Manual

Benchmarking integration of single-cell RNA-seq differential analysis

Hai C. T. Nguyen[1†], Bukyung Baik[1†], Sora Yoon[1,2], Hae-Ock Lee, Taesung Park[3], Dougu Nam[1,4*]

[1]Department of Biological Sciences, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea

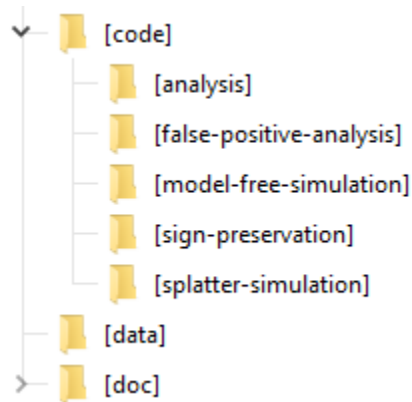[2]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104

[3]Department of Statistics, Seoul National University, Seoul, 08826, Republic of Korea

[4]Department of Mathematical Sciences, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea

†These authors contributed equally to this work.

*To whom correspondence should be addressed. Tel: +82-52-217-2525; Fax: +82-52-217-2639; Email: dougnam@unist.ac.kr

1. Github folder structure

```
[code]
    [analysis]
    [false-positive-analysis]
    [model-free-simulation]
    [sign-preservation]
    [splatter-simulation]
[data]
[doc]
```

- **'code'** contains the core analysis R scripts for this study

  o **'analysis'** contains scripts for all interesting methods and integration approaches that were benchmarked in this study.

  o **'false-positive-analysis'** contains Splatter scripts for generating null dataset for false positive analysis.

  o **'model-free-simulation'** contains scripts for generating batch effected dataset for false positive analysis.

  o **'sign-preservation'** contains scripts for checking the ratio of DE genes that preserved their signs during batch correction procedures.

  o **'splatter-simulation'** contains Splatter scripts for simulating 2-batch and 4-batch datasets.

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

- **'doc'** contains R markdown tutorials for 'model-free simulation' and 'incomplete association simulation' and this manual.
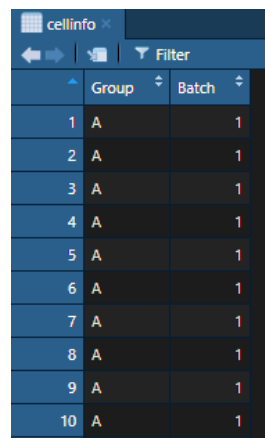
2. Analysis code input-output

R gb_process

R GBC_run_mast_cov	R GBC-run_limma_trend_False_cov
R GBC_run_mast	R GBC-run_limma_trend_False
R GBC-run_zinbwave_edger_cov	R GBC-run_limma_trend_Combat_false
R GBC-run_zinbwave_edger	R GBC-run_limma_trend_Combat
R GBC-run_zinbwave_deseq2_cov	R GBC-run_limma_trend
R GBC-run_zinbwave_deseq2	R GBC-run_limma
R GBC-run_zinbwave	R GBC-run_edger_cov
R GBC-run_seurat3	R GBC-run_edger_DetRate_cov
R GBC-run_scmerge	R GBC-run_edger_DetRate
R GBC-run_pseudobulk_edger	R GBC-run_edger
R GBC-run_limma_voom_cov	R GBC-run_deseq2_cov
R GBC-run_limma_voom	R GBC-run_deseq2
R GBC-run_limma_trend_scMerge_False	R GBC-run_combat
R GBC-run_limma_trend_scMerge	R GBC-run_MNN
R GBC-run_limma_trend_mnnCorrect_False	R GBC-run_DEGs_zinbwave_auc
R GBC-run_limma_trend_mnnCorrect	R GBC-run_DEGs_from_Seurat_auc
R GBC-run_limma_trend_cov	R GBC-Seurat_DEG_analysis_auc

- **'gb_process':** a list of all testing methods to run multiple scripts at the same time.

- **'GBC-run_<method_name>':** script to specifically test a particular method

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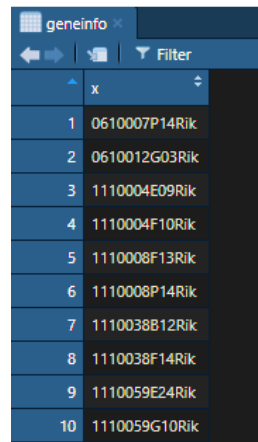
- a data frame of cell descriptions (group, batch, … information)

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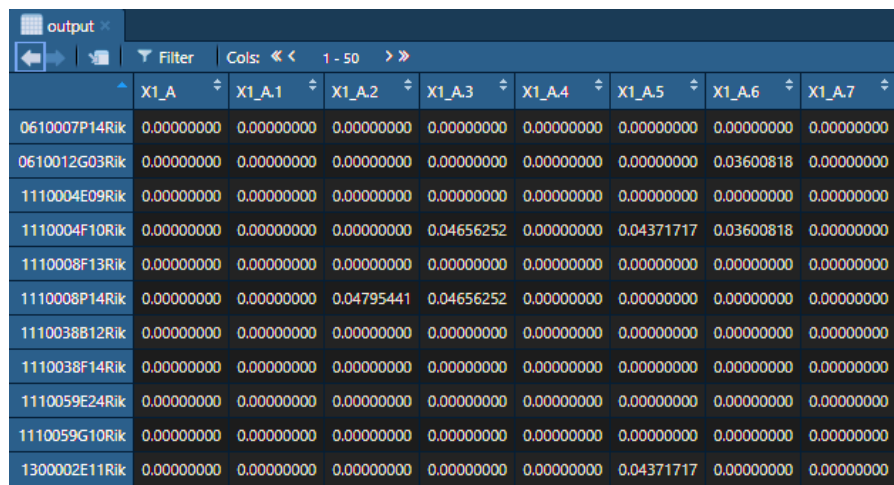
- **output:**

  - 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
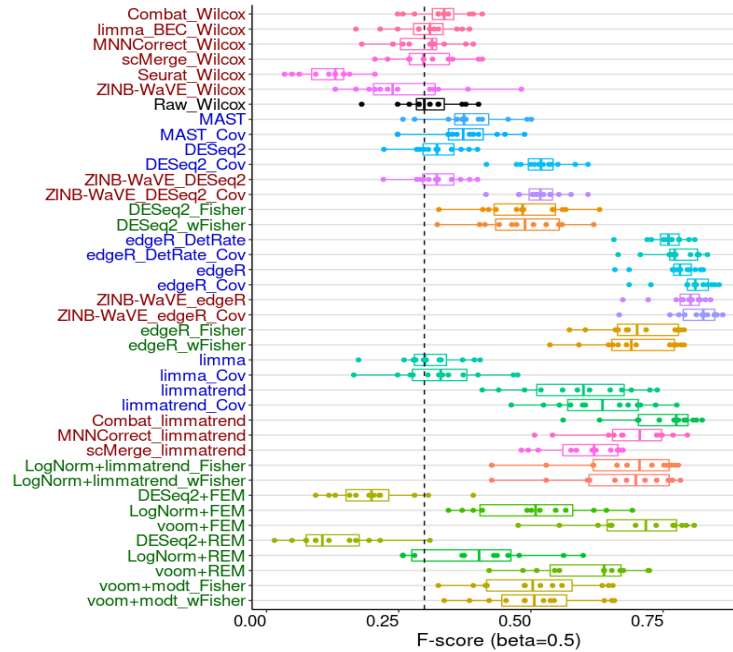
- ○ **'GBC-meta_PR':**

  - ▪ Aggerate all output results and illustrate the AUPR curve