	1.379971e-01	4.504801e-01	0.575286590
1110038B12Rik	5.304236e-04	7.953754e-03	1.536010377
1110038F14Rik	4.149591e-02	2.073660e-01	-0.725644048
1110059E24Rik	2.805607e-01	6.566887e-01	-0.412511845
1110059G10Rik	4.577728e-02	2.218105e-01	-0.708648964

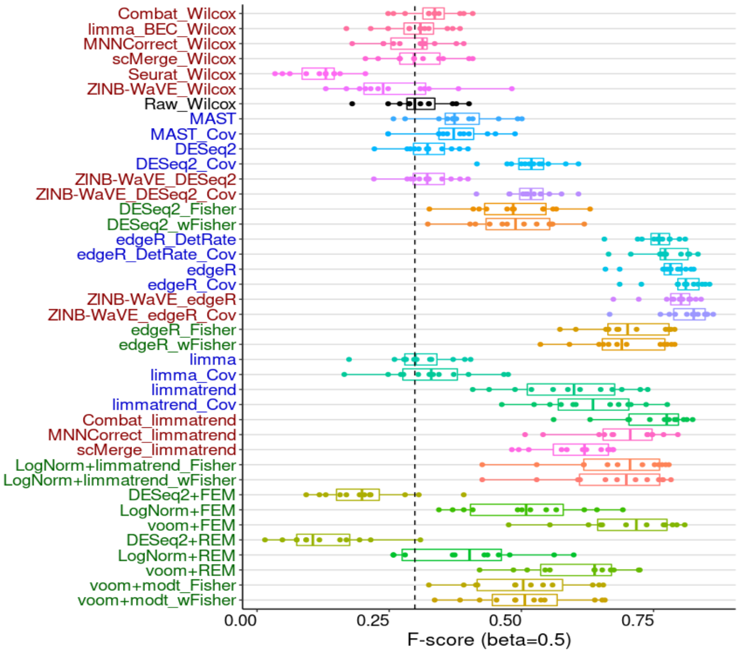
- meta-analysis methods: a data frame of gene ranking analysis



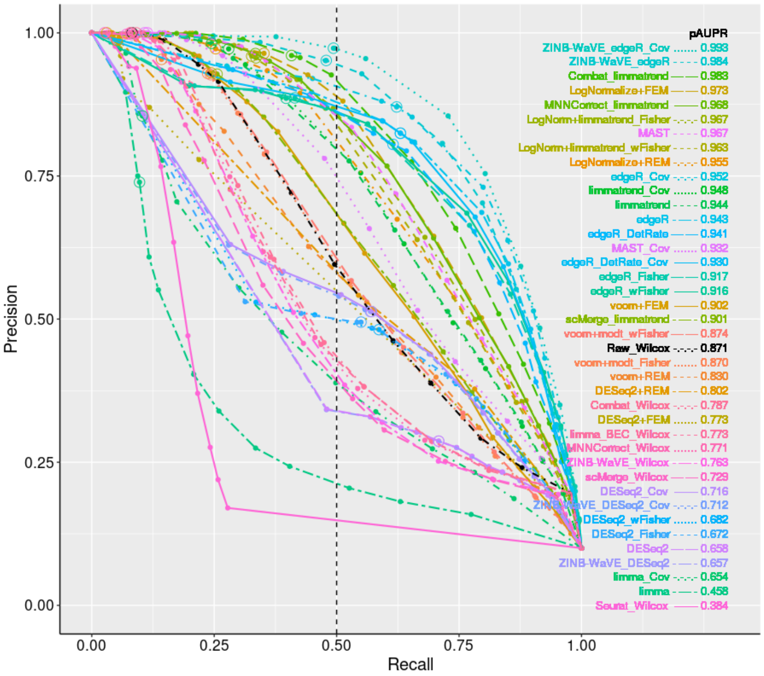
Name	Type	Value
MetaDE.Res$`voom+FEM`	list [5] (S3: MetaDE.ES)	List of length 5
mu.hat	double [2609]	0.1070 -0.0710 -0.1316 -0.1213 0.0643 0.0697 ...
mu.var	double [2609]	0.00920 0.00920 0.00919 0.00919 0.00918 0.00919 ...
zval	double [2609]	1.116 -0.741 -1.373 -1.266 0.671 0.727 ...
pval	double [2609]	0.8677 0.2294 0.0849 0.1028 0.7489 0.7664 ...
FDR	double [2609 x 1]	0.939 0.592 0.416 0.433 0.881 0.893 ...

3. Visualization

- **‘GBC-meta’:**
  - Aggerate all output results and visualize F-beta performance



- **‘GBC-meta_PR’:**
  - Aggerate all output results and illustrate the AUPR curve


Precision
Recall
pAUPR
ZINB-WaVE_edgeR_Cov 0.993
ZINB-WaVE_edgeR - - - - 0.984
Combat_limmatrend — — 0.983
LogNormalize+FEM — — 0.973
MNNCorrect_limmatrend —— 0.968
LogNorm+limmatrend_Fisher -.-.-.- 0.967
MAST - - - - 0.967
LogNorm+limmatrend_wFisher - - - - 0.963
LogNormalize+REM -.-.-.- 0.955
edgeR_Cov -.-.-.- 0.952
limmatrend_Cov 0.948
limmatrend - - - - 0.944
edgeR - — - 0.943
edgeR_DetRate — — 0.941
MAST_Cov 0.932
edgeR_DetRate_Cov —— 0.930
edgeR_Fisher — — 0.917
edgeR_wFisher —— 0.916
voom+FEM —— 0.902
scMerge_limmatrend - — - 0.901
voom+modt_wFisher — — 0.874
Raw_Wilcox -.-.-.- 0.871
voom+modt_Fisher 0.870
voom+REM - - - - 0.830
DESeq2+REM - — - 0.802
Combat_Wilcox -.-.-.- 0.787
DESeq2+FEM 0.773
limma_BEC_Wilcox - - - - 0.773
MNNCorrect_Wilcox 0.771
ZINB-WaVE_Wilcox - — - 0.763
scMerge_Wilcox — — 0.729
DESeq2_Cov —— 0.716
ZINB-WaVE_DESeq2_Cov -.-.-.- 0.712
DESeq2_wFisher 0.682
DESeq2_Fisher - - - - 0.672
DESeq2 — — 0.658
ZINB-WaVE_DESeq2 - — - 0.657
limma_Cov -.-.-.- 0.654
limma - — - 0.458
Seurat_Wilcox —— 0.384