Ensemble of similarity metrics with a multiview self-organizing map improves cell clustering in highly multiplexed imaging cytometry data

Elijah Willie[1,2], Pengyi Yang[1,2,3,5], and Ellis Patrick[1,2,3,4]

[1]Sydney Precision Data Science Centre, The University of Sydney, NSW, Australia.

[2]School of Mathematics and Statistics, The University of Sydney, NSW, Australia.

[3]Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong SAR, China.

[4]Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, NSW 2006, Australia.

[5]Computational Systems Biology Group, Children's Medical Research Institute, The University of Sydney, Westmead, NSW 2145, Australia.

[*]To whom correspondence should be addressed. Email: Ellis Patrick, ellis.patrick@sydney.edu.au.

Abstract

Highly multiplexed in situ imaging cytometry assays have enabled researchers to scrutinize cellular systems at an unprecedented level. With the capability of these assays to simultaneously profile the spatial distribution and molecular features of many cells, unsupervised machine learning, and in particular clustering algorithms, have become indispensable for identifying cell types and subsets based on these molecular features. However, the most widely used clustering approaches applied to these novel technologies were developed for cell suspension technologies and may not be optimal for in situ imaging assays. In this work, we systematically evaluated the performance of various similarity metrics used to quantify the similarity between cells when clustering. Our results demonstrate that performance in cell clustering varies significantly when

different similarity metrics were used. Lastly, we propose FuseSOM, an ensemble clustering algorithm employing hierarchical multi-view learning of similarity metrics and self-organizing maps (SOM). Using a stratified subsampling analysis framework, FuseSOM exhibits superior clustering performance compared to the current state-of-the-art clustering approaches for in situ imaging cytometry data analysis.

Introduction

Over the last decade, there has been a significant increase in the resolution at which researchers can scrutinize biological systems. This increase in resolution has been driven mainly by technological advancements. Such technological advancements have enabled the simultaneous measurements of multiple molecular features of biological systems at the single-cell level. Highly multiplexed in situ imaging cytometry assays [3] are one such suite of technologies. These technologies can currently measure approximately 50 features with extremely high throughput (millions of cells). With the increase in measurement capabilities, researchers can now answer many biological questions with unprecedented complexity.

Defining functionally distinct cell types is one recurring task in analyzing highly multiplexed imaging data. Historically, for low throughput imaging data, defining cell types was accomplished by manual gating methods [50, 1, 24]. Manual gating is when cell types are demarcated manually using scatter plots comparing the expression values of two features across all cells and requires specialist domain knowledge. This approach can become labor-intensive if multiple pairs of features are required to isolate a cell type and if multiple types need to be gated. To alleviate this, semi-supervised methods have been developed that leverage domain knowledge to build classifiers [1, 20]. However, these methods require specialist expertise and predefined cell types to be fully exploited.

With the increase in the abundance of biological data, clustering has grown in popularity since it aids in identifying groups of cells using all measured features. Clustering also has the added advantage that, when applied naively, it does not require prior knowledge of the cell types present in an experiment. There have been many clustering algorithms developed [39] and when applied to cytometry data, these are very effective at identifying known and novel cell types [40, 49]. In addition, various clustering methods have been specifically developed to analyze multiplex imaging cytometry datasets using different clustering paradigms. These include density estimation methods such as ClusterX [38], and Assense [46], graph-based methods such as

PhenoGraph [21], and partitioning methods such as FlowSOM [56], and Xshift [44]. In practice, these methods are often employed to generate a set of candidate clusters which are typically further refined and annotated based on biological features such as key marker expression, cell localization, and expert domain knowledge.

Clustering algorithms partition a set of points based on similarity or dissimilarity, so it is essential to consider how to best represent the similarities between points in a dataset to demarcate clusters. Optimising the similarity metric used for cell clustering is critical, as different ways of quantifying similarity will inevitably generate different clusters. These measures are typically referred to as distance metrics. The most common metric is the Euclidean distance, which is used in many clustering algorithms, including FlowSOM, Xshift, ClusterX, and PhenoGraph. However, a recent study [18] showed that distance-based metrics such as Euclidean perform less competitive to correlation-based metrics such as Pearson or Spearman correlation when clustering single-cell RNA sequencing (scRNA-seq) data. Multiplexed imaging data is similar to scRNA-seq data; therefore, evaluating the performance of using different similarity metrics when defining cell types in multiplexed imaging data may lead to the development of better clustering algorithms for such data.

In this work, we systematically assess the performance of various distance- and correlation-based metrics and their combinations. We leverage this assessment to develop a clustering algorithm called FuseSOM that uses self-organizing maps (SOM) and combines multiple similarity metrics through multi-view ensemble learning and hierarchical clustering to identify cell types in multiplexed in situ imaging cytometry assays. We apply FuseSOM to 15 imaging datasets and compare its performance to state-of-the-art methods across various evaluation criteria. Overall, we demonstrate the impact of similarity metrics on clustering cells from multiplexed in situ imaging cytometry data and propose FuseSOM, a multi-view ensemble clustering method, for robust and accurate analysis of such data.

Results

Correlation-based metrics perform better for clustering

To assess the impact of similarity metrics on clustering performance, we performed hierarchical clustering on a MIBI-TOF dataset [26] using either correlation-based metrics (Pearson, Spearman, and Cosine) or distance-based metrics (Euclidean, Manhattan, and Maximum). To assess

performance, we compared the clusters from the hierarchical clustering to the manually curated cell type labels identified in the manuscript (**Figure 2**). On average, correlation-based metrics outperform distance-based metrics by 8.0% for ARI, 10.7% for NMI, 1.70% for FM-index, and 8.0% for F-Measure. Furthermore, there is evidence that clustering performance is improved when using correlation-based metrics compared to distance-based metrics, as seen across three evaluation metrics.

To provide a comprehensive assessment of the performance of the similarity metrics, we quantified the clustering performance of the metrics on 15 multiplexed in situ imaging cytometry datasets (**Table 1**). These datasets were chosen as each had some manual intervention when cell type labels were defined. Each dataset was randomly subsampled to 20K cells five times, and each subset was clustered using hierarchical clustering with all the similarity metrics. Finally, the average was taken across the five subsets. Across all 15 datasets, correlation-based metrics consistently outperformed distance-based metrics (**Figure 3**), more accurately recapitulating the manually curated cell-type labels from their original publications. These results show the efficacy of correlation-based metrics in hierarchical clustering.

Next, we assessed if the performance differences between correlation and Euclidean distance are consistent across multiple clustering methods. We reconfigured PhenoGraph [21], FlowSOM [56], and K-means clustering to use correlation instead of euclidean. As previously, the 15 datasets were randomly subsetted to 20K cells five times, and after clustering, the average score was taken across all evaluation metrics. While there does not appear to be a significant benefit in clustering performance for PhenoGraph, the overall PhenoGraph performance for both distance measures is worse when compared to the correlation-based distances (**Figure 4**) and (**Supplementary Figure S1**). For K-means, hierarchical clustering, and FlowSOM, we observe Pearson correlation performing better than Euclidean distance across all evaluation metrics. See (**Figure 4**) and (**Supplementary Figure S1**).

