RIA: a novel Regression-based Imputation Approach for single-cell RNA sequencing

The 11th IEEE International Conference on Knowledge and Systems Engineering (KSE 2019)
Da Nang, Vietnam

Authors

Bang Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA

Duc Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA

Hung Nguyen
Computer Science & Engineering
University of Nevada, Reno
Reno, USA

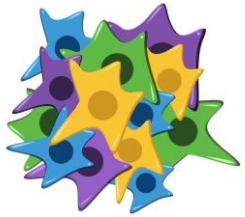
Tin Nguyen
Computer Science & Engineering
University of Nevada, Reno
Reno, USA

Nam Sy Vo
Computational Biomedicine
Vingroup Big Data Institute

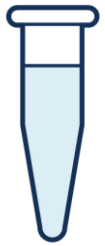
https://bioinformatics.cse.unr.edu

Background

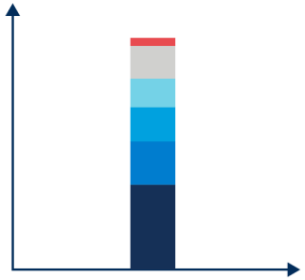
Single – Cell RNA Sequencing



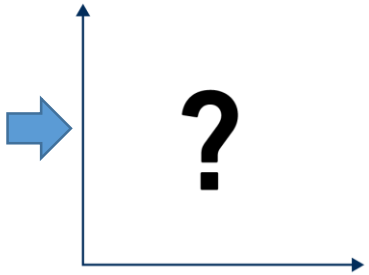
