Sequence motifs

汇报人：Lilian

2021/11/12

nature biotechnology

Published: April 2006

What are DNA sequence motifs?

Patrik D'haeseleer

Nature Biotechnology **24**, 423–425 (2006) | Cite this article

Background

- **胃癌以及ceRNAs**

  - ✓ 胃癌(GC)是最常见的癌症之一，在世界范围内死亡率很高。转移是众多复杂过程的最终结果，是临床实践中的主要挑战，也是胃癌致死和复发的主要来源。
  - ✓ 近年，研究开始描述ncRNA可能对GC具有调控作用。这些ncRNA与肿瘤的发生、发展、侵袭、转移以及耐药性密切相关。越来越多的证据揭示了 miRNA 以及LncRNAs在癌基因和抑癌基因的转录后调控中不可或缺的功能..
  - ✓ 有研究表明lncRNA可能作为竞争性内源性RNA（ceRNA)发挥作用，并通过竞争性地结合它们共同的miRNAs与mRNA进行串扰

主要工作

- ✓ 对6对胃癌组织和非肿瘤癌旁组织进行分析，选取异常表达的mRNA、miRNA和lncRNA进行深入分析；

文献来源

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Comprehensive network of miRNA-induced intergenic interactions and a biological role of its core in cancer

Vladimir V. Galatenko[1,2,3], Alexey V. Galatenko[1], Timur R. Samatov[2,7], Andrey A. Turchinovich[4], Maxim Yu. Shkurnikov[5], Julia A. Makarova[5,6] & Alexander G. Tonevitsky[2,5]

Received: 15 June 2017
Accepted: 16 January 2018
Published online: 05 February 2018

MicroRNAs (miRNAs) are a family of short noncoding RNAs that posttranscriptionally regulate

Background

- **miRNA的生物学背景**

In humans, more than 80% of intronic miRNA genes have a sense orientation with respect to their host genes. Therefore, most human intragenic miRNAs are co-transcribed with their host genes and subsequently released at the splicing stage.

——Berezikov, E., E. & Lai, E. C. Mammalian mirtron genes. Mol. Cell. 28, 328–336 (2007).

- ✓ 人类的内含子miRNAs占所有基因内miRNAs13的85%以上
- ✓ 内含子microRNA通常与宿主基因的转录方向相同；
- ✓ 对于人类，超过80%的内含子miRNA基因与其宿主基因之间具有定向关系。

- **宿主基因与内含子型miRNA之间的功能关系类型**

  - ✓ **miRNA靶向的基因，**
  - ✓ **其产物是宿主基因产物的下游效应物，**
  - ✓ **与宿主基因对抗的基因，**
  - ✓ **与宿主基因属于同一途径的基因**

主要工作

"We hypothesized that **regulatory network motifs involving intragenic miRNAs could simultaneously function in a cell as interconnected parts of the entire mechanism of gene expression regulation.** Therefore, in the present study, we did not focus on individual regulatory motifs but, for the first time, **constructed and analyzed the entire network of intergenic interactions induced by intragenic miRNAs**."

1. 构建了以基因为导向的网络，其边缘与miRNA相对应，并代表一个基因被另一个基因中的miRNA靶向；
2. 确定了所构建网络的凝聚核心。核心包含21或12个基因(分别适用于所有基因内mirna或仅含内含子意义的mirna)，参与关键的细胞过程，包括DNA复制、转录、蛋白质稳态和细胞代谢。
3. 鉴定了仅由核心基因组成的基因表达标记，用于乳腺癌和结直肠癌的高效复发预后

Network construction and analysis.

