Is UMAP accurate?
Addressing some fair and unfair criticism

Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden

NBIS AI and IO Seminar Series, 14.02.2025

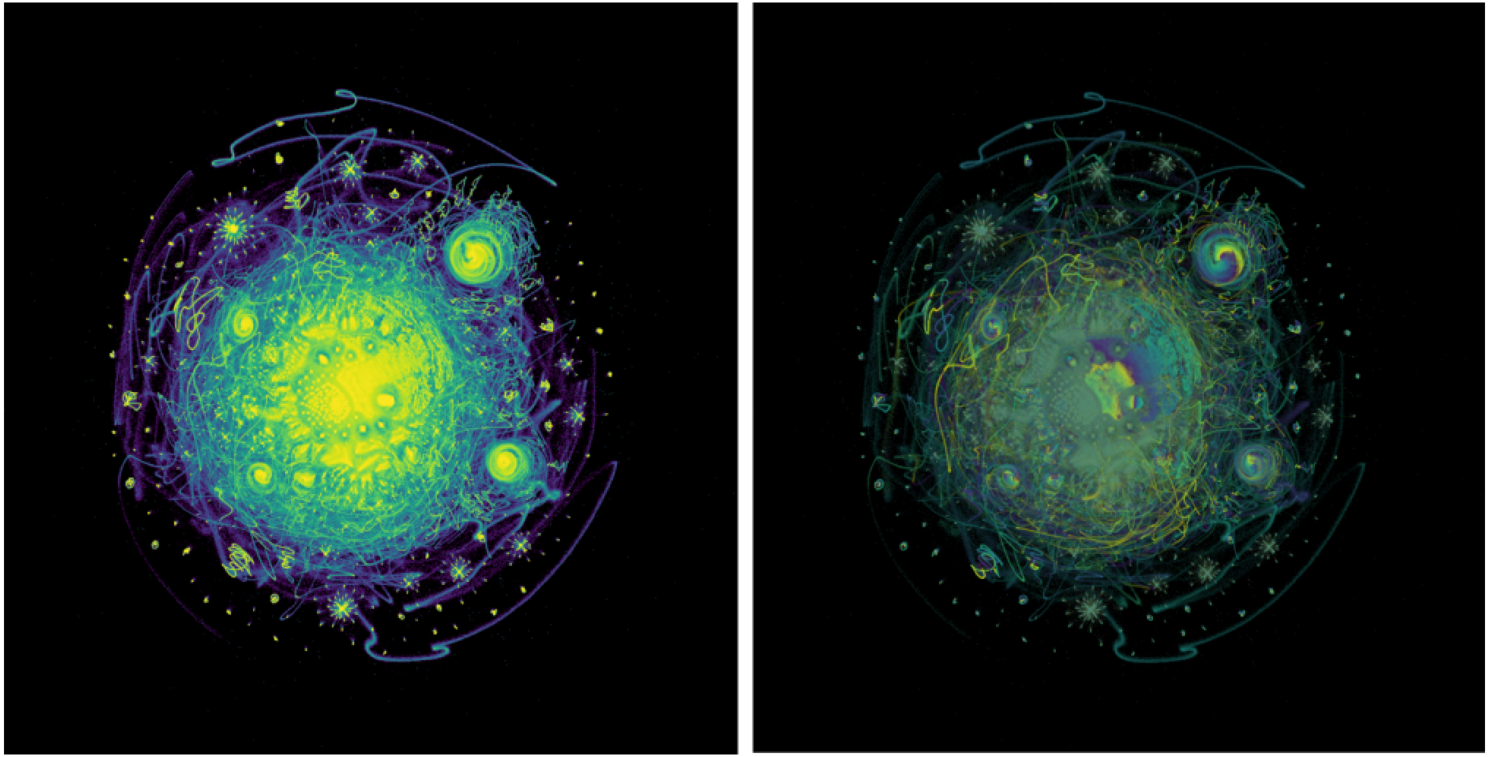


Image adapted from McInnes et al. 2018

@NikolayOskolkov

@oskolkov.bsky.social

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

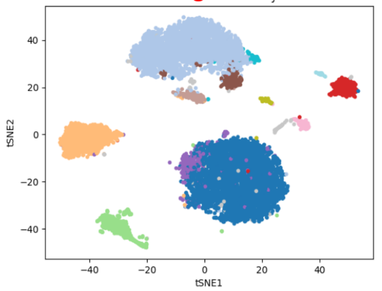
Brief introduction: who am I

2007 PhD in theoretical physics

2011 medical genetics at Lund University

2016 working at NBIS SciLifeLab, Sweden

Single cell



Biomedical data integration

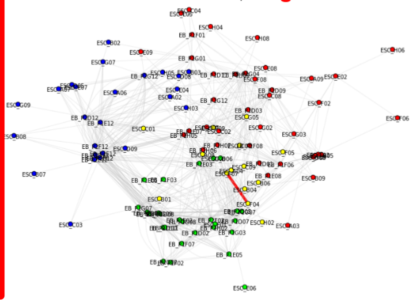
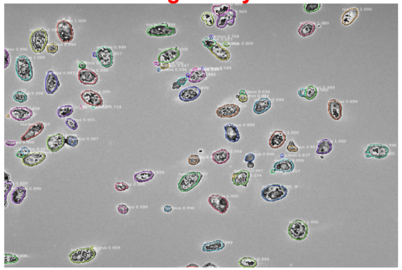
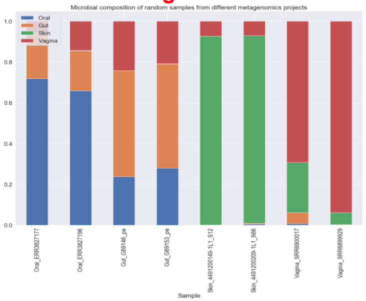


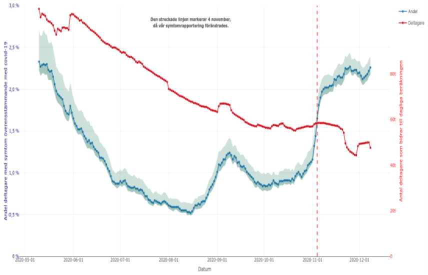
Image analysis



Metagenomics



Epidemiology

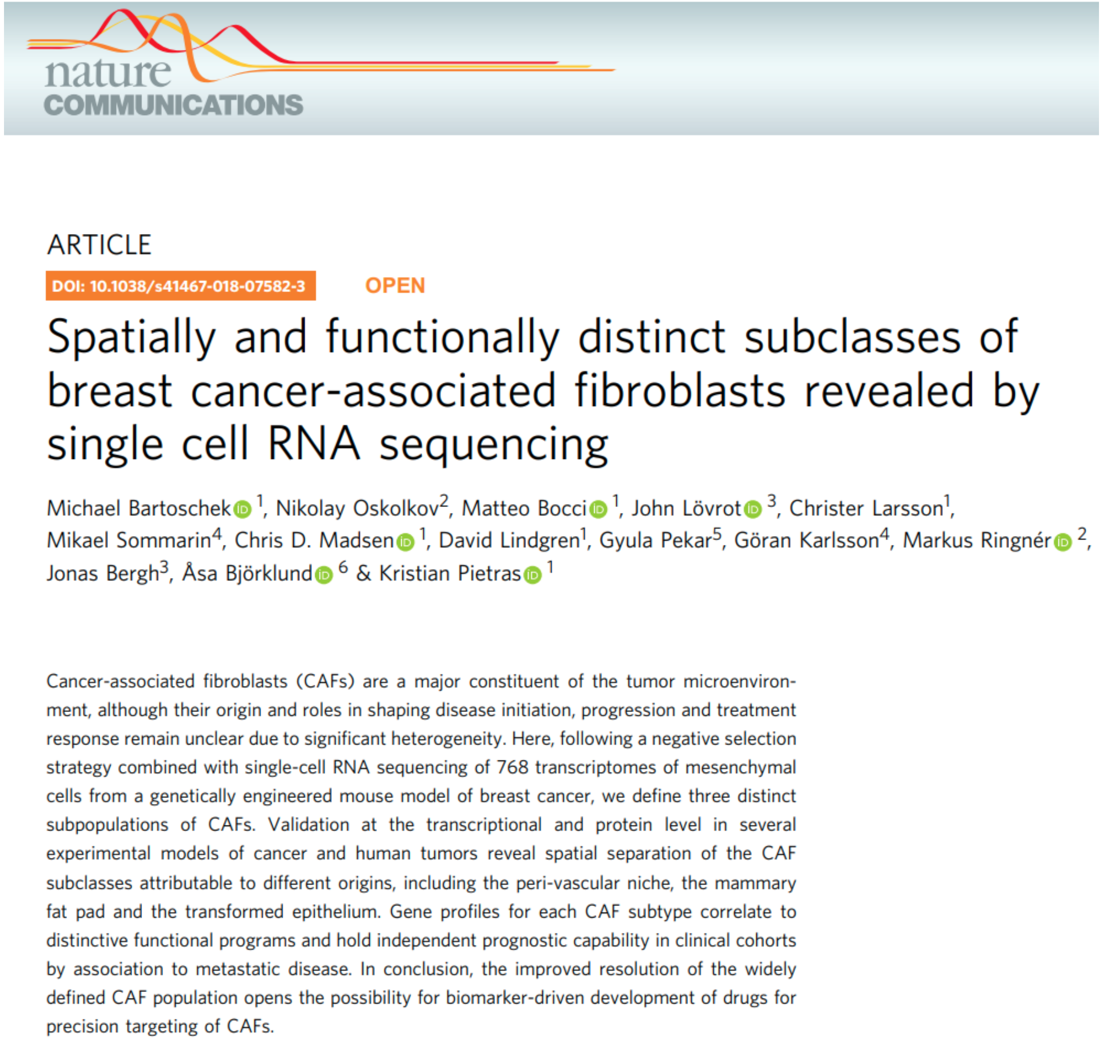


Ancient microbiome

Lund University

SciLifeLab

My single cell papers using tSNE / UMAP



nature COMMUNICATIONS

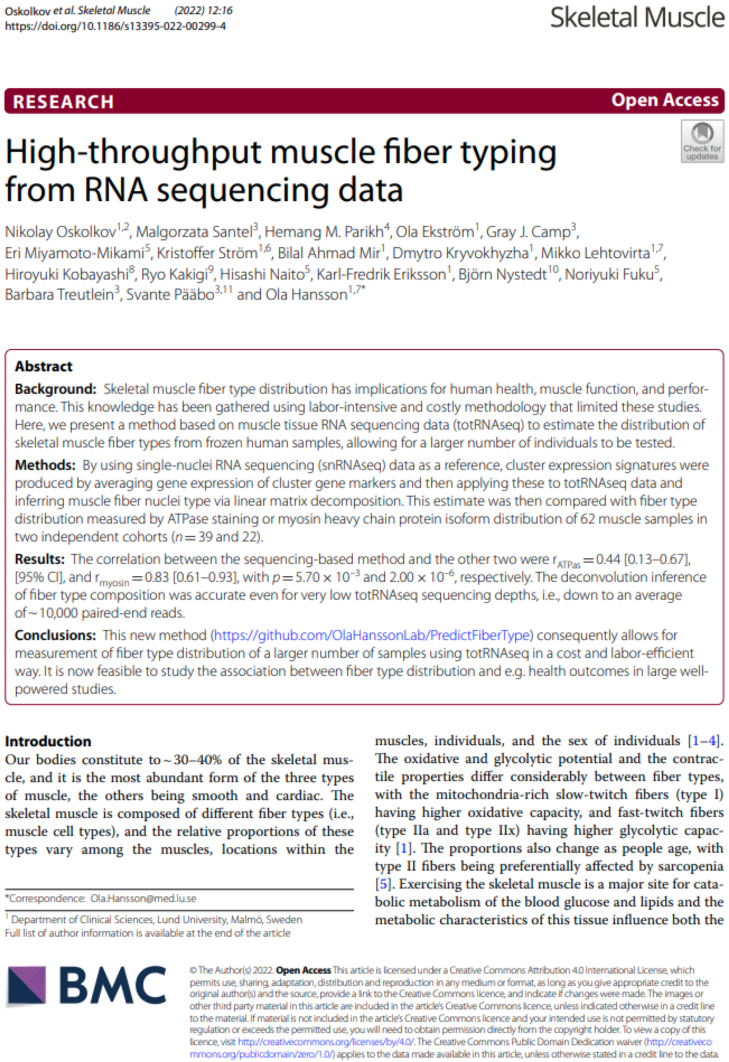
ARTICLE

DOI: 10.1038/s41467-018-07582-3 OPEN

Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing

Michael Bartoschek[1], Nikolay Oskolkov[2], Matteo Bocci[1], John Lövrot[3], Christer Larsson[1], Mikael Sommarin[4], Chris D. Madsen[1], David Lindgren[1], Gyula Pekar[5], Göran Karlsson[4], Markus Ringnér[2], Jonas Bergh[3], Åsa Björklund[6] & Kristian Pietras[1]

Cancer-associated fibroblasts (CAFs) are a major constituent of the tumor microenvironment, although their origin and roles in shaping disease initiation, progression and treatment response remain unclear due to significant heterogeneity. Here, following a negative selection strategy combined with single-cell RNA sequencing of 768 transcriptomes of mesenchymal cells from a genetically engineered mouse model of breast cancer, we define three distinct subpopulations of CAFs. Validation at the transcriptional and protein level in several experimental models of cancer and human tumors reveal spatial separation of the CAF subclasses attributable to different origins, including the peri-vascular niche, the mammary fat pad and the transformed epithelium. Gene profiles for each CAF subtype correlate to distinctive functional programs and hold independent prognostic capability in clinical cohorts by association to metastatic disease. In conclusion, the improved resolution of the widely defined CAF population opens the possibility for biomarker-driven development of drugs for precision targeting of CAFs.



Oskolkov *et al. Skeletal Muscle* (2022) 12:16
https://doi.org/10.1186/s13395-022-00299-4

Skeletal Muscle

RESEARCH Open Access

High-throughput muscle fiber typing from RNA sequencing data

Nikolay Oskolkov[1,2], Malgorzata Santel[3], Hemang M. Parikh[4], Ola Ekström[1], Gray J. Camp[3], Eri Miyamoto-Mikami[5], Kristoffer Ström[1,6], Bilal Ahmad Mir[1], Dmytro Kryvokhyzha[1], Mikko Lehtovirta[1,7], Hiroyuki Kobayashi[8], Ryo Kakigi[9], Hisashi Naito[5], Karl-Fredrik Eriksson[1], Björn Nystedt[10], Noriyuki Fuku[5], Barbara Treutlein[3], Svante Pääbo[3,11] and Ola Hansson[1,7*]

Abstract

Background: Skeletal muscle fiber type distribution has implications for human health, muscle function, and performance. This knowledge has been gathered using labor-intensive and costly methodology that limited these studies. Here, we present a method based on muscle tissue RNA sequencing data (totRNAseq) to estimate the distribution of skeletal muscle fiber types from frozen human samples, allowing for a larger number of individuals to be tested.

Methods: By using single-nuclei RNA sequencing (snRNAseq) data as a reference, cluster expression signatures were produced by averaging gene expression of cluster gene markers and then applying these to totRNAseq data and inferring muscle fiber nuclei type via linear matrix decomposition. This estimate was then compared with fiber type distribution measured by ATPase staining or myosin heavy chain protein isoform distribution of 62 muscle samples in two independent cohorts ($n = 39$ and 22).

Results: The correlation between the sequencing-based method and the other two were $r_{ATPas} = 0.44$ [0.13–0.67], [95% CI], and $r_{myosin} = 0.83$ [0.61–0.93], with $p = 5.70 \times 10^{-3}$ and 2.00×10^{-6}, respectively. The deconvolution inference of fiber type composition was accurate even for very low totRNAseq sequencing depths, i.e., down to an average of ~ 10,000 paired-end reads.

Conclusions: This new method (https://github.com/OlaHanssonLab/PredictFiberType) consequently allows for measurement of fiber type distribution of a larger number of samples using totRNAseq in a cost and labor-efficient way. It is now feasible to study the association between fiber type distribution and e.g. health outcomes in large well-powered studies.