Tissue



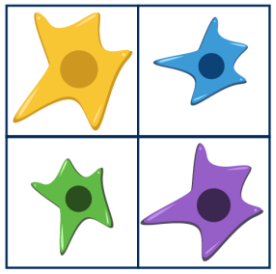
Bulk-Cell analysis



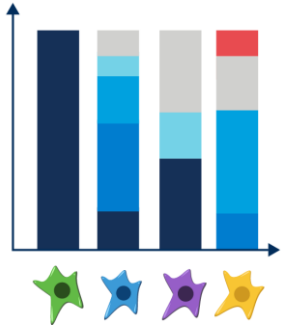
Average expression from all cells



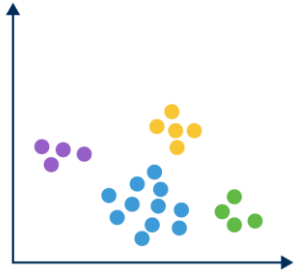
No cellular heterogeneity revealed



Single-Cell Analysis

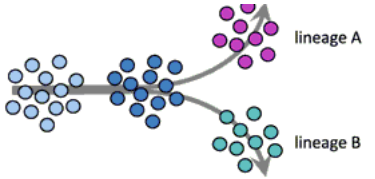


Each cell-type has a unique expression profile

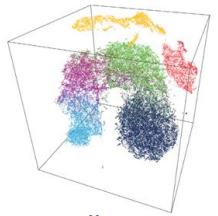


Reveal heterogeneity and subpopulation expression variability

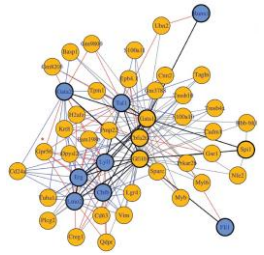
Application



Cell lineage analysis



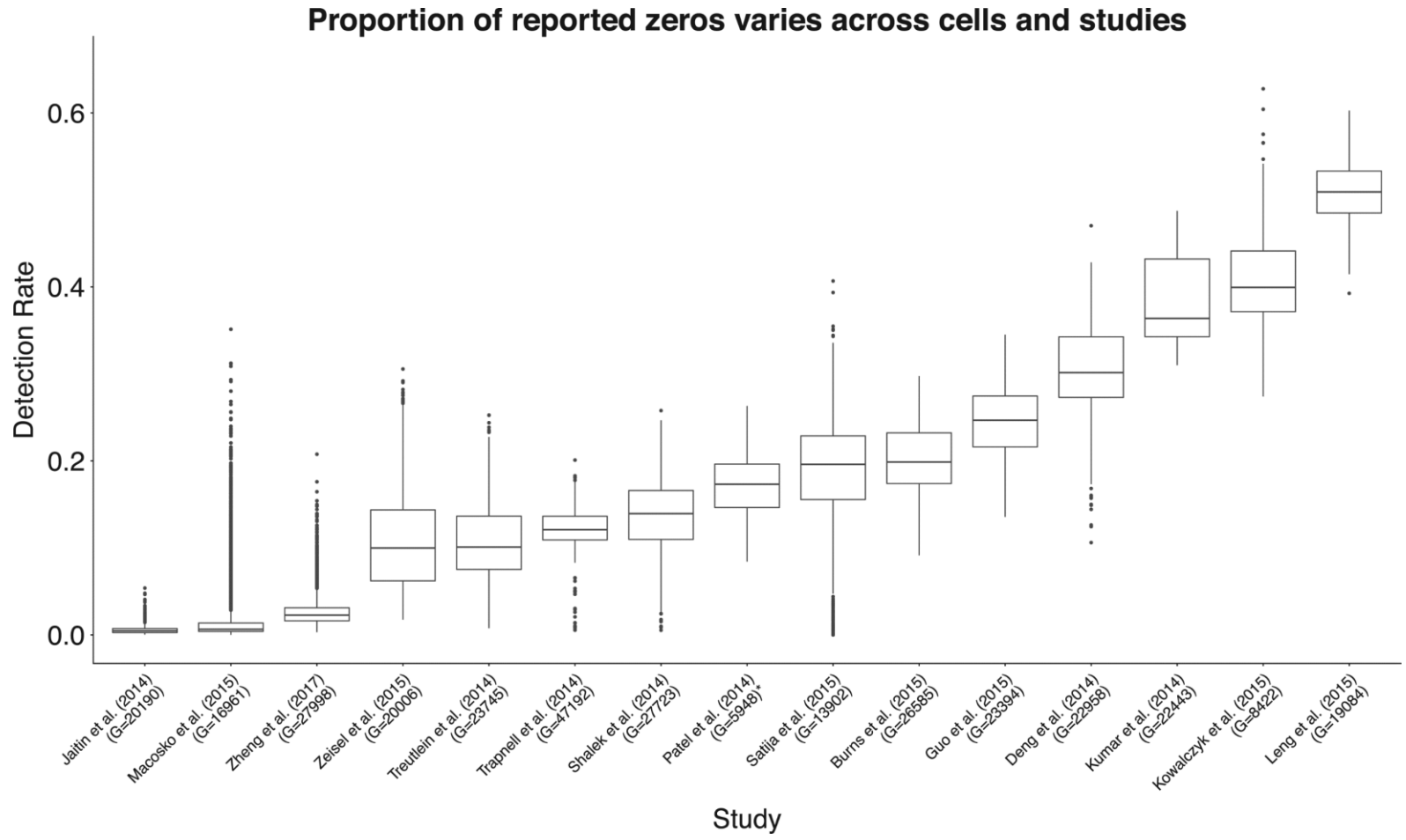
Cell type analysis



scRNAseq network inference

Source: 10xgenomics.com

Dropout Events

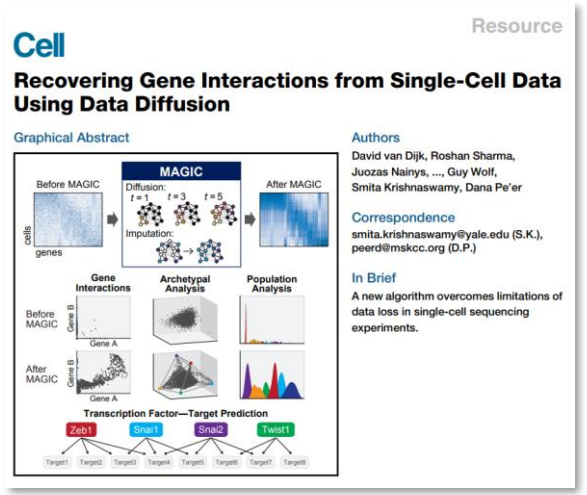


Dropout introduces zero expression value of genes in datasets.

