scMerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo-replication

Yingxin Lin[1], Shila Ghazanfar[1,2†], Kevin Y. X. Wang[1†], Johann A. Gagnon-Bartsch[3], Kitty K. Lo[1], Xianbin Su[5], Ze-Guang Han[5], John T. Ormerod[1], Terence P. Speed[4], Pengyi Yang[1,2*], Jean Yee Hwa Yang[1,2*]

[1] School of Mathematics and Statistics, University of Sydney, [2] Charles Perkins Centre University of Sydney, [3] Department of Statistics, University of Michigan,
[4] Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research,
[6] Key Laboratory of Systems Biomedicine and Collaborative Innovation Center of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University,

scMerge Framework

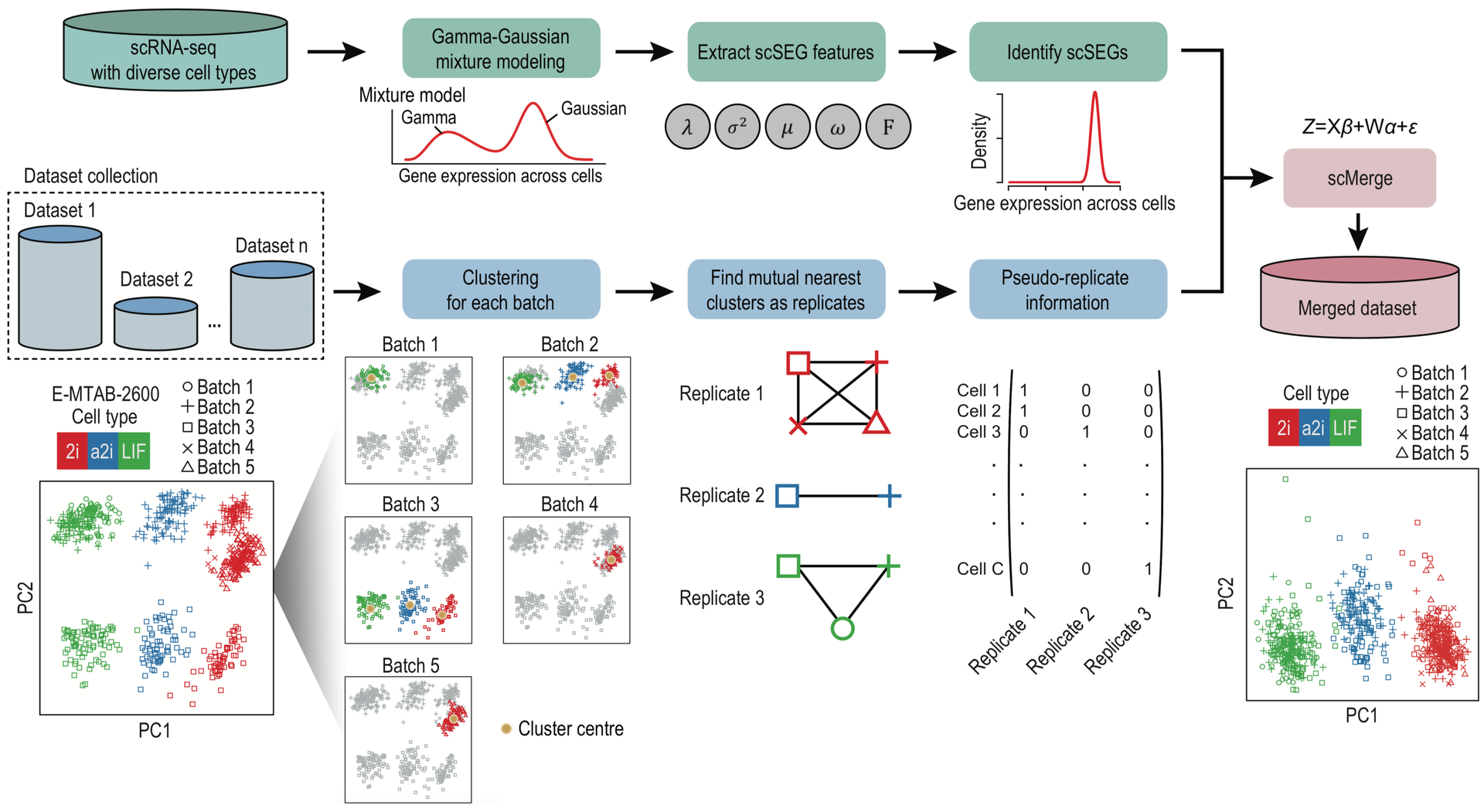
Concerted examination of multiple collections of single cell RNA-seq data promises further biological insights that cannot be uncovered with individual datasets.