Introduction

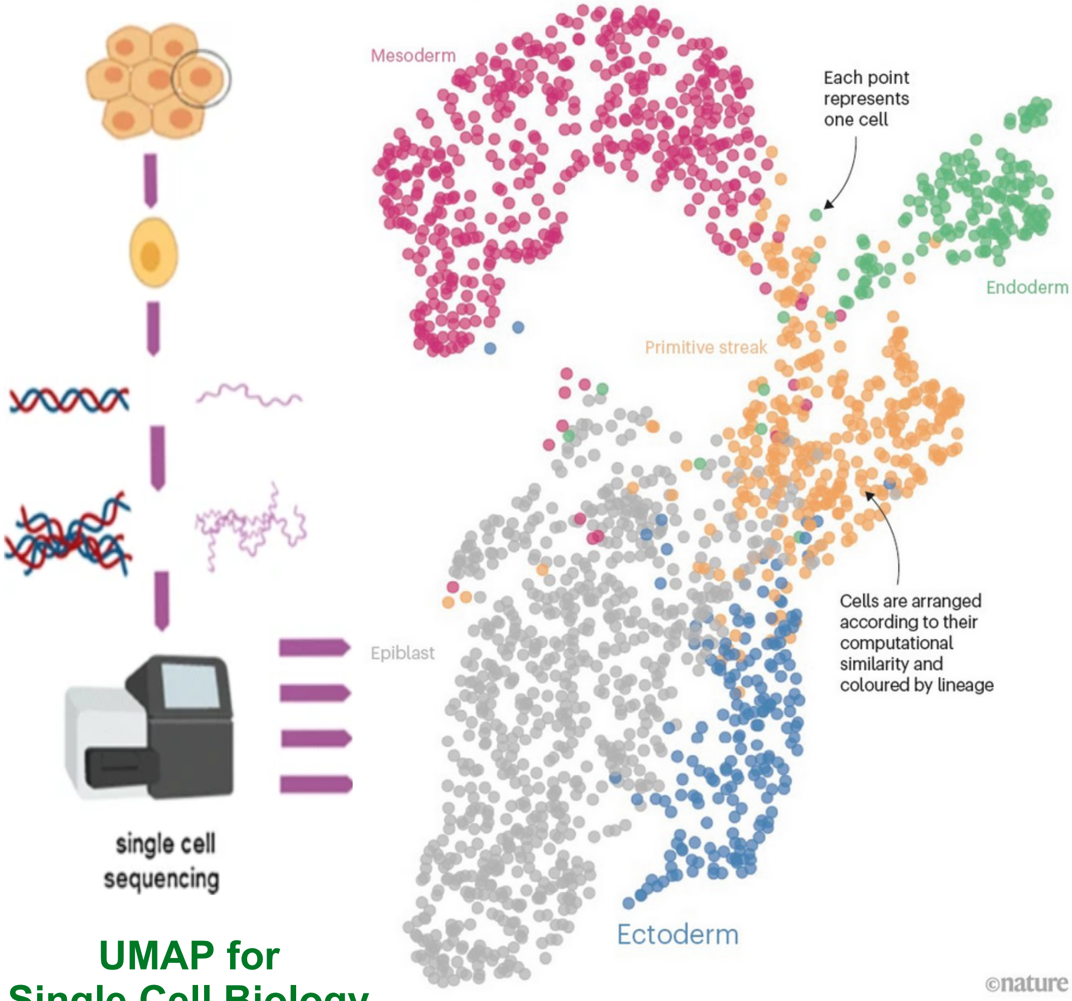
Our bodies constitute to ~ 30–40% of the skeletal muscle, and it is the most abundant form of the three types of muscle, the others being smooth and cardiac. The skeletal muscle is composed of different fiber types (i.e., muscle cell types), and the relative proportions of these types vary among the muscles, locations within the muscles, individuals, and the sex of individuals [1–4]. The oxidative and glycolytic potential and the contractile properties differ considerably between fiber types, with the mitochondria-rich slow-twitch fibers (type I) having higher oxidative capacity, and fast-twitch fibers (type IIa and type IIx) having higher glycolytic capacity [1]. The proportions also change as people age, with type II fibers being preferentially affected by sarcopenia [5]. Exercising the skeletal muscle is a major site for catabolic metabolism of the blood glucose and lipids and the metabolic characteristics of this tissue influence both the

*Correspondence: Ola.Hansson@med.lu.se

[1] Department of Clinical Sciences, Lund University, Malmö, Sweden
Full list of author information is available at the end of the article

tSNE / UMAP: problem formulation

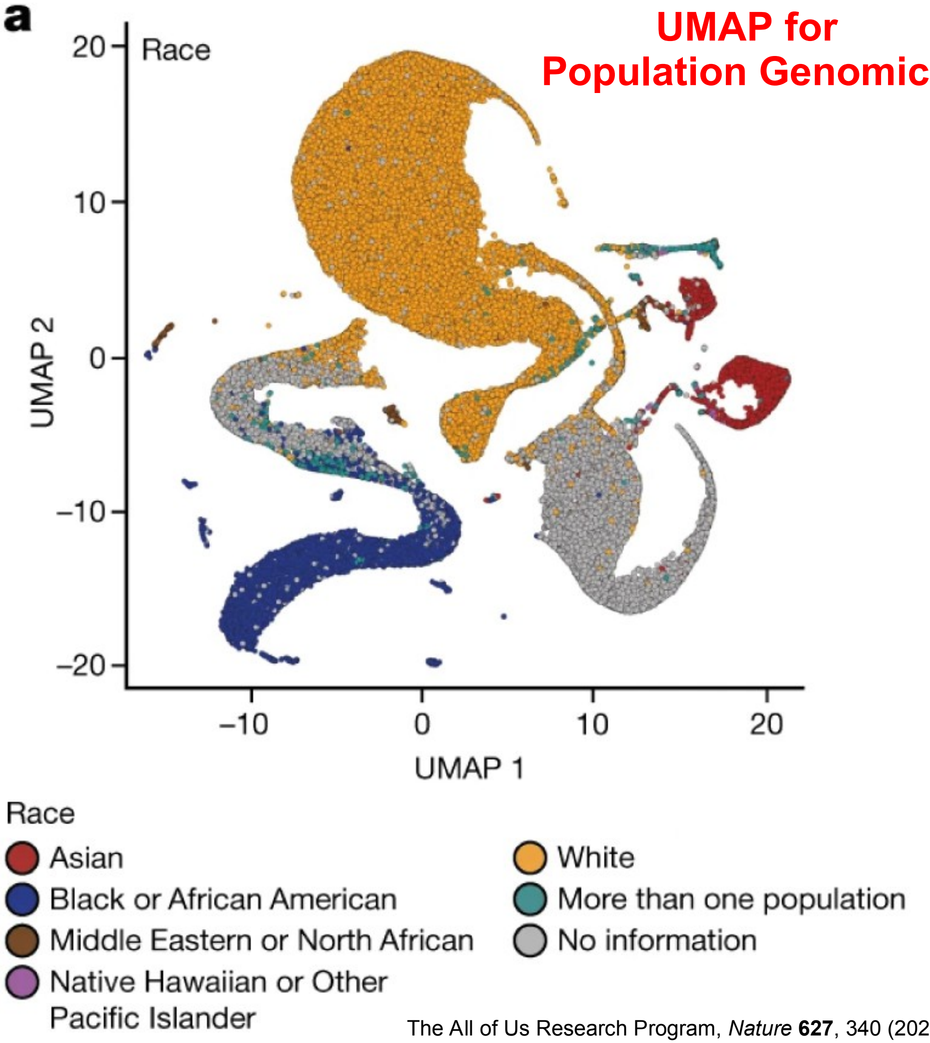
UMAP: Single Cell vs. PopGen



UMAP for Single Cell Biology

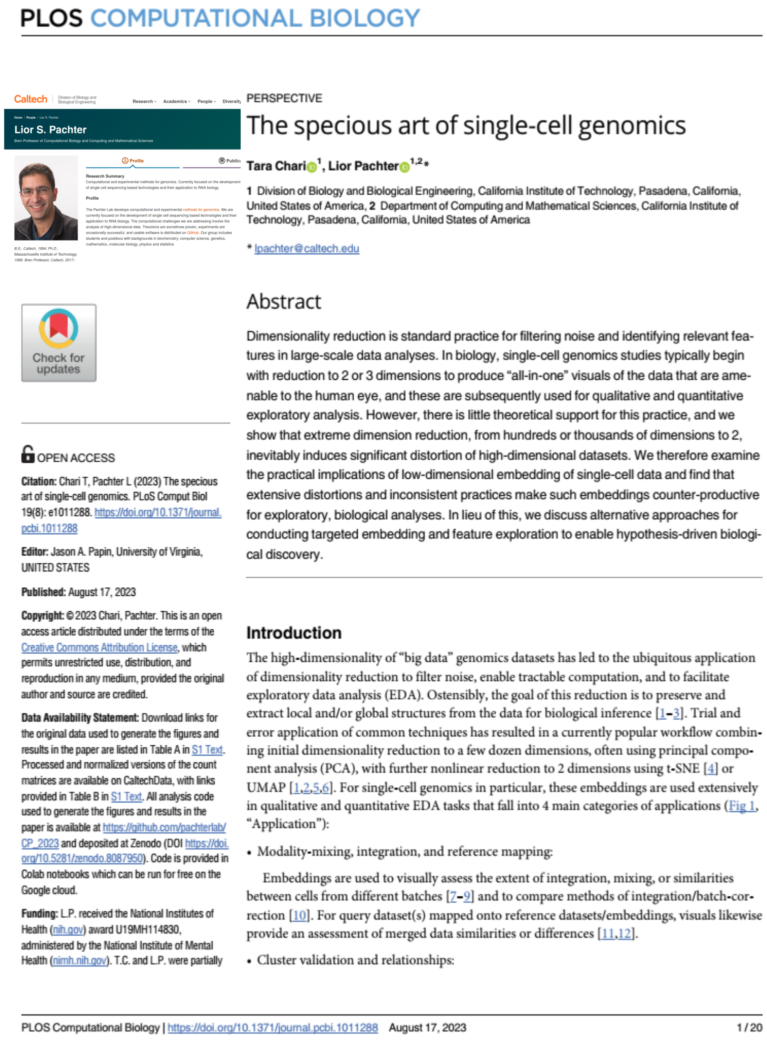
https://www.nature.com/articles/d41586-021-01994-w

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   Publica

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-.

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

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

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

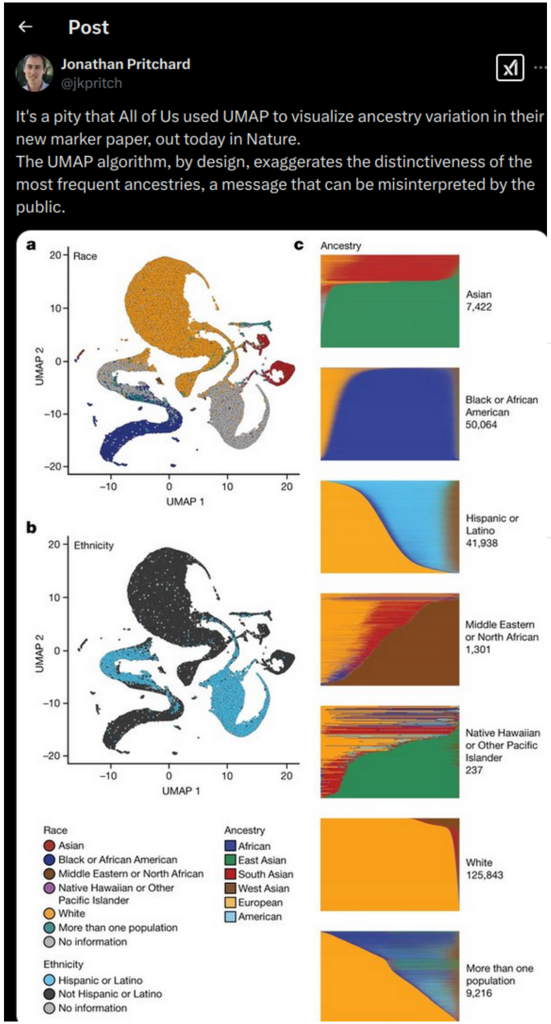
The high-dimensionality of "big data" genomics datasets has led to the ubiquitous application of dimensionality reduction to filter noise, enable tractable computation, and to facilitate exploratory data analysis (EDA). Ostensibly, the goal of this reduction is to preserve and extract local and/or global 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:

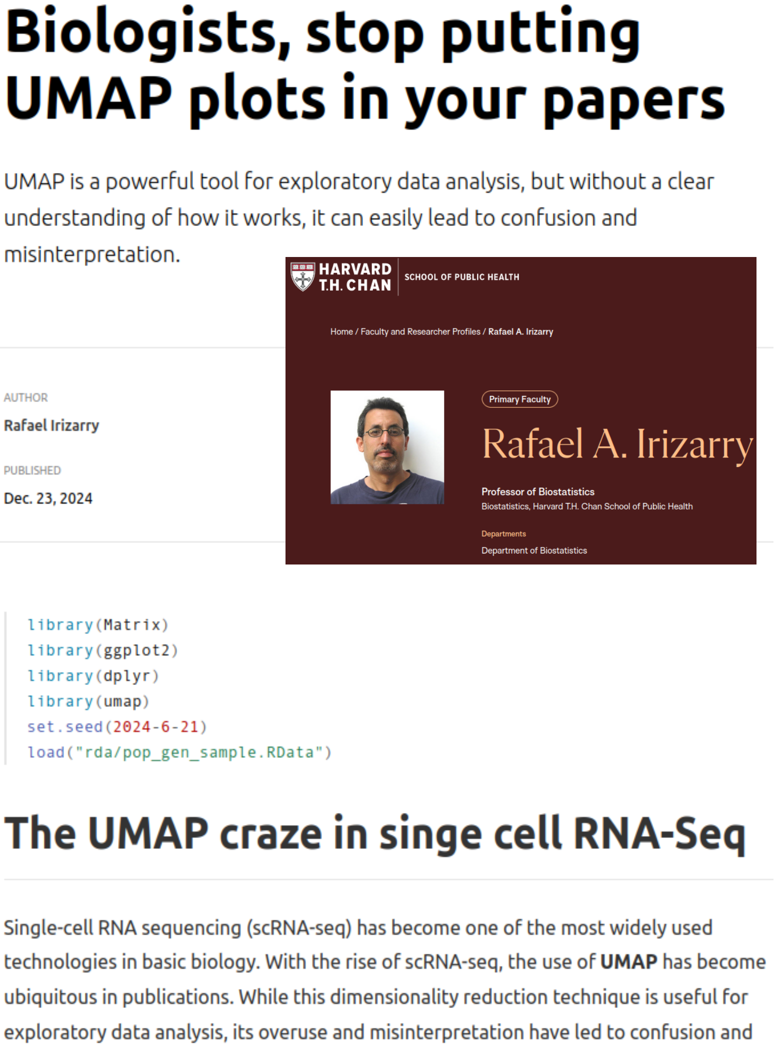
PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1011288   August 17, 2023   1 / 20