Single – Cell RNAseq Imputation State of The Art

MAGIC [van Dijk et al., 2017]:

- **Citation:** 123
- **Lab:** Krishnaswamy Lab - Yale
- Using Markov Affinity-based Graph Imputation


Resource

Cell

Recovering Gene Interactions from Single-Cell Data Using Data Diffusion

Graphical Abstract

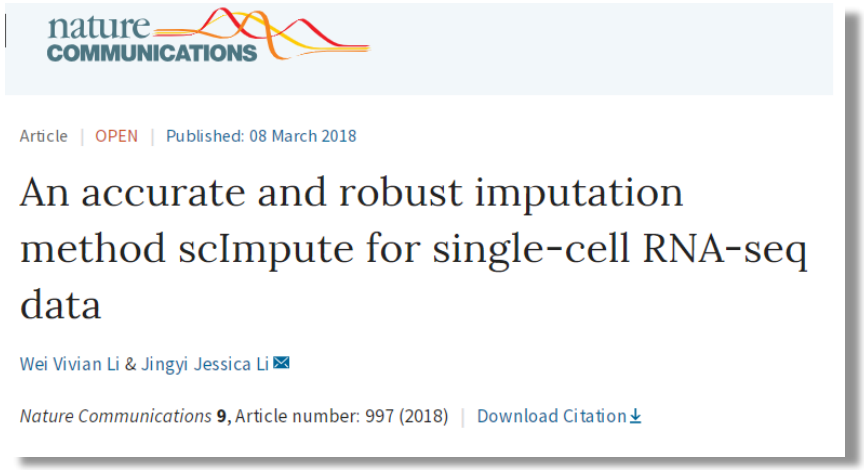
Authors
David van Dijk, Roshan Sharma, Juozas Nainys, ..., Guy Wolf, Smita Krishnaswamy, Dana Pe'er

Correspondence
smita.krishnaswamy@yale.edu (S.K.), peerd@mskcc.org (D.P.)

In Brief
A new algorithm overcomes limitations of data loss in single-cell sequencing experiments.

scImpute [Kwak et al., 2017]:

- **Citation:** 104
- **Lab: UCLA**
- Using statistical approach.


nature COMMUNICATIONS

Article | OPEN | Published: 08 March 2018

An accurate and robust imputation method scImpute for single-cell RNA-seq data

Wei Vivian Li & Jingyi Jessica Li

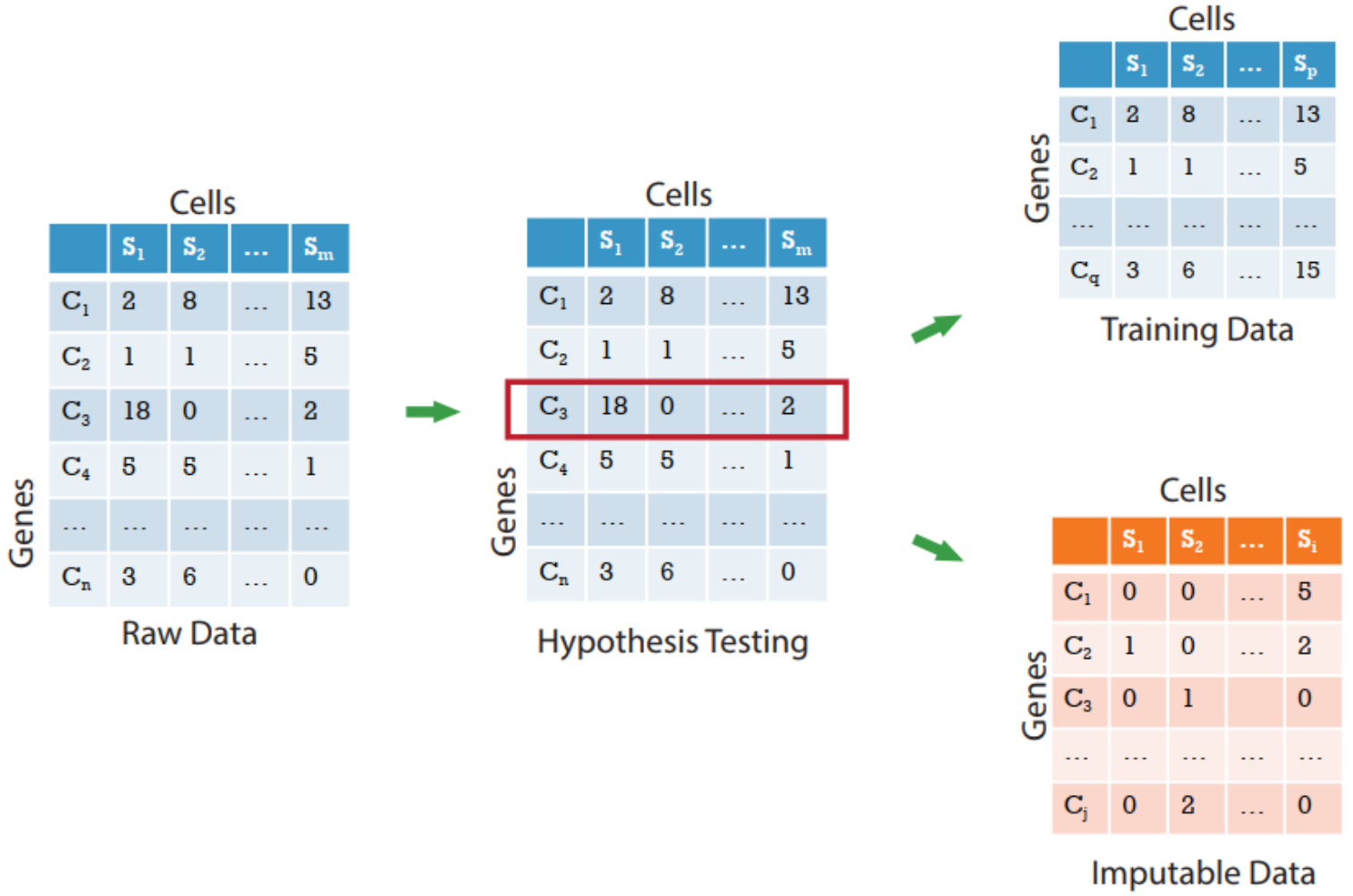
Nature Communications **9**, Article number: 997 (2018) | Download Citation

