Determining which distance metrics are ideal within a mass cytometry data analysis pipeline

Tyler J. Burns, Axel R. Schulz, Pawel Durek, Henrik E. Mei

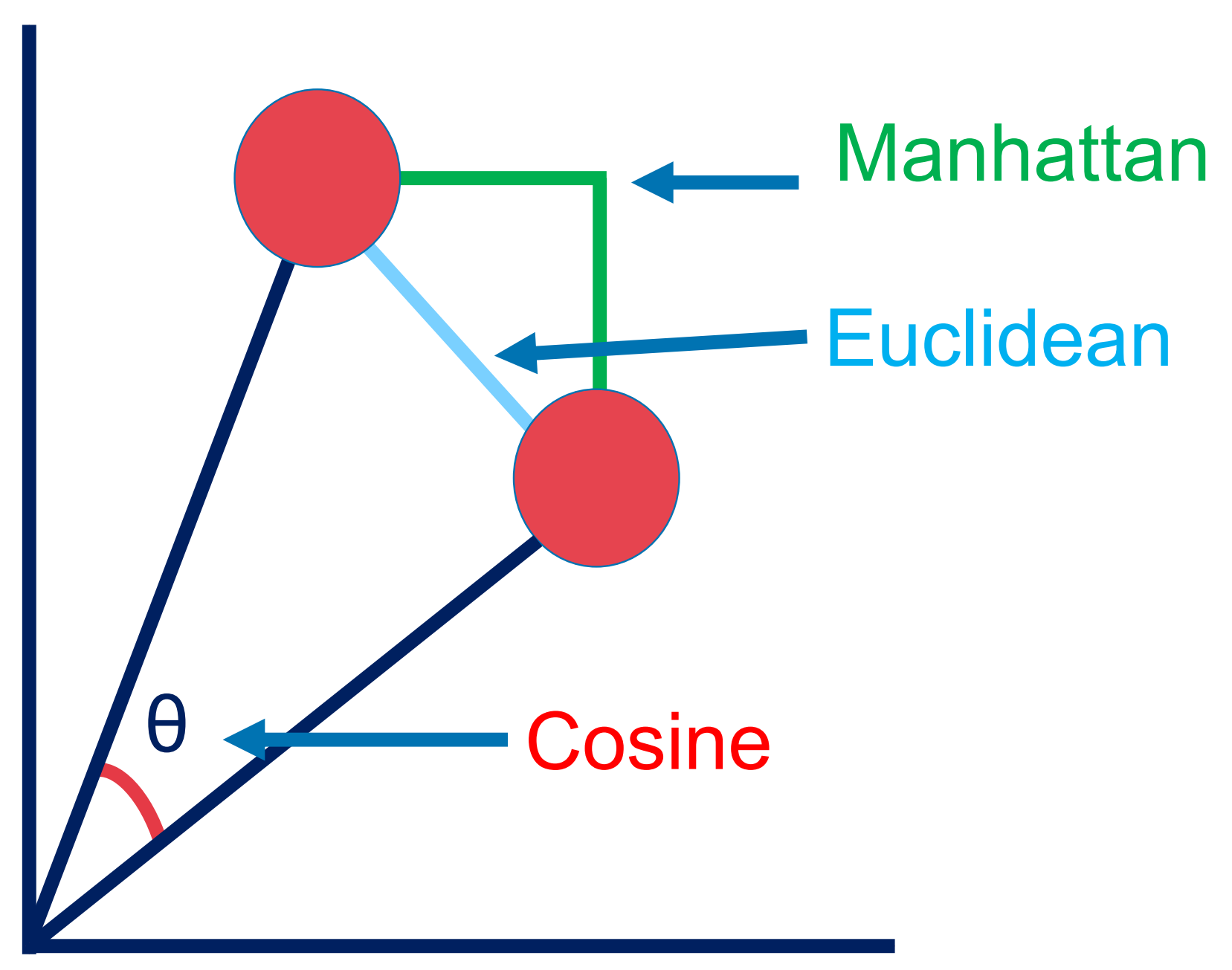
Deutsches Rheuma-Forschungszentrum Berlin (DRFZ), a Leibniz Institute, Berlin, Germany
tyler.burns@drfz.de, GitHub: tjburns08

DRFZ BERLIN
Deutsches Rheuma-Forschungszentrum
Ein Institut der Leibniz-Gemeinschaft

Introduction

Due to the rise of high-dimensional single cell technologies in the past few years, there has been an increasing number of both computational methods and workflows to analyze the new wealth of data. However, non-intuitive properties of high-dimensional space can give rise to analysis artifacts, collectively known of as the "curse of dimensionality." Increasing dimensions differentially affect the performance of distance metrics, and there is no clear consensus about which distance metrics to use for which analysis strategies. While the influence of many tool-specific parameters has been evaluated, we study here the impact of commonly used distance metrics on the outcome of dimensionality reduction and clustering.