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")
```

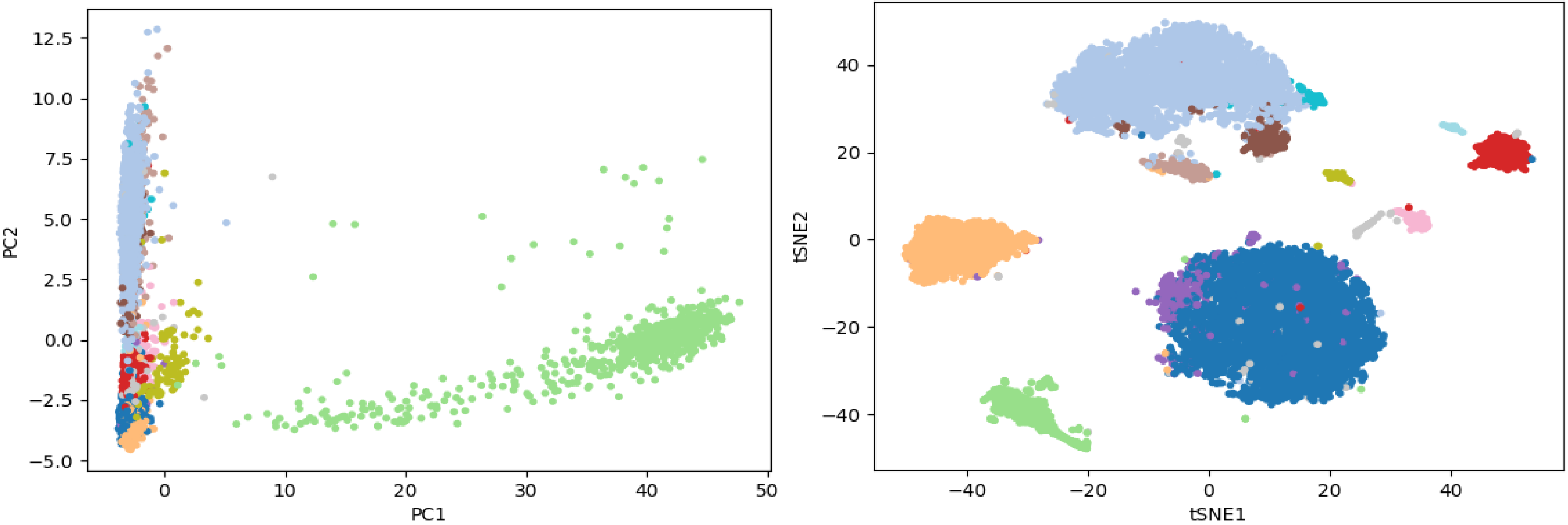
The UMAP craze in singe cell RNA-Seq

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

How and why tSNE / UMAP became popular in Single Cell?

How it started (at least in Single Cell)

3k Peripheral Blood Mononuclear Cells (PBMC) available from 10X Genomics



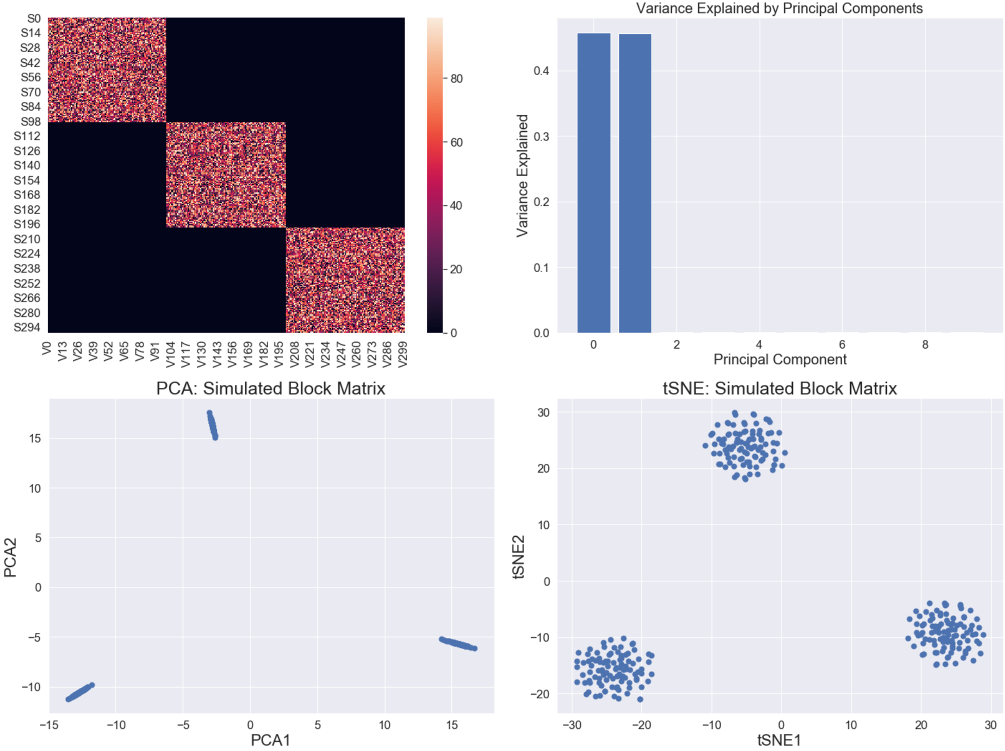
Two principal components (PCs) seem to be insufficient to fully reveal heterogeneity in single cell gene expression data.

Solution: use more PCs or tSNE / UMAP

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

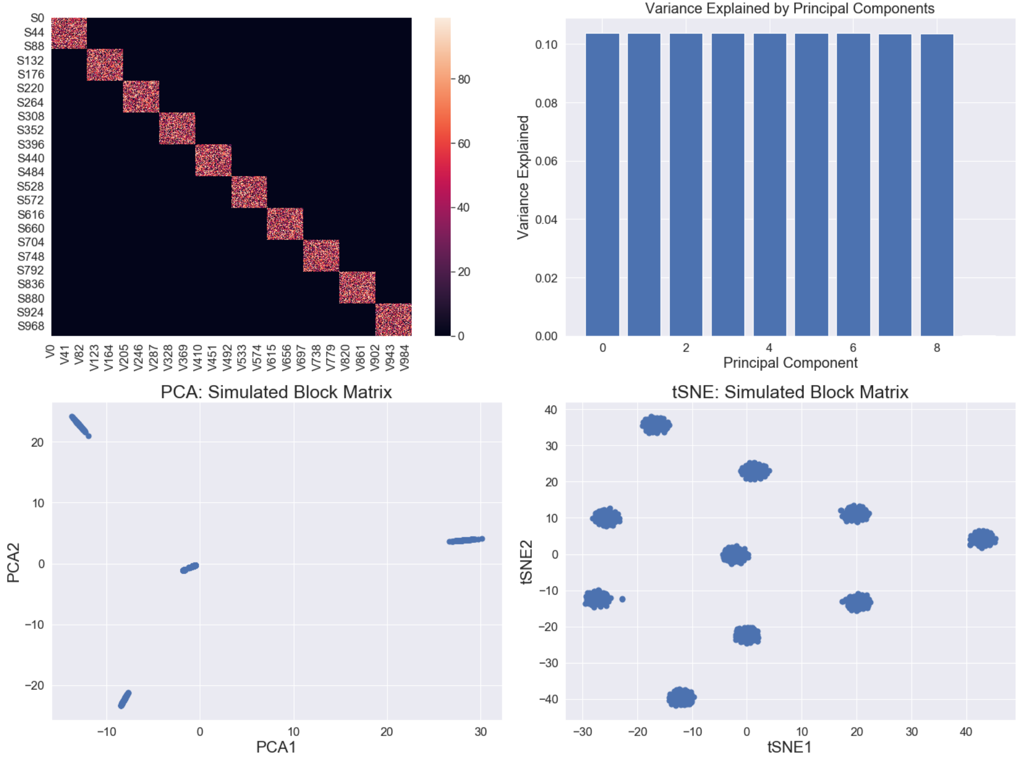
PCA vs. tSNE: when data complexity grows

Three classes of data points



PCA and tSNE tell the same story

Ten classes of data points

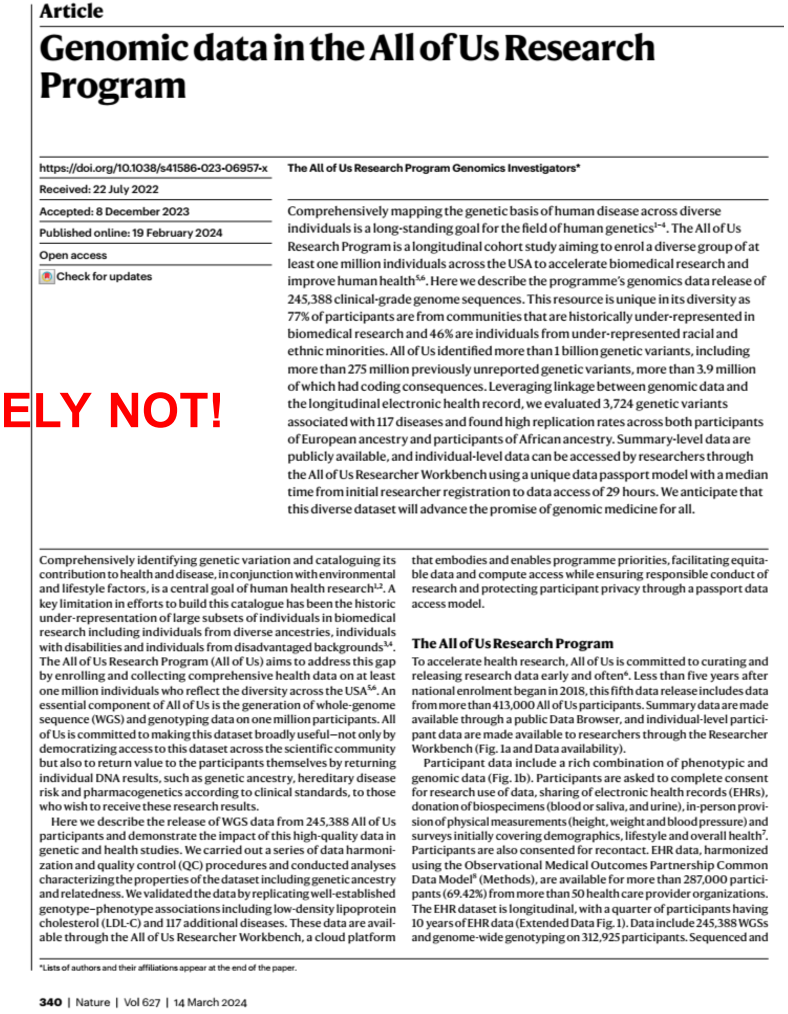


tSNE is more informative than PCA

Oskolkov et al., unpublished

If I was a naive PopGen person, may I say: UMAP is like PCA but with sharper clusters?

ABSOLUTELY NOT!



Article

Genomic data in the All of Us Research Program

https://doi.org/10.1038/s41586-023-06957-x

Received: 22 July 2022

Accepted: 8 December 2023

Published online: 19 February 2024

Open access

Check for updates

The All of Us Research Program Genomics Investigators*

Comprehensively mapping the genetic basis of human disease across diverse individuals is a long-standing goal for the field of human genetics[1–4]. The All of Us Research Program is a longitudinal cohort study aiming to enrol a diverse group of at least one million individuals across the USA to accelerate biomedical research and improve human health[5,6]. Here we describe the programme's genomics data release of 245,388 clinical-grade genome sequences. This resource is unique in its diversity as 77% of participants are from communities that are historically under-represented in biomedical research and 46% are individuals from under-represented racial and ethnic minorities. All of Us identified more than 1 billion genetic variants, including more than 275 million previously unreported genetic variants, more than 3.9 million of which had coding consequences. Leveraging linkage between genomic data and the longitudinal electronic health record, we evaluated 3,724 genetic variants associated with 117 diseases and found high replication rates across both participants of European ancestry and participants of African ancestry. Summary-level data are publicly available, and individual-level data can be accessed by researchers through the All of Us Researcher Workbench using a unique data passport model with a median time from initial researcher registration to data access of 29 hours. We anticipate that this diverse dataset will advance the promise of genomic medicine for all.

Comprehensively identifying genetic variation and cataloguing its contribution to health and disease, in conjunction with environmental and lifestyle factors, is a central goal of human health research[1,2]. A key limitation in efforts to build this catalogue has been the historic under-representation of large subsets of individuals in biomedical research including individuals from diverse ancestries, individuals with disabilities and individuals from disadvantaged backgrounds[3,4]. The All of Us Research Program (All of Us) aims to address this gap by enrolling and collecting comprehensive health data on at least one million individuals who reflect the diversity across the USA[5,6]. An essential component of All of Us is the generation of whole-genome sequence (WGS) and genotyping data on one million participants. All of Us is committed to making this dataset broadly useful—not only by democratizing access to this dataset across the scientific community but also to return value to the participants themselves by returning individual DNA results, such as genetic ancestry, hereditary disease risk and pharmacogenetics according to clinical standards, to those who wish to receive these research results.

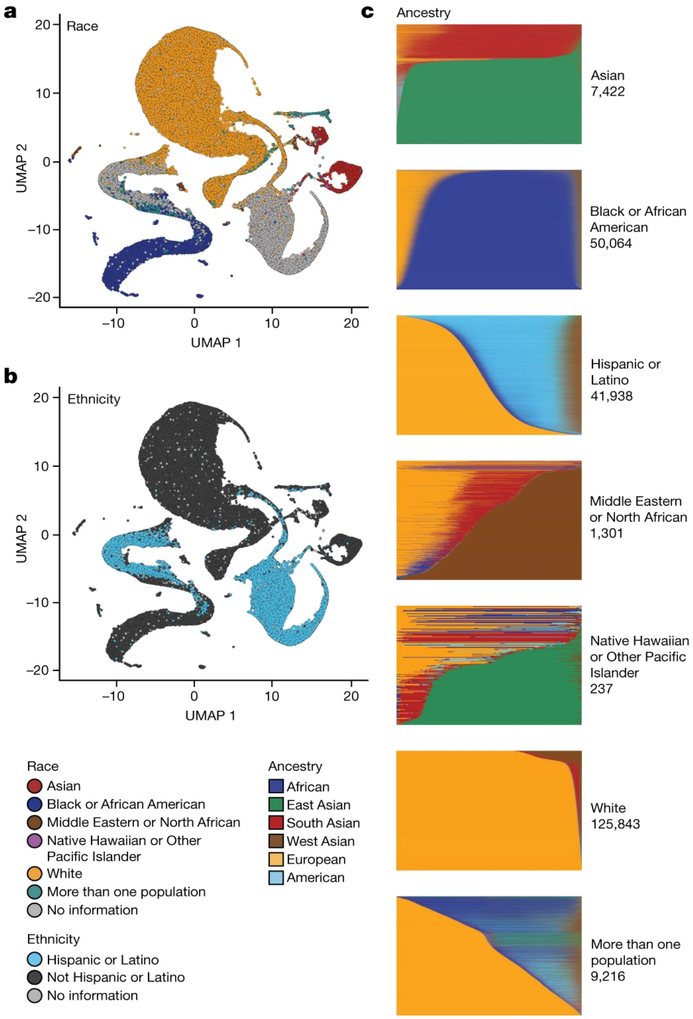
Here we describe the release of WGS data from 245,388 All of Us participants and demonstrate the impact of this high-quality data in genetic and health studies. We carried out a series of data harmonization and quality control (QC) procedures and conducted analyses characterizing the properties of the dataset including genetic ancestry and relatedness. We validated the data by replicating well-established genotype–phenotype associations including low-density lipoprotein cholesterol (LDL-C) and 117 additional diseases. These data are available through the All of Us Researcher Workbench, a cloud platform that embodies and enables programme priorities, facilitating equitable data and compute access while ensuring responsible conduct of research and protecting participant privacy through a passport data access model.