Multi-view ensemble learning provides a stronger signal for clustering

Given the performance differences between the similarity metrics, we next wondered if combining multiple metrics using strategies such as multi-view ensemble learning [7] would further improve performance. To evaluate the efficacy of combining multiple distance metrics for clustering, we performed a comparison study combining various combinations of Pearson, Spearman, Cosine, and the Euclidean distance. First, all possible combinations of metrics were used to clus-

ter the prototypes generated by the SOM algorithm. Of all the combinations, combining Pearson, Spearman, Cosine, and Euclidean consistently provided the best score across all four evaluation metrics (**Figure 5**). The combination of Pearson, Spearman, and Cosine also provided strong results, which imply that the addition of Euclidean does not markedly improve overall clustering performance.

FuseSOM combines self-organizing maps with multi-view ensemble of distances

Here, we introduce FuseSOM for the clustering of highly multiplexed imaging data. FuseSOM leverages all the ideas already discussed by combining a correlation-based Self Organizing Map and multiview hierarchical clustering to define cell types robustly (**Figure 1**).

Next, we compared FuseSOM with FlowSOM. FlowSOM is currently one of the most popular clustering methods for multiplexed imaging datasets. Before clustering, each dataset was processed and stratified (see Methods). Each stratified set was then clustered using FuseSOM and FlowSOM, and all evaluation metrics were computed on the results. Finally, the average score was taken for each metric across all five substratified sets (**Figure 6**). FuseSOM shows a superior performance across all metrics, which implies that multiview ensemble of correlation-based metrics does provide a more robust signal when clustering. The performance gain is particularly evident when looking at ARI and NMI, with average differences in scores being 32% for ARI, 27% for NMI, 10% for FM-Index, and 9.0% for FMeasure.

FuseSOM provides various methods for estimating the number of clusters

Several methods for estimating the number of clusters have been implemented in FuseSOM. The relative error (RE) between the predicted number of clusters and the number of clusters used in the corresponding manuscripts was used to assess the accuracy of the cluster estimation methods. The Jump and Discriminant method tends to overestimate the actual number of clusters while the others tend to underestimate the actual number of clusters (**Figure** 7). Next, FuseSOM was used to cluster the cells in the datasets using the number of clusters estimated by each method. For both ARI and NMI, the Jump and Discriminant methods appear to have the highest average score across all datasets (**Figure 7**). This demonstrates that the Discriminant and Jump methods have the greatest average accuracy in predicting the number of clusters chosen

in the original manuscripts. This is further reflected in the clustering performance based on the predicted number of clusters.

Running time and memory usage

To investigate the running time and memory usage of FuseSOM, we applied FuseSOM to the Hoch [15] dataset. This dataset contains 800K cells across 41 markers. Next, clustering was performed in increments of 100K cells (8 clustering solutions in total) to gauge how memory and running are affected by an increasing number of cells. Clustering was performed on an 11th Gen Intel® Core™ i7-1165G7 @ 2.80GHz with four cores and 32 GB of memory. Memory usage scales linearly with the size of the data (**Supplmentary Figure S2a**). Doubling the size of the data requires twice as much memory. FuseSOM can process a dataset with 800K cells using only 8GB of memory, implying that FuseSOM can scale to datasets with more than a million cells with moderate memory usage. FuseSOM is also efficient, being able to process 800K cells in under 5 minutes (**Supplementary Figure S2b**). The computational time scales sub-linearly as the size of the dataset increases. For example, the running time is not doubled by doubling the data size.

Discussion

In this work, we provided a comparative analysis of the performance of various similarity metrics for clustering highly multiplexed in situ imaging cytometry assays. Using multiple clustering methods across multiple similarity metrics, we demonstrated that the choice of similarity metric could affect the clustering performance of highly multiplexed cytometry in situ imaging data with correlation-based metrics on average out-performing distance-based metrics. Finally, we leveraged correlation-based metrics to develop a novel clustering algorithm called FuseSOM. In short, FuseSOM combines correlation and euclidean-based distances to identify cell types robustly. In this work, we have successfully illustrated the impact of the choice of distances when clustering. We have also combined these ideas into a user-friendly tool called FuseSOM that is fast, robust, and freely available. We have also demonstrated FuseSOM's ability to recover semi-supervised cell-type annotations across various datasets from differing imaging technologies with reasonable accuracy. Doing so has demonstrated FuseSOM's superior performance for unbiased identification of cell types in highly multiplexed in situ imaging cytometry data relative to current

state-of-the-art clustering algorithms.

A recent study [18] demonstrated that for single-cell RNA-seq datasets, correlation-based metrics outperformed distance-based metrics. This point was further illustrated by [57], where many distance metrics were benchmarked for clustering scRNA-seq data. The authors showed the significance of the impact of distance metrics on clustering and that correlation-based distances tend to perform better on average. Throughout this work, we show a similar trend with data from different imaging technologies. In this work, the correlation-based metric outperformed the euclidean-based metric across multiple clustering algorithms designed to use distance-based metrics, including Euclidean and Manhattan distance. These algorithms include Hierarchical clustering, K-means, FlowSOM, and PhenoGraph. One of the reasons why correlation-based metrics perform better is their invariance to the scaling of the data. Distance-based metrics such as Euclidean and Manhattan are sensitive to the scaling of the data and therefore are susceptible to changes in the expression of markers across images or even different regions of the tissue imaged. However, correlation-based metrics such as Pearson and Spearman are scale-invariant and, therefore, could be less susceptible to changes in the expression of markers driven by technical artifacts. As correlation-based metrics only consider the relative expression between markers, we suspect this makes them more robust and therefore are more accurate in capturing expression trends in highly multiplexed in situ imaging cytometry data.

For many clustering algorithms, the number of clusters is a parameter that must be chosen before executing the algorithm. Some algorithms, including graph and density based, can estimate the number of clusters as part of the algorithm. However, even for these methods, other parameters indirectly affect the number of clusters. For example, the size of the neighborhood in density-based methods can affect the number of clusters. Expert domain or prior knowledge about the data can be used to estimate the number of clusters for clustering algorithms that do not estimate the number of clusters. However, this may only sometimes be the case, and various methods exist to help automate this. While multiple methods exist for scRNA-seq data [58], there is yet to be a generally accepted way to estimate the number of clusters for multiplexed imaging cytometry data. Furthermore, selecting a suitable number of clusters is accepted as more of an art than a science since different applications or scientific questions may impact the number of clusters. For example, when looking for rare cell types in biological data, one might over-cluster the data to find smaller populations of cells. However, this may only be suitable for some individuals since those interested in prominent cell types will use fewer clusters. Therefore, the data was over-clustered in most of the datasets used for benchmarking in this study. After over-clustering,

expert domain knowledge was used to reduce the number of clusters so that the resulting clusters were biologically sensible. All of these factors are still evident for FuseSOM. We expect that clustering will always benefit from manual intervention and scrutiny to decide the number of appropriate clusters for a dataset. The methods provided for cluster estimation in FuseSOM are solely data-dependent and do not rely on any prior biological domain knowledge. As such, they may only sometimes provide reasonable estimations for the correct number of clusters. However, the results provide a suitable starting point that may be further expanded upon.