Distance metrics for high-dimensional data



Here, we tested three of the most commonly used distance metrics in flow and mass cytometry analysis: Manhattan, Euclidean, and Cosine. We also use the distance of binarized (zero vs nonzero) coordinates as an internal inferior control.

Distance metric	Properties	Employed in CyTOF algorithms
Euclidean	Intuitive, easy to visualize.	Flow-SOM clustering (1), t-SNE (2)
Manhattan	Simple summation of each element. Assumes independency of dimensions. Maximizes distance.	SPADE (3)
Cosine	Considers correlation between markers. Disregards vector magnitude.	X-Shift clustering (4)

Results

Euclidean and Cosine distance return more similar nearest neighborhoods than Manhattan distance for a 44 dimensional PBMC dataset

We used a 10,000 PBMC mass cytometry data set from a healthy human donor using a panel of 44 markers that define the major and many minor cell subsets. We analyzed the similarity of K-nearest neighborhood identities per cell across our chosen distance metrics using the Sconify Bioconductor package (5).

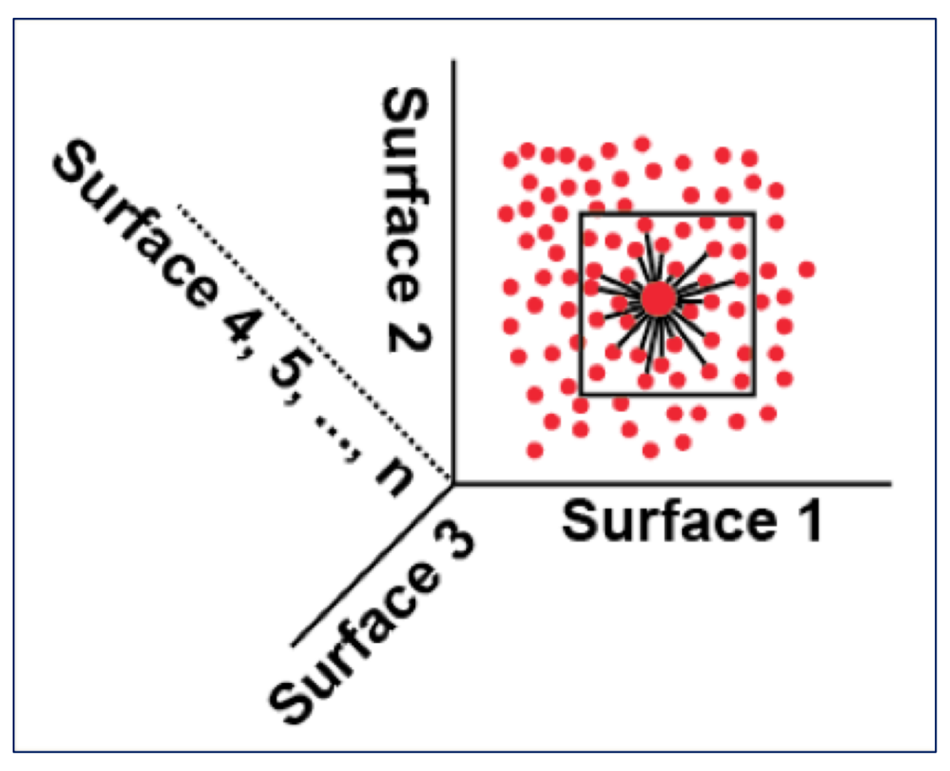
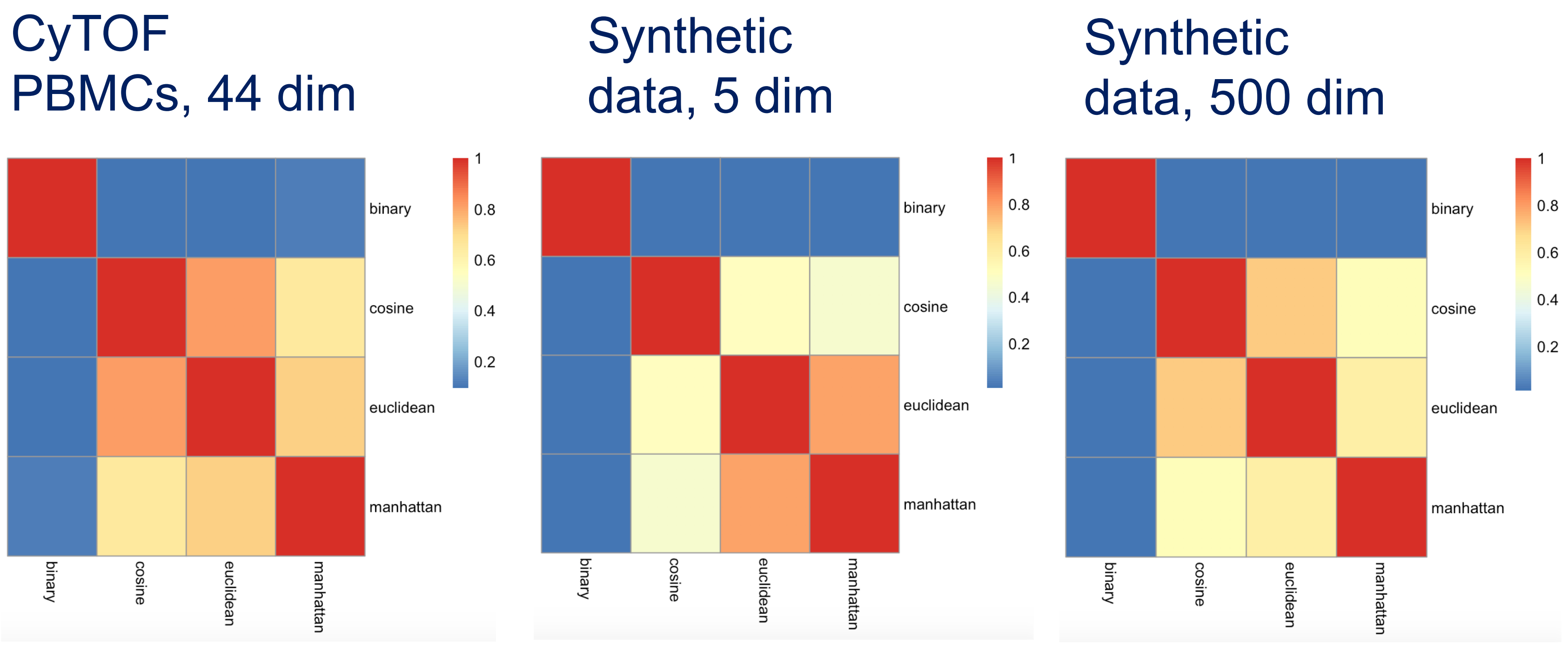


Figure 1: KNN identity similarity across distance metrics. KNN identities with n/100 nearest neighbors (n = number of cells) are compared across distance metrics. Incomplete and variable overlap exist between KNN identities based on different distance metrics. Synthetic data suggests that this is dimensionality-dependent.



Results

t-SNE is visually robust to different distance metrics in human PBMC data

Kullback-Leibler divergence ($\overline{\text{KL}}$), mean from 10 runs

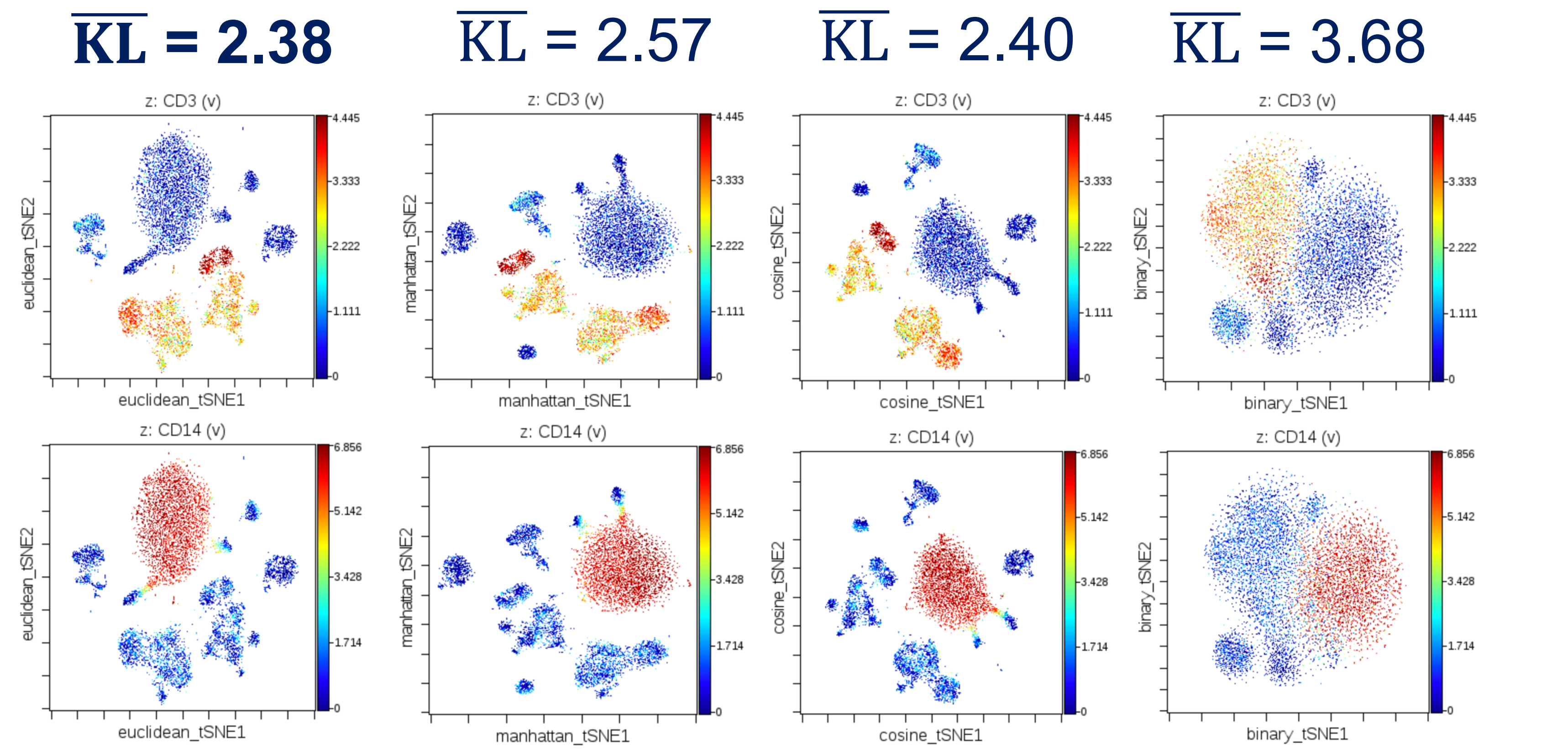
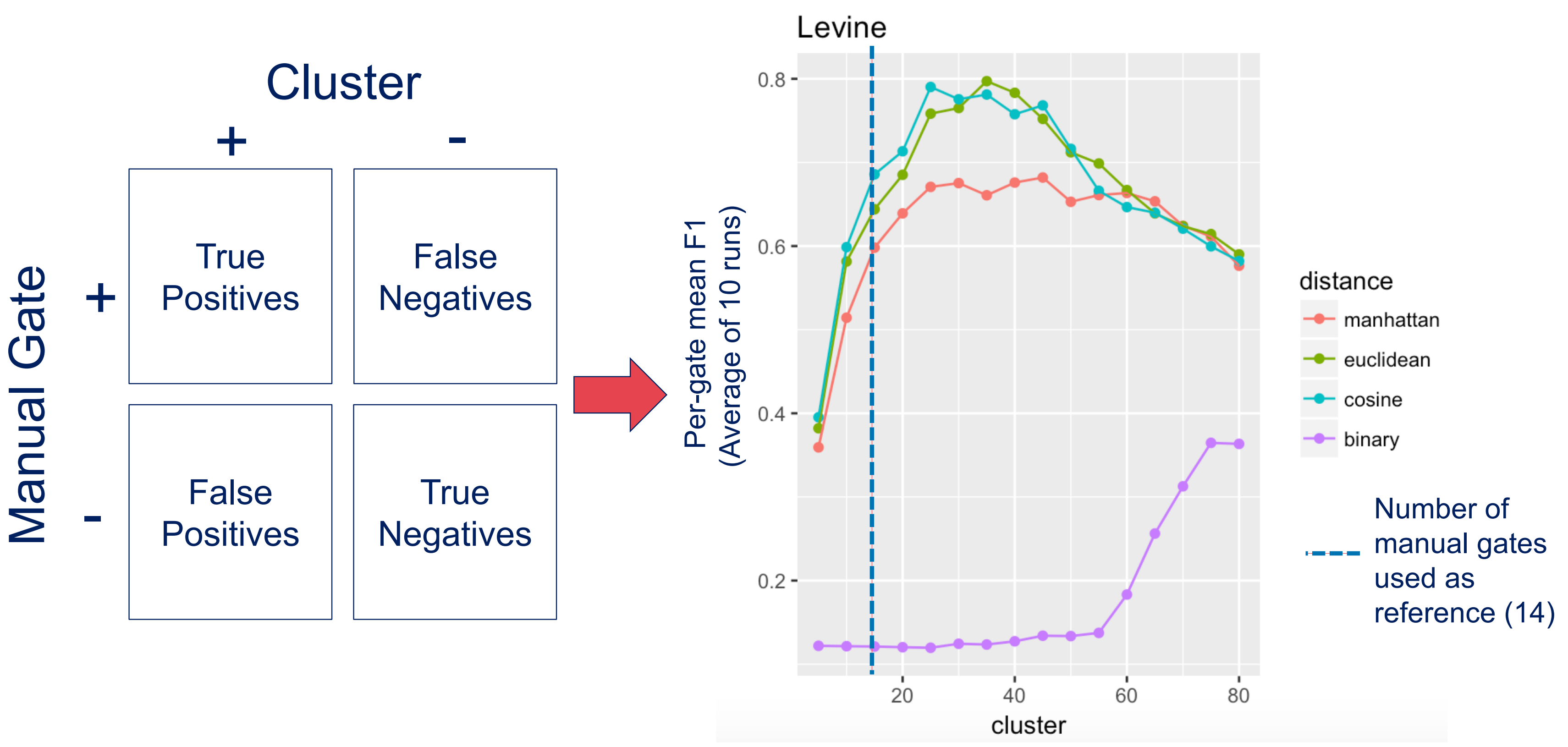