The All of Us Research Program

To accelerate health research, All of Us is committed to curating and releasing research data early and often[6]. Less than five years after national enrolment began in 2018, this fifth data release includes data from more than 413,000 All of Us participants. Summary data are made available through a public Data Browser, and individual-level participant data are made available to researchers through the Researcher Workbench (Fig. 1a and Data availability).

Participant data include a rich combination of phenotypic and genomic data (Fig. 1b). Participants are asked to complete consent for research use of data, sharing of electronic health records (EHRs), donation of biospecimens (blood or saliva, and urine), in-person provision of physical measurements (height, weight and blood pressure) and surveys initially covering demographics, lifestyle and overall health[7]. Participants are also consented for recontact. EHR data, harmonized using the Observational Medical Outcomes Partnership Common Data Model[8] (Methods), are available for more than 287,000 participants (69.42%) from more than 50 health care provider organizations. The EHR dataset is longitudinal, with a quarter of participants having 10 years of EHR data (Extended Data Fig. 1). Data include 245,388 WGSs and genome-wide genotyping on 312,925 participants. Sequenced and

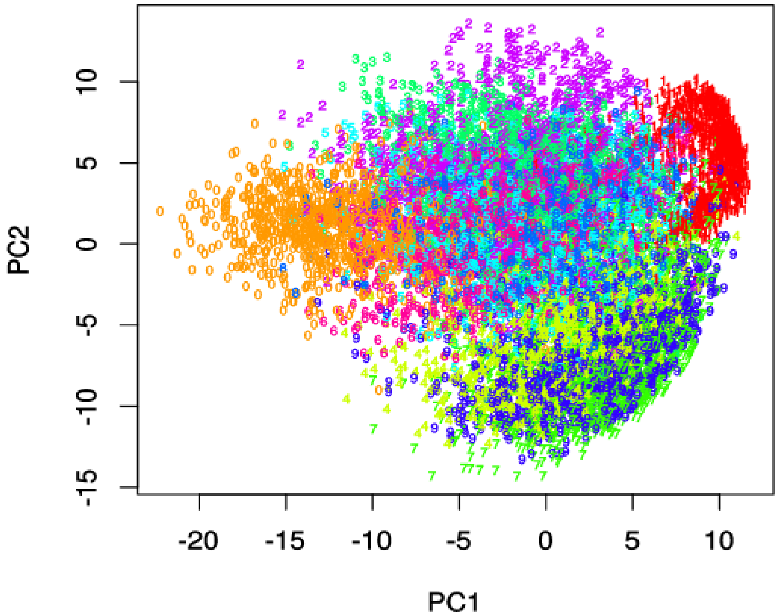
*Lists of authors and their affiliations appear at the end of the paper.



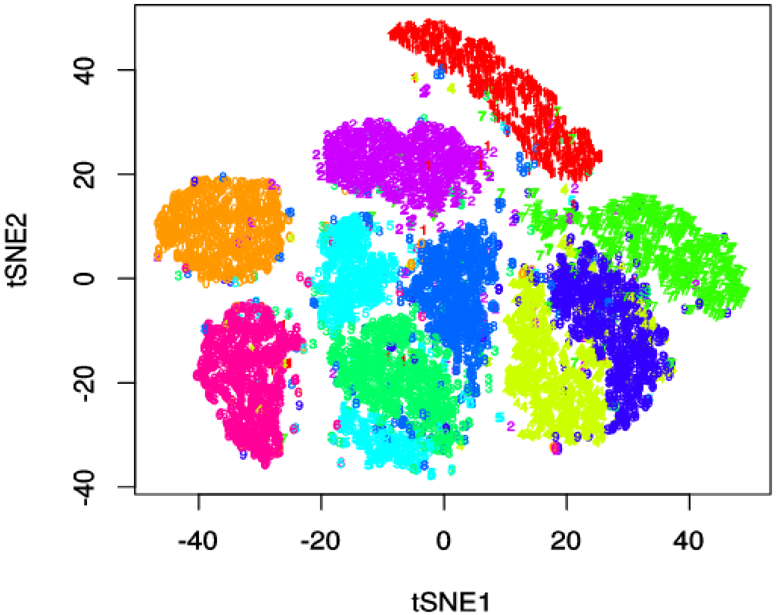
Fundamentals of linear and non-linear dimension reduction

Dimension reduction: more than visualization

PCA PLOT WITH PRCOMP

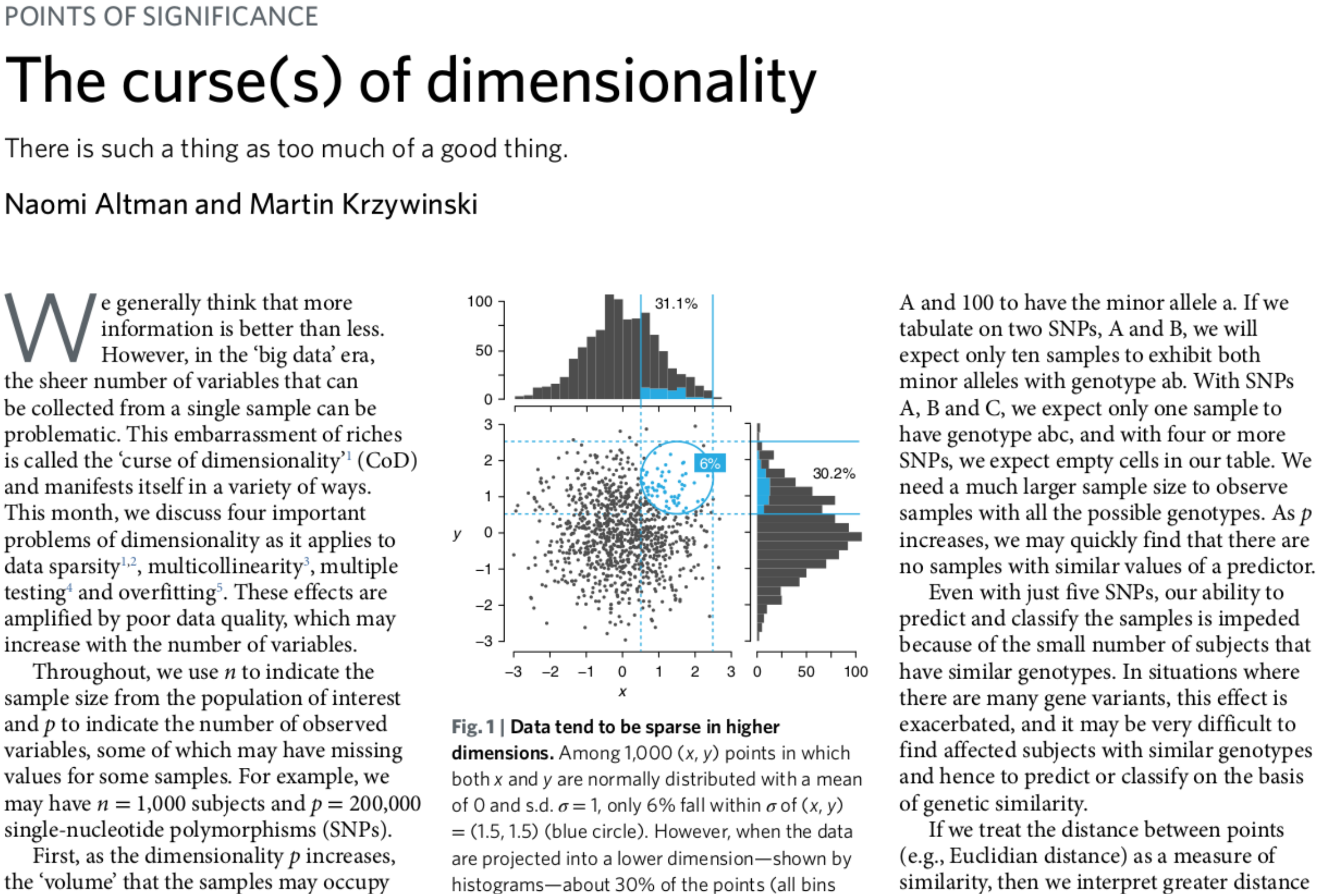


tSNE MNIST



The goal of dimension reduction is not only visualization but also reducing dimensions

Literature on the Curse of Dimensionality



POINTS OF SIGNIFICANCE

The curse(s) of dimensionality

There is such a thing as too much of a good thing.

Naomi Altman and Martin Krzywinski

We generally think that more information is better than less. However, in the 'big data' era, the sheer number of variables that can be collected from a single sample can be problematic. This embarrassment of riches is called the 'curse of dimensionality'[1] (CoD) and manifests itself in a variety of ways. This month, we discuss four important problems of dimensionality as it applies to data sparsity[1,2], multicollinearity[3], multiple testing[4] and overfitting[5]. These effects are amplified by poor data quality, which may increase with the number of variables.

Throughout, we use n to indicate the sample size from the population of interest and p to indicate the number of observed variables, some of which may have missing values for some samples. For example, we may have $n = 1{,}000$ subjects and $p = 200{,}000$ single-nucleotide polymorphisms (SNPs).

First, as the dimensionality p increases, the 'volume' that the samples may occupy

Fig. 1 | Data tend to be sparse in higher dimensions. Among 1,000 (x, y) points in which both x and y are normally distributed with a mean of 0 and s.d. $\sigma = 1$, only 6% fall within σ of (x, y) = (1.5, 1.5) (blue circle). However, when the data are projected into a lower dimension—shown by histograms—about 30% of the points (all bins

A and 100 to have the minor allele a. If we tabulate on two SNPs, A and B, we will expect only ten samples to exhibit both minor alleles with genotype ab. With SNPs A, B and C, we expect only one sample to have genotype abc, and with four or more SNPs, we expect empty cells in our table. We need a much larger sample size to observe samples with all the possible genotypes. As p increases, we may quickly find that there are no samples with similar values of a predictor.

Even with just five SNPs, our ability to predict and classify the samples is impeded because of the small number of subjects that have similar genotypes. In situations where there are many gene variants, this effect is exacerbated, and it may be very difficult to find affected subjects with similar genotypes and hence to predict or classify on the basis of genetic similarity.

If we treat the distance between points (e.g., Euclidian distance) as a measure of similarity, then we interpret greater distance

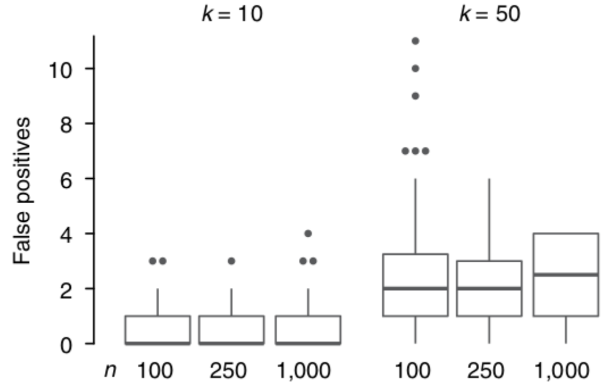
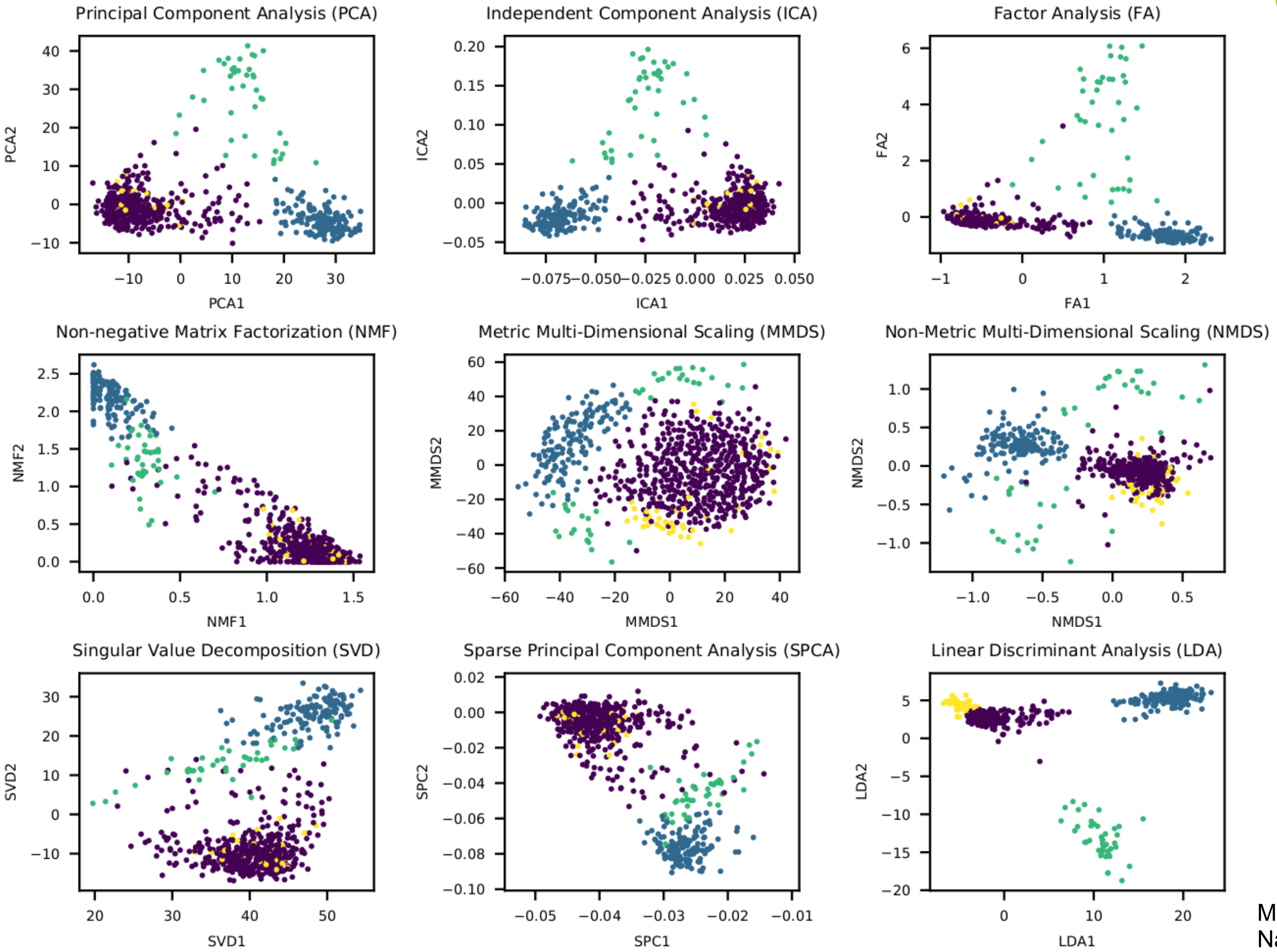