Methods

Imaging datasets collection and processing

To benchmark the performance of FuseSOM on imaging datasets from various technologies, we selected a set of 15 datasets generated using different imaging technologies. In addition, we selected datasets with human intervention in manually gating cell populations or merging biologically similar clusters. Datasets were sourced from major databases, including Zenodo, Figshare, and Mendeley. When available, we used the version data that had been processed as described in the original manuscript. We also used the same markers for clustering as described in the manuscript. The imaging technologies used included Co-detection by indexing(CODEX) [4] (four datasets), Imaging Mass Cytometry(IMC)[13] (six datasets), Multiplexed ion beam imaging by the time of flight (MIBI-TOF) [17] (four datasets), and sequential Fluorescence In Situ Hybridization (seqFISH) [9] (one dataset). Next, we employed a substratification framework for each dataset to evaluate clustering methods. This framework takes an imaging dataset as input, and five stratified samples are returned. Substratification is done to account for possible variability in clustering results. This framework takes as input and returns five stratified datasets. Stratification is done by selecting 80% of cells from each annotated class. See [18]

Similarity metrics

Six types of metrics across two classes that are predominantly used across machine learning clustering literature were used in this study. The two classes include correlation-based and distance-based. The distance-based metrics were Euclidean, Manhattan, and Maximum distance, while the correlation-based metrics included Pearson correlation, Spearman correlation, and Cosine similarity. More formally, let x_{im} and x_{jm} denote the expression of a marker $m = 1, ..., M$ in

cell $i = 1, ..., N$ and cell $j = 1, ..., N$, where G and N are the total number of markers and cells, respectively. Led $D = d_{ij}$ be a distance matrix where d_{ij} represents the distance between $cell_i$ and $cell_j$. We can then define the distance-based metrics as follows:

Euclidean distance,

$$d_{ij} = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \tag{1}$$

Manhattan distance,

$$d_{ij} = \sum_{m=1}^{M} |x_{im} - x_{jm}| \tag{2}$$

Maximum distance,

$$d_{ij} = \max_{m} |x_{im} - x_{jm}| \tag{3}$$

Similarly, the correlation-based metrics can be defined as follows:

Pearson distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M}(x_{im} - \bar{x}_i)(x_{jm} - \bar{x}_j)}{\sqrt{\sum_{m=1}^{M}(x_{im} - \bar{x}_i)^2}\sqrt{\sum_{m=1}^{M}(x_{jm} - \bar{x}_j)^2}}\right)} \tag{4}$$

Spearman distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M}(r_{im} - \bar{r}_i)(r_{jm} - \bar{r}_j)}{\sqrt{\sum_{m=1}^{M}(r_{im} - \bar{r}_i)^2}\sqrt{\sum_{m=1}^{M}(r_{jm} - \bar{r}_j)^2}}\right)} \tag{5}$$

Cosine distance

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M} x_{im}x_{jm}}{\sqrt{\sum_{m=1}^{M} x_{im}^2}\sqrt{\sum_{m=1}^{M} x_{jm}^2}}\right)} \tag{6}$$

where r_{ij} is the rank of marker m in $cell_i$, $\bar{x_i}$ is the mean expression of $cell_i$, $\bar{x_j}$ is the mean expression of $cell_j$, $\bar{r_i}$ is the mean expression rank of $cell_i$, and $\bar{r_j}$ is the mean expression rank of $cell_j$,

The hierarchical clustering and self-organizing map algorithms

We used the agglomerative hierarchical clustering algorithm with varying similarity metrics to investigate the effect of similarity metrics on clustering algorithms. Agglomerative clustering is a type of clustering where each data point starts as its cluster. At every step, two clusters are merged using a linkage function. This step repeats until a single cluster comprises the entire

dataset. There are many options for a linkage function. We opted to use the average linkage function [32], which merges two clusters based on the average distance between pairs of points between two clusters. This algorithm outputs a tree-like structure called a dendrogram which can be cut at any level to generate k clusters. We also used the FlowSOM algorithm proposed by [56]. This algorithm uses a Self Organizing Map (SOM) architecture to generate prototypes which are then clustered using hierarchical consensus clustering using euclidean distances to generate meta-clusters. These meta-clusters are then projected back on the original dataset to obtain clusters.

FuseSOM

Here the FuseSOM algorithm is described. The algorithm starts by generating a Self Organizing Map. A Self Organizing Map is a type of dimensionality reduction algorithm that maps points in a high dimensional space $(d > 2)$ to a lower dimensional space $(d = 2)$. The SOM architecture preserves the topological relationships between points when reduced to a lower dimension. The SOM also provides a set of points called prototypes which are representations of points in the higher dimensional space. Like cluster centers in the k-means algorithm, many points can be mapped to a single prototype. For a more thorough treatment of SOMs, see [29]. The *YASOMI* package [41] was obtained and modified to implement the self-organizing map used in FuseSOM.

In this work, we generate a SOM, and the prototypes are used for clustering. After clustering, the clusters are projected back to the original data to classify the original data points. The SOM algorithm requires a 2-d grid(x,y) size, which determines the number of prototypes. Grid sizes of varying shapes are allowed. However, square grids are typically used. To estimate the size of the grid for a dataset, we use the method described in [33]. This method computes the number of eigenvalues of a covariance matrix significantly different from the Tracy-Widom distribution [55].

Next, multiview integration combines the Pearson correlation distance, Cosine distance, Spearman correlation, and Euclidean distance between the prototypes to generate a final distance matrix for clustering. The method described in [28] is adopted to combine the four transformed matrices. The *analogue* [48] package is used to perform the multiview integration. The *psych* [36] package transforms the similarity matrices to distance matrices. Finally, to generate final cluster labels using the integrated distance matrices, hierarchical clustering using the average linkage function is used. The *FCPS* package [53] was used for hierarchical clustering.

Estimating the number of clusters

For most clustering algorithms, the number of clusters k is an important hyperparameter that must be set. To this end, many methods have been developed to help practitioners choose an appropriate number for their dataset. We have included well-known methods for estimating the number of clusters as part of the FuseSOM package. These methods include the Gap statistic [54], the Slope statistic [12], the Jump statistic [52], the Silhouette statistic [42], and the within-cluster distance (WCD). We also developed a method for estimating the number of clusters based on maximum clusterability discriminant projection pursuit. To accomplish this, we couple hierarchical clustering with discriminant analysis [10, 30, 43] and multimodality testing [2, 14, 47] to estimate the number of clusters. First, we generate a dendrogram using hierarchical clustering with average linkage. Next, for each node in the resulting tree, we project the two classes onto a line such that both classes are well separated. See (**Supplementary Figure S3**). The Dip test for multimodality testing [14] is then applied to the distribution of the points along this line. See (**Figure 7**). The method described in [27] controls the family-wise error rate (FWER). Finally, the number of nodes with significant p-values is returned as the number of clusters.

Evaluation metrics

To evaluate clustering performance for clustering solutions generated across methods, we used a set of methods; the Adjusted Rand Index (ARI) [51], Normalized Mutual Information (NMI) [19], Fowlkes-Mallows index (FM-Index) [11], and the F-Measure [16]. All these metrics take values between 0 and 1, with 0 being no similarity and 1 being perfect similarity.

Data availability

Publicly available data were used for all evaluations. All data were downloaded as described in the originating manuscripts.