Method: RIA a novel Regression-based Imputation Approach for single-cell RNA sequencing

RIA

- Identifying genes that are impacted by dropouts
- Impute using genes with high confidence

Hypothesis Testing and Dropout Identification

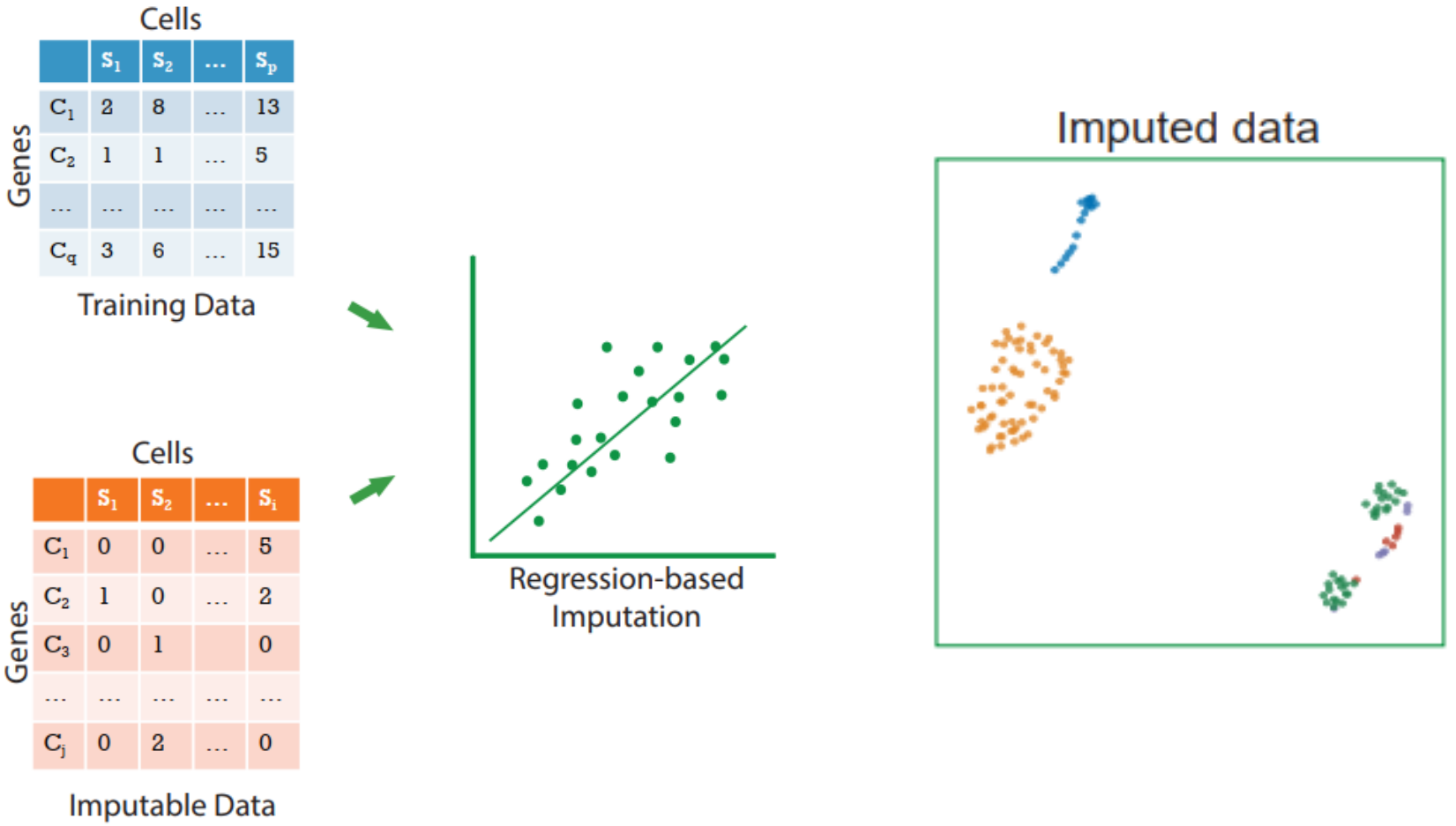


	S_1	S_2	...	S_m
C_1	2	8	...	13
C_2	1	1	...	5
C_3	18	0	...	2
C_4	5	5	...	1
...	...	...	...	...
C_n	3	6	...	0

	S_1	S_2	...	S_p
C_1	2	8	...	13
C_2	1	1	...	5
...	...	...	...	...
C_q	3	6	...	15

	S_1	S_2	...	S_i
C_1	0	0	...	5
C_2	1	0	...	2
C_3	0	1		0
...	...	...	...	...
C_j	0	2	...	0

Generalized Linear Regression Model



RIA Overview