Methods

miRIAD

基因内pre- miRNA及其各自宿主基因的标识符列表

↓

pre- mirna与剪切后的成熟mRNA的对应关系

DIANA-TarBase v7.042 and miRTarBase

成熟mirna的有效靶标列表

↓

成熟mRNA与靶基因的对应关系

多个边和环，以及孤立的节点，
在分析之前被删除

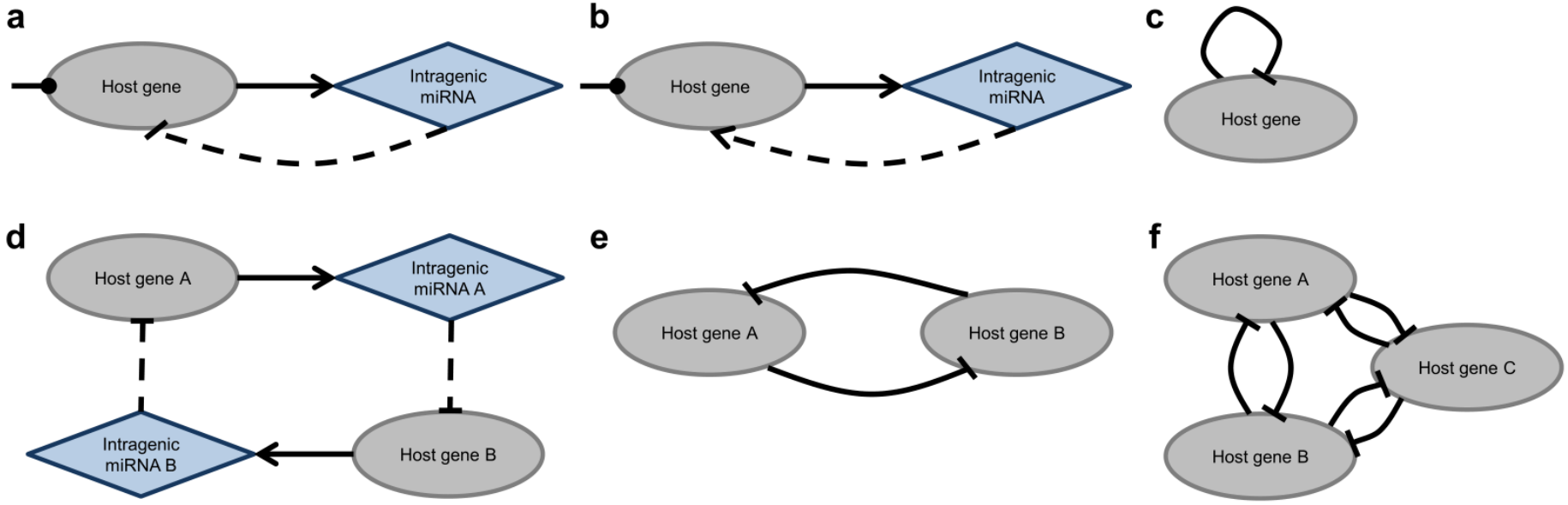
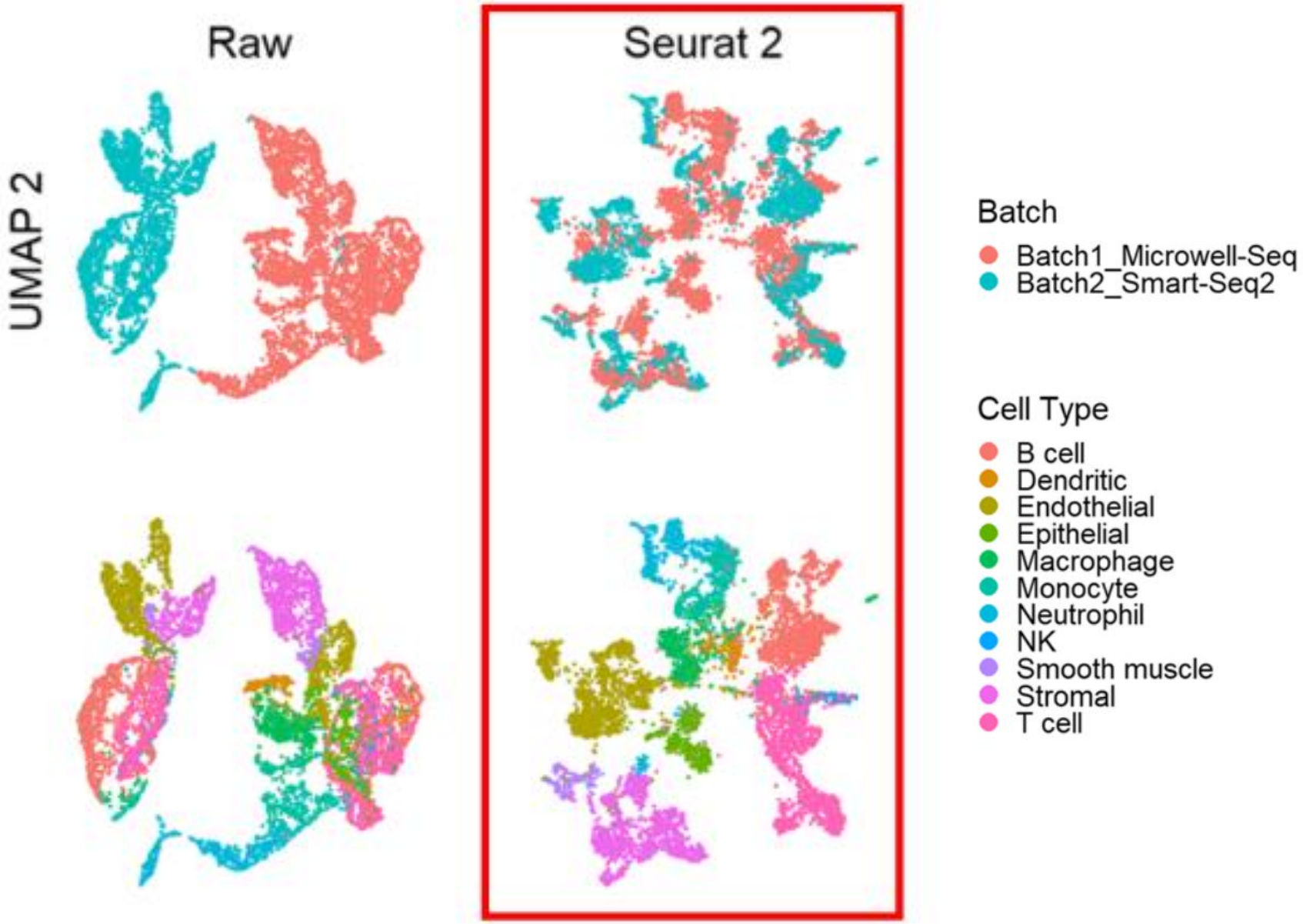
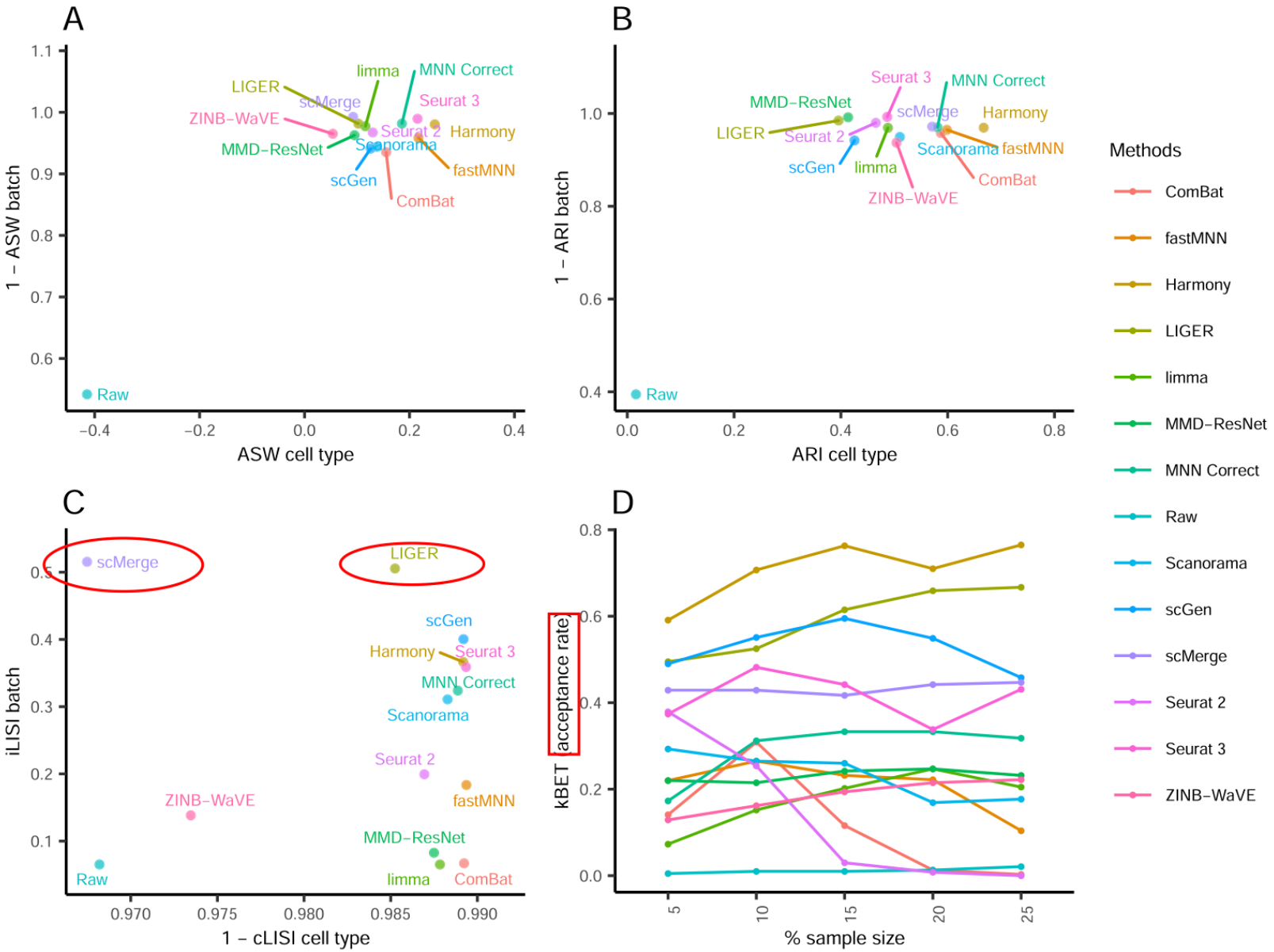


Figure 1. Regulatory network motifs involving intragenic miRNAs. (**a**) A self-regulatory negative feedback loop. (**b**) A self-regulatory positive feedback loop. (**c**) Representation of a self-regulatory feedback loop in the constructed network of intergenic interactions induced by intragenic miRNAs (note that loops are removed prior to the analysis of the network). (**d**) A pair of genes mutually targeting each other via their intragenic miRNAs. (**e**) Representation of miRNA-induced intergenic interactions shown in panel (d) in the constructed network. (**f**) A three-node sub-network in which each pair of nodes is mutually (bidirectionally) connected.