Code availability

The *FuseSOM* R package is available on Bioconductor and is available under the GPL-3 license. All code for analysis done can be found at Github

Acknowledgments

The authors thank all their colleagues, particularly at The University of Sydney, Sydney Precision Data Science Centre and Charles Perkins Centre for their support and intellectual engagement.

The following sources of funding for each author are gratefully acknowledged: the AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong to EP and PY. Australian Research Council Discovery Early Career Researcher Award (DE200100944) funded by the Australian Government to EP; National Health and Medical Research Council (NHMRC) Investigator Grant (1173469) to P.Y.; and the University of Sydney Postgraduate Excellence Award for E.W.; The funding sources had no impact on the study design; in the collection, analysis, and interpretation of data, the writing of the manuscript, and in the decision to submit the manuscript for publication.

Author contributions

EP and EW conceived and designed the study. EP and EW led the method development and guided the evaluation data analysis. EW curated the imaging data, implemented all data analytics, and developed the corresponding R code. All authors wrote, read, reviewed the manuscript, and approved the final version.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Tamim Abdelaal et al. "Predicting Cell Populations in Single Cell Mass Cytometry Data". en. In: *Cytometry Part A* 95.7 (July 2019), pp. 769–781. ISSN: 1552-4922, 1552-4930. DOI: 10.1002/cyto.a.23738.

[2] Jose Ameijeiras-Alonso, Rosa M. Crujeiras, and Alberto Rodríguez-Casal. "Mode testing, critical bandwidth and excess mass". en. In: *TEST* 28.3 (Sept. 2019), pp. 900–919. ISSN: 1133-0686, 1863-8260. DOI: 10.1007/s11749-018-0611-5.

[3] Heeva Baharlou et al. "Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies". In: *Frontiers in Immunology* 10 (Nov. 2019), p. 2657. ISSN: 1664-3224. DOI: 10.3389/fimmu.2019.02657.

[4] Sarah Black et al. "CODEX multiplexed tissue imaging with DNA-conjugated antibodies". en. In: *Nature Protocols* 16.8 (Aug. 2021), pp. 3802–3835. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/s41596-021-00556-8.

[5] Michele Bortolomeazzi et al. "Immunogenomics of Colorectal Cancer Response to Checkpoint Blockade: Analysis of the KEYNOTE 177 Trial and Validation Cohorts". en. In: *Gastroenterology* 161.4 (Oct. 2021), pp. 1179–1193. ISSN: 00165085. DOI: 10.1053/j.gastro.2021.06.064.

[6] Maria Brbić et al. *Annotation of Spatially Resolved Single-cell Data with STELLAR*. en. preprint. Bioinformatics, Nov. 2021. DOI: 10.1101/2021.11.24.469947.

[7] Yue Cao et al. "Ensemble deep learning in bioinformatics". In: *Nature Machine Intelligence* 2.9 (2020), pp. 500–508.

[8] Nicolas Damond et al. "A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry". en. In: *Cell Metabolism* 29.3 (Mar. 2019), 755–768.e5. ISSN: 15504131. DOI: 10.1016/j.cmet.2018.11.014.

[9] Chee-Huat Linus Eng et al. "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". en. In: *Nature* 568.7751 (Apr. 2019), pp. 235–239. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1049-y.

[10] Kamran Etemad and Rama Chellappa. "Discriminant analysis for recognition of human face images". en. In: *Journal of the Optical Society of America A* 14.8 (Aug. 1997), p. 1724. ISSN: 1084-7529, 1520-8532. DOI: 10.1364/JOSAA.14.001724.

[11] E. B. Fowlkes and C. L. Mallows. "A Method for Comparing Two Hierarchical Clusterings". en. In: *Journal of the American Statistical Association* 78.383 (Sept. 1983), pp. 553–569. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1983.10478008.

[12] André Fujita, Daniel Y. Takahashi, and Alexandre G. Patriota. "A non-parametric method to estimate the number of clusters". en. In: *Computational Statistics & Data Analysis* 73 (May 2014), pp. 27–39. ISSN: 01679473. DOI: 10.1016/j.csda.2013.11.012.

[13] Charlotte Giesen et al. "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry". en. In: *Nature Methods* 11.4 (Apr. 2014), pp. 417–422. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2869.

[14] J. A. Hartigan and P. M. Hartigan. "The Dip Test of Unimodality". In: *The Annals of Statistics* 13.1 (Mar. 1985). ISSN: 0090-5364. DOI: 10.1214/aos/1176346577.

[15] Tobias Hoch et al. "Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy". en. In: *Science Immunology* 7.70 (Apr. 2022), eabk1692. ISSN: 2470-9468. DOI: 10.1126/sciimmunol.abk1692.

[16] G. Hripcsak. "Agreement, the F-Measure, and Reliability in Information Retrieval". en. In: *Journal of the American Medical Informatics Association* 12.3 (Jan. 2005), pp. 296–298. ISSN: 1067-5027, 1527-974X. DOI: 10.1197/jamia.M1733.

[17] Leeat Keren et al. "MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure". en. In: *Science Advances* 5.10 (Oct. 2019), eaax5851. ISSN: 2375-2548. DOI: 10.1126/sciadv.aax5851.

[18] Taiyun Kim et al. "Impact of similarity metrics on single-cell RNA-seq data clustering". en. In: *Briefings in Bioinformatics* 20.6 (Nov. 2019), pp. 2316–2326. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bby076.

[19] Tarald Kvålseth. "On Normalized Mutual Information: Measure Derivations and Properties". en. In: *Entropy* 19.11 (Nov. 2017), p. 631. ISSN: 1099-4300. DOI: 10.3390/e19110631.

[20] Hao-Chih Lee et al. "Automated cell type discovery and classification through knowledge transfer". en. In: *Bioinformatics* 33.11 (June 2017). Ed. by Jonathan Wren, pp. 1689–1695. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btx054.

[21] Jacob H. Levine et al. “Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis”. In: *Cell* 162.1 (July 2015), pp. 184–197. DOI: 10.1016/j.cell.2015.05.047.

[22] Candace C. Liu et al. “Reproducible, high-dimensional imaging in archival human tissue by multiplexed ion beam imaging by time-of-flight (MIBI-TOF)”. en. In: *Laboratory Investigation* 102.7 (July 2022), pp. 762–770. ISSN: 0023-6837, 1530-0307. DOI: 10.1038/s41374-022-00778-8.

[23] T. Lohoff et al. “Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis”. en. In: *Nature Biotechnology* 40.1 (Jan. 2022), pp. 74–85. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-021-01006-2.

[24] Florian Mair et al. “The end of gating? An introduction to automated analysis of high dimensional cytometry data: Highlights”. en. In: *European Journal of Immunology* 46.1 (Jan. 2016), pp. 34–43. ISSN: 00142980. DOI: 10.1002/eji.201545774.

[25] Febe van Maldegem et al. “Characterisation of tumour microenvironment remodelling following oncogene inhibition in preclinical studies with imaging mass cytometry”. en. In: *Nature Communications* 12.1 (Dec. 2021), p. 5906. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26214-x.

[26] Erin F. McCaffrey et al. “The immunoregulatory landscape of human tuberculosis granulomas”. en. In: *Nature Immunology* 23.2 (Feb. 2022), pp. 318–329. ISSN: 1529-2908, 1529-2916. DOI: 10.1038/s41590-021-01121-x.