Hypothesis testing approach

Generalized linear model

Results

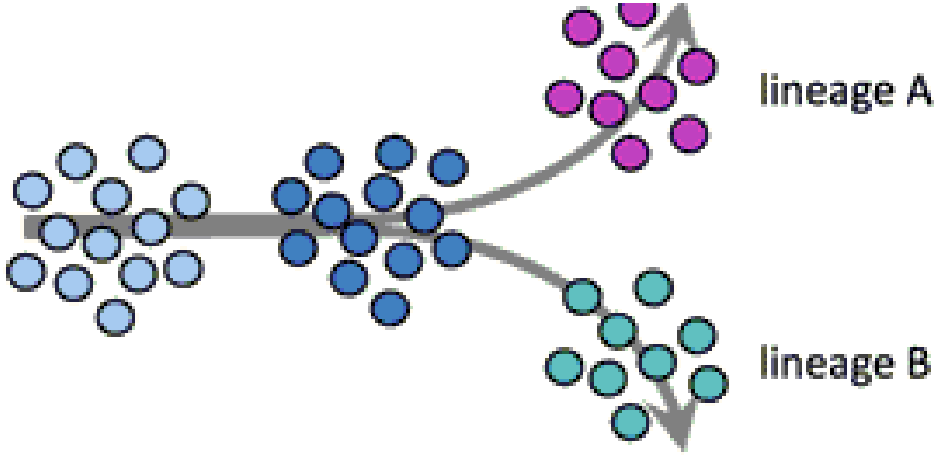
Data an Evaluation Metrics

- MAGIC
- scImpute

TABLE I
SINGLE-CELL DATA OBTAINED FROM NIH GEO

Dataset	Accession ID	Size	K	Organism	Protocol
Biase[27]	GSE57249	49	4	Mouse Embryo	SMARTer
Yan[28]	GSE36552	90	6	Human Embryo	Tang
Goolam[29]	E-MTAB-3321	124	5	Mouse Embryo	Smart-Seq2
Deng[30]	GSE45719	268	6	Mouse Embryo	Smart-Seq2
Zeisel[31]	GSE60361	3,005	9	Mouse Brain	STRT-Seq

Cell type clustering



Time-trajectory inference

RIA Improves the Cells Clustering Results

TABLE II
COMPARISONS USING ADJUSTED RAND INDEX (ARI).