Scenario 1: identical cell types, different technologies

Data collection


Raw
Seurat 2
UMAP 2
Batch
Batch1_Microwell-Seq
Batch2_Smart-Seq2
Cell Type
B cell
Dendritic
Endothelial
Epithelial
Macrophage
Monocyte
Neutrophil
NK
Smooth muscle
Stromal
T cell


A
1 – ASW batch
ASW cell type
Raw
LIGER
limma
MNN Correct
scMerge
Seurat 3
ZINB–WaVE
Seurat 2
Harmony
MMD–ResNet
Scanorama
fastMNN
scGen
ComBat
B
1 – ARI batch
ARI cell type
Raw
Seurat 3
MNN Correct
MMD–ResNet
scMerge
Harmony
LIGER
Seurat 2
Scanorama
fastMNN
scGen
limma
ComBat
ZINB–WaVE
C
iLISI batch
1 – cLISI cell type
scMerge
LIGER
scGen
Harmony
Seurat 3
MNN Correct
Scanorama
Seurat 2
fastMNN
ZINB–WaVE
MMD–ResNet
Raw
limma
ComBat
D
kBET (acceptance rate)
% sample size
Methods
ComBat
fastMNN
Harmony
LIGER
limma
MMD–ResNet
MNN Correct
Raw
Scanorama
scGen
scMerge
Seurat 2
Seurat 3
ZINB–WaVE

Evaluation functions

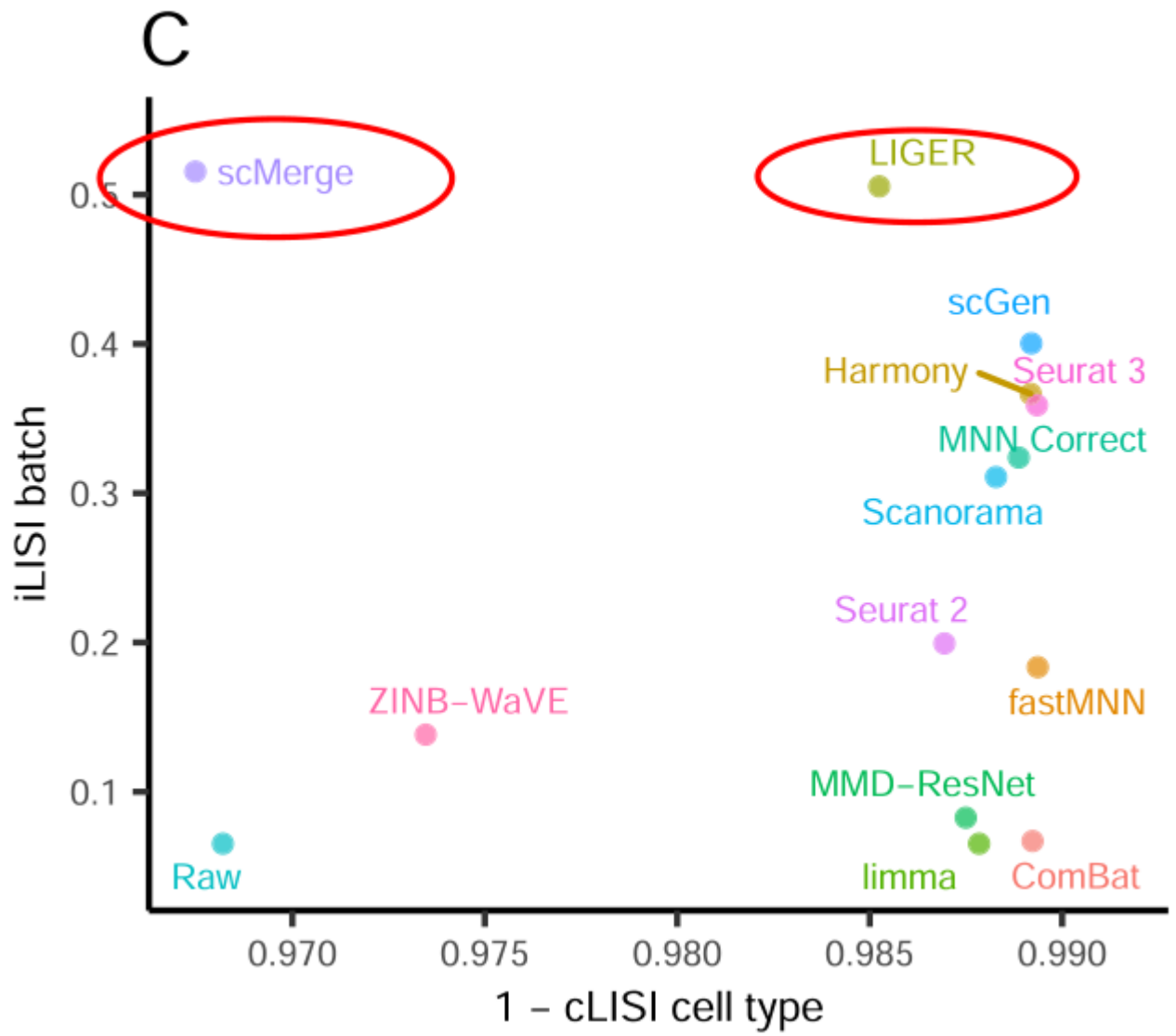
评价批次混合以及细胞分类效果

Local inverse Simpson's index (LISI)

- In the case of LISI integration (iLISI) to measure batch mixing, the index is computed for batch labels, and **a score close to the expected number of batches denotes good mixing.**
- For cell type LISI (cLISI), the index is computed for all cell type labels, and **a score close to 1 denotes that the clusters contain pure cell types**.
- For combined assessment of cell type purity and batch mixing, the harmonic mean of cLISI and iLISI was computed to obtain the F1 score as described by Lin et al.

$$F1_{LISI} = \frac{2(1-cLISI_{norm})(iLISI_{norm})}{1-cLISI_{norm}+iLISI_{norm.}}$$