[27] Nicolai Meinshausen. “Hierarchical testing of variable importance”. In: *Biometrika* 95.2 (June 2008), pp. 265–278. ISSN: 0006-3444. DOI: 10.1093/biomet/asn007. eprint: https://academic.oup.com/biomet/article-pdf/95/2/265/623449/asn007.pdf.

[28] Willem Melssen, Ron Wehrens, and Lutgarde Buydens. “Supervised Kohonen networks for classification problems”. en. In: *Chemometrics and Intelligent Laboratory Systems* 83.2 (Sept. 2006), pp. 99–113. ISSN: 01697439. DOI: 10.1016/j.chemolab.2006.02.003.

[29] Dubravko Miljkovic. “Brief review of self-organizing maps”. In: *2017 40th International Convention on Information and Communication Technology, Electronics and Microelec-*

tronics (MIPRO). Opatija, Croatia: IEEE, May 2017, pp. 1061–1066. ISBN: 9789532330908. DOI: 10.23919/MIPRO.2017.7973581.

[30] Mozhgan Mokari, Hoda Mohammadzade, and Benyamin Ghojogh. "Recognizing Involuntary Actions from 3D Skeleton Data Using Body States". en. In: *Scientia Iranica* 0.0 (May 2018), pp. 0–0. ISSN: 2345-3605. DOI: 10.24200/sci.2018.20446.

[31] Dan Moldoveanu et al. "Spatially mapping the immune landscape of melanoma using imaging mass cytometry". en. In: *Science Immunology* 7.70 (Apr. 2022), eabi5072. ISSN: 2470-9468. DOI: 10.1126/sciimmunol.abi5072.

[32] Frank Nielsen. "Hierarchical Clustering". In: *Introduction to HPC with MPI for Data Science*. Cham: Springer International Publishing, 2016, pp. 195–211. ISBN: 9783319219028 9783319219035. DOI: 10.1007/978-3-319-21903-5_8.

[33] Nick Patterson, Alkes L Price, and David Reich. "Population Structure and Eigenanalysis". In: *PLOS Genetics* 2.12 (Dec. 2006), pp. 1–20. DOI: 10.1371/journal.pgen.0020190.

[34] Darci Phillips et al. "Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma". en. In: *Nature Communications* 12.1 (Dec. 2021), p. 6726. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26974-6.

[35] André F. Rendeiro et al. "The spatial landscape of lung pathology during COVID-19 progression". en. In: *Nature* 593.7860 (May 2021), pp. 564–569. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03475-6.

[36] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.2.9. Northwestern University. Evanston, Illinois, 2022.

[37] Tyler Risom et al. "Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma". en. In: *Cell* 185.2 (Jan. 2022), 299–310.e18. ISSN: 00928674. DOI: 10.1016/j.cell.2021.12.023.

[38] Alex Rodriguez and Alessandro Laio. "Clustering by fast search and find of density peaks". en. In: *Science* 344.6191 (June 2014), pp. 1492–1496. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1242072.

[39] Lior Rokach. “A survey of Clustering Algorithms”. en. In: *Data Mining and Knowledge Discovery Handbook*. Ed. by Oded Maimon and Lior Rokach. Boston, MA: Springer US, 2009, pp. 269–298. ISBN: 9780387098227 9780387098234. DOI: 10.1007/978-0-387-09823-4_14.

[40] Tom Ronan, Zhijie Qi, and Kristen M. Naegle. “Avoiding common pitfalls when clustering biological data”. en. In: *Science Signaling* 9.432 (June 2016). ISSN: 1945-0877, 1937-9145. DOI: 10.1126/scisignal.aad1932.

[41] Fabrice Rossi. *yasomi: Yet Another Self Organising Map Implementation*. R package version 0.3/r39. 2012.

[42] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. en. In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. ISSN: 03770427. DOI: 10.1016/0377-0427(87)90125-7.

[43] Ali-Akbar Samadani et al. “Perception and Generation of Affective Hand Movements”. en. In: *International Journal of Social Robotics* 5.1 (Jan. 2013), pp. 35–51. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-012-0169-4.

[44] Nikolay Samusik et al. “Automated mapping of phenotype space with single-cell data”. en. In: *Nature Methods* 13.6 (June 2016), pp. 493–496. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3863.

[45] Christian M. Schürch et al. “Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front”. en. In: *Cell* 182.5 (Sept. 2020), 1341–1359.e19. ISSN: 00928674. DOI: 10.1016/j.cell.2020.07.005.

[46] Karthik Shekhar et al. “Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)”. en. In: *Proceedings of the National Academy of Sciences* 111.1 (Jan. 2014), pp. 202–207. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1321405111.

[47] B. W. Silverman. “Using Kernel Density Estimates to Investigate Multimodality”. en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 43.1 (Sept. 1981), pp. 97–99. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1981.tb01155.x.

[48] G. L. Simpson and J. Oksanen. *analogue: Analogue and weighted averaging methods for palaeoecology*. R package version 0.17-6. 2021.

[49] R. Sloutsky et al. “Accounting for noise when clustering biological data”. en. In: *Briefings in Bioinformatics* 14.4 (July 2013), pp. 423–436. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbs057.

[50] Matthew H. Spitzer and Garry P. Nolan. “Mass Cytometry: Single Cells, Many Features”. en. In: *Cell* 165.4 (May 2016), pp. 780–791. ISSN: 00928674. DOI: 10.1016/j.cell.2016.04.019.

[51] Douglas Steinley. “Properties of the Hubert-Arable Adjusted Rand Index.” en. In: *Psychological Methods* 9.3 (2004), pp. 386–396. ISSN: 1939-1463, 1082-989X. DOI: 10.1037/1082-989X.9.3.386.

[52] Catherine A Sugar and Gareth M James. “Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach”. en. In: *Journal of the American Statistical Association* 98.463 (Sept. 2003), pp. 750–763. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214503000000666.

[53] Michael C. Thrun and Quirin Stier. “Fundamental clustering algorithms suite”. en. In: *SoftwareX* 13 (Jan. 2021), p. 100642. ISSN: 23527110. DOI: 10.1016/j.softx.2020.100642.

[54] Robert Tibshirani, Guenther Walther, and Trevor Hastie. “Estimating the number of clusters in a data set via the gap statistic”. en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423. ISSN: 13697412. DOI: 10.1111/1467-9868.00293.

[55] Craig A. Tracy and Harold Widom. “Level-spacing distributions and the Airy kernel”. en. In: *Communications in Mathematical Physics* 159.1 (Jan. 1994), pp. 151–174. ISSN: 0010-3616, 1432-0916. DOI: 10.1007/BF02100489.

[56] Sofie Van Gassen et al. “FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data: FlowSOM”. en. In: *Cytometry Part A* 87.7 (July 2015), pp. 636–645. ISSN: 15524922. DOI: 10.1002/cyto.a.22625.

[57] Ebony Rose Watson et al. “How does the structure of data impact cell–cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data”. en. In: *Briefings in Bioinformatics* 23.6 (Nov. 2022), bbac387. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbac387.

[58] Lijia Yu et al. “Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data”. In: *Genome Biology* 23.1 (2022), pp. 1–21.