Dataset	Adjusted Rand Index			
	Raw	RIA	scImpute	MAGIC
Biase	0.558	**0.711**	-0.009	0.154
Yan	0.558	**0.573**	0.507	0.029
Goolam	0.501	**0.914**	0.321	0.197
Deng	0.549	**0.815**	0.229	0.483
Zeisel	0.738	**0.768**	0.689	0.289

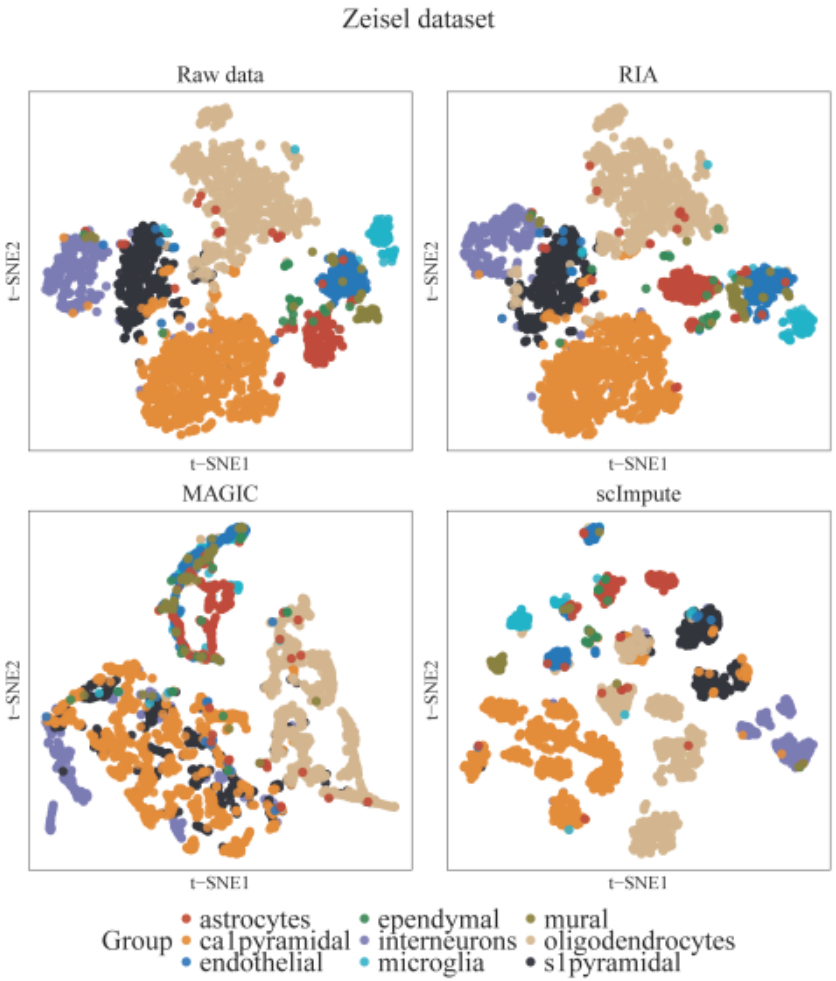
TABLE III
COMPARISON USING JACCARD INDEX

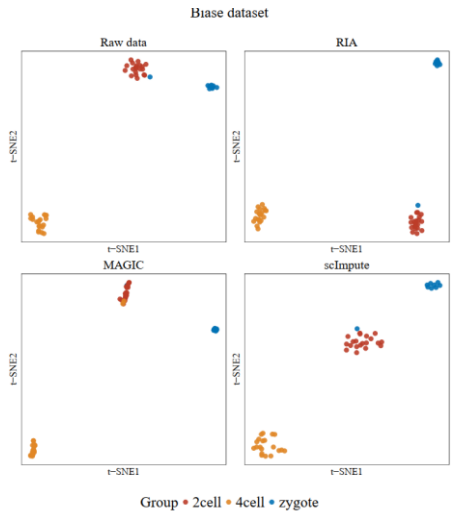
Dataset	Jaccard Index			
	Raw	RIA	scImpute	MAGIC
Biase	0.589	**0.708**	0.339	0.289
Yan	0.498	**0.498**	0.473	0.146
Goolam	0.496	**0.892**	0.375	0.312
Deng	0.524	**0.781**	0.395	0.518
Zeisel	0.651	**0.683**	0.605	0.285