Fig. 3 | The number of false positives increases with each additional predictor. The box plots show the number of false positive regression-fit P values (tested at $\alpha = 0.05$) of 100 simulated multiple regression fits on various numbers of samples ($n = 100$, 250 and 1,000) in the presence of one true predictor and $k = 10$ and 50 extraneous uncorrelated predictors. Box plots show means (black center lines), 25th and 75th percentiles (box edges), and minimum and maximum values (whiskers). Outliers (dots) are jittered.

Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018 Jun;15(6):399-400. doi: 10.1038/s41592-018-0019-x. PMID: 29855577.

Correcting for multiple testing does not solve the problem of too many false-positive hits

Linear dimensionality reduction



Principal Component Analysis (PCA)

Independent Component Analysis (ICA)

Factor Analysis (FA)

Non-negative Matrix Factorization (NMF)

Metric Multi-Dimensional Scaling (MMDS)

Non-Metric Multi-Dimensional Scaling (NMDS)

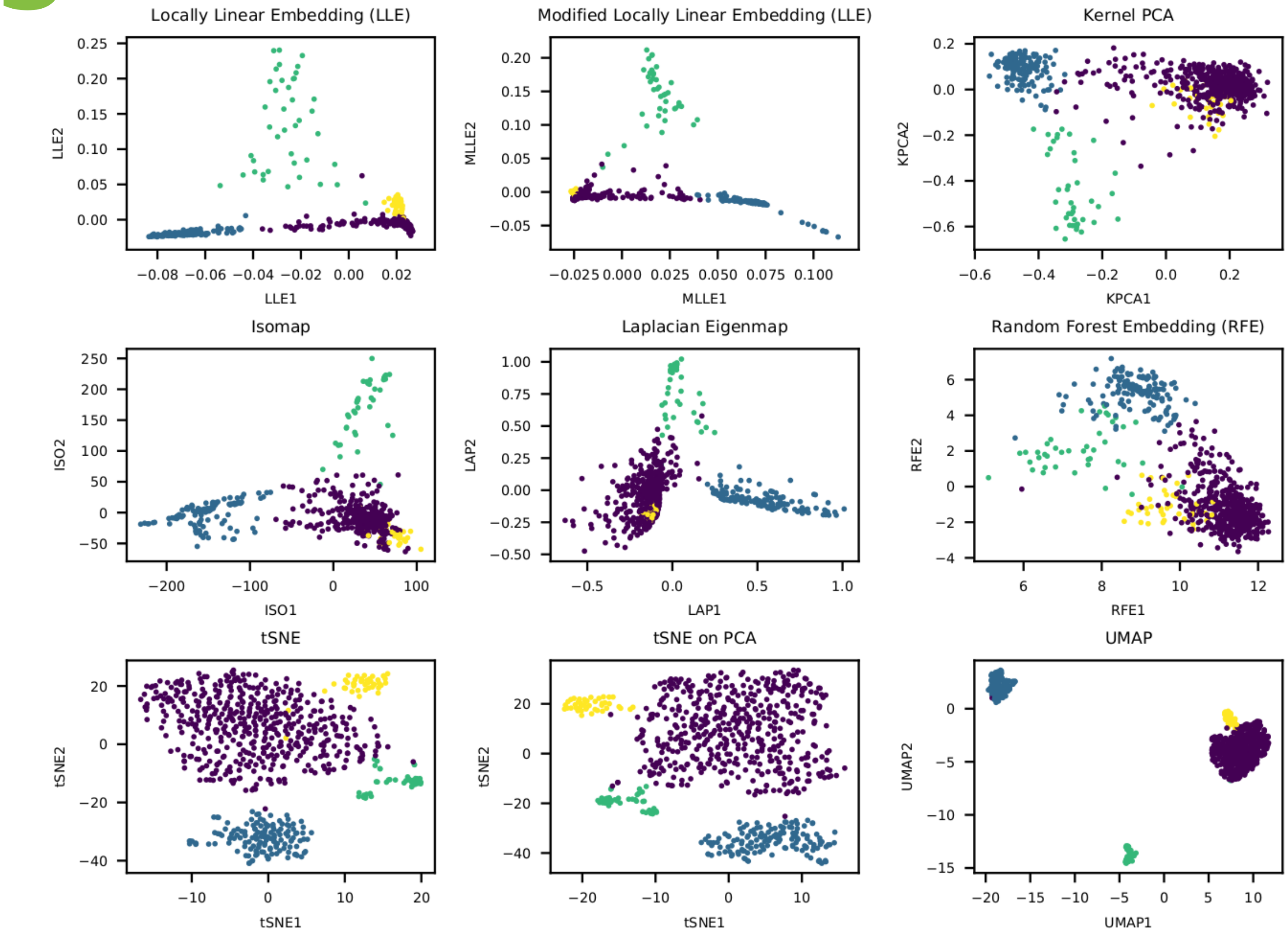
Singular Value Decomposition (SVD)

Sparse Principal Component Analysis (SPCA)

Linear Discriminant Analysis (LDA)

M. Bartoschek, N. Oskolkov et al.,
Nature Communications 2018

Non-linear dimensionality reduction



Locally Linear Embedding (LLE)

Modified Locally Linear Embedding (LLE)

Kernel PCA

Isomap

Laplacian Eigenmap

Random Forest Embedding (RFE)

tSNE

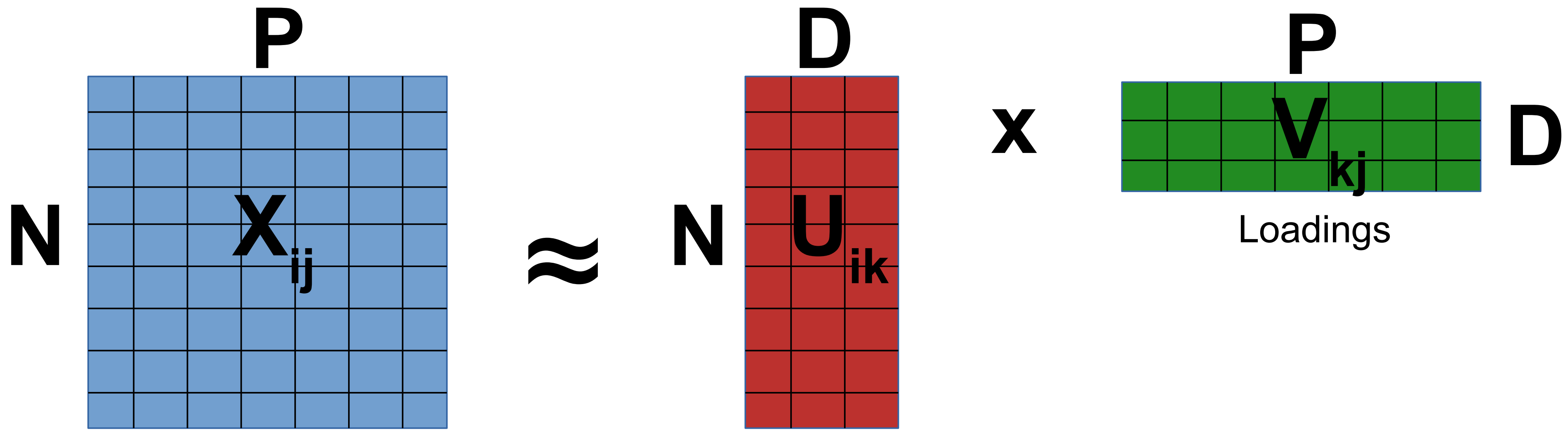
tSNE on PCA

UMAP

M. Bartoschek, N. Oskolkov et al.,
Nature Communications 2018

Linear dimension reduction: matrix factorization

$$\mathbf{X_{ij}} \approx \mathbf{U_{ik}}\mathbf{V_{kj}}$$



Data

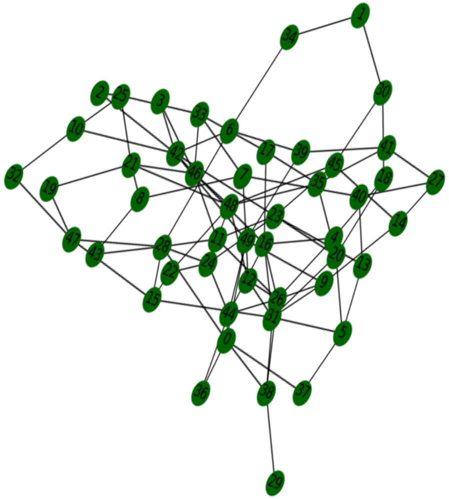
Low-dimensional data representation (embeddings)

$$\mathrm{Loss} = \sum_{i=1}^{N}\sum_{j=1}^{P}(\mathbf{X_{ij}} - \mathbf{U_{ik}}\mathbf{V_{kj}})^2$$

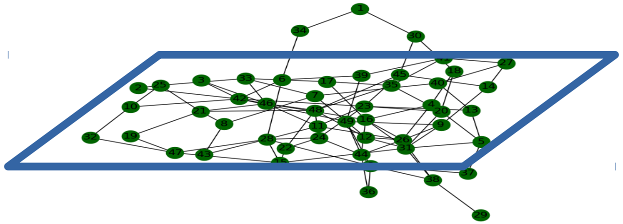
Non-linear dimension reduction: neighborhood graph

1) Construct high-dimensional graph

p_{ij}

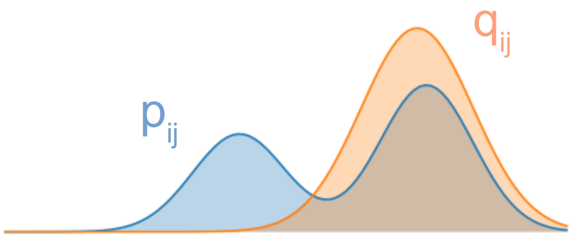


2) Construct low-dimensional graph



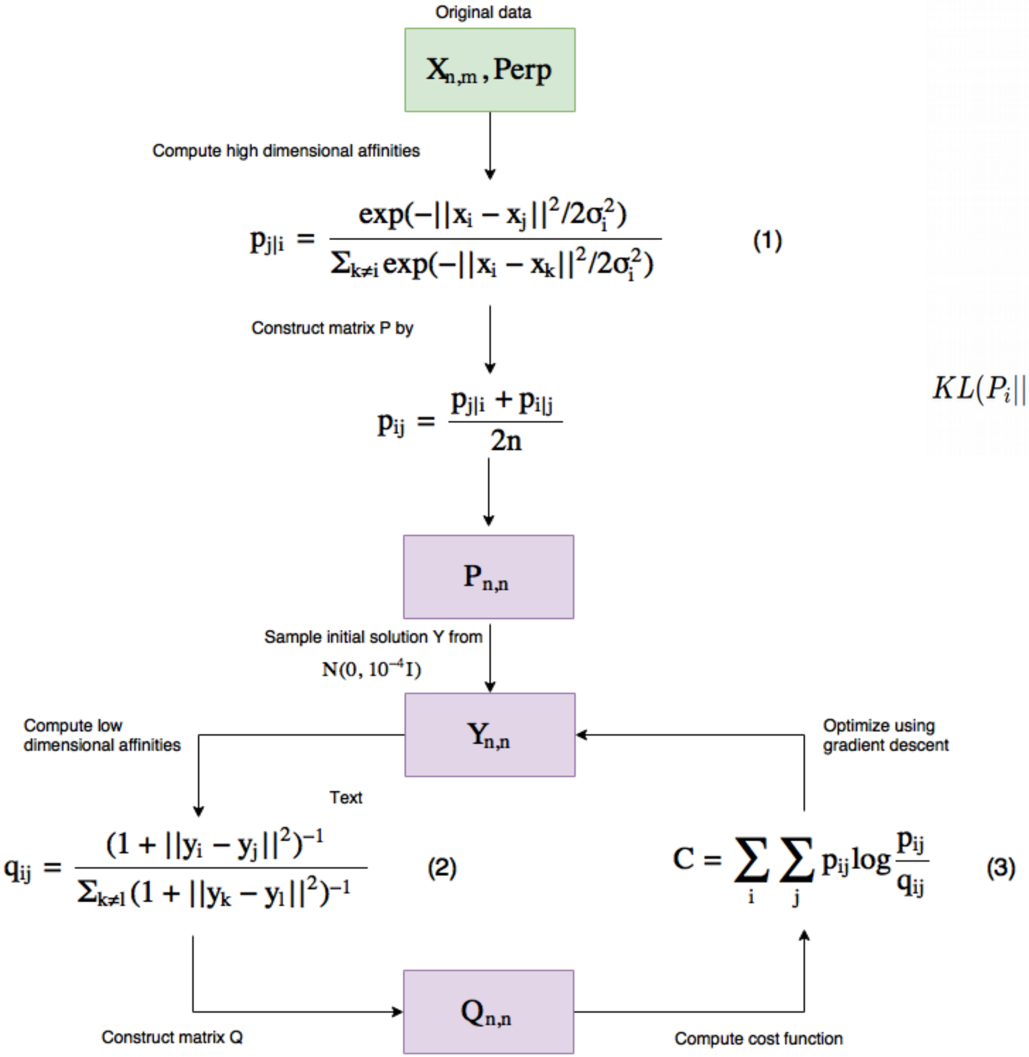
q_{ij}

3) Collapse the graphs together



Kullback-Leibler divergence

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

Text

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

Construct matrix Q

$Q_{n,n}$

Compute cost function

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

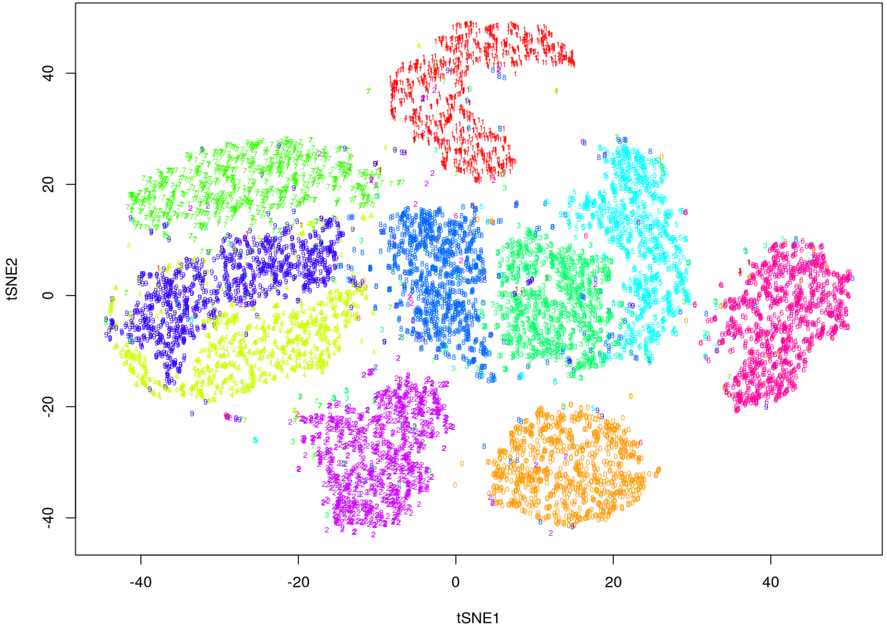
Optimize using gradient descent

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

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

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

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



Limitations of tSNE and promise of UMAP

tSNE does not scale for large data sets?

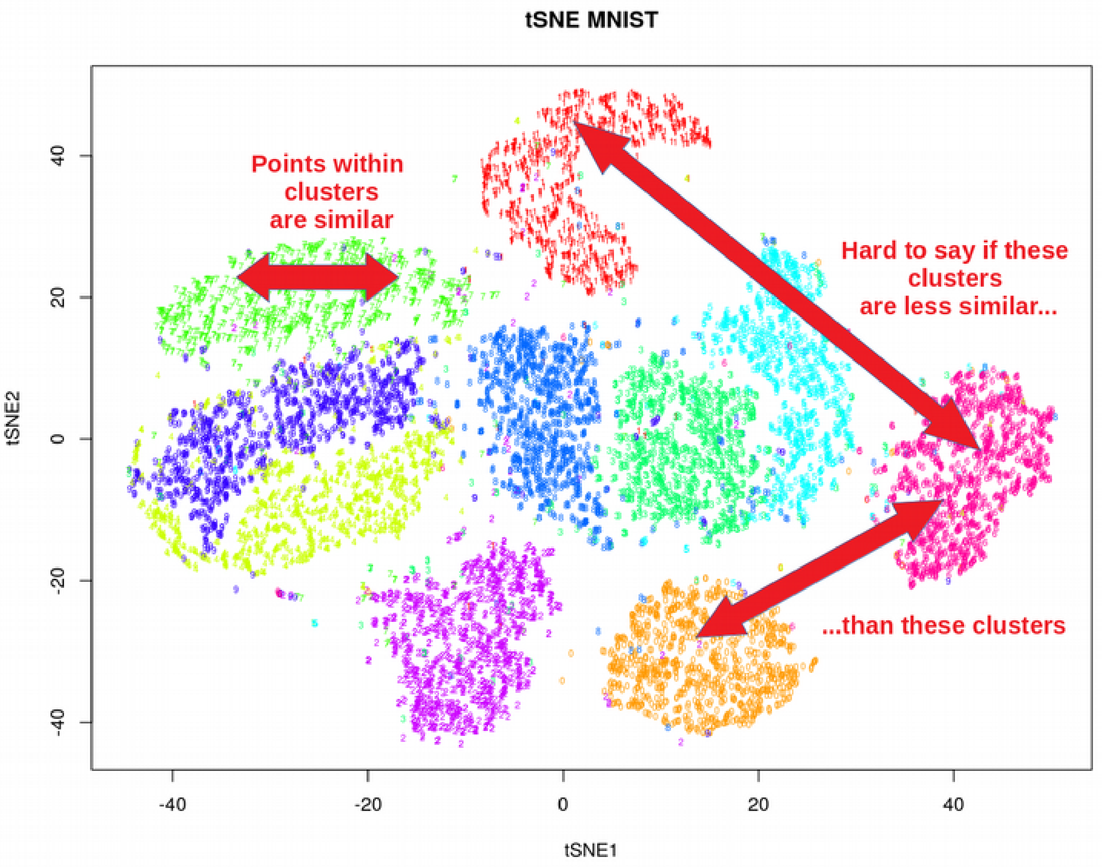
tSNE does not preserve global structure?

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?



Here is the caveat for PopGen analysis:
meaningless inter-cluster distances hinder interpretation of functional (genetic) relation between the clusters

How is UMAP different from tSNE

UMAP uses local connectivity for high-dim probabilities

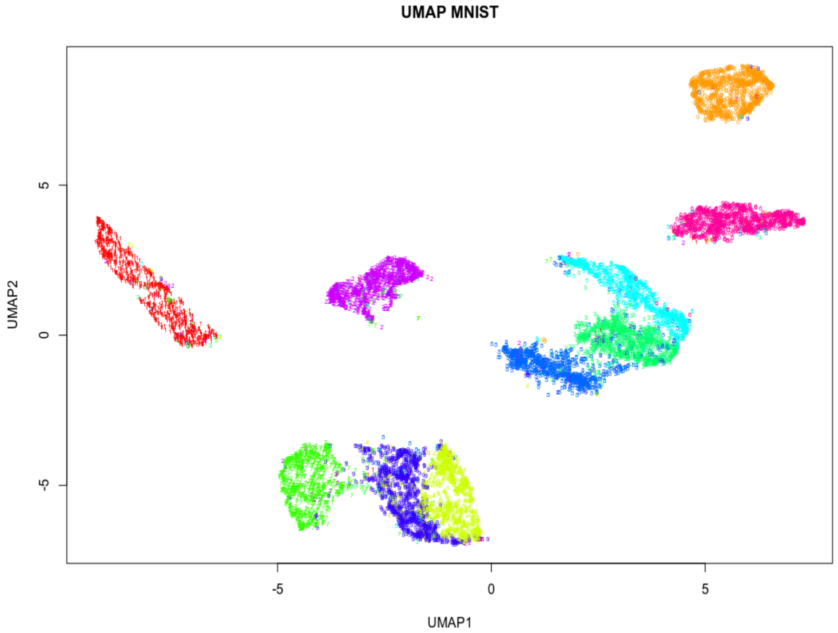
$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}$$

UMAP does not normalize probabilities (speed-up)

UMAP can deliver a number of components for clustering

UMAP uses Laplacian Eigenmap for initialization

UMAP uses Cross-Entropy (not KL) as cost function



$$CE(X,Y) = \sum_i \sum_j \left[p_{ij}(X) \log\left(\frac{p_{ij}(X)}{q_{ij}(Y)}\right) + (1 - p_{ij}(X)) \log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right]$$

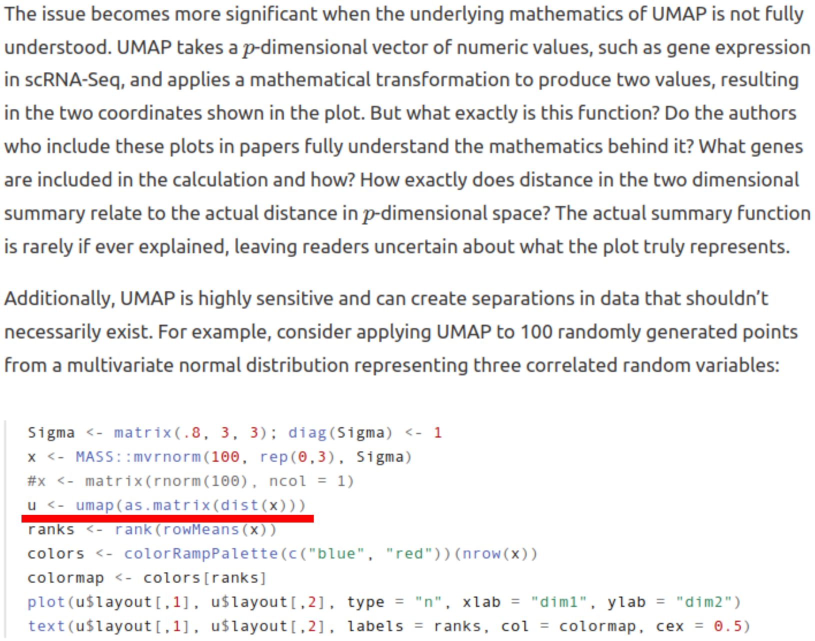
This is similar to tSNE cost function

This term is UMAP specific

Going through the post of Rafael Irizarry:

Point1: UMAP makes artificial clusters!

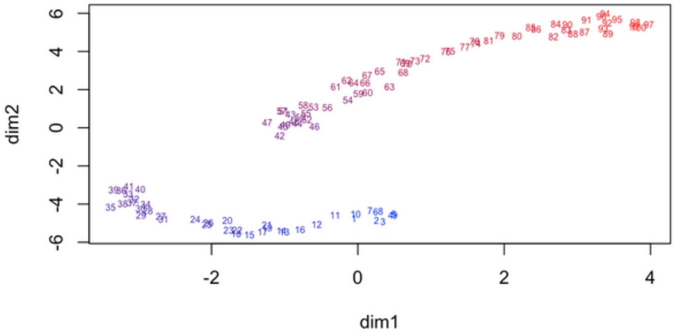
Are we looking at 3- or 100-dimensional data?

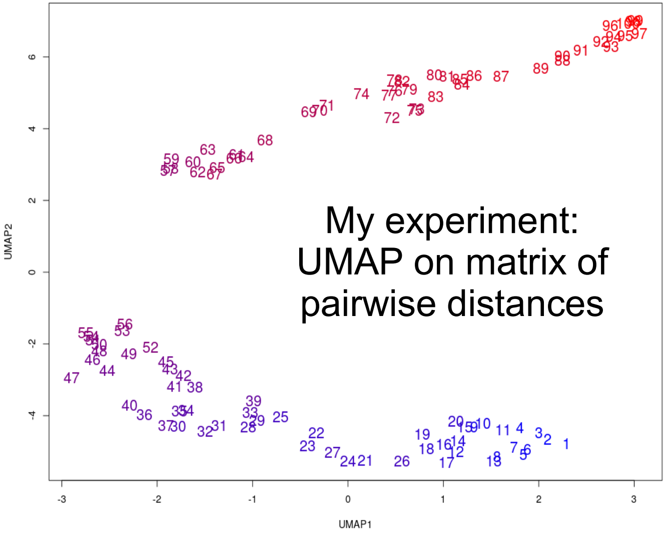


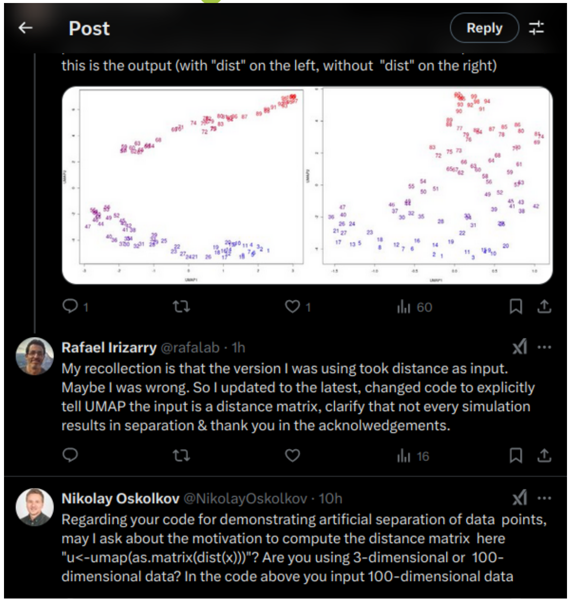
The issue becomes more significant when the underlying mathematics of UMAP is not fully understood. UMAP takes a p-dimensional vector of numeric values, such as gene expression in scRNA-Seq, and applies a mathematical transformation to produce two values, resulting in the two coordinates shown in the plot. But what exactly is this function? Do the authors who include these plots in papers fully understand the mathematics behind it? What genes are included in the calculation and how? How exactly does distance in the two dimensional summary relate to the actual distance in p-dimensional space? The actual summary function is rarely if ever explained, leaving readers uncertain about what the plot truly represents.

Additionally, UMAP is highly sensitive and can create separations in data that shouldn't necessarily exist. For example, consider applying UMAP to 100 randomly generated points from a multivariate normal distribution representing three correlated random variables:

```
Sigma <- matrix(.8, 3, 3); diag(Sigma) <- 1
x <- MASS::mvrnorm(100, rep(0,3), Sigma)
#x <- matrix(rnorm(100), ncol = 1)
u <- umap(as.matrix(dist(x)))
ranks <- rank(rowMeans(x))
colors <- colorRampPalette(c("blue", "red"))(nrow(x))
colormap <- colors[ranks]
plot(u$layout[,1], u$layout[,2], type = "n", xlab = "dim1", ylab = "dim2")
text(u$layout[,1], u$layout[,2], labels = ranks, col = colormap, cex = 0.5)
```




My experiment: UMAP on matrix of pairwise distances



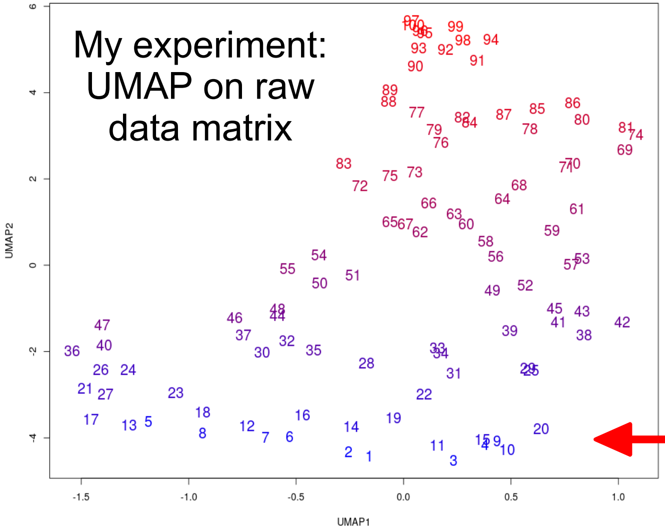
this is the output (with "dist" on the left, without "dist" on the right)

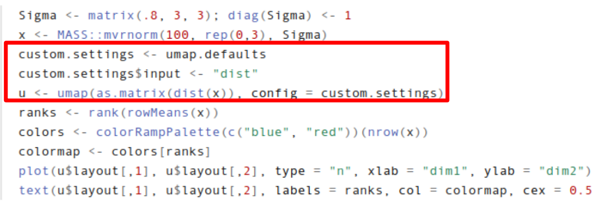
Rafael Irizarry @rafalab · 1h

My recollection is that the version I was using took distance as input. Maybe I was wrong. So I updated to the latest, changed code to explicitly tell UMAP the input is a distance matrix, clarify that not every simulation results in separation & thank you in the acknowledgements.

Nikolay Oskolkov @NikolayOskolkov · 10h

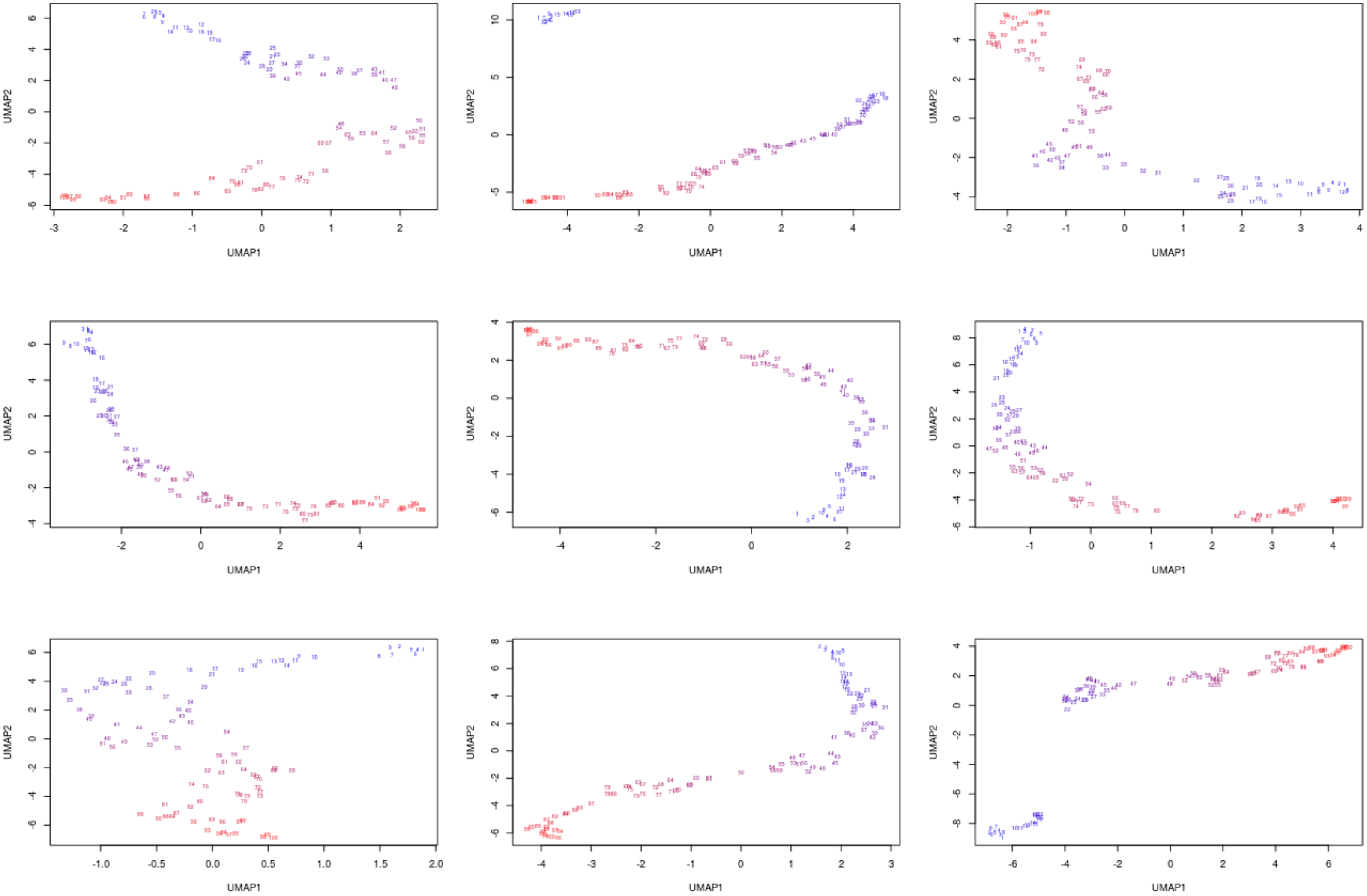
Regarding your code for demonstrating artificial separation of data points, may I ask about the motivation to compute the distance matrix here "u<-umap(as.matrix(dist(x)))"? Are you using 3-dimensional or 100-dimensional data? In the code above you input 100-dimensional data


My experiment: UMAP on raw data matrix

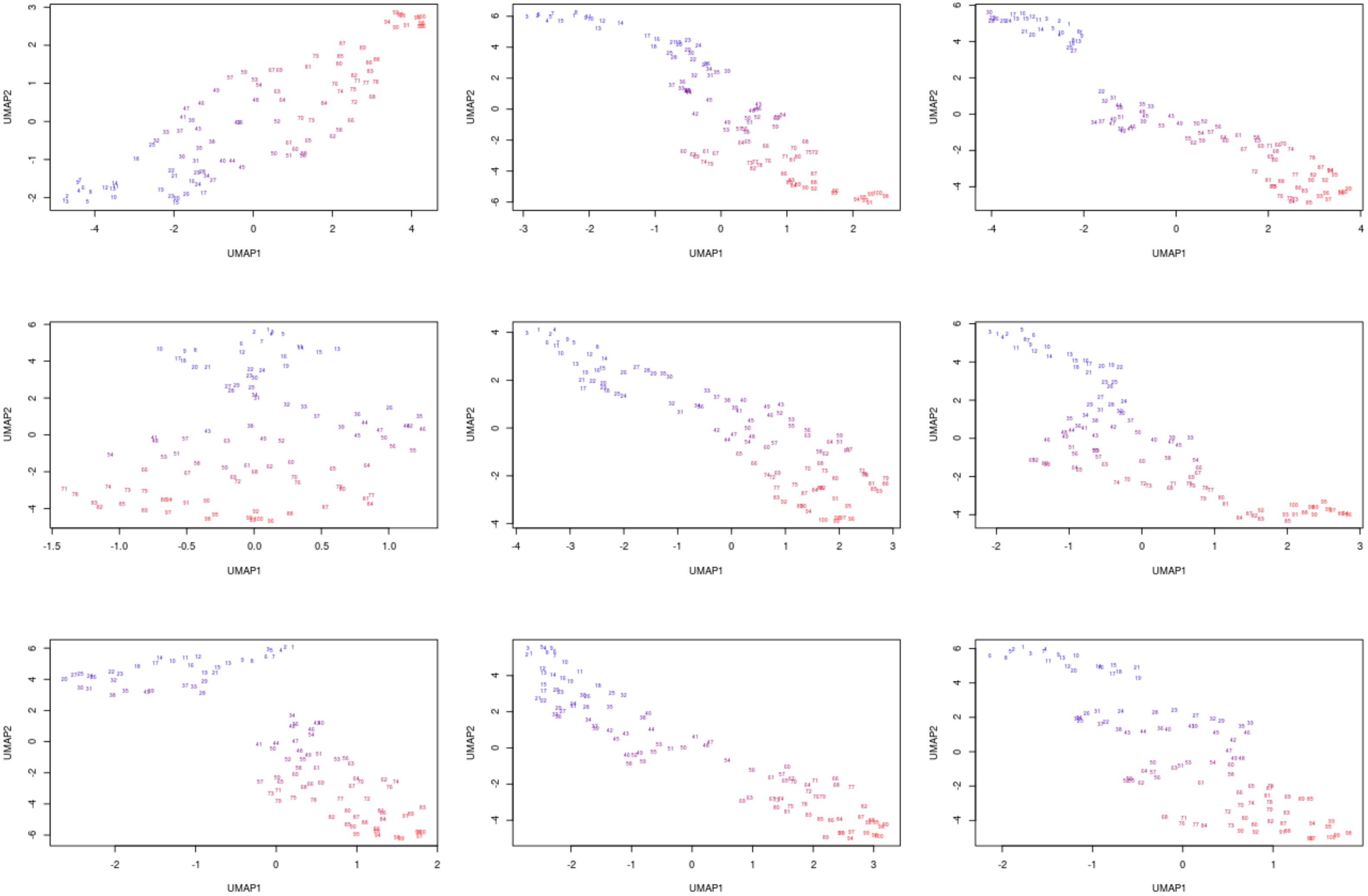