Figure legends

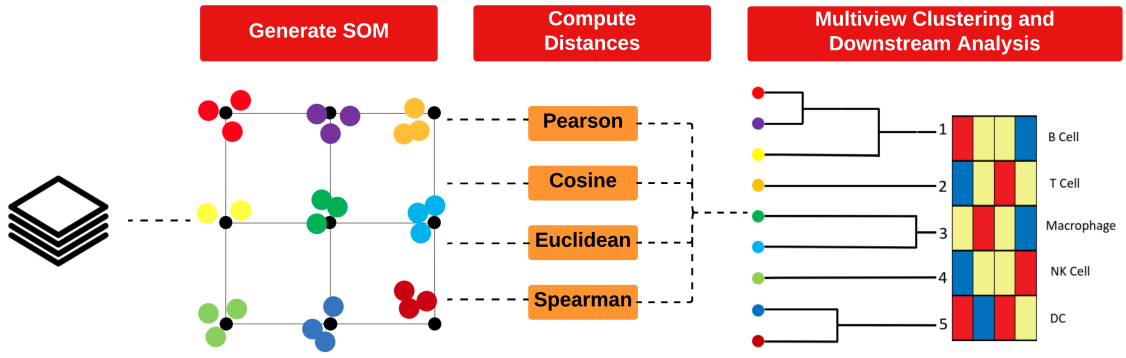


Figure 1: Overview of FuseSOM: This new scalable algorithm uses (i) A Self Organizing Map to reduce the dimension of the data while preserving its topological structure; (ii) A multiview integration of various similarity metrics to capture all relevant signals; and (iii) Hierarchical clustering to generate a clustering solution for further downstream analysis.

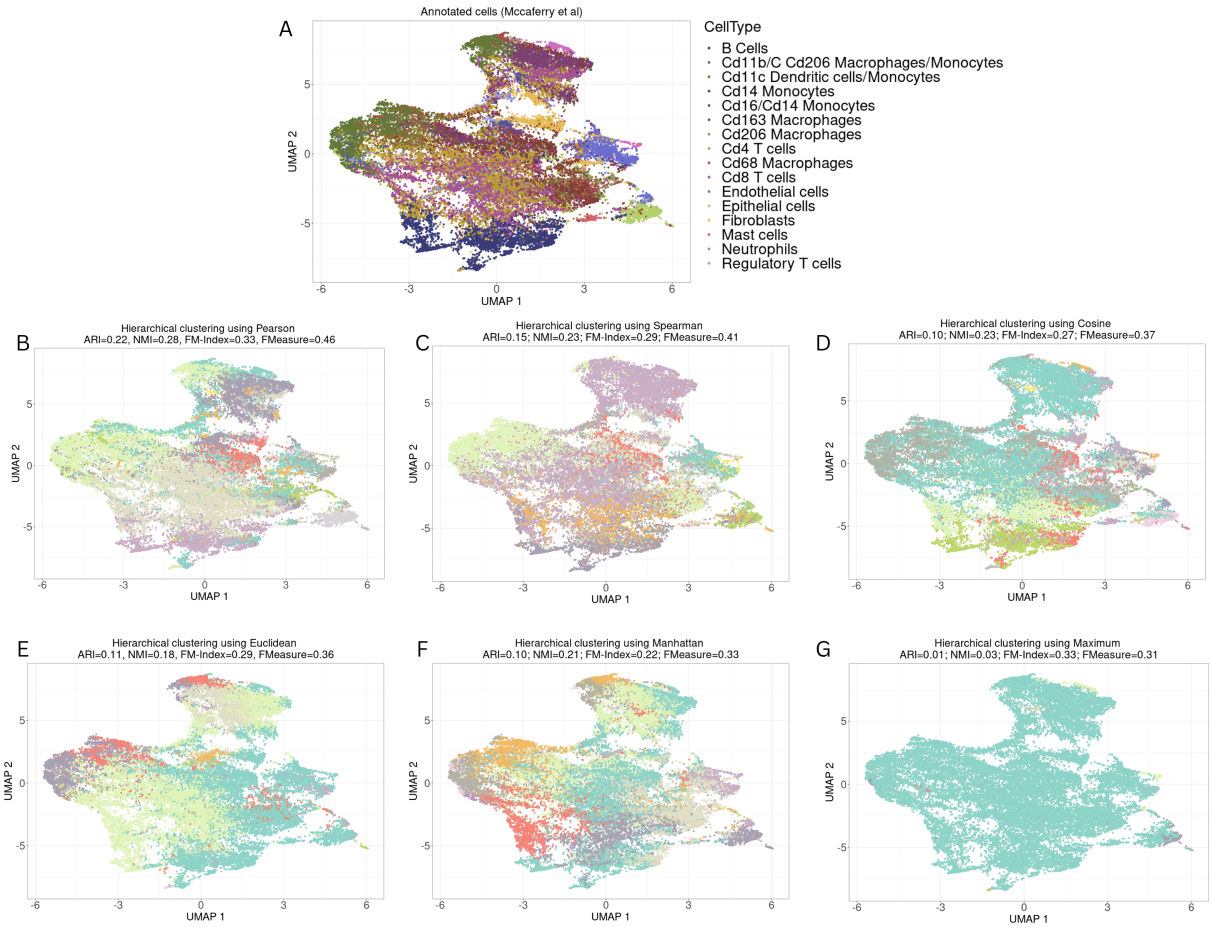


Figure 2: UMAP visualization of cells from a sample imaging dataset. (A) Cells colored by annotations from original study [26]. (B) Hierarchical clustering using Pearson's Correlation and concordance quantified by ARI, NMI, FM-Index, and FMeasure. (C) Hierarchical clustering using Cosine's Distance (D) Hierarchical clustering using Spearman's Correlation. (E) Hierarchical clustering using Euclidean distance. (F) Hierarchical clustering using Manhattan distance. (G) Hierarchical clustering using Maximum distance.