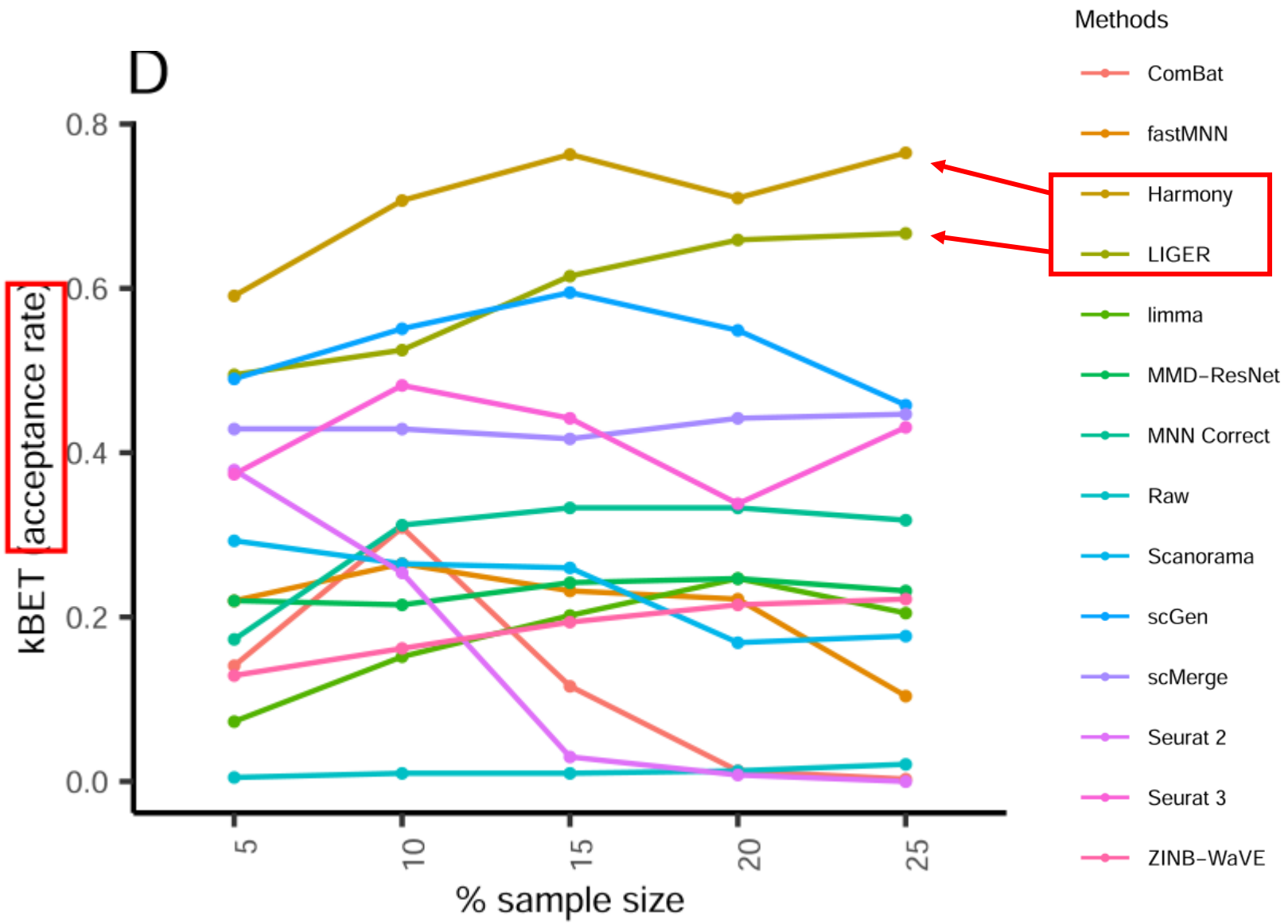
对于综合效果而言， iLISI越高， cLISI越低，F1值越大，效果越好。


C
scMerge
LIGER
scGen
Harmony
Seurat 3
MNN Correct
Scanorama
Seurat 2
fastMNN
ZINB-WaVE
MMD-ResNet
Raw
limma
ComBat
0.5
0.4
0.3
0.2
0.1
iLISI batch
0.970
0.975
0.980
0.985
0.990
1 – cLISI cell type

k-Nearest neighbor batch-effect test (kBET)

The null hypothesis of all batches being well mixed is not rejected if the local distribution is sufficiently similar to the global distribution. The fraction of rejections ranges from 0 to 1. **If the fraction of rejections is close to zero, this signifies that the batches are well mixed.**

通过kBET获得拒绝率，拒绝率越低，混合效果越好。

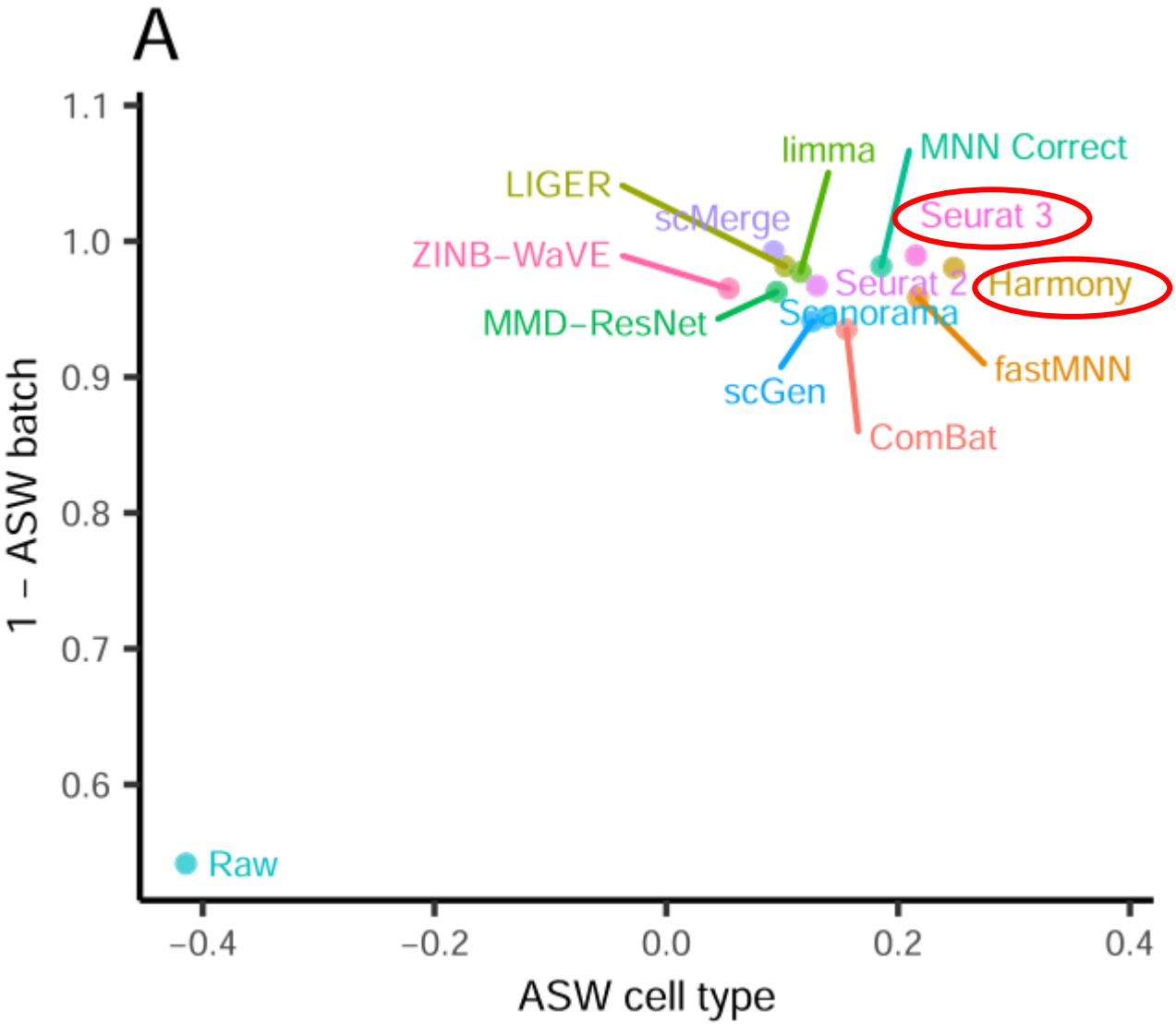

D
kBET (acceptance rate)
0.8
0.6
0.4
0.2
0.0
5
10
15
20
25
% sample size
Methods
ComBat
fastMNN
Harmony
LIGER
limma
MMD-ResNet
MNN Correct
Raw
Scanorama
scGen
scMerge
Seurat 2
Seurat 3
ZINB-WaVE

Average silhouette width (ASW)

- The silhouette score of a data point is computed by subtracting its average distance to other members in the same cluster from its average distance to all members of the neighboring clusters, and then dividing by the larger of the two values.
- For **combined assessment of cell type purity and batch mixing**, we calculated the harmonic mean of batch and cell type ASW scores to obtain the F1 score:

$$F1_{ASW} = \frac{2(1-ASW_{batch\_norm})(ASW_{cell\_type\_norm})}{1-ASW_{batch\_norm} + ASW_{cell\_type\_norm}}$$

对于综合效果而言，批次ASW分数越小，细胞类型ASW分数越高，F1值越大，效果越好。


A
1.1
1.0
0.9
0.8
0.7
0.6
1 – ASW batch
−0.4
−0.2
0.0
0.2
0.4
ASW cell type
Raw
LIGER
limma
MNN Correct
scMerge
Seurat 3
ZINB–WaVE
Seurat 2
Harmony
MMD–ResNet
Scanorama
fastMNN
scGen
ComBat

