Laleh Haghverdi[1,2], John C Marioni[2,3,4], Wolfgang Huber[1]

(1) European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.
(2) European Bioinformatics Institute (EBI-EMBL), Hinxton, UK.
(3) Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
(4) Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

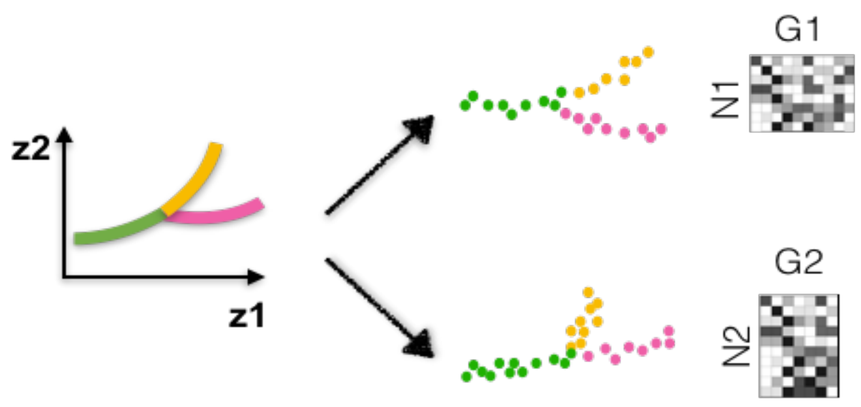
A Soft Alignment of t-SNEs

Abstract

As single-cell profiling at multiple molecular levels becomes increasingly prevalent in many labs, data integration has become central to single-cell data analysis. One example of such data integration is the need/ability to combine two different measurement modes, collected on cells that are sampled from the same biological system. Here, we present a novel cell-alignment strategy, which uses the similarity between the shapes of the data views (i.e., structure of cell populations with respect to each other, which is quantitatively captured in the cells pairwise similarity matrix) to gradually align them in a common non-linear latent space, provided by t-distributed stochastic neighbour embedding (t-SNE) [1], hence performing a soft alignment of t-SNEs (SATSNE). Optionally, our method can use available information on any shared features collected for both data views to leverage the alignment quality.

Concept

Data sets (views) collected from the same biological system encode similar shapes, in spite of difference in individual cells and features assayed in each view.



- Shape information is encoded in the cells pairwise similarity matrix **P,** which t-SNE takes as input.
- t-SNE's gradient descent approach for optimisation of data embedding allows gradual alignment interventions during the optimization iterations.

Key features

- With **Annealing**, i.e., starting with a large perplexity parameter of t-SNE and decreasing it gradually, in order to preserve the global structure of the data manifold in the embedding space,
- Data embedding in a **fixed-size box**, in order to ensure coupling of the two data views rather than chase and escape iterations,
- Periodically alternating between **uncoupled and coupled modes** of t-SNE, in order to allow escaping from suboptimal couplings,
- Identifications of mutual nearest neighbours (MNNs) [2] between the data views using **a few shared features**, in order to guide and limit possible cell-cell couplings in the coupled t-SNE periods.

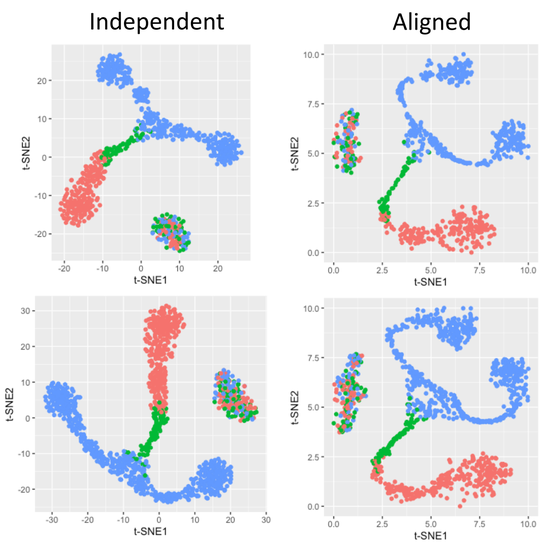
Discussion

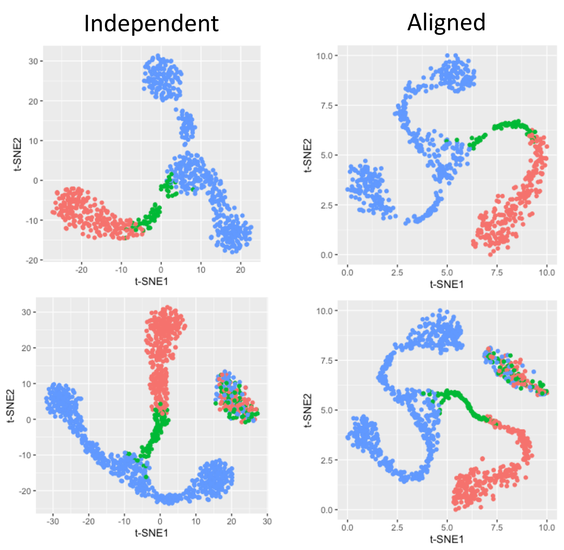
- SATSNE facilitates alignment of data views with different cell samples and different feature sets.
- Alignment of data with complex shapes (e.g. several clusters, symmetry in the shape) is possible in presence of a few shared features.
- Can be used for integration of rather mature information levels of omics data (e.g. mRNA, protein, surface marker levels), generally, whenever unsupervised cells clustering is sensible.