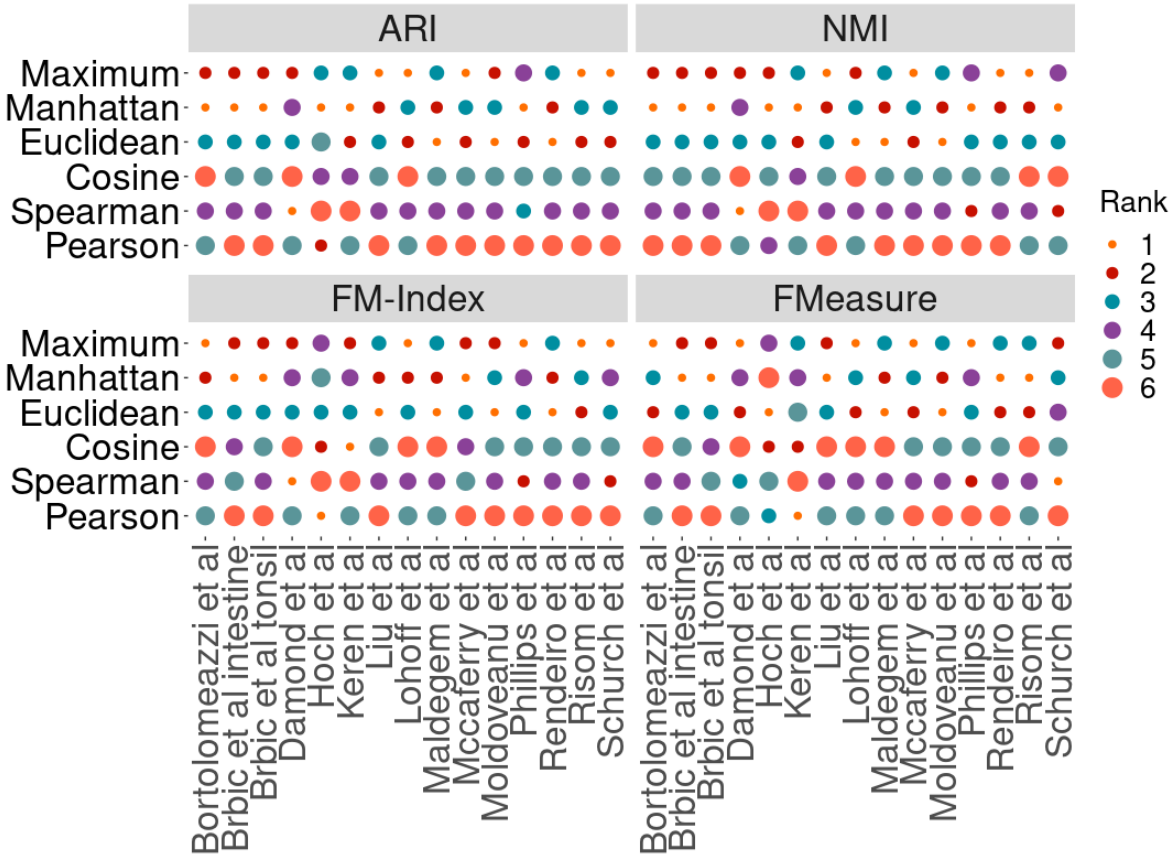


Figure 3: Benchmarking similarity metrics on agglomerative clustering of 15 multiplexed imaging datasets. Each dataset was subsetted to 20K cells five times, and the average clustering score was recorded. Results were ranked in descending order across all similarity metrics and datasets by each evaluation metric. A larger circle size indicates better performance. Correlation-based metrics are consistently ranked higher than distance-based metrics across most datasets.

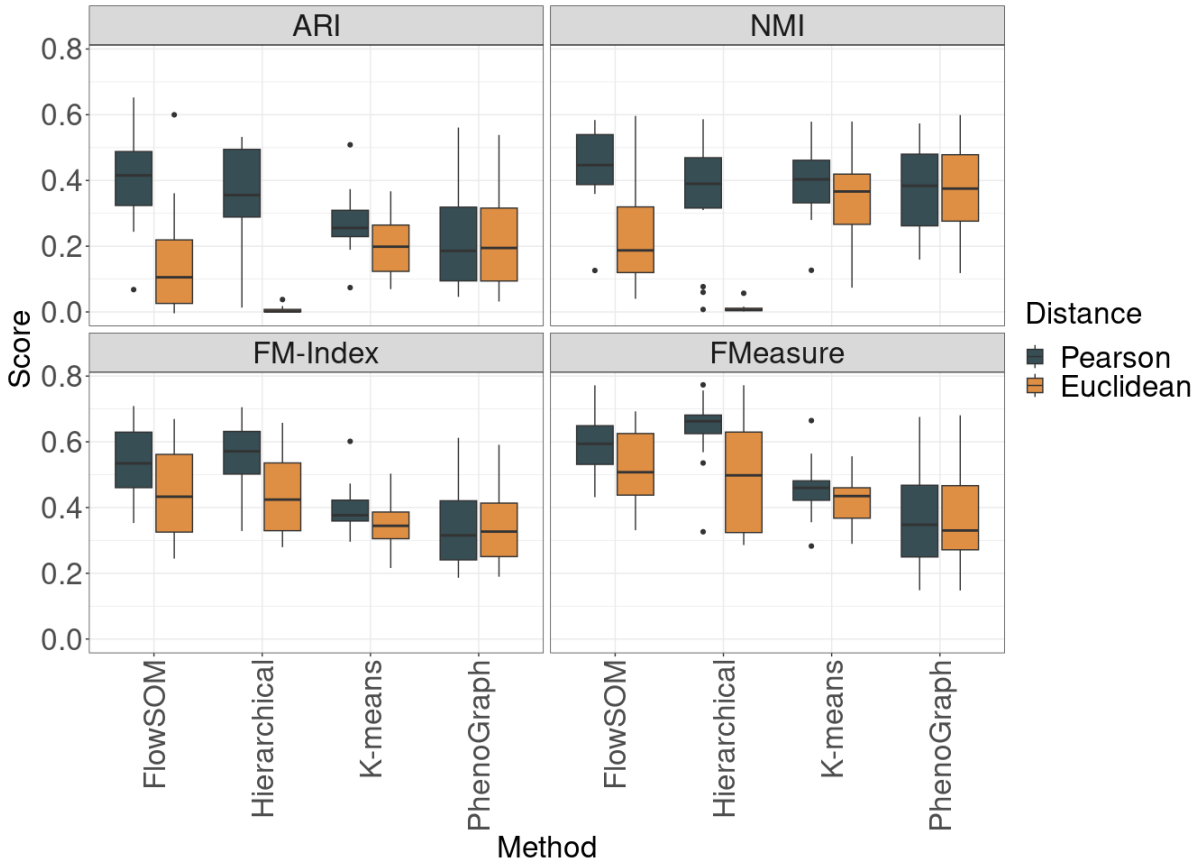


Figure 4: Boxplots of clustering performance of four clustering methods using Pearson correlation and euclidean across four evaluation metrics (ARI, NMI, FM-Index, and FMeasure). For three of the four methods, Pearson on average performs better than euclidean for clustering. This is not evident for PhenoGraph.