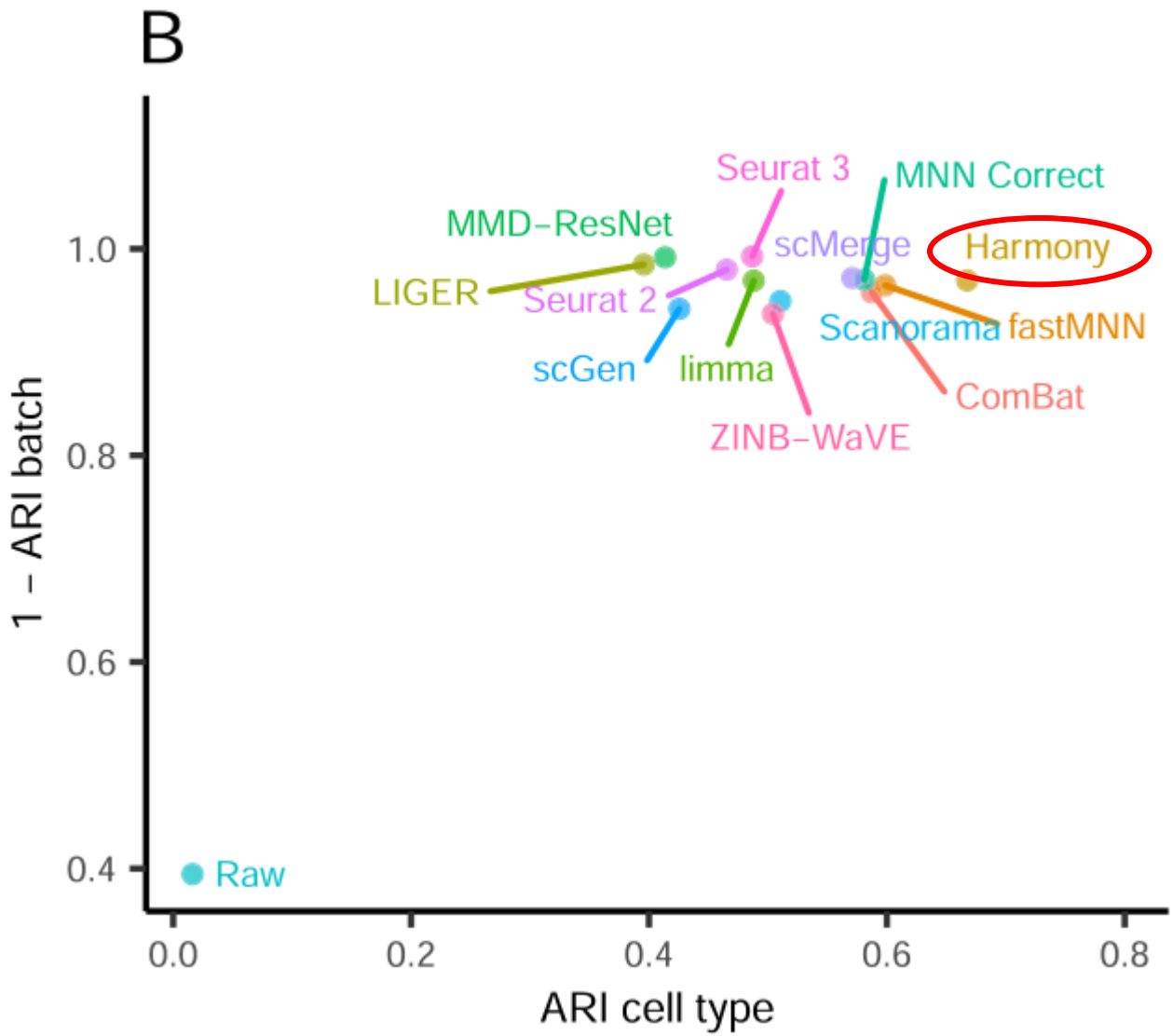
Adjusted rand index (ARI)

- The ARI measures the percentage of matches between two label lists, corrected for chance.

- a combined F1 score was obtained for each batch correction method by computing the harmonic mean of the ARI scores:

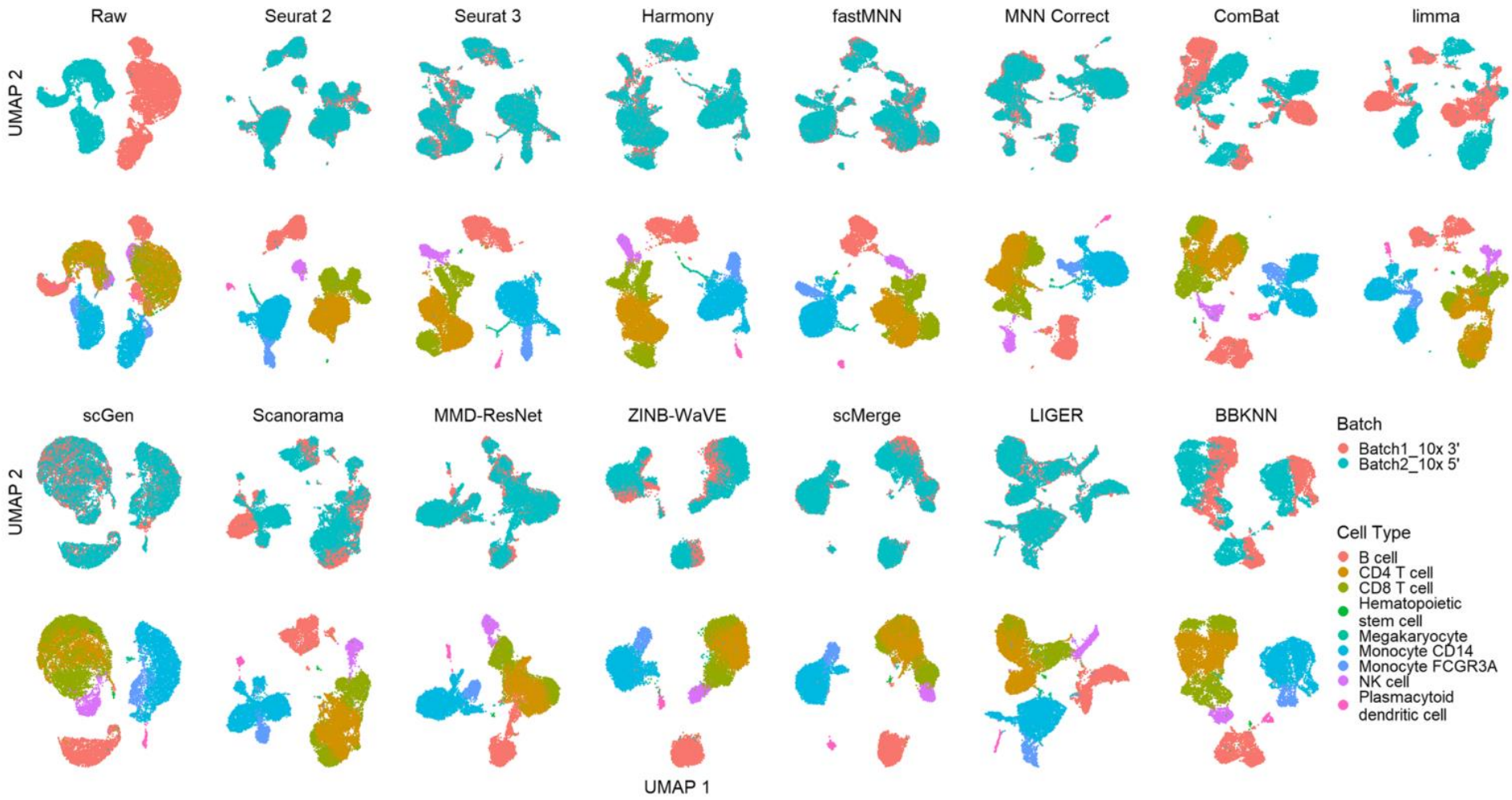
$$F1_{ARI} = \frac{2(1-ARI_{batch\_norm})(ARI_{cell\_type\_norm})}{1-ARI_{batch\_norm} + ARI_{cell\_type\_norm}}$$