The **scMerge** algorithm consists of three key components:

(i) the identification of ***stably expressed genes (scSEGs)*** via a Gamma-Gaussian mixture model (Ghazanfar et al. 2016) for use as 'negative controls' for estimating unwanted factors;

(ii) the construction of ***pseudo-replicates*** to estimate the effects of unwanted factors;

(iii) the adjustment of the datasets with unwanted variation using a ***fastRUVIII*** model with the negative control genes and pseudo-replicates, resulting in a single merged dataset.

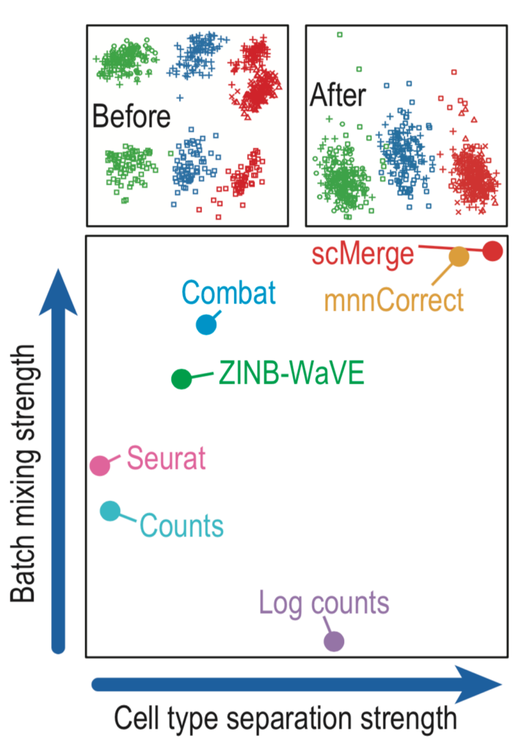


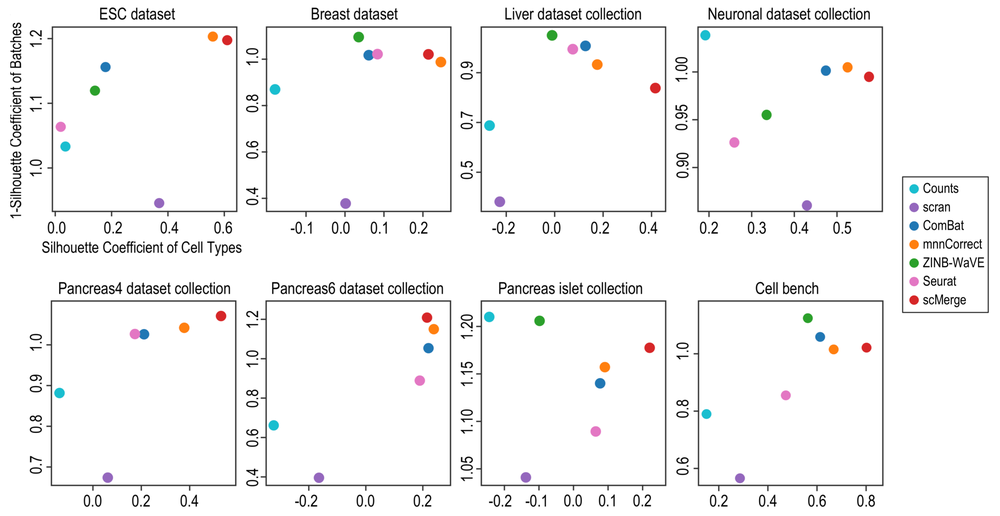
Dataset collections

Name	Accession	Organism	Tissue/organ source	# of cell types	# of cells	Protocol
mESC	E-MTAB-2600	Mouse	mESC	3	704	SMART-seq/C1
Breast	GSE113197	Human	Breast cancer	3	24520	10x Chromium
Liver	GSE87795 GSE90047 GSE87038 GSE96981	Mouse	Liver	8	1249	SMART-seq/C1 SMART-seq2 SMART-seq2 SMART-seq/C1
Neuronal	SRP065920 GSE75413	Mouse	Neuronal	2	145	SMART-seq2 STRT-seq
Pancreas6 / Pancreas4	GSE81608 GSE83139 GSE86469 E-MTAB-5061	Human	Pancreas	6	4566	SMART-seq/C1 SMART-seq/C1 SMART-seq/C1 SMART-seq2
Pancreas6	GSE85241	Human	Pancreas	6	1773	Cel-seq2
Pancreas6	GSE84133	Human/mouse	Pancreas Islets	13	8569	inDrop
CellBench	NA	Human	Adenocarcinoma cell lines	3	1401	Cel-seq2, Drop-seq, & 10x Chromium
Embryogenesis	GSE45719 GSE57249 E-MTAB-3321	Mouse	Embryogenesis	10+	2144	SMART-seq SMARTer SMART-seq2
Embryogenesis	GSE44183 E-MTAB-3929	Human/mouse	Embryogenesis	10+	2144	Tang et al., 2010 SMART-seq2
Embryogenesis	GSE36552 GEO66507	Human	Embryogenesis	10+	2144	Tang et al., 2010 SMARTer

Dataset collections that are used in this study.