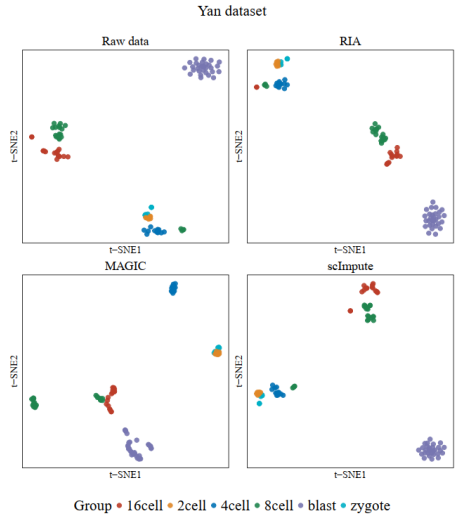
TABLE IV
COMPARISON USING PURITY INDEX

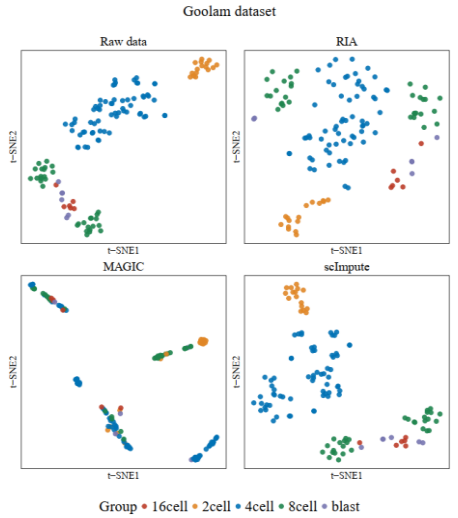
Dataset	Purity Index			
	Raw	RIA	scImpute	MAGIC
Biase	0.795	**0.836**	0.449	0.612
Yan	0.711	**0.778**	0.733	0.467
Goolam	0.822	**0.952**	0.693	0.621
Deng	0.805	**0.839**	0.627	0.750
Zeisel	0.876	**0.893**	0.840	0.668

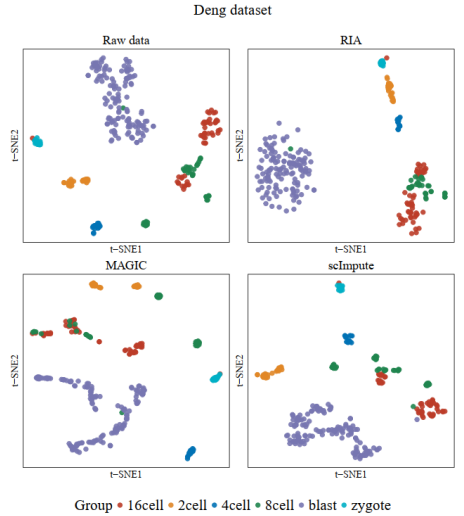
RIA Preserves Original Transcriptome Landscape