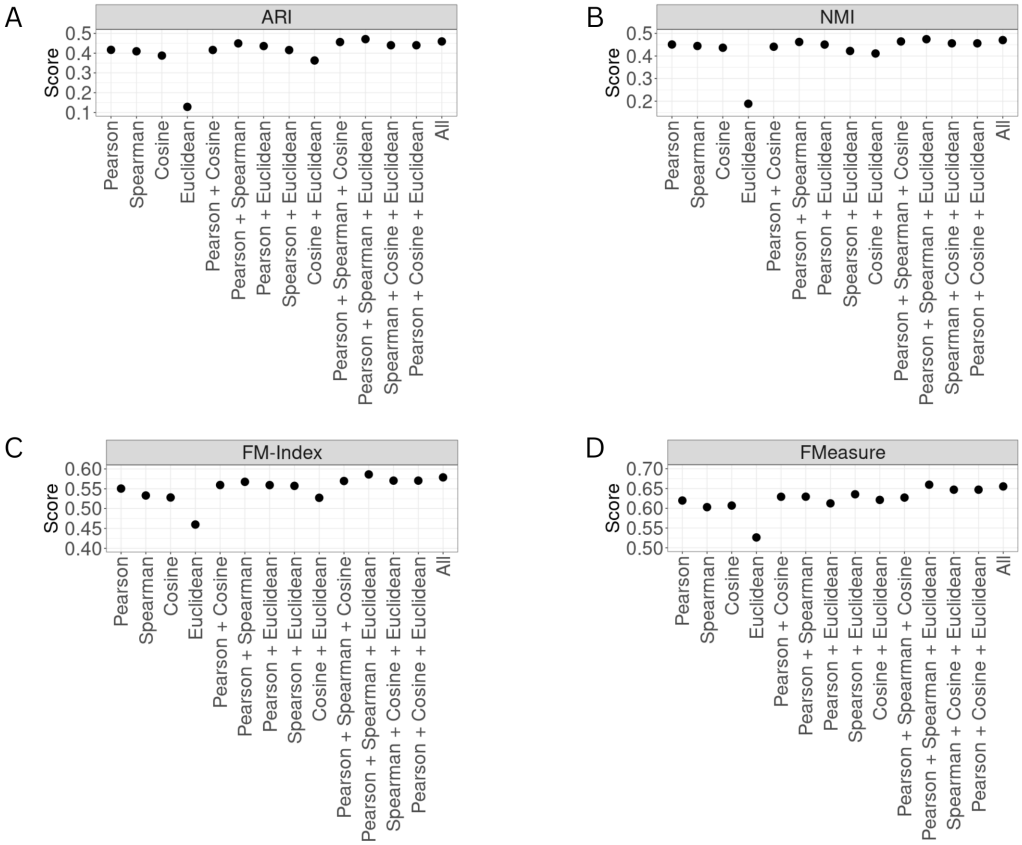


Figure 5: Scatter plots of average clustering performance of all combinations of Pearson, Cosine, Spearman, and Euclidean. Across all performance metrics, there is evidence that combining all distances provides a better signal for clustering.

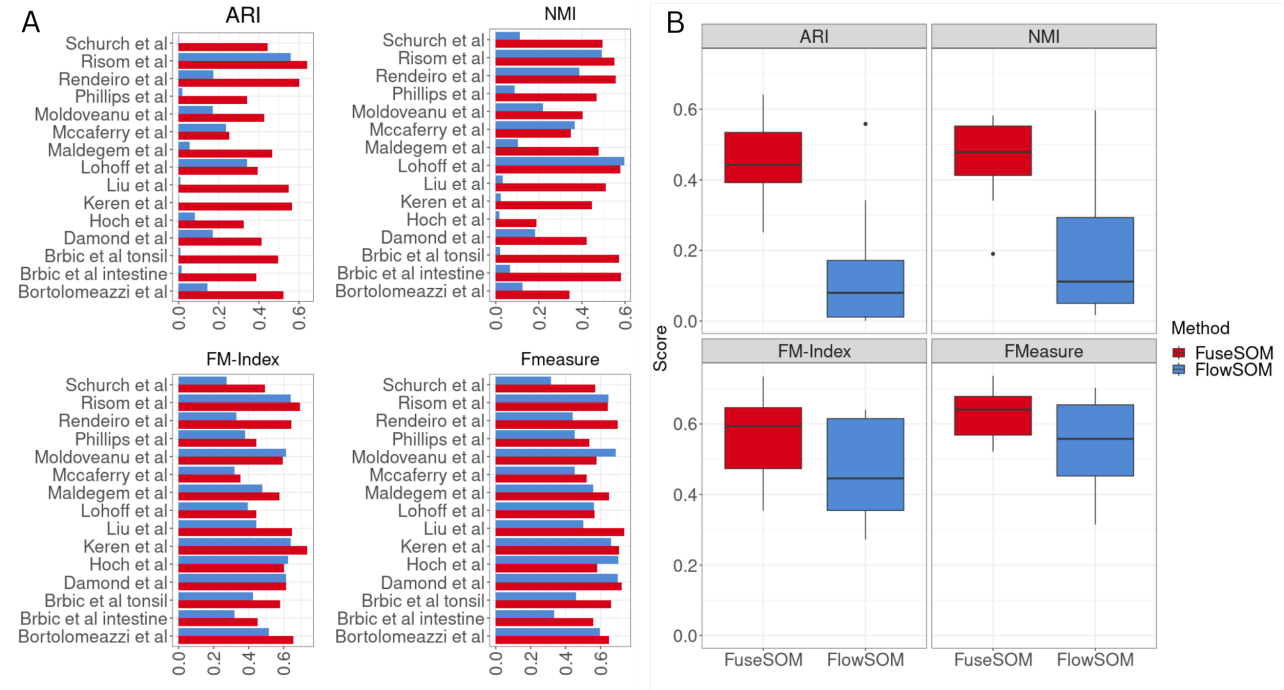


Figure 6: (A) Bar plots of average clustering performance for FuseSOM and FlowSOM showed for all 15 datasets across four performance metrics (ARI, NMI, FM-Index, and FMeasure). For a majority of the datasets, FuseSOM outperforms FlowSOM. (B) Boxplots for the average clustering performance across all datasets. For all four evaluation metrics, FuseSOM shows an overall higher average score.

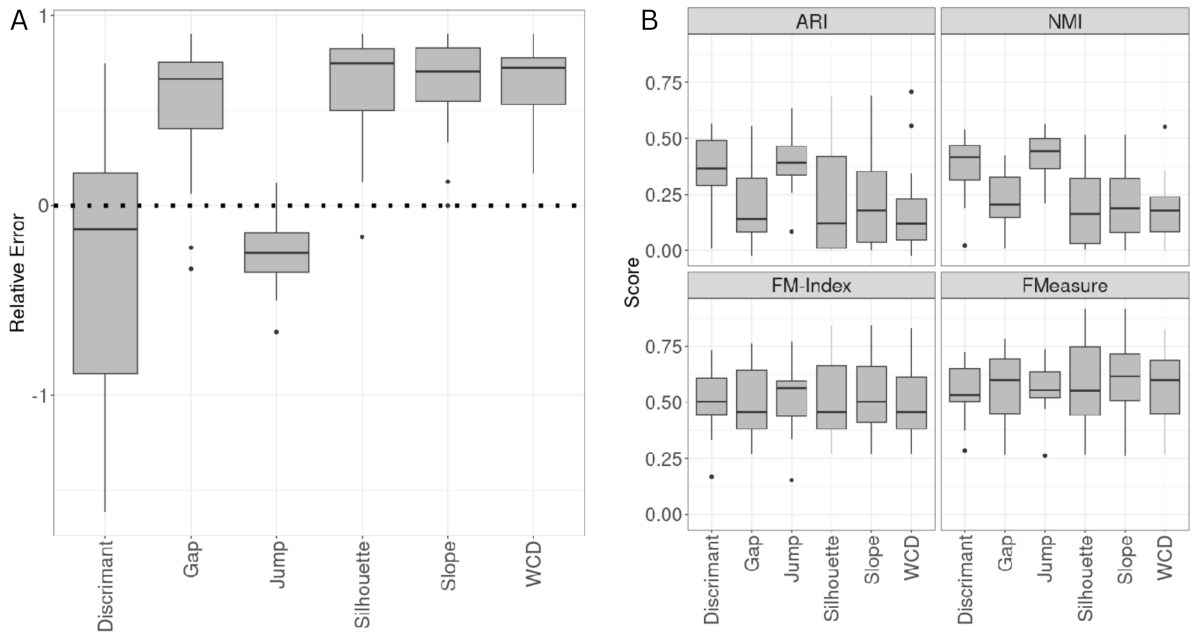