```
Sigma <- matrix(.8, 3, 3); diag(Sigma) <- 1
x <- MASS::mvrnorm(100, rep(0,3), Sigma)
custom.settings <- umap.defaults
custom.settings$input <- "dist"
u <- umap(as.matrix(dist(x)), config = custom.settings)
ranks <- rank(rowMeans(x))
colors <- colorRampPalette(c("blue", "red"))(nrow(x))
colormap <- colors[ranks]
plot(u$layout[,1], u$layout[,2], type = "n", xlab = "dim1", ylab = "dim2")
text(u$layout[,1], u$layout[,2], labels = ranks, col = colormap, cex = 0.5)
```

UMAP on matrix of pairwise distances (N=100)

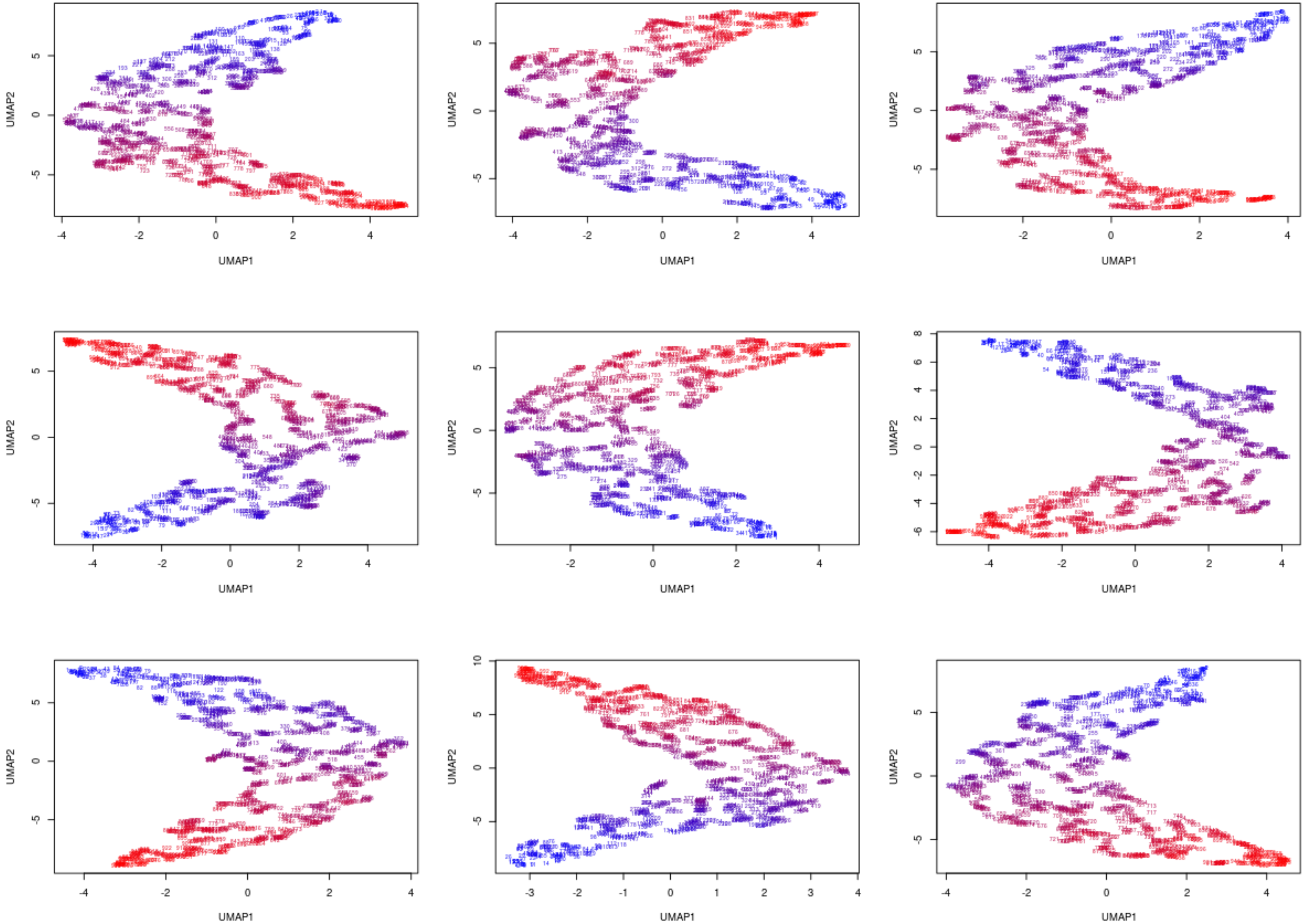


UMAP on raw data matrix (N=100)

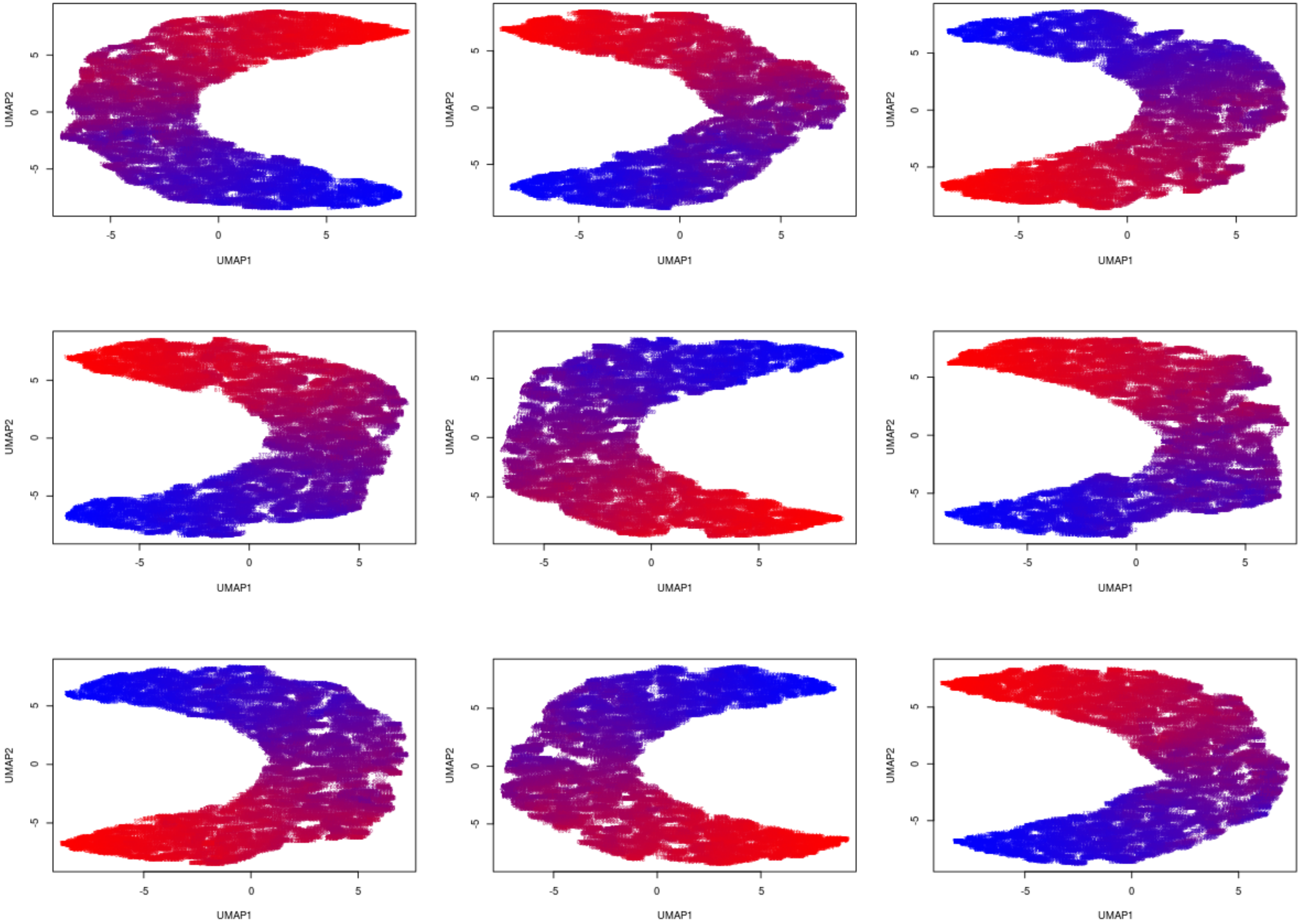


Is N=100 OK for statistics?

UMAP on raw data matrix (N=1000)



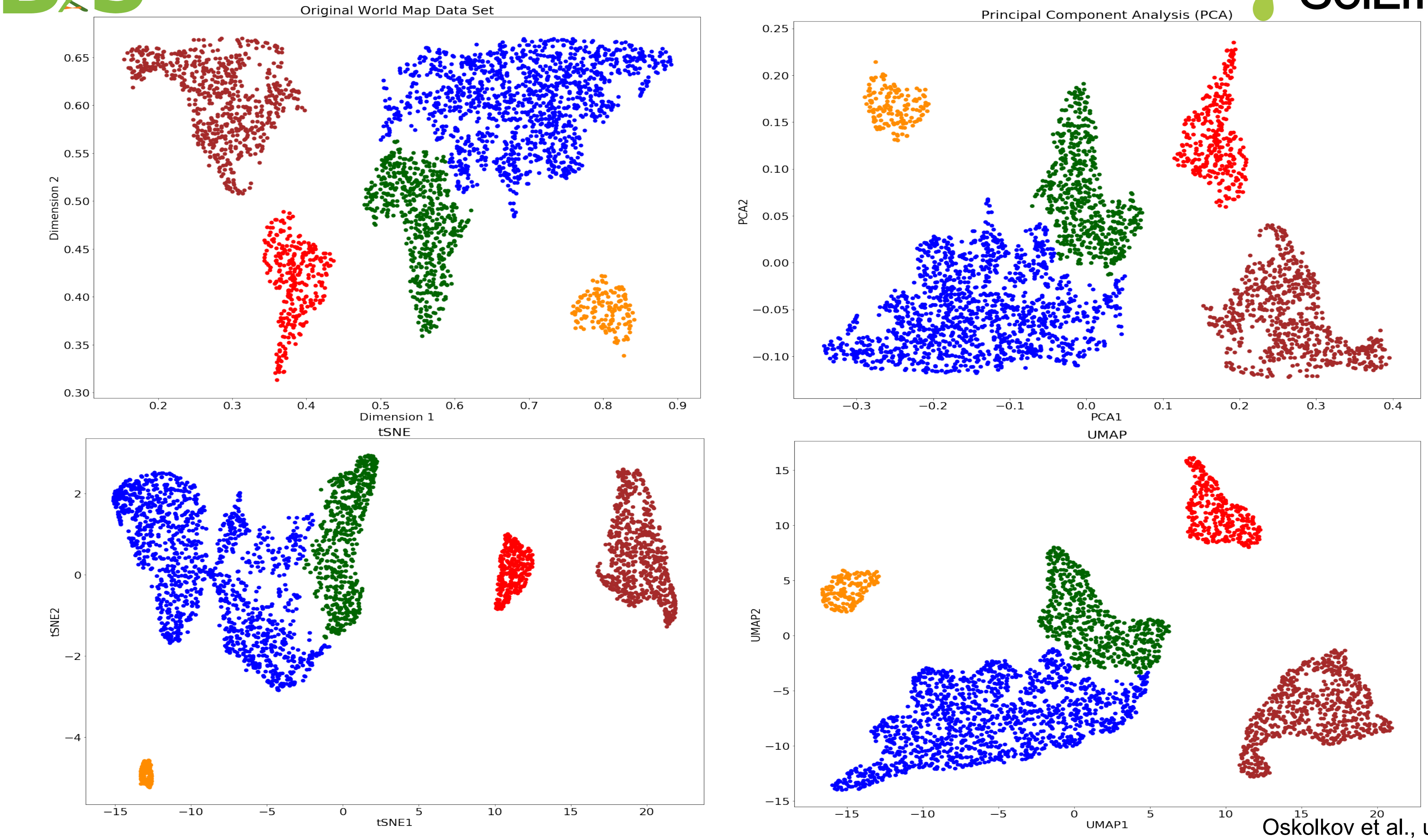
UMAP on raw data matrix (N=10 000)



Going through the post of Rafael Irizarry:

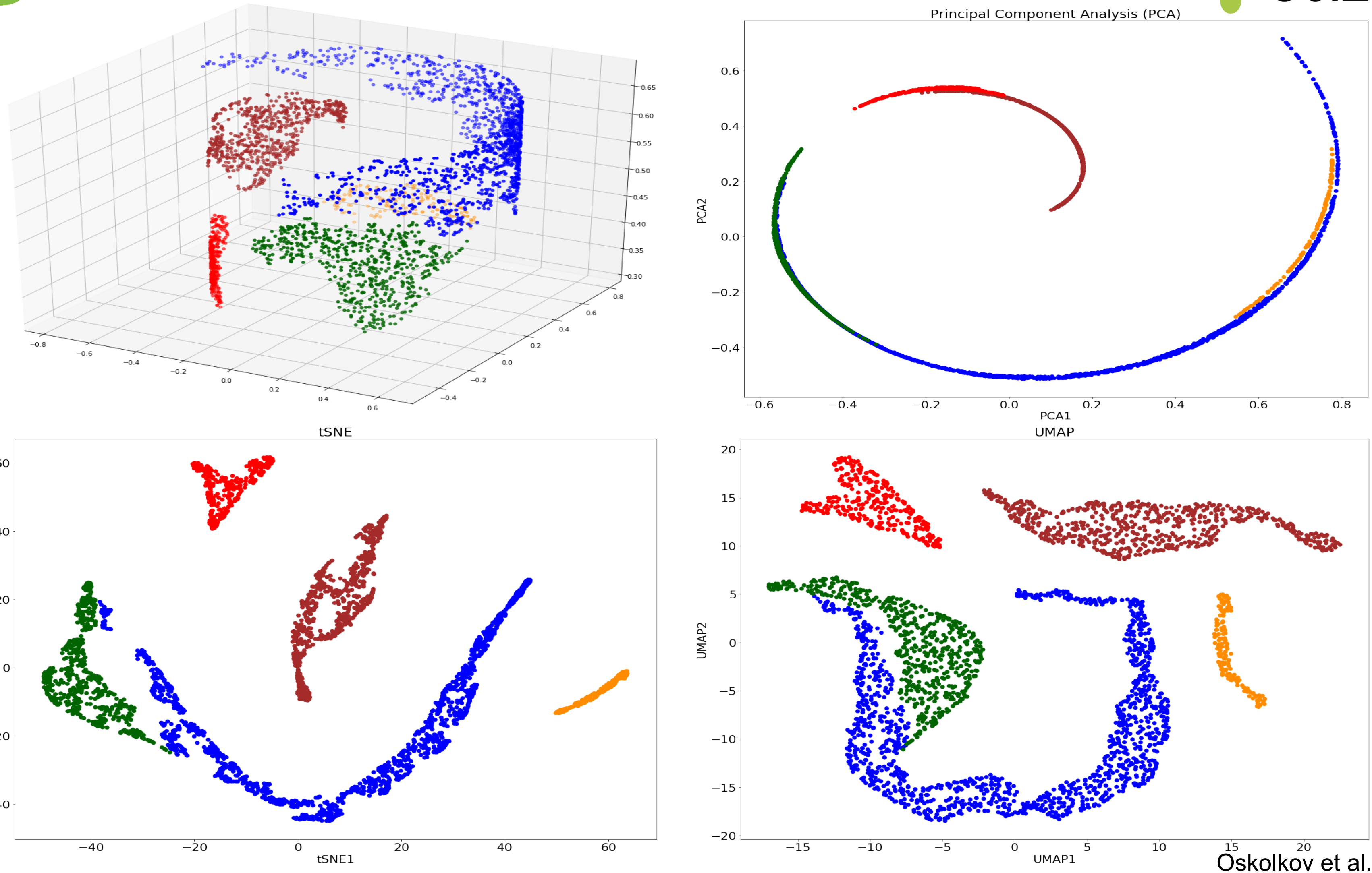
Point2: There are synthetic datasets where tSNE / UMAP fail!

PCA works fine on a linear manifold



Oskolkov et al., unpublished

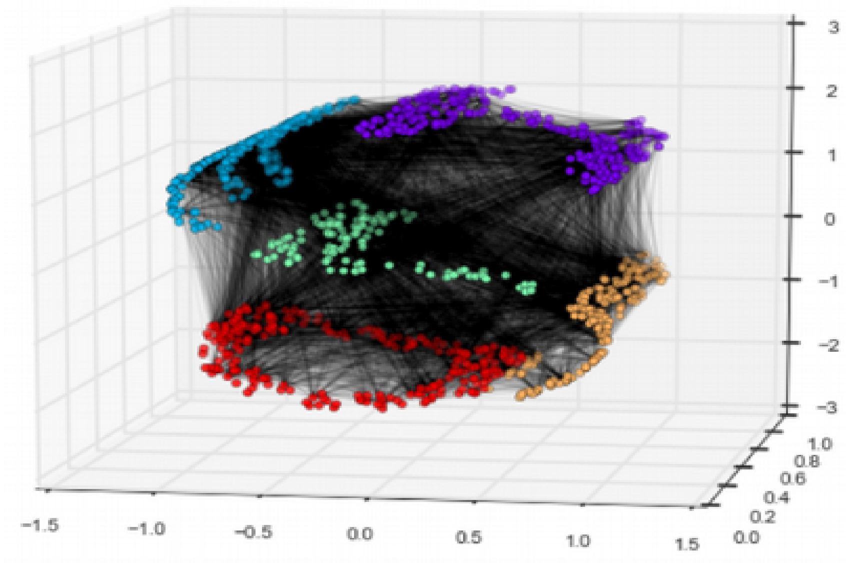
PCA vs. tSNE vs. UMAP on non-linear manifold



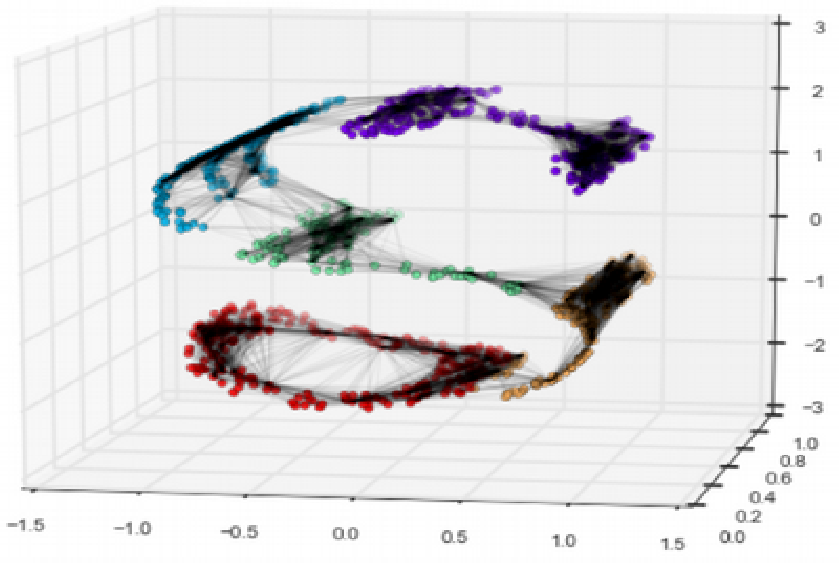
Oskolkov et al., unpublished

Swiss Roll: global vs. local distance preservation

PCA / MDS



Neighborhood graphs



“Who cares about Swiss rolls when you can embed complex real-world data nicely?”

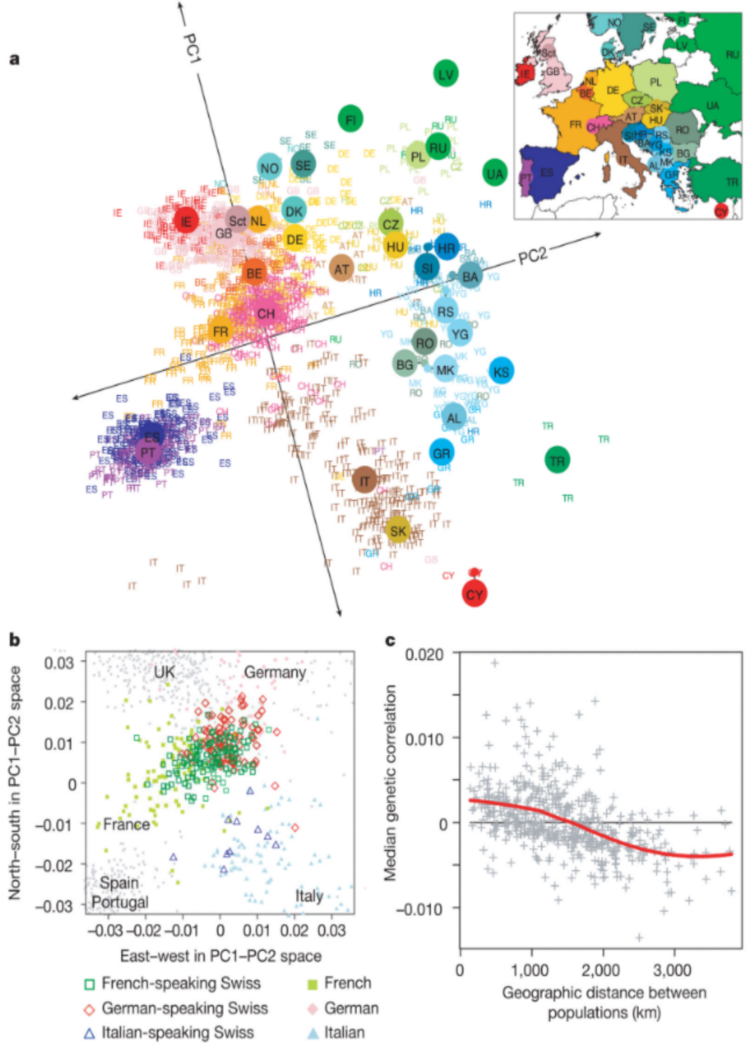
Laurens van der Maaten (author of tSNE)

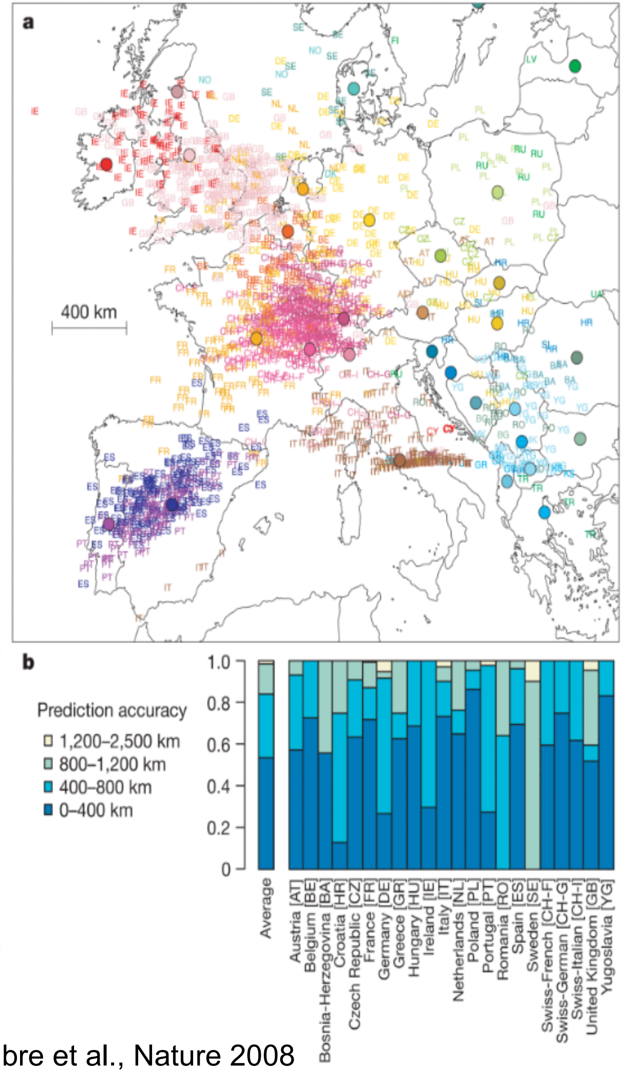
https://lvdmaaten.github.io/tsne/

Going through the post of Rafael Irizarry:

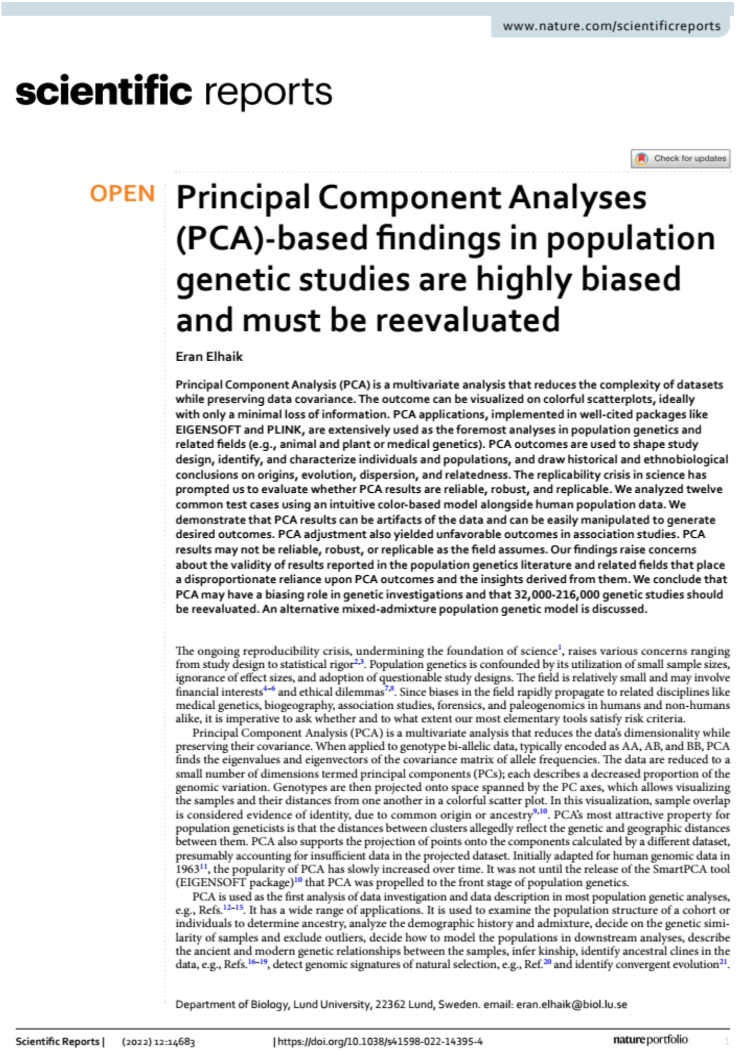
Point3: PCA better than UMAP for PopGen!

PCA golden standard in PopGen: still criticized





Novembre et al., Nature 2008



www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated

Eran Elhaik

Principal Component Analysis (PCA) is a multivariate analysis that reduces the complexity of datasets while preserving data covariance. The outcome can be visualized on colorful scatterplots, ideally with only a minimal loss of information. PCA applications, implemented in well-cited packages like EIGENSOFT and PLINK, are extensively used as the foremost analyses in population genetics and related fields (e.g., animal and plant or medical genetics). PCA outcomes are used to shape study design, identify, and characterize individuals and populations, and draw historical and ethnobiological conclusions on origins, evolution, dispersion, and relatedness. The replicability crisis in science has prompted us to evaluate whether PCA results are reliable, robust, and replicable. We analyzed twelve common test cases using an intuitive color-based model alongside human population data. We demonstrate that PCA results can be artifacts of the data and can be easily manipulated to generate desired outcomes. PCA adjustment also yielded unfavorable outcomes in association studies. PCA results may not be reliable, robust, or replicable as the field assumes. Our findings raise concerns about the validity of results reported in the population genetics literature and related fields that place a disproportionate reliance upon PCA outcomes and the insights derived from them. We conclude that PCA may have a biasing role in genetic investigations and that 32,000-216,000 genetic studies should be reevaluated. An alternative mixed-admixture population genetic model is discussed.

The ongoing reproducibility crisis, undermining the foundation of science[1], raises various concerns ranging from study design to statistical rigor[2,3]. Population genetics is confounded by its utilization of small sample sizes, ignorance of effect sizes, and adoption of questionable study designs. The field is relatively small and may involve financial interests[4–6] and ethical dilemmas[7,8]. Since biases in the field rapidly propagate to related disciplines like medical genetics, biogeography, association studies, forensics, and paleogenomics in humans and non-humans alike, it is imperative to ask whether and to what extent our most elementary tools satisfy risk criteria.