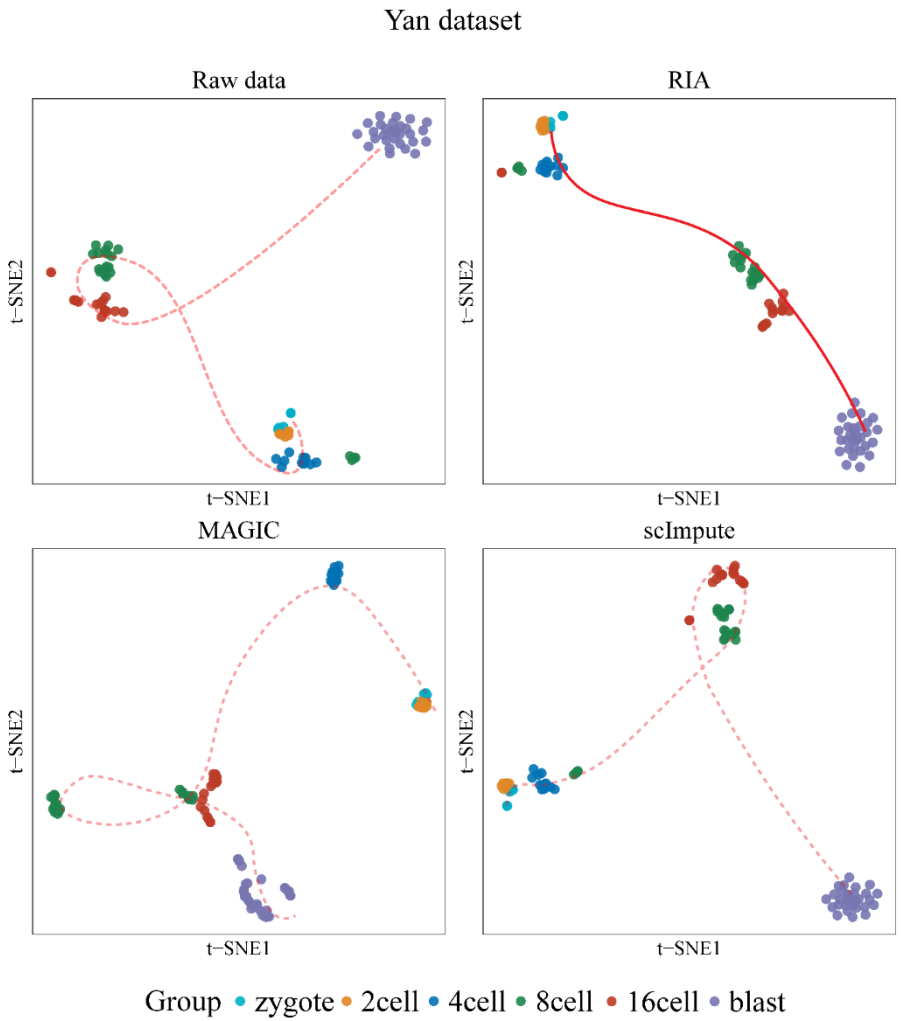


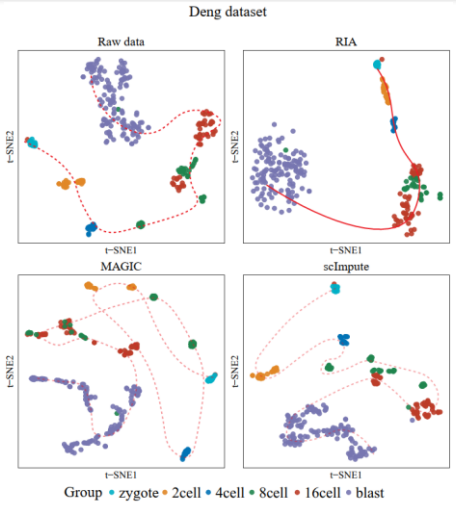


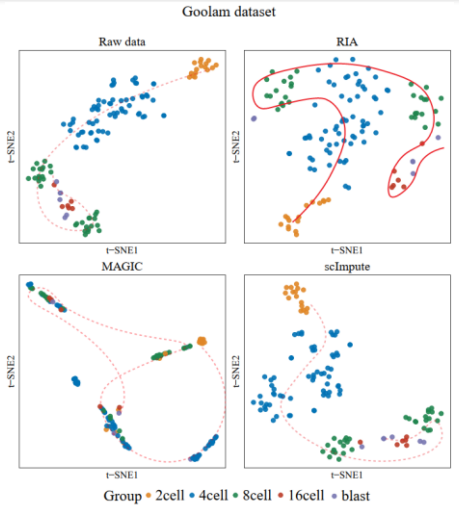


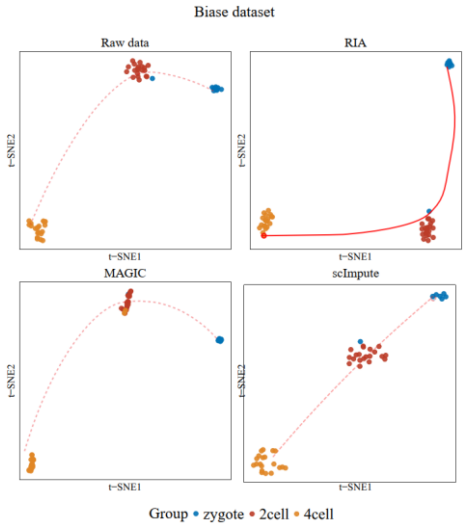


RIA Improves Time Trajectory Inference for TSCAN









Conclusion

- RIA presents a new method to recover missing values caused by dropout events in scRNA-seq data.
- RIA uses a statistical hypothesis testing to identify the set of genes that are likely to be affected by dropouts.
- RIA imputes missing values by using highly correlated genes that share similar biological characteristics.
- RIA dramatically outperforms existing state-of-the-art approaches in improving the identification of cell populations.
- RIA is able to recover temporal trajectories in embryonic development stages.
- RIA is a fast method that can impute thousands of cells with tens of thousands of genes in minutes.

Acknowledgements

People in Bioinformatics Lab at University of Nevada Reno

This work was partially supported by the National Aeronautics and Space Administration (NASA) under Grant Number 80NSSC19M0170

References

[1]. Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, *11*(7), 740.

Q & A

Thank you!