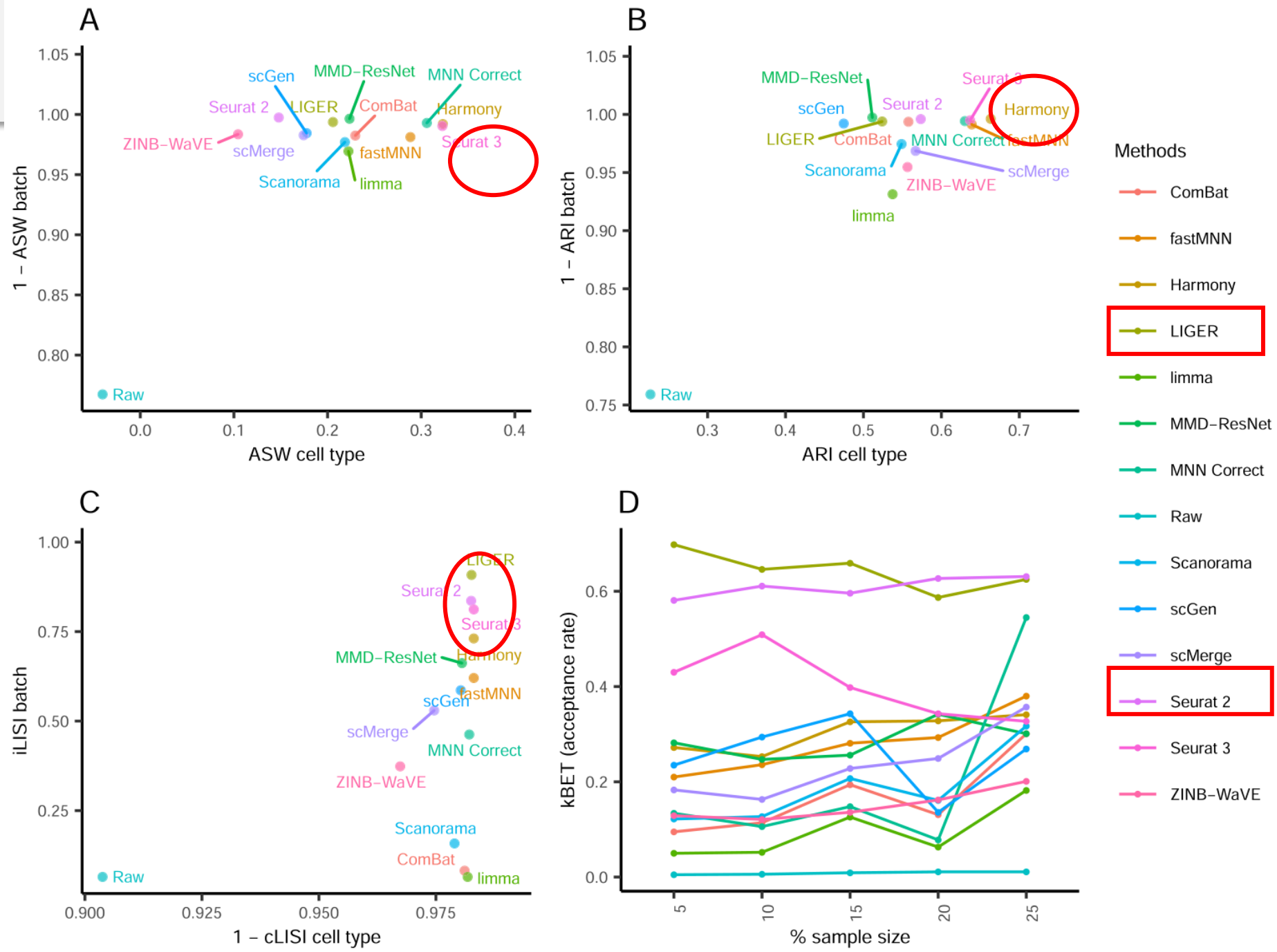
F1分数越高，ARI批次混合分数越低，ARI细胞型混合分数越高。


B
Seurat 3
MNN Correct
MMD-ResNet
scMerge
Harmony
LIGER
Seurat 2
Scanorama
fastMNN
scGen
limma
ComBat
ZINB-WaVE
Raw
1.0
0.8
0.6
0.4
1 – ARI batch
0.0
0.2
0.4
0.6
0.8
ARI cell type

天津大学
Tianjin University

5	Human Peripheral Blood Mononuclear Cell	2	15,476	10x 3' 10x 5'




A
1 – ASW batch
ASW cell type
scGen
MMD–ResNet
MNN Correct
Seurat 2
LIGER
ComBat
Harmony
ZINB–WaVE
scMerge
fastMNN
Seurat 3
Scanorama
limma
Raw
B
1 – ARI batch
ARI cell type
MMD–ResNet
Seurat 2
scGen
Harmony
LIGER
ComBat
MNN Correct
fastMNN
Scanorama
scMerge
ZINB–WaVE
limma
Raw
C
iLISI batch
1 – cLISI cell type
LIGER
Seurat 2
Seurat 3
MMD–ResNet
Harmony
fastMNN
scGen
scMerge
MNN Correct
ZINB–WaVE
Scanorama
ComBat
limma
Raw
D
kBET (acceptance rate)
% sample size
Methods
ComBat
fastMNN
Harmony
LIGER
limma
MMD–ResNet
MNN Correct
Raw
Scanorama
scGen
scMerge
Seurat 2
Seurat 3
ZINB–WaVE

Scenario 1: identical cell types, different technologies

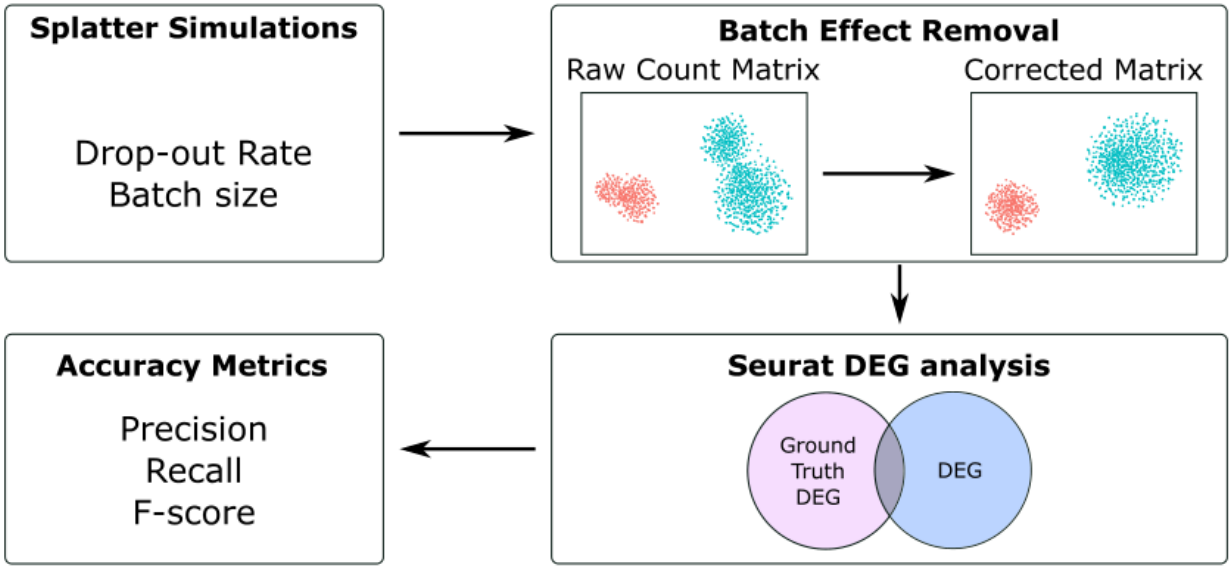
Conclusions

For both datasets, **Harmony** was the top method, and **Seurat 3** ranked second and third once. Based on these results, both methods are highly recommended for datasets with common cell types. Though **LIGER** was only ranked third for dataset 5 and tied at fourth place with fastMNN for dataset 2, it was a consistent performer and thus also a competitive method worth considering.

Scenario 5: simulation

实际单细胞测序过程中，不可避免的存在在测序过程中RNA捕获或扩增失败，失败的比率即“drop-out rate”

A. Simulation workflow



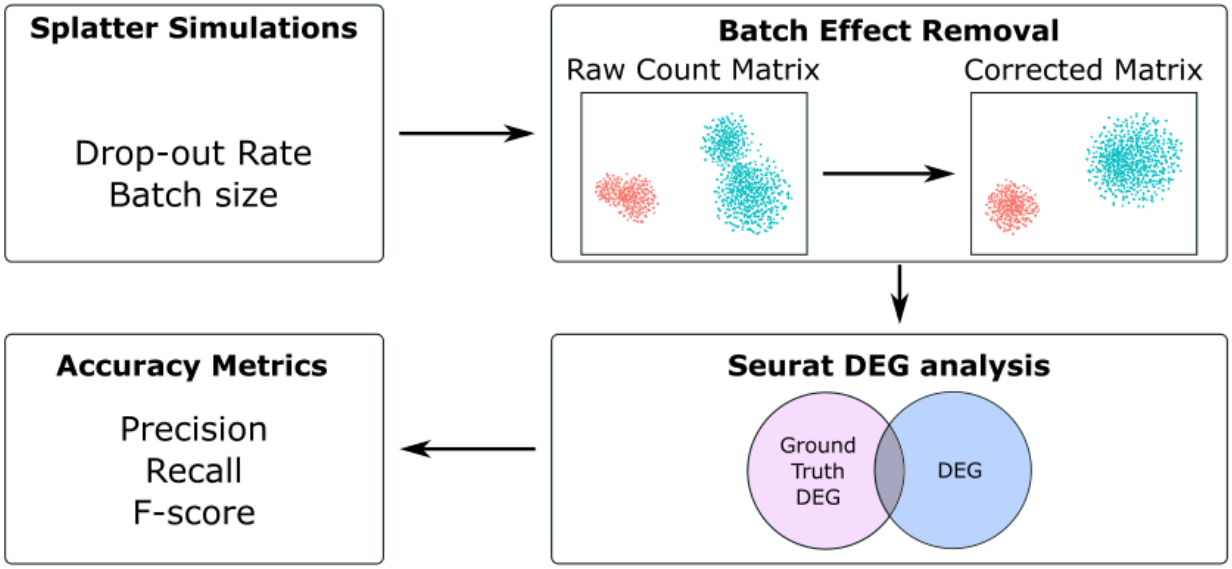
B. Use cases

Simulation	Drop-out	Number of cells		Remarks
		Batch 1	Batch 2	
1	0.05	500	900	unbalanced number of cells in 2 batches, small drop-out
2	0.25	500	900	unbalanced number of cells in 2 batches large drop-out
3	0.05	500	450	balanced number of cells in 2 batches, small drop-out
4	0.25	500	450	balanced number of cells in 2 batches, large drop-out
5	0.05	80	400	small number of cells in batch 1, small drop-out
6	0.25	80	400	small number of cells in batch 1, large drop-out

Scenario 5: simulation

实际单细胞测序过程中，不可避免的存在在测序过程中RNA捕获或扩增失败，失败的比率即“drop-out rate”

A. Simulation workflow



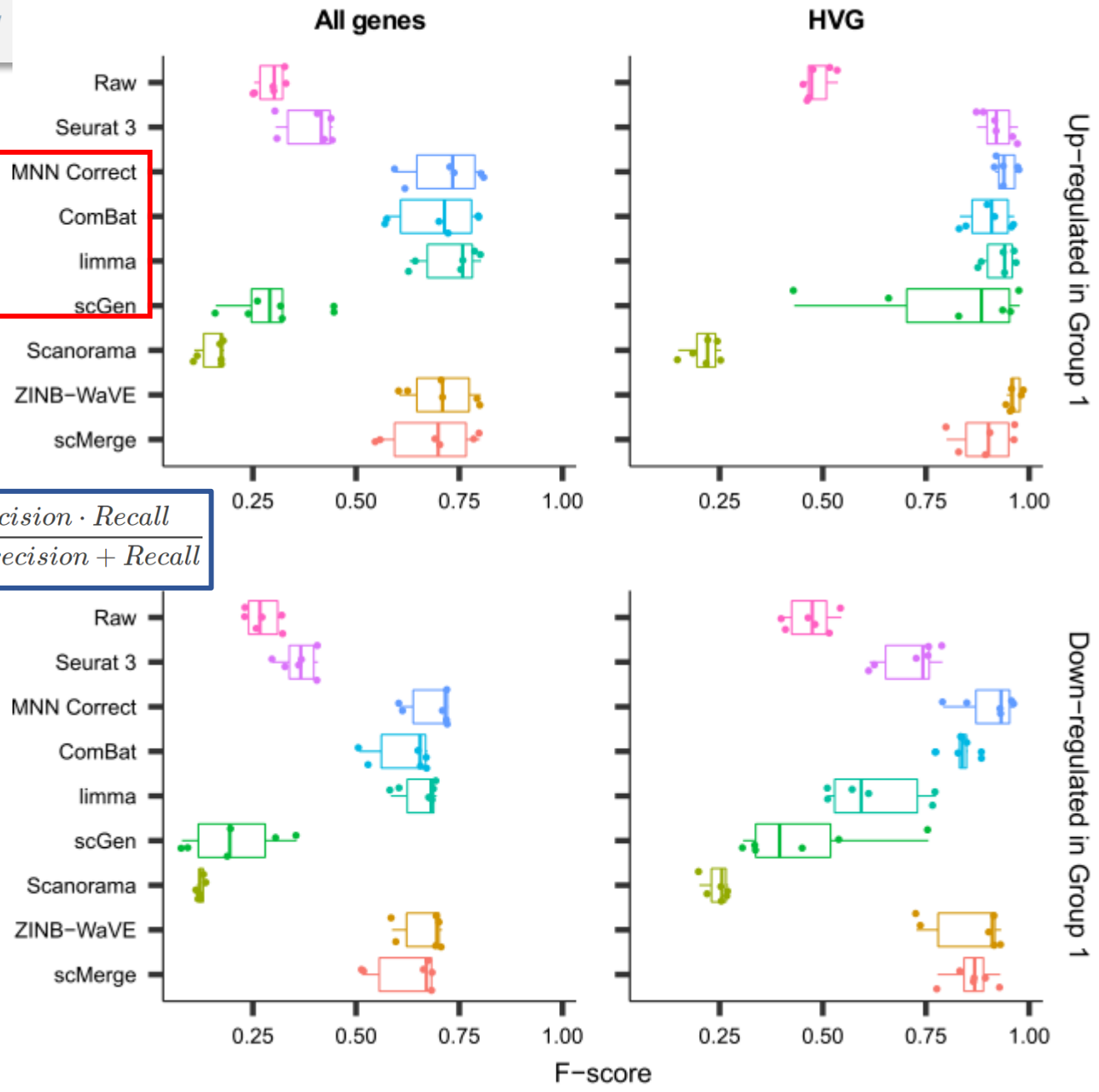
B. Use cases

Simulation	Drop-out	Number of cells		Remarks
		Batch 1	Batch 2	
1	0.05	500	900	unbalanced number of cells in 2 batches, small drop-out
2	0.25	500	900	unbalanced number of cells in 2 batches large drop-out
3	0.05	500	450	balanced number of cells in 2 batches, small drop-out
4	0.25	500	450	balanced number of cells in 2 batches, large drop-out
5	0.05	80	400	small number of cells in batch 1, small drop-out
6	0.25	80	400	small number of cells in batch 1, large drop-out

