SIMLR: a tool for large-scale genomic analyses by multi-kernel learning

Bo Wang[*1], Daniele Ramazzotti[*1,2], Luca De Sano[3], Junjie Zhu[4], Emma Pierson[1], and Serafim Batzoglou[1]

[1]Department of Computer Science, Stanford University, Stanford, CA, USA
[2]Department of Pathology, Stanford University, Stanford, CA, USA
[3]Department of Informatics, University of Milano-Bicocca, Milan, Italy
[4]Department of Electrical Engineering, Stanford University, Stanford, CA, USA

Abstract

Motivation: We here present SIMLR (Single-cell Interpretation via Multi-kernel LeaRning), an open-source tool that implements a novel framework to learn a sample-to-sample similarity measure from expression data observed for heterogenous samples. SIMLR can be effectively used to perform tasks such as dimension reduction, clustering, and visualization of heterogeneous populations of samples. SIMLR was benchmarked against state-of-the-art methods for these three tasks on several public datasets, showing it to be scalable and capable of greatly improving clustering performance, as well as providing valuable insights by making the data more interpretable via better a visualization.
Availability and Implementation: SIMLR is available on GitHub in both R and MATLAB implementations. Furthermore, it is also available as an R package on bioconductor.org.

The recent development of high resolution single-cell RNA-seq (scRNA-seq) technologies increases the availability of high throughput gene expression measurements of individual cells. This allows us to dissect previously unknown heterogeneity and functional diversity among cell populations [1]. In this line of work recent efforts (see [2, 3, 4]) have demonstrated that *de novo* cell type discovery of functionally distinct cell sub-populations is possible via unbiased analysis of all transcriptomic information provided by

[*] Equal contributors. Correspondance to bowang87@stanford.edu or daniele.ramazzotti@stanford.edu.

scRNA-seq data. However, such analysis heavily relies on the accurate assessment of pairwise cell-to-cell similarities, which poses unique challenges such as outlier cell populations, transcript amplification noise, and dropout events (*i.e.,* zero expression measurements due to sampling or stochastic transcriptional activities) [5].

Recently, new single-cell platforms such as DropSeq [6] and GemCode single-cell technology [7] have enabled a dramatic increase in throughput to hundreds of thousands of cells. While such technological advances may add additional power for *de novo* discovery of cell populations, they also increase computational burdens for traditional unsupervised learning methods.

To address all of the aforementioned challenges, SIMLR was originally proposed in [8] as a novel framework capable of learning an appropriate cell-to-cell similarity metric from the input single-cell data. However, although originally proposed for the analysis of single-cell data, SIMLR can be effectively adopted in the broader task of studying biological data describing heterogeneous populations including but not limited to single-cell analysis (see below and the Supplemenatary Materials). The learned similarities in fact can be exploited for multible tasks, such as effective dimension reduction, clustering, and visualization. SIMLR provides a more scalable analytical framework, which works on hundreds of thousands of samples without any loss of accuracy in dissecting heterogeneity.

SIMLR is available in both R and MATLAB implementations. The framework is capable of learning similarities among gene expression data within an heterogeneous populations of samples, which have been shown to capture different representations of the data. To this end, the approach combines multiple Gaussian kernels in an optimization framework, which can be efficiently solved by a simple iterative procedure. Moreover, SIMLR addresses the challenge of high levels of noise and dropout events by employing a rank constraint and graph diffusion in the learned similarity [9]. See Figure 1 for an overview of the framework.

In the tool are provided both a standard implementation and a large-scale extension of SIMLR together with two examples to test the methods on the datasets by [10] for the standard SIMLR and [11] for the large-scale extension (see Supplementary Material for details). SIMLR can accurately analyze both datasets within minutes on a single core laptop.

Moreover, we also report in the Supplementary Materials a complementary example of usage of SIMLR to study heterogeneity in a cancer dataset. Specifically, we consider the data from [12] and we report how our framework can also be applied in this context, first by estimating the number of populations as discussed in [9] and then by learning a patient-to-patient similarity which may allow, e.g., to effectively stratify the tumors.

An other advantage of SIMLR is that the learned similarities can be

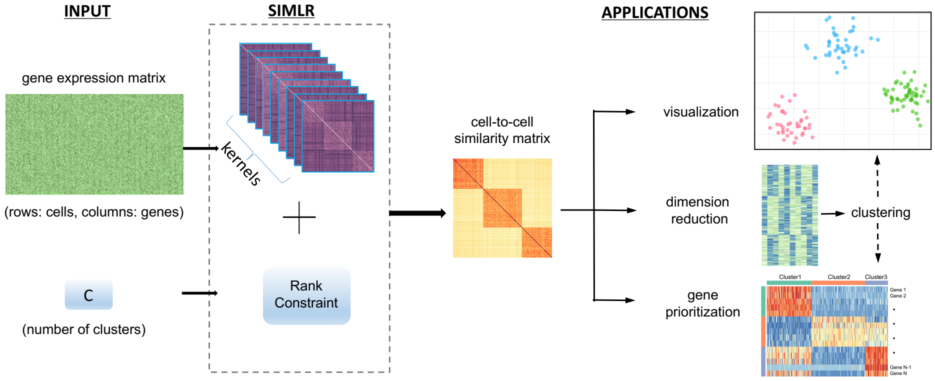

Figure 1: We start with an input matrix with gene expression observations for a set of genes. SIMLR is then capable of learning a set of sample-to-sample similarities by estimating multiple kernels, with the assumptions of the presence of C separable populations within the data. To this extent, SIMLR constraints the similarity matrix to have an approximate block-diagonal structure with C blocks where the samples of the same populations to be more similar. The learned similarities can be used for multiple tasks; they can be used for visualization, reduce the dimension of the data, cluster the populations into subgroups and prioritize the most variable genes that explain the differences across the populations.

efficiently adopted for multiple downstream applications. Some applications include prioritizing genes by ranking their concordance with the similarity and creating low-dimensional representations of the samples by transforming the input into a stochastic neighbor embedding framework, all of which is implemented in our software. We refer to the Supplementary Material for detailed use cases of the tool and to [8] for a detailed description of the method and for several applications on genomic data from public datasets.

In conclusions, SIMLR can infer similarities that can be used to perform dimension reduction, clustering, and visualization in different contexts, with the goal of better understandying the underlying heterogeneity of the studied phenomenon. While the multiple-kernel learning framework has obvious advantages on heterogeneous datasets, where several clusters coexist, we also believe that this approach, together with its visualization framework, may also be valuable for data that does not contain clear clusters, such as cell populations that contain cells spanning a continuum or a developmental pathway.

References

[1] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

[2] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053–1058, 2014.

[3] Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145–153, 2015.

[4] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason CH Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, 2015.

[5] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):241, 2015.

[6] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[7] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 2017.

[8] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel- based similarity learning. *Nature Methods*, 2017.

[9] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[10] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[11] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[12] Cancer Genome Atlas Research Network et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med*, 2015(372):2481–2498, 2015.