Evaluation





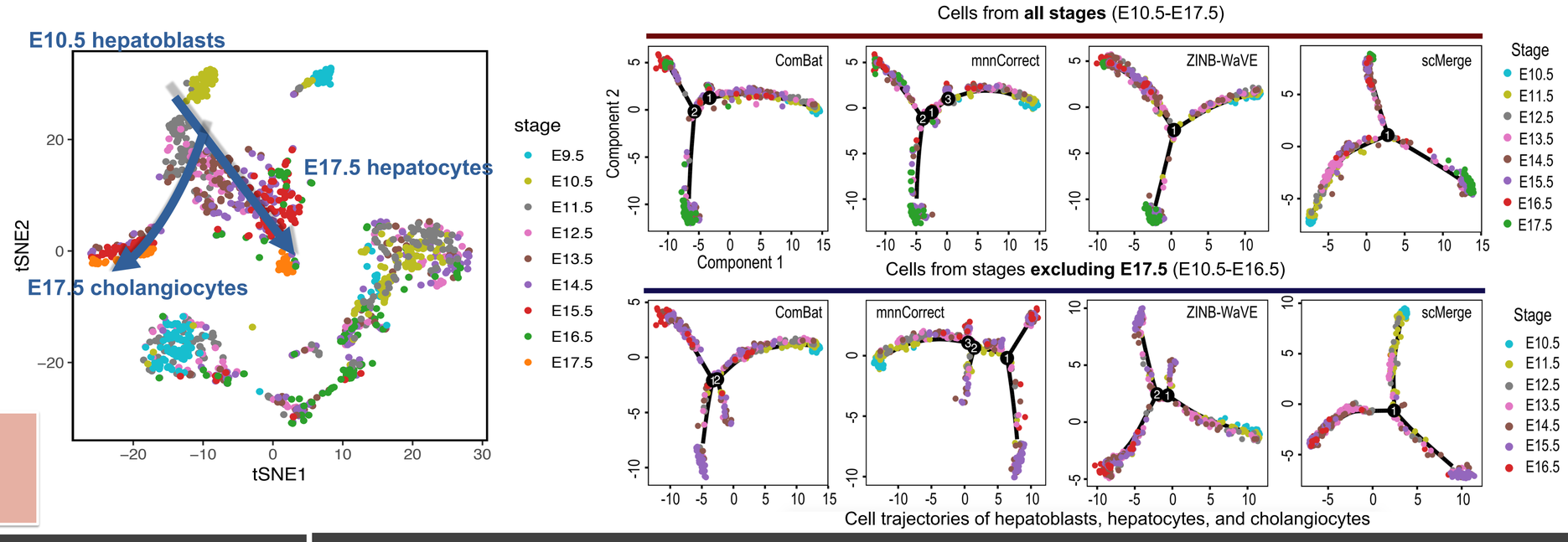
Results: scMerge performed favorably in terms of maintaining strong biological signal and reducing unwanted variation such as batch and/or data-specific noise in seven data collections that cover different tissues, species and protocols.

Case study: Integration of four mouse fetal liver single-cell RNA-seq datasets

Summary: To illustrate the capability of scMerge to enable further downstream analyses, we examined the stability when faced with incomplete data.

Approach: We reconstructed the cell trajectories, using Monocle 2, of hepatoblasts, hepatocytes, and cholangiocytes for both the full Liver data collection and for a subset of the original Liver data collection; with E17.5 time point of GSE90047 removed.

Results: scMerge was most consistent with the full Liver data collection and agrees with current literature.



Acknowledgement

The authors thank all their colleagues, particularly at The University of Sydney, School of Mathematics and Statistics, for their support and intellectual engagement. The following sources of funding for each author, and for the manuscript preparation, are gratefully acknowledged: Australian Research Council Discovery Project grant (DP170100654) to JYHY, JTO, PY; Discovery Early Career Researcher Award (DE170100759) to PY; Australia NHMRC Career Developmental Fellowship (APP1111338) to JYHY and the Judith and David Coffey Life Lab at the Charles Perkins Centre, The University of Sydney to SG. Australian Postgraduate Award to KW. Research Training Program Tuition Fee Offset and Stipend Scholarship to YL. NHMRC Program Grant (1054618) to TPS. J.G. was supported by the National Science Foundation under grant no. DMS-1646108. SJTU-USYD Translate Medicine Fund-Systems Biomedicine AF6260003.

Further information

The scMerge R package and more case studies are available at https://sydneybiox.github.io/scMerge

Reference:

Ghazanfar et al. BMC Systems Biology 2016, 10(Suppl 5):127 doi: 10.1186/s12918-016-0370-4
Lin et al. BioRxiv 2018, doi: https://doi.org/10.1101/393280

Contact: yingxin.lin@sydney.edu.au / jean.yang@sydney.edu.au