Simulations

Cells' capture time used as a shared feature.

a) A mixture of a branching (temporal) process and a (stationary) Gaussian population in **two views.**



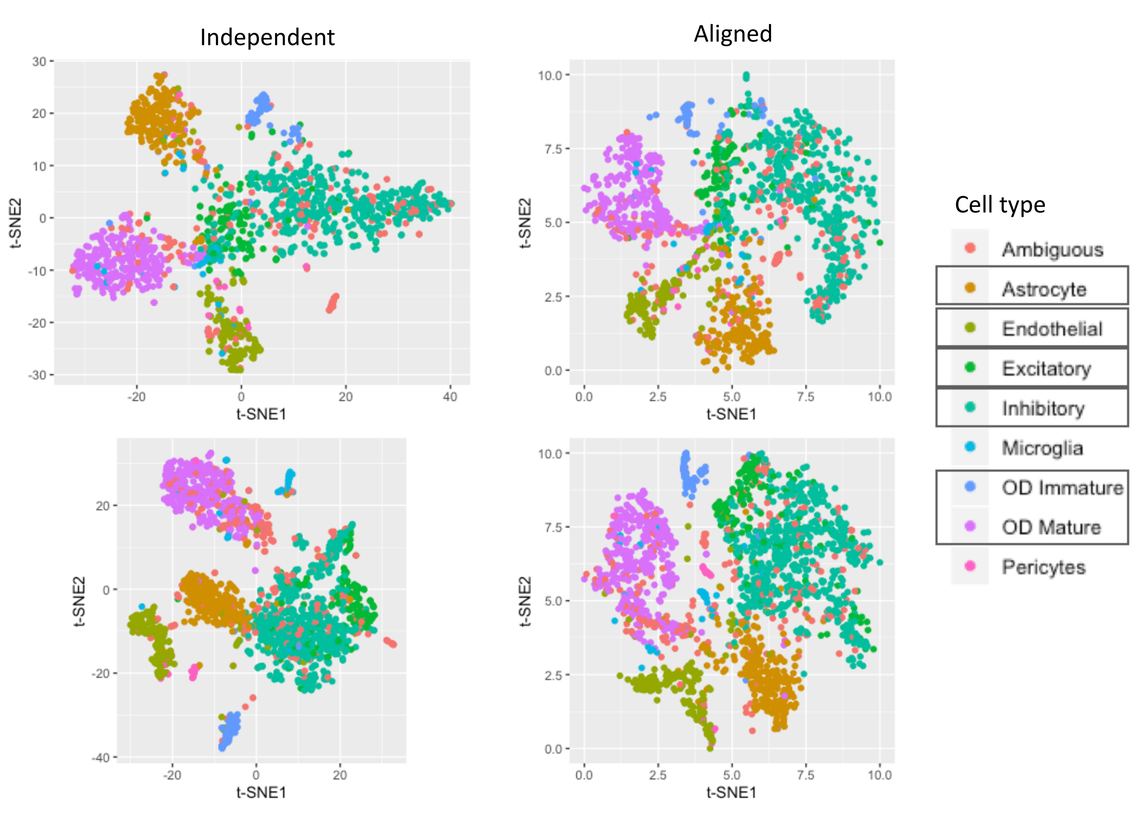


b) The Gaussian population **missing** in view 2,

c) The branching event **not resolved** by the feature set in view 2.

Mouse hypothalamic preoptic cells in situ

- Moffitt *et al.* [3] assayed 160 mRNAs in 3039 cells using multiplexed error robust fluorescence in situ hybridization (MERRFISH) technology [4].
- We split the data into two views with nonoverlapping gene sets (each with 80 genes) and 1500 and 2000 cells respectively.
- Used five major cell types (rectangles in the legend) memberships as shared features.



[1] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.
[2] Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, *36*(5), 421.
[3] Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez, J. D., ... & Zhuang, X. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, *362*(6416), eaau5324.
[4] Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, *348*(6233), aaa6090.

GitHub link:
https://github.com/LalehHaghverdi/SATSNE2019
Email: laleh.haghverdi@embl.de