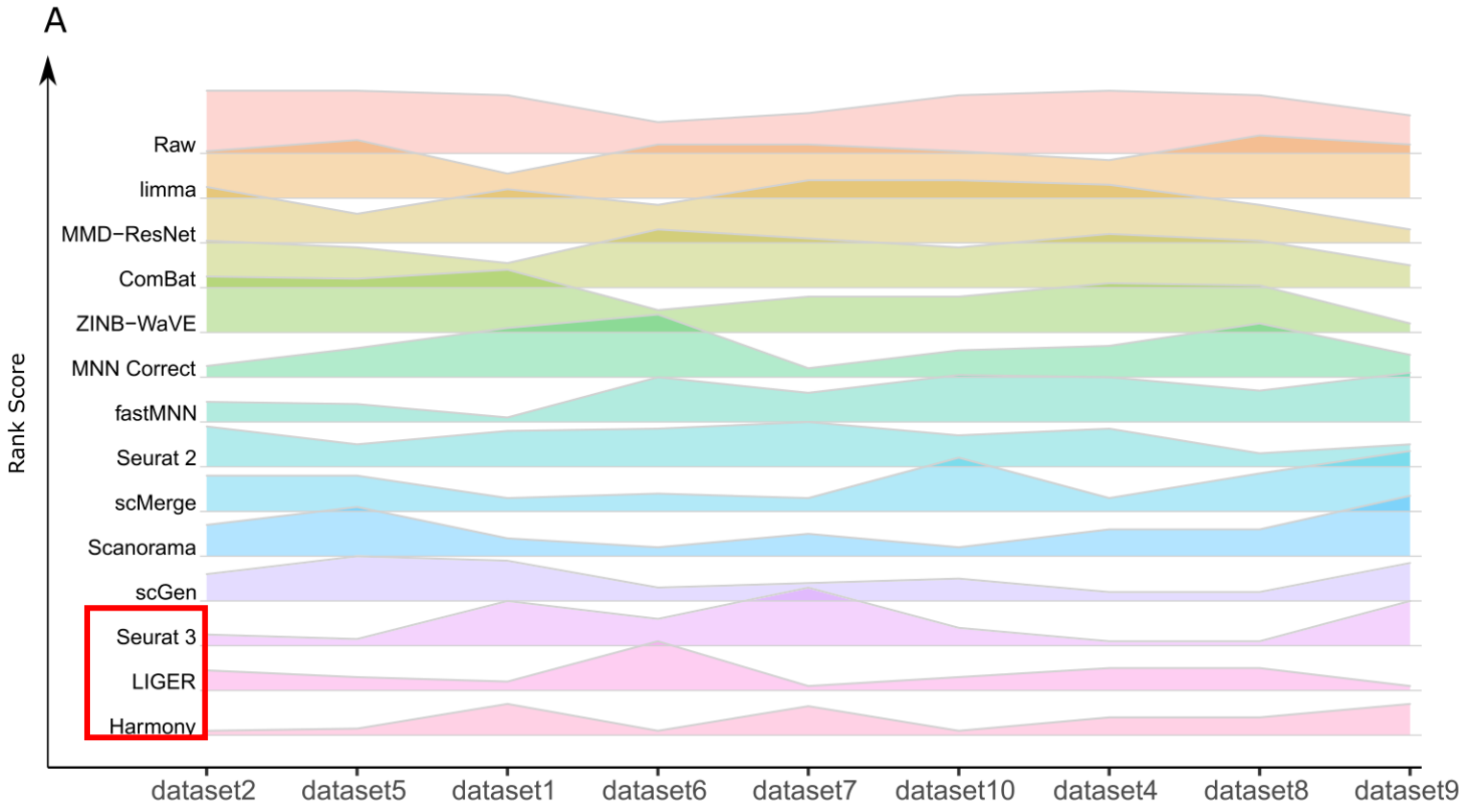
C. Results

Scenario 5

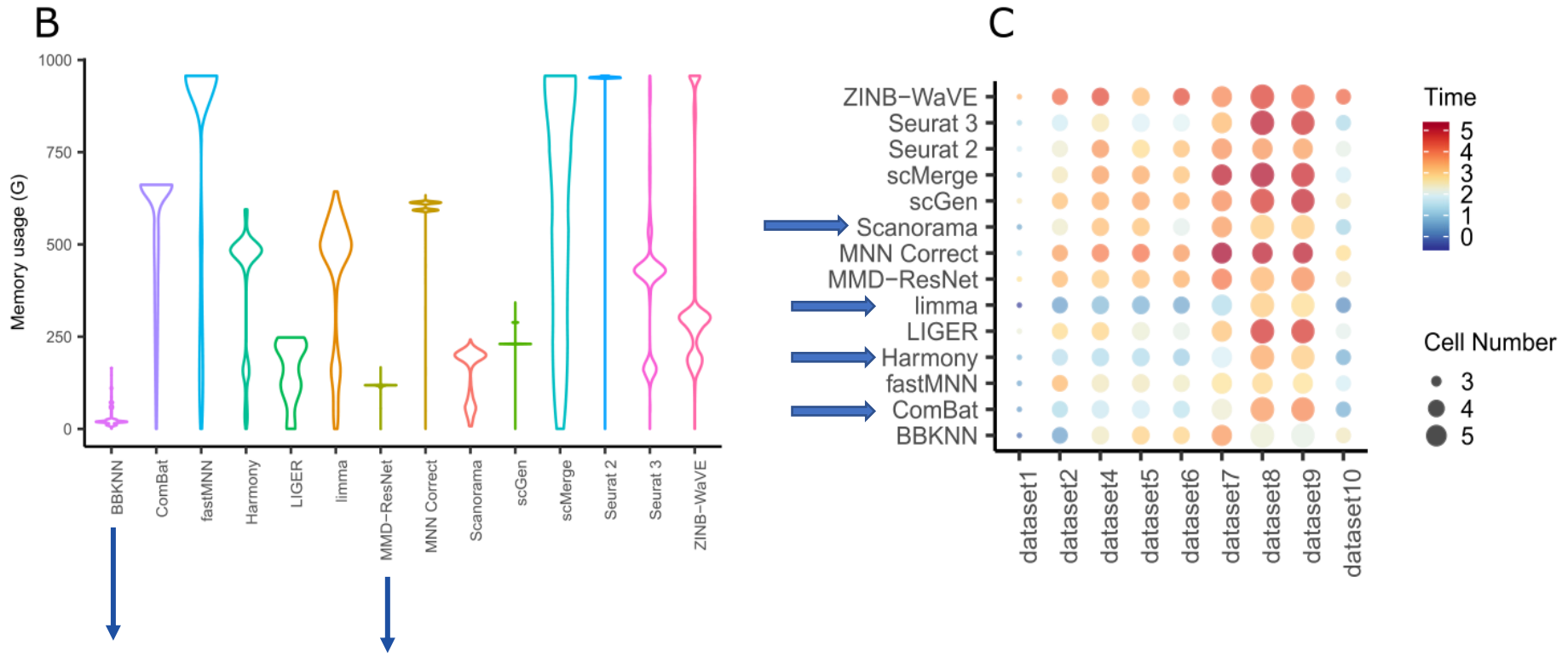
$$F - Score = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$



Conclusions



Conclusions



BBKNN和MMD-ResNet的峰值内存使用量最低(≤170 GB)

Conclusions

- ✓ 通过14中批次矫正的方法和5种情景，作者发现综合考虑，在进行单细胞的批次处理的时候，**LIGER，Harmony以及Seurat3**都能够很好的进行批次处理。
- ✓ **Harmony**无论是对常见的细胞类型以及不同的技术和运行时间上都是非常不错的方法，同样LIGER也是，尤其在未知细胞类型的时候。
- ✓ 对于LIGER来说主要的缺陷是运行时间长
- ✓ Seurat3也能处理大数据集，但是其运行时间比LIGER还要长20%-50%
- ✓ 对于进行下游的DEG分析，作者推荐scMerge进行批次矫正。