Figure 2: t-SNE maps generated using distance matrices as input with the Rtsne package. Maps are colored by CD14 and CD3 expression to highlight monocytes and T cells. Only binary distance, our internal inferior control, has a visually different map with a higher KL divergence.

For one public pre-gated dataset, Euclidean and Cosine distance outperform Manhattan distance for Flow-SOM meta-clustering

Figure 3: We build upon the per-gate mean F1 scores reported by Weber and Robinson (6) using the "Levine_2015_marrow_32" dataset by altering the Flow-SOM meta-clustering algorithm to utilize distance metrics other than the default Euclidean distance as input.



Conclusions and future directions

Employing different distance metrics yields different KNN identities. Dimensionality reduction by t-SNE appears to be very robust against varying the distance metric, while the accuracy of Flow-SOM meta-clustering as tested against manual gating results differs depending on the chosen distance metric. The difficulties in interpreting t-SNE plots and the limited accordance between manual gating and clustering results mandate further research into alternative distance metrics and analysis methods to enhance information retrieval from high-dimensional single cell cytometric studies.

References

1. Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry*, *87*(7), 636–645. http://doi.org/10.1002/cyto.a.22625

2. Amir, E.-A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., et al. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, *31*(6), 545–552. http://doi.org/10.1038/nbt.2594

3. Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., et al. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, *29*(10), 886–891. http://doi.org/10.1038/nbt.1991

4. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L., & Nolan, G. P. (2016). Automated mapping of phenotype space with single-cell data. *Nature Methods*, *13*(6), 493–496. http://doi.org/10.1038/nmeth.3863

5. Burns T.J., Sconify: Group CyTOF data into overlapping k-nearest neighborhoods for enhanced single-cell visualizations (2018). Bioconductor.

6. Weber, L. M., & Robinson, M. D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry*, *89*(12), 1084–1096. http://doi.org/10.1002/cyto.a.23030

Funding

Deutsche Forschungsgemeinschaft (DFG) grant Me 3644/5-1