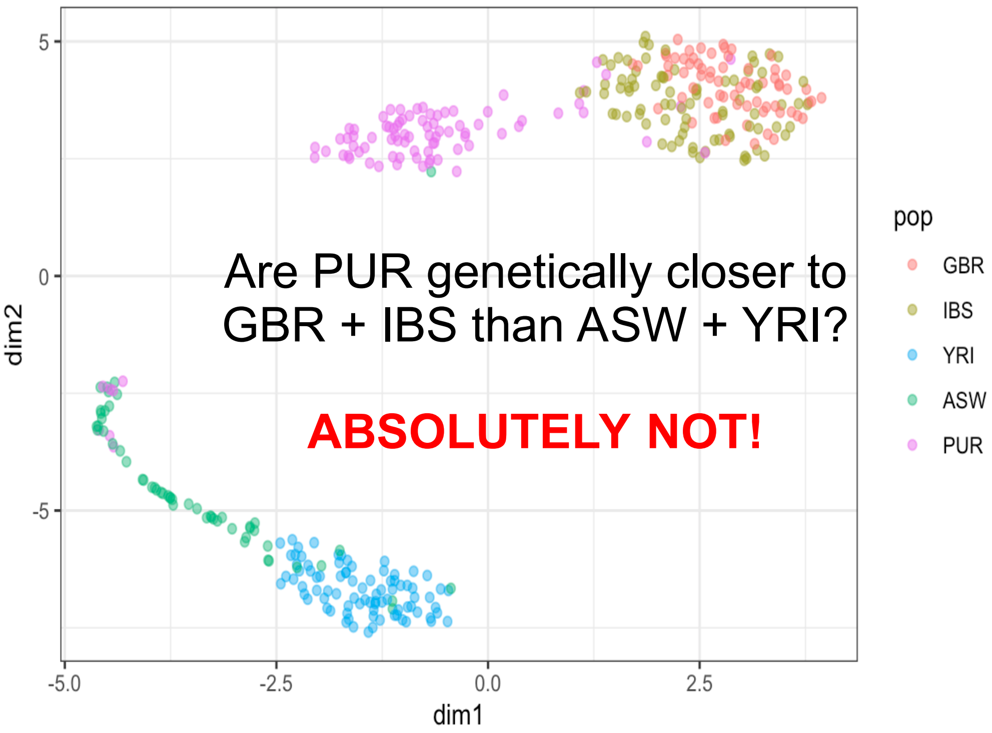
Principal Component Analysis (PCA) is a multivariate analysis that reduces the data's dimensionality while preserving their covariance. When applied to genotype bi-allelic data, typically encoded as AA, AB, and BB, PCA finds the eigenvalues and eigenvectors of the covariance matrix of allele frequencies. The data are reduced to a small number of dimensions termed principal components (PCs); each describes a decreased proportion of the genomic variation. Genotypes are then projected onto space spanned by the PC axes, which allows visualizing the samples and their distances from one another in a colorful scatter plot. In this visualization, sample overlap is considered evidence of identity, due to common origin or ancestry[9,10]. PCA's most attractive property for population geneticists is that the distances between clusters allegedly reflect the genetic and geographic distances between them. PCA also supports the projection of points onto the components calculated by a different dataset, presumably accounting for insufficient data in the projected dataset. Initially adapted for human genomic data in 1963[11], the popularity of PCA has slowly increased over time. It was not until the release of the SmartPCA tool (EIGENSOFT package)[10] that PCA was propelled to the front stage of population genetics.

PCA is used as the first analysis of data investigation and data description in most population genetic analyses, e.g., Refs.[12–15]. It has a wide range of applications. It is used to examine the population structure of a cohort or individuals to determine ancestry, analyze the demographic history and admixture, decide on the genetic similarity of samples and exclude outliers, decide how to model the populations in downstream analyses, describe the ancient and modern genetic relationships between the samples, infer kinship, identify ancestral clines in the data, e.g., Refs.[16–19], detect genomic signatures of natural selection, e.g., Ref.[20] and identify convergent evolution[21].

Department of Biology, Lund University, 22362 Lund, Sweden. email: eran.elhaik@biol.lu.se

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

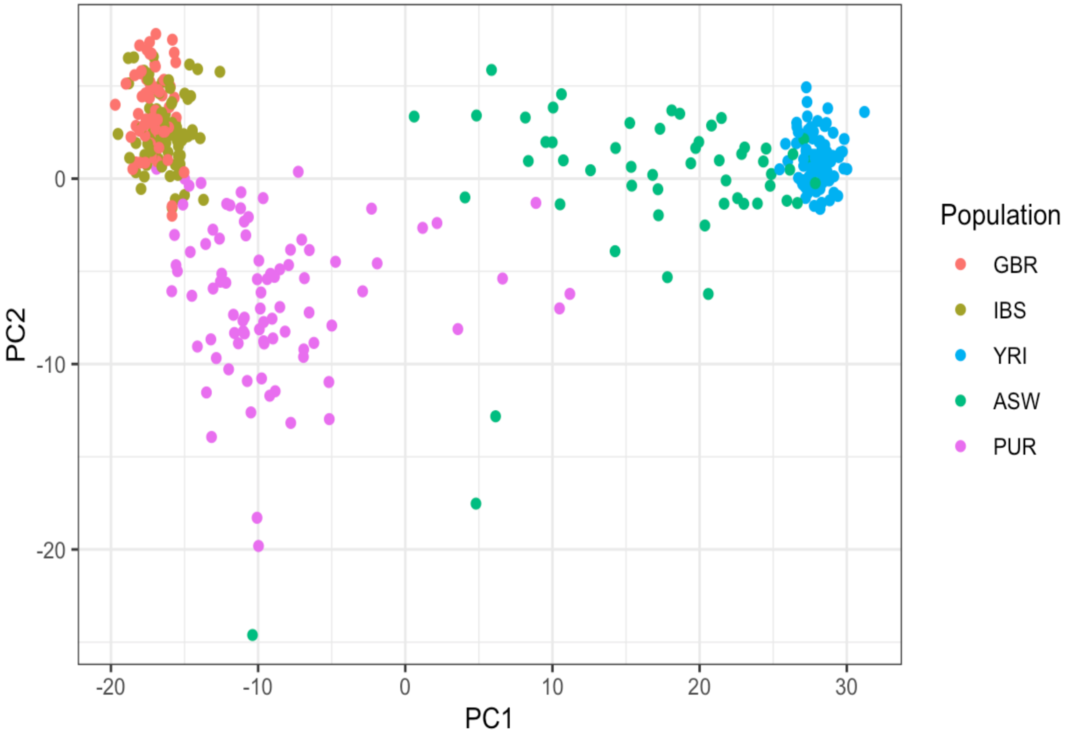
UMAP



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

ABSOLUTELY NOT!

PCA

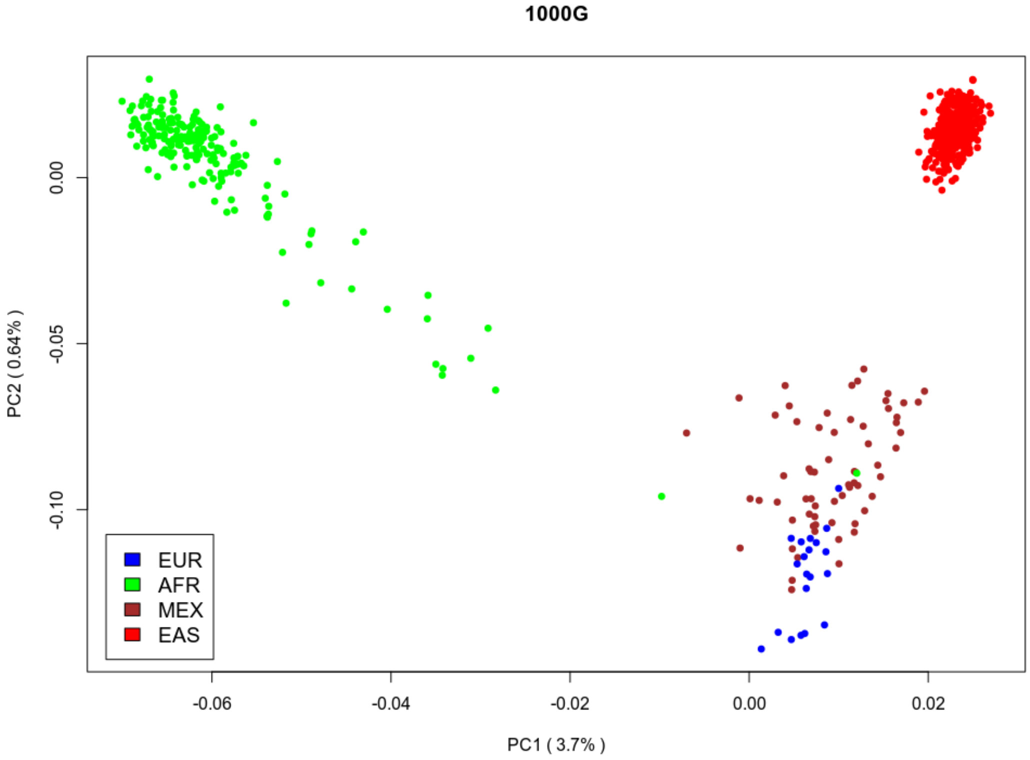


- Because of their meaningless inter-cluster distances tSNE / UMAP are less useful for population genomics than PCA.

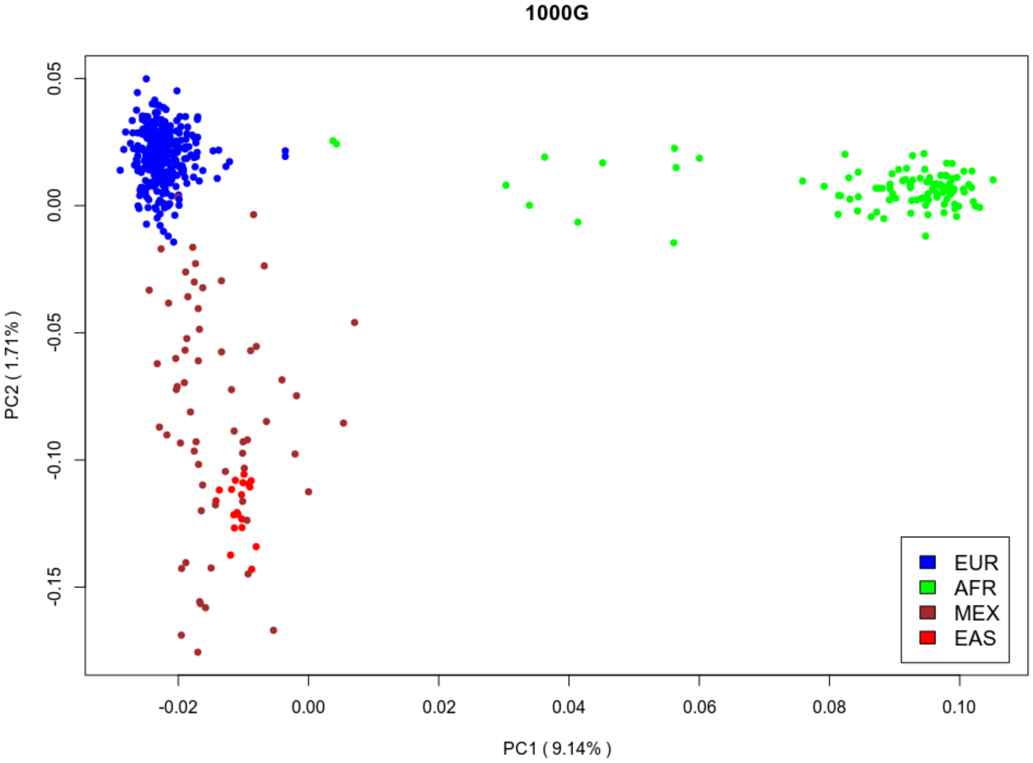
- The goal of tSNE / UMAP is to **discover clusters**, which is sufficient for Single Cell Biology but not for PopGen.

- 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!

PCA has a known pitfall: uneven sampling of populations



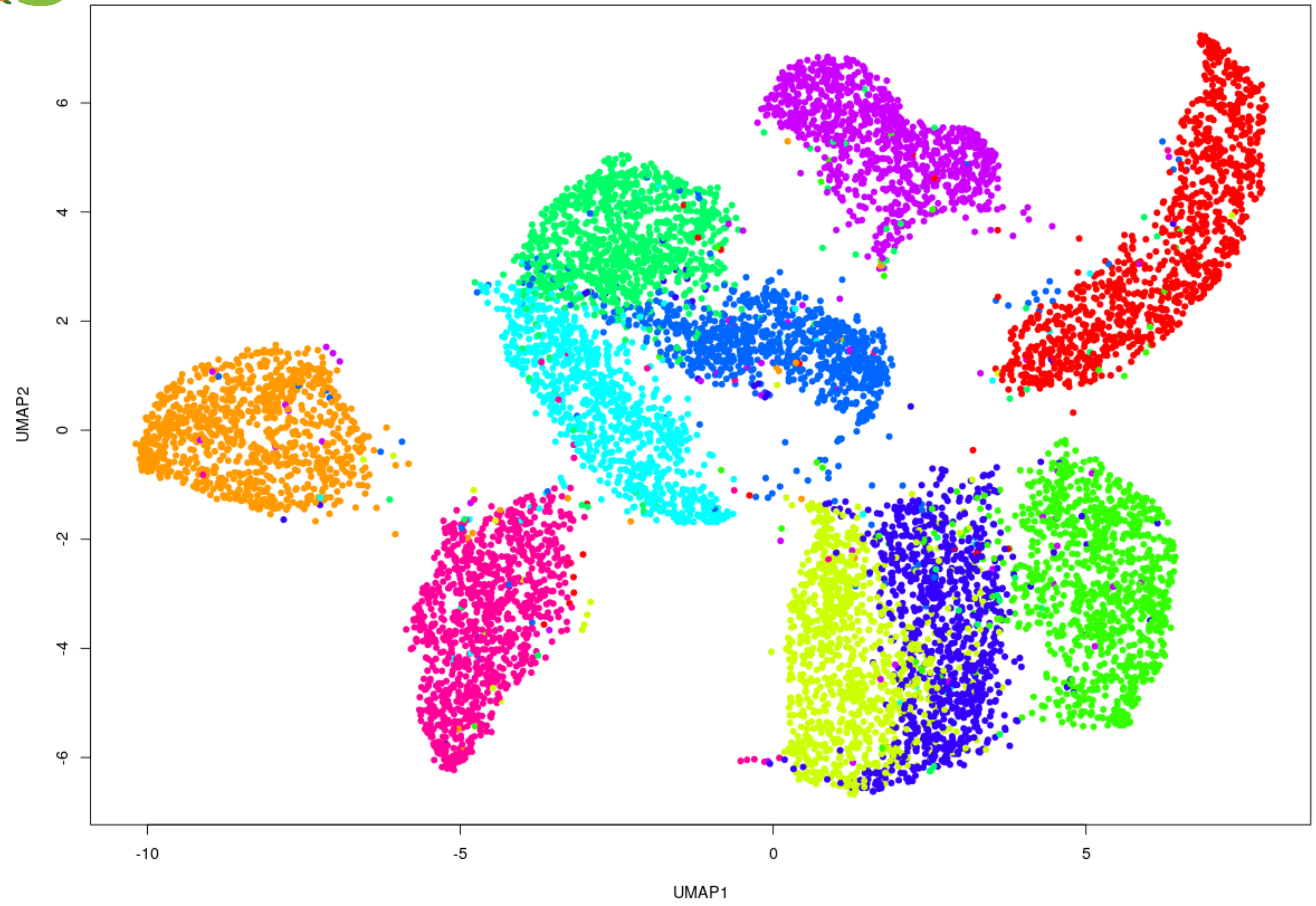
Downsampled Europeans



Downsampled Asians

Final words before you go

UMAP does not know anything about the labels



UMAP is just
a model, please
validate it!

Validation is a
main criterion
of success

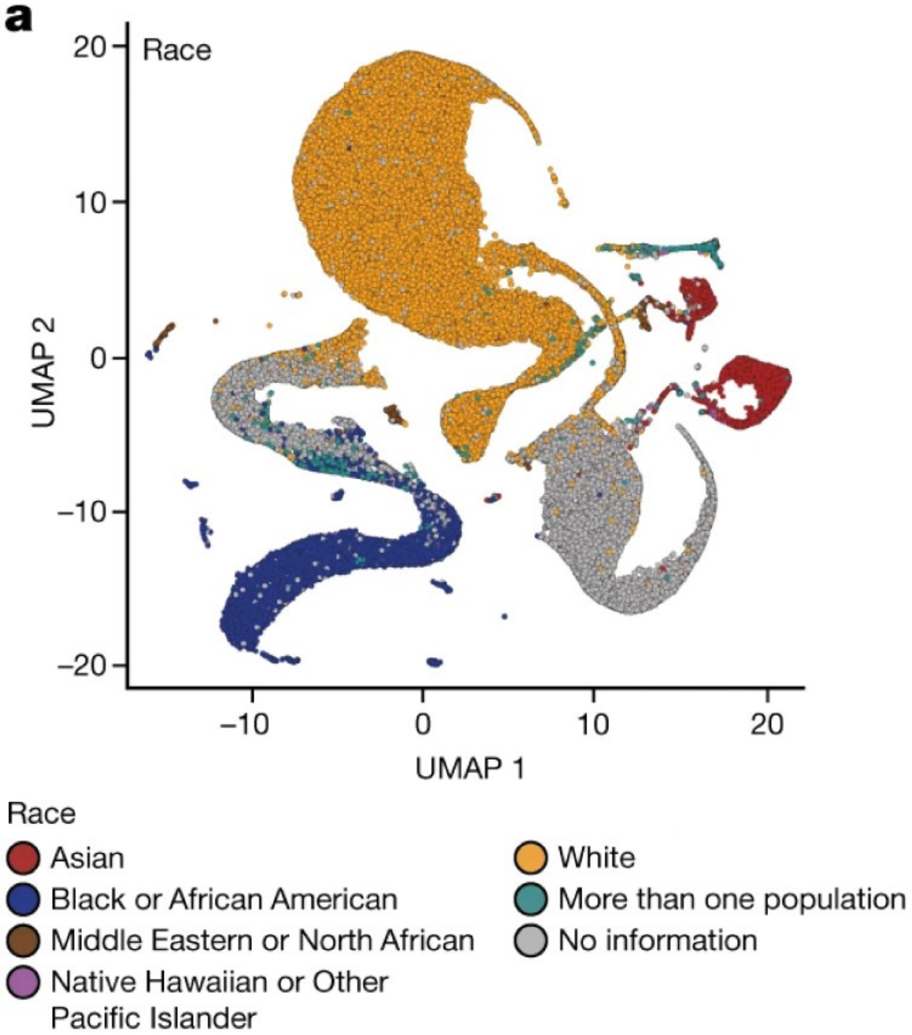
Take home: is tSNE / UMAP accurate?

1. The clusters are (mostly) not fake

tSNE / UMAP is accurate for discovering clusters (exploring data heterogeneity), and this is good enough for single cell biology, but not necessarily interesting for population genomics.

2. The inter-cluster distances are (mostly) fake

tSNE / UMAP is not accurate for exploring genetic or functional relatedness between clusters, and this is (probably) the main interest of population genomics.



NBIS

National Bioinformatics Infrastructure Sweden (NBIS)

SciLifeLab

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET