Non-Linear Dimensionality Reduction in R

Nikolay Oskolkov, MRG Group Leader, LIOS, Riga, Latvia

R course, 09.02.2026

@NikolayOskolkov

@oskolkov.bsky.social

Personal homepage:
https://nikolay-oskolkov.com

Image generated by ChatGPT

Session content

Topics we'll cover in this session:

1) Neighborhood graph principle behind non-linear dimension reduction

2) Overview of tSNE and UMAP algorithms

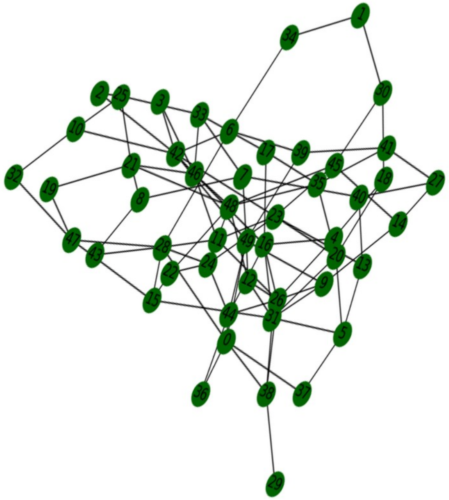
3) Limitations of tSNE and UMAP algorithms

4) Comparing pros and cons of PCA and UMAP applications to cell biology and population genetics studies

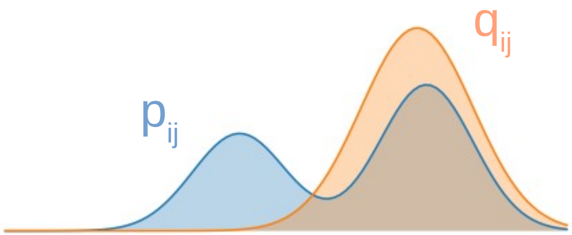
Non-linear dimension reduction: neighborhood graph

1) Construct high-dimensional graph

p_{ij}

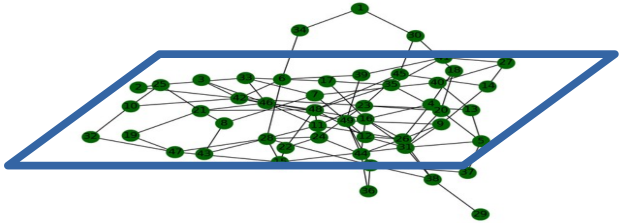


3) Collapse the graphs together



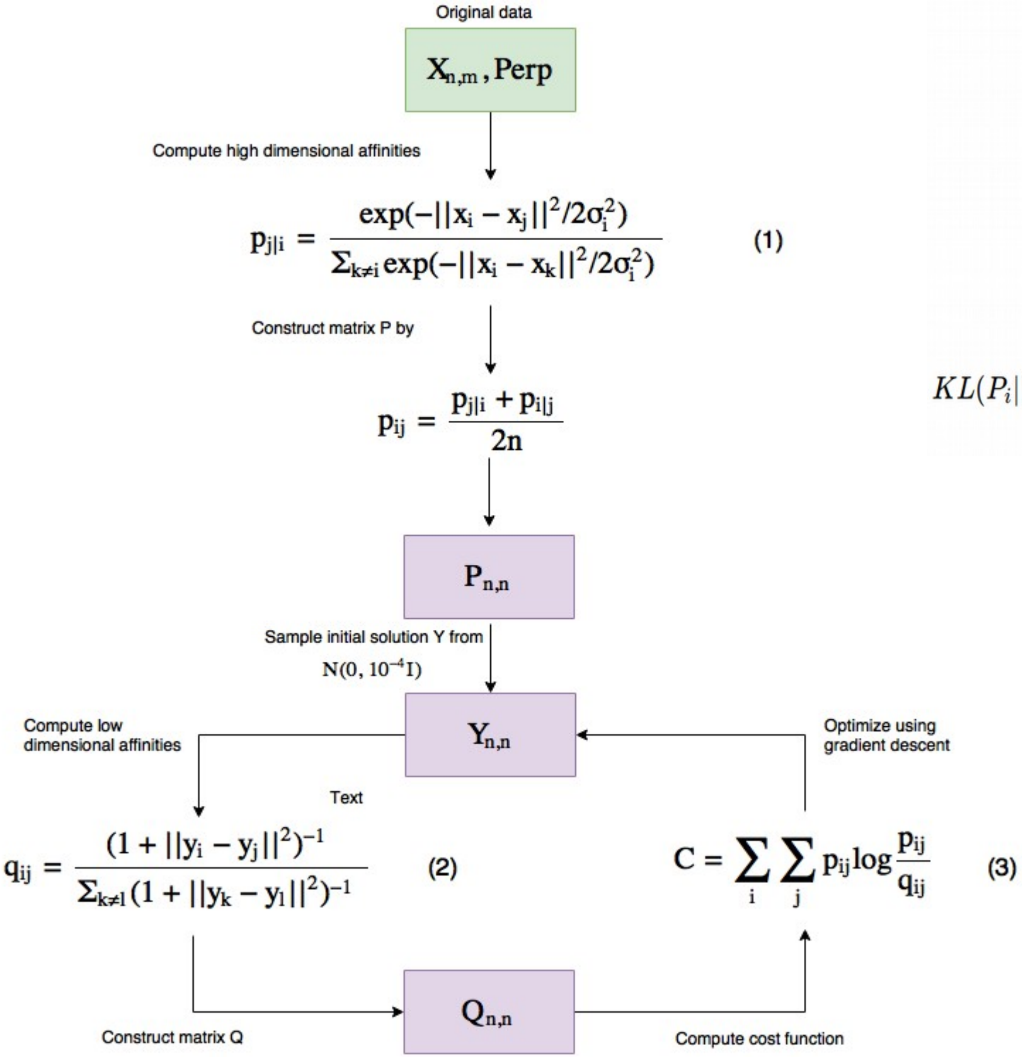
Kullback-Leibler divergence

2) Construct low-dimensional graph



q_{ij}

tSNE dimension reduction algorithm



Original data

$X_{n,m}$, Perp

Compute high dimensional affinities

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\Sigma_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)} \qquad (1)$$

Construct matrix P by

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

$P_{n,n}$

Sample initial solution Y from $N(0, 10^{-4}I)$

$Y_{n,n}$

Compute low dimensional affinities

Optimize using gradient descent

Text

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\Sigma_{k\neq l}(1 + ||y_k - y_l||^2)^{-1}} \qquad (2)$$

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \qquad (3)$$

$Q_{n,n}$

Construct matrix Q

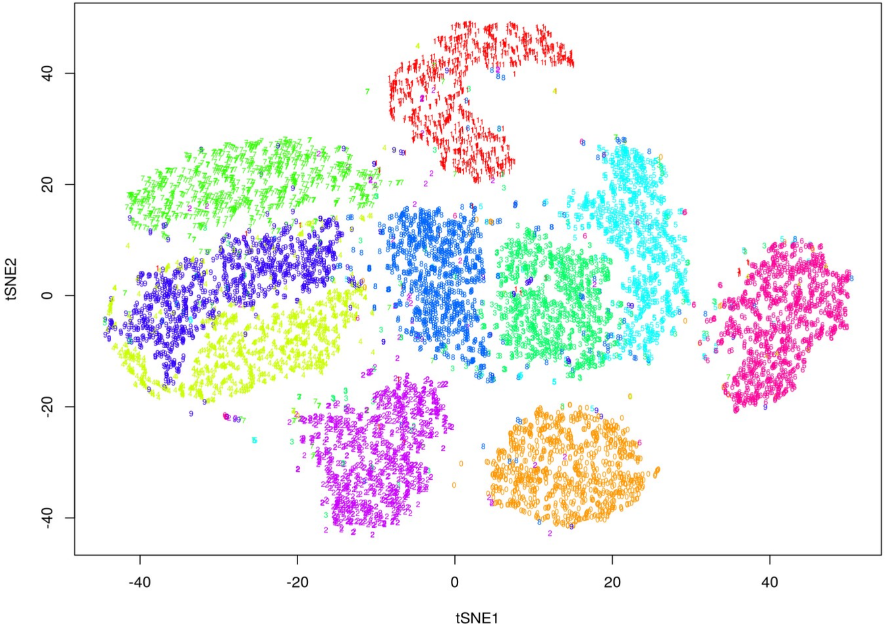
Compute cost function

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-||x_i - x_k||^2/2\sigma_i^2\right)}, \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \qquad (1)$$

$$\mathrm{Perplexity} = 2^{-\sum_j p_{j|i}\log_2 p_{j|i}} \qquad (2)$$

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k\neq l}\left(1 + ||y_k - y_l||^2\right)^{-1}} \qquad (3)$$

$$KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \qquad \frac{\partial KL}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + ||y_i - y_j||^2\right)^{-1} \qquad (4)$$



Limitations of tSNE and promise of UMAP

tSNE does not scale for large data sets?

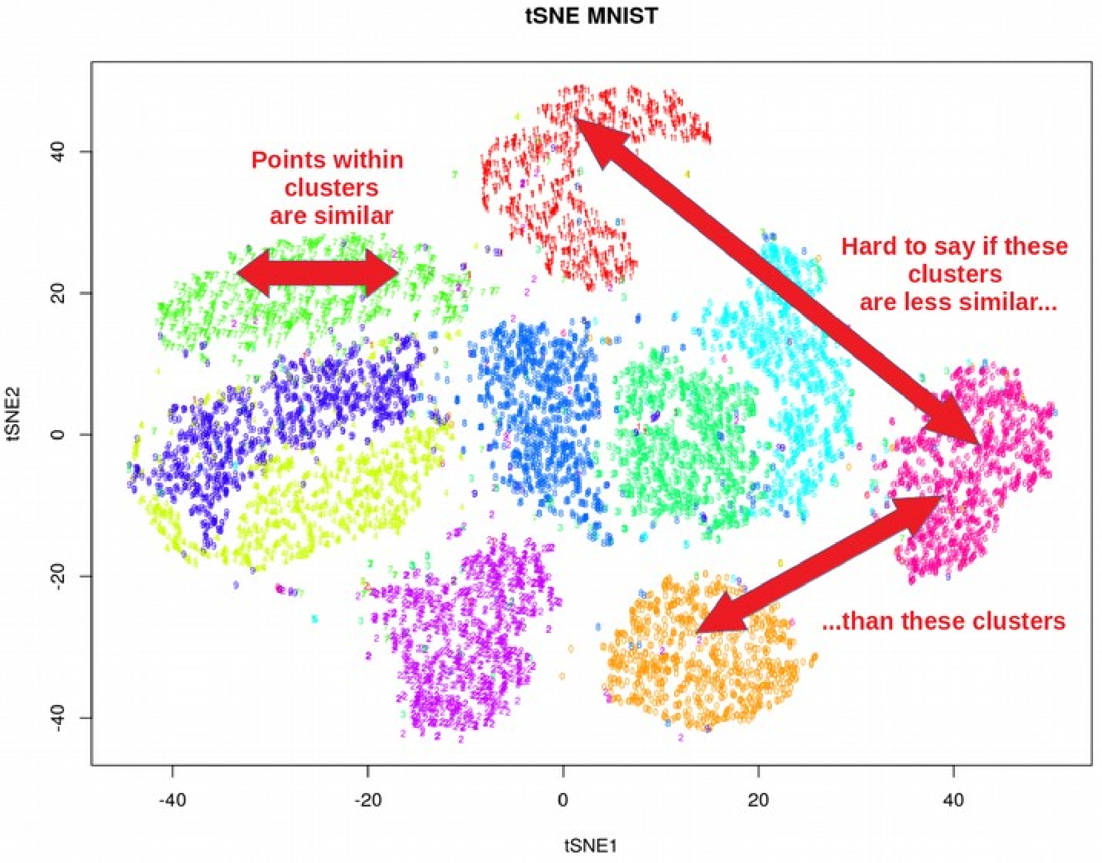
tSNE does not preserve global structure?

tSNE can only embed into 2-3 dims?

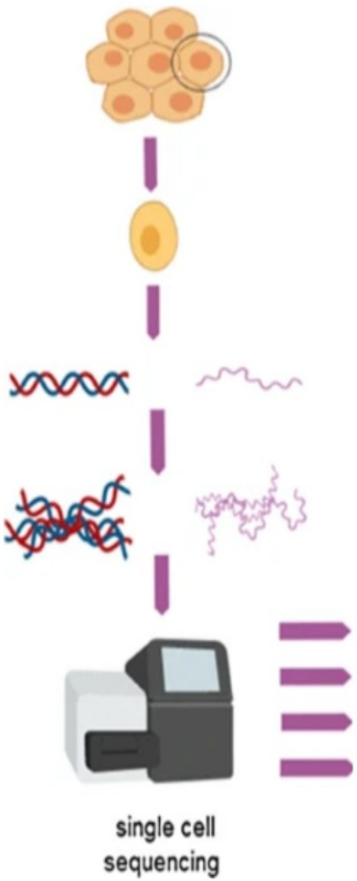
tSNE performs non-parametric mapping (no variance explained statistics)?

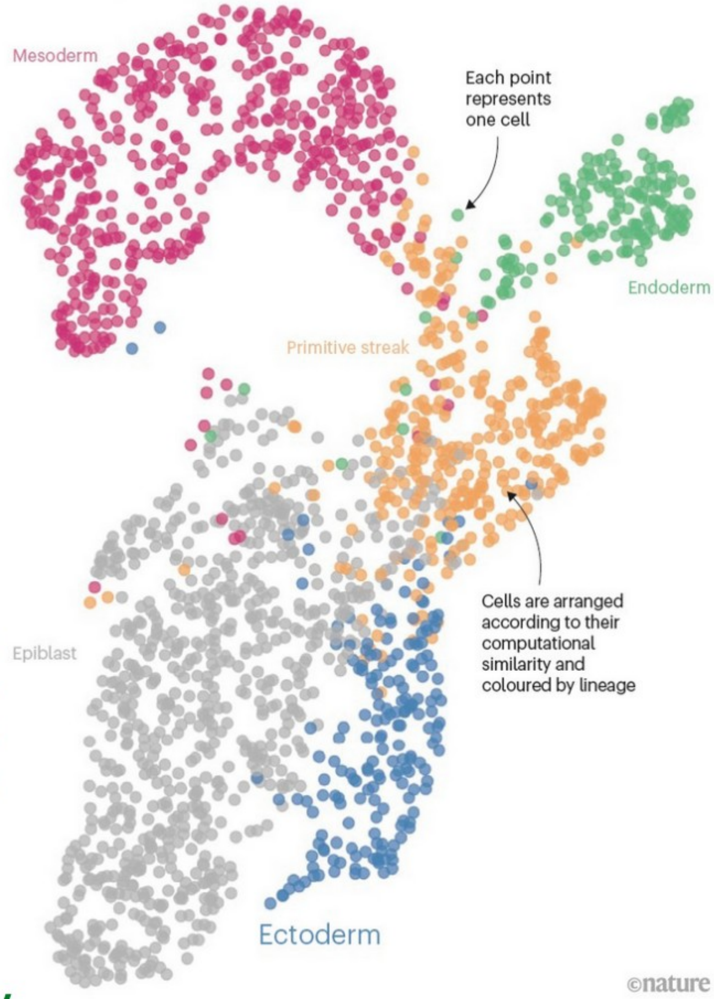
tSNE can not work with high-dimensional data directly (PCA needed)?

tSNE uses too much RAM at large perp?



UMAP: Single Cell vs. PopGen



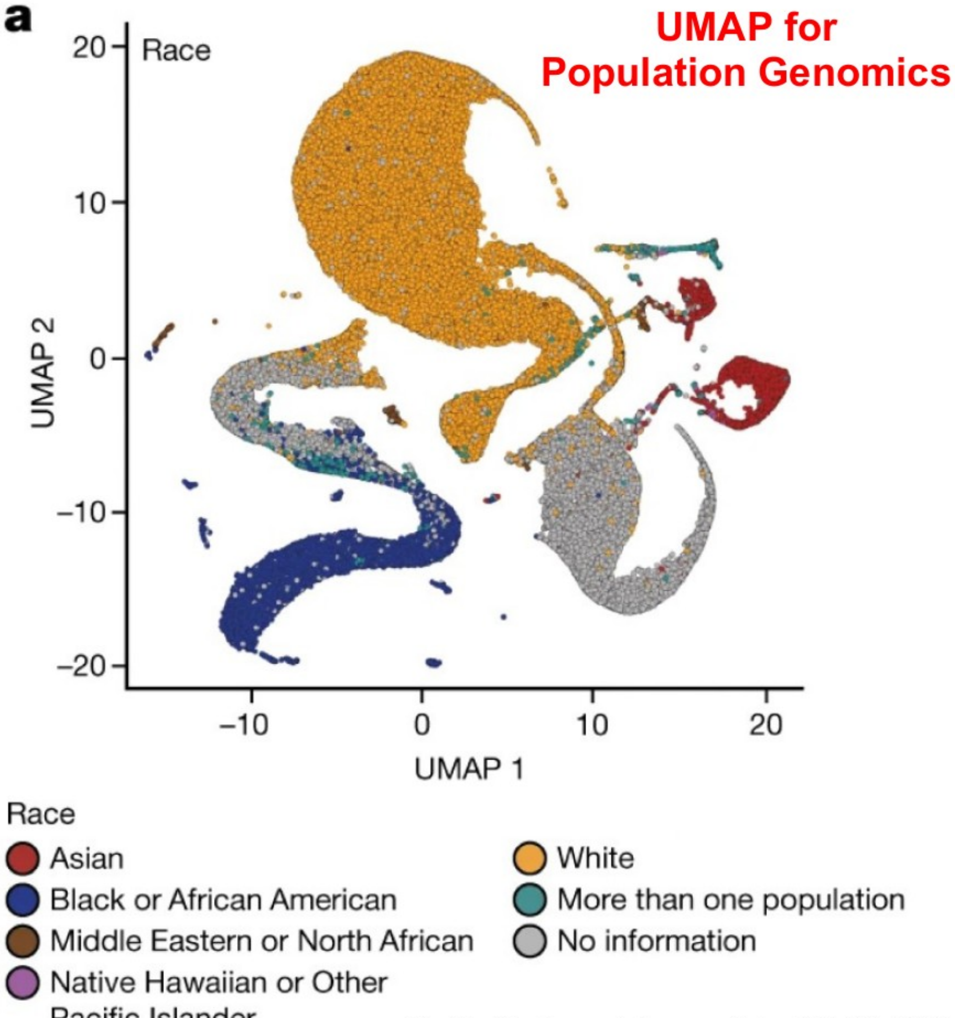


Each point represents one cell

Cells are arranged according to their computational similarity and coloured by lineage

UMAP for Single Cell Biology

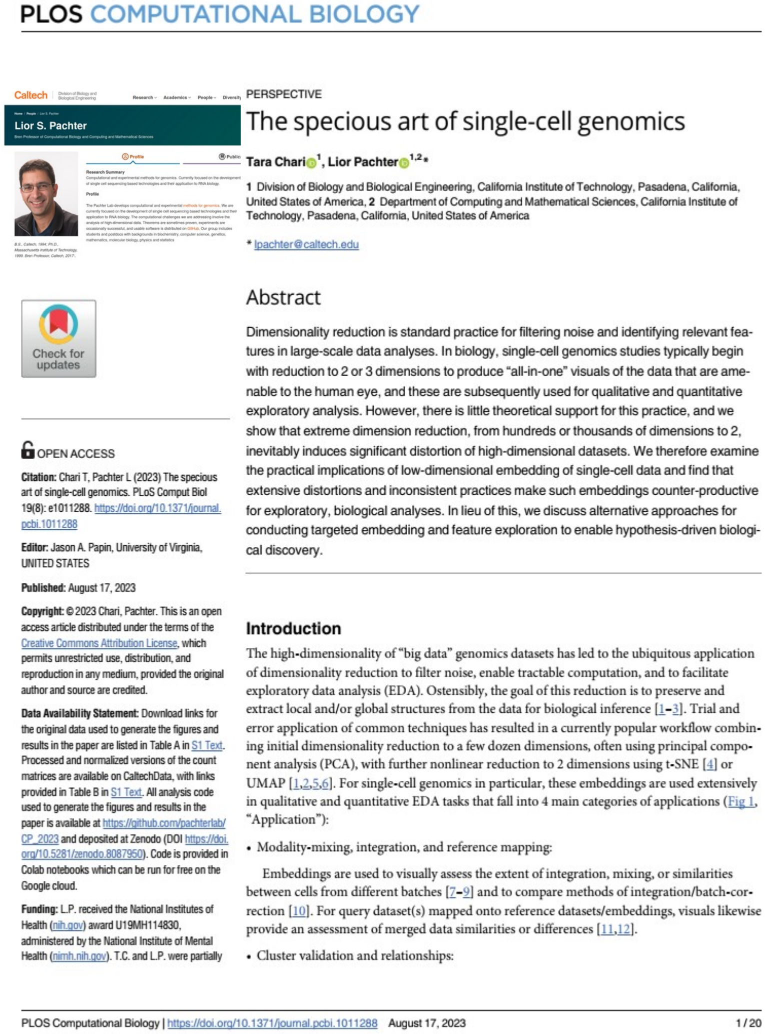
https://www.nature.com/articles/d41586-021-01994-w



UMAP for Population Genomics

The All of Us Research Program, *Nature* **627**, 340 (2024)

UMAP (and Single Cell) Criticism



PLOS COMPUTATIONAL BIOLOGY

Caltech | Division of Biology and Biological Engineering

Research ⌄ Academics ⌄ People ⌄ Diversity

Home / People / Lior S. Pachter

Lior S. Pachter

Bren Professor of Computational Biology and Computing and Mathematical Sciences

Profile | Public

Research Summary

Computational and experimental methods for genomics. Currently focused on the development of single cell sequencing based technologies and their application to RNA biology.

Profile

The Pachter Lab develops computational and experimental methods for genomics. We are currently focused on the development of single cell sequencing based technologies and their application to RNA biology. The computational challenges we are addressing involve the analysis of high-dimensional data. Theorems are sometimes proven, experiments are occasionally successful, and usable software is distributed on GitHub. Our group includes students and postdocs with backgrounds in biochemistry, computer science, genetics, mathematics, molecular biology, physics and statistics

B.S., Caltech, 1994; Ph.D., Massachusetts Institute of Technology, 1999. Bren Professor, Caltech, 2017-.

Check for updates

OPEN ACCESS

Citation: Chari T, Pachter L (2023) The specious art of single-cell genomics. PLoS Comput Biol 19(8): e1011288. https://doi.org/10.1371/journal.pcbi.1011288

Editor: Jason A. Papin, University of Virginia, UNITED STATES

Published: August 17, 2023

Copyright: © 2023 Chari, Pachter. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Download links for the original data used to generate the figures and results in the paper are listed in Table A in S1 Text. Processed and normalized versions of the count matrices are available on CaltechData, with links provided in Table B in S1 Text. All analysis code used to generate the figures and results in the paper is available at https://github.com/pachterlab/CP_2023 and deposited at Zenodo (DOI https://doi.org/10.5281/zenodo.8087950). Code is provided in Colab notebooks which can be run for free on the Google cloud.

Funding: L.P. received the National Institutes of Health (nih.gov) award U19MH114830, administered by the National Institute of Mental Health (nimh.nih.gov). T.C. and L.P. were partially

PERSPECTIVE

The specious art of single-cell genomics

Tara Chari[1], Lior Pachter[1,2]*

1 Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, United States of America, 2 Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, United States of America

* lpachter@caltech.edu

Abstract

Dimensionality reduction is standard practice for filtering noise and identifying relevant features in large-scale data analyses. In biology, single-cell genomics studies typically begin with reduction to 2 or 3 dimensions to produce "all-in-one" visuals of the data that are amenable to the human eye, and these are subsequently used for qualitative and quantitative exploratory analysis. However, there is little theoretical support for this practice, and we show that extreme dimension reduction, from hundreds or thousands of dimensions to 2, inevitably induces significant distortion of high-dimensional datasets. We therefore examine the practical implications of low-dimensional embedding of single-cell data and find that extensive distortions and inconsistent practices make such embeddings counter-productive for exploratory, biological analyses. In lieu of this, we discuss alternative approaches for conducting targeted embedding and feature exploration to enable hypothesis-driven biological discovery.

Introduction

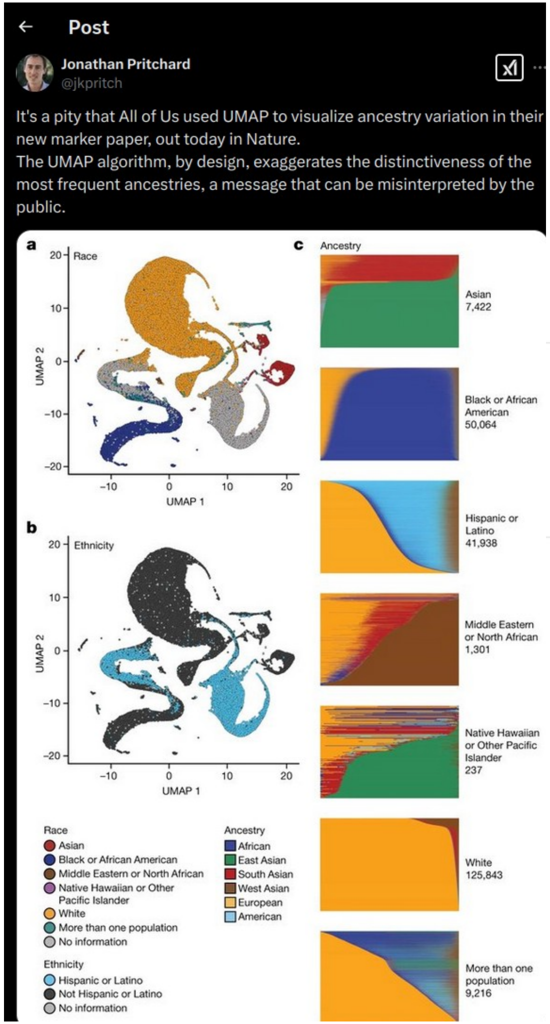
The high-dimensionality of "big data" genomics datasets has led to the ubiquitous application of dimensionality reduction to filter noise, enable tractable computation, and to facilitate exploratory data analysis (EDA). Ostensibly, the goal of this reduction is to preserve and extract local and/or global structures from the data for biological inference [1–3]. Trial and error application of common techniques has resulted in a currently popular workflow combining initial dimensionality reduction to a few dozen dimensions, often using principal component analysis (PCA), with further nonlinear reduction to 2 dimensions using t-SNE [4] or UMAP [1,2,5,6]. For single-cell genomics in particular, these embeddings are used extensively in qualitative and quantitative EDA tasks that fall into 4 main categories of applications (Fig 1, "Application"):

- Modality-mixing, integration, and reference mapping:

Embeddings are used to visually assess the extent of integration, mixing, or similarities between cells from different batches [7–9] and to compare methods of integration/batch-correction [10]. For query dataset(s) mapped onto reference datasets/embeddings, visuals likewise provide an assessment of merged data similarities or differences [11,12].

- Cluster validation and relationships:

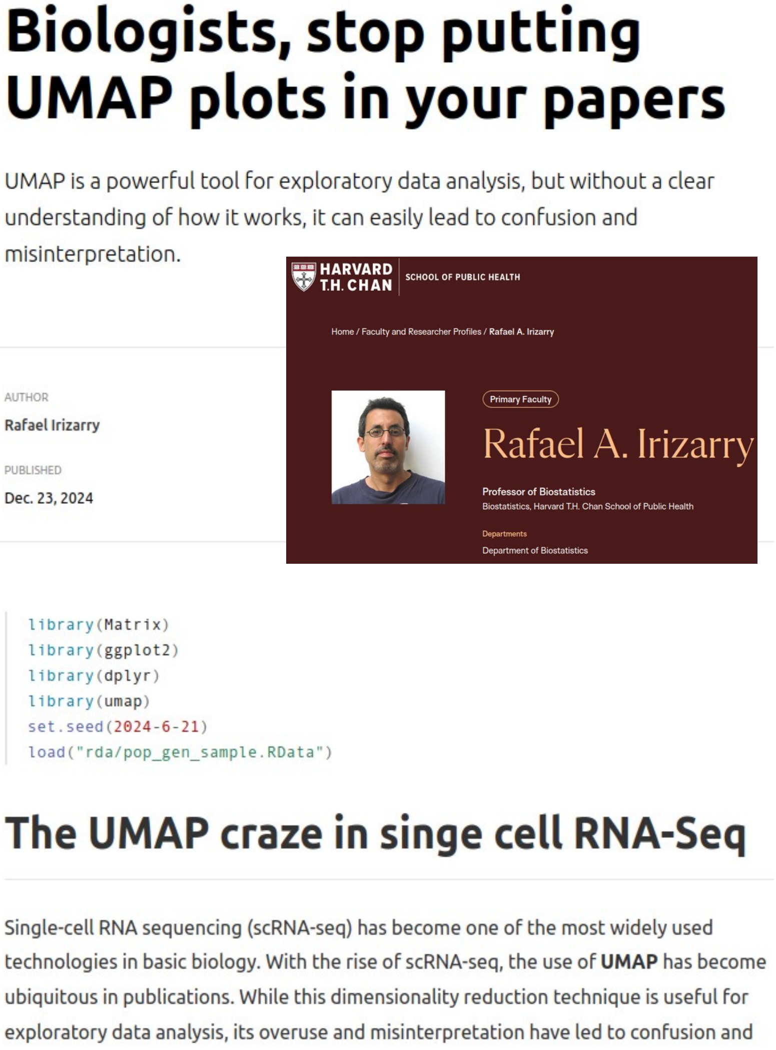
PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1011288 August 17, 2023 1 / 20



← Post

Jonathan Pritchard
@jkpritch

It's a pity that All of Us used UMAP to visualize ancestry variation in their new marker paper, out today in Nature.
The UMAP algorithm, by design, exaggerates the distinctiveness of the most frequent ancestries, a message that can be misinterpreted by the public.



Biologists, stop putting UMAP plots in your papers

UMAP is a powerful tool for exploratory data analysis, but without a clear understanding of how it works, it can easily lead to confusion and misinterpretation.

AUTHOR

Rafael Irizarry

PUBLISHED

Dec. 23, 2024

HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

Home / Faculty and Researcher Profiles / Rafael A. Irizarry

Primary Faculty

Rafael A. Irizarry

Professor of Biostatistics

Biostatistics, Harvard T.H. Chan School of Public Health

Departments

Department of Biostatistics

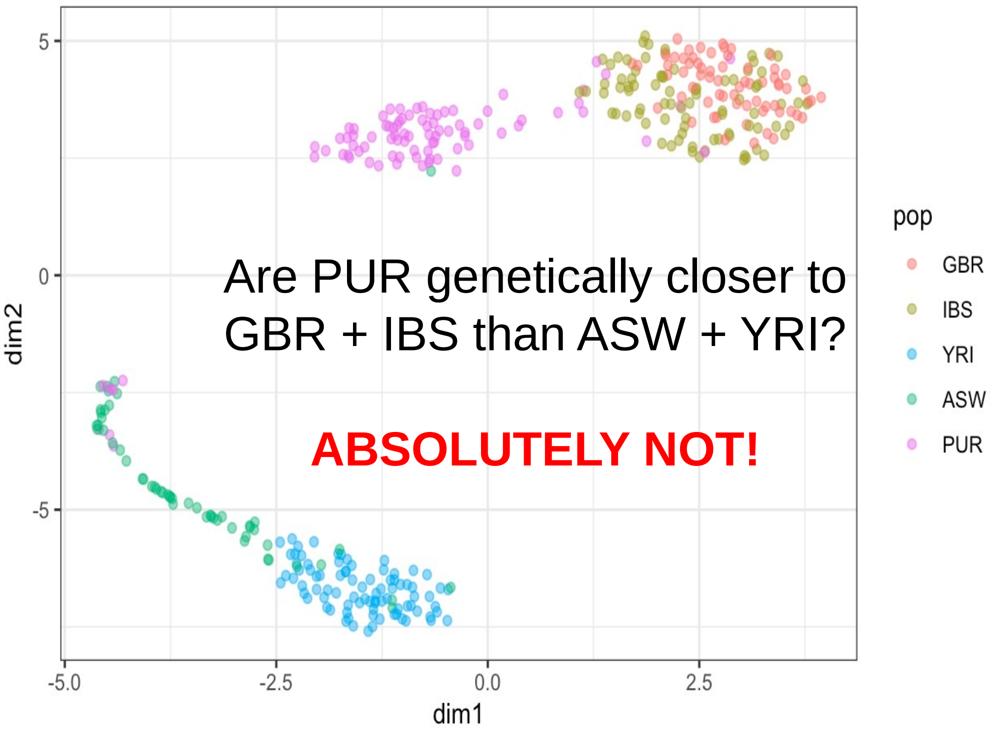
```
library(Matrix)
library(ggplot2)
library(dplyr)
library(umap)
set.seed(2024-6-21)
load("rda/pop_gen_sample.RData")
```

The UMAP craze in singe cell RNA-Seq

Single-cell RNA sequencing (scRNA-seq) has become one of the most widely used technologies in basic biology. With the rise of scRNA-seq, the use of **UMAP** has become ubiquitous in publications. While this dimensionality reduction technique is useful for exploratory data analysis, its overuse and misinterpretation have led to confusion and

HapMap3: UMAP vs PCA (Rafael Irizarry's post)

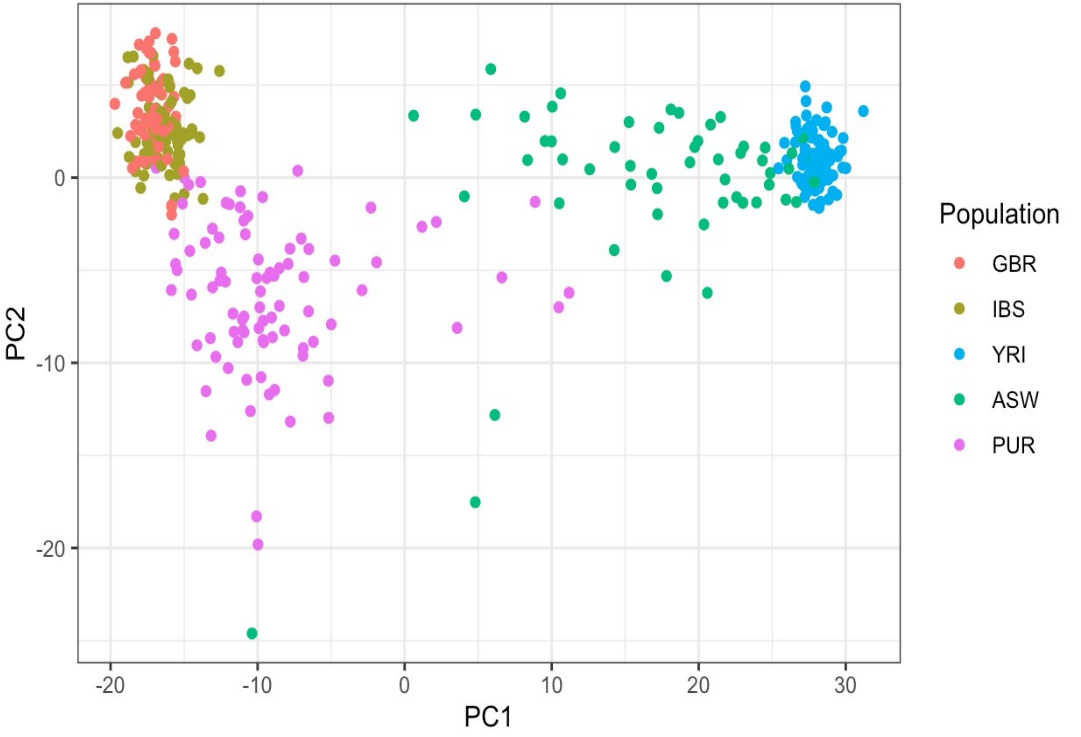
UMAP



Are PUR genetically closer to GBR + IBS than ASW + YRI?

ABSOLUTELY NOT!

PCA



- Because of their meaningless inter-cluster distances tSNE / UMAP are less useful for population genomics than PCA
- The goal of tSNE / UMAP is to **discover clusters**, which is sufficient for Single Cell Biology but not for PopGen.
- In PopGen we generally do not discover clusters, we have an idea about e.g. human populations, and the aim is often to explore the **genetic relatedness** between the populations, a task UMAP can absolutely not solve!

Session summary

Take home messages of the session:

1) Neighborhood graph is the key of non-linear dimension reduction

2) Kullback-Leibler divergence is the objective function of tSNE

3) Inter-cluster distances in tSNE and UMAP are meaningless

4) tSNE and UMAP are appropriate for cell biology but not for human genetics applications where PCA is more accurate and informative

Acknowledgments: LIOS + TARGETWISE

TARGET WISE

Latvian Institute of Organic Synthesis

2027
National Development Plan

Funded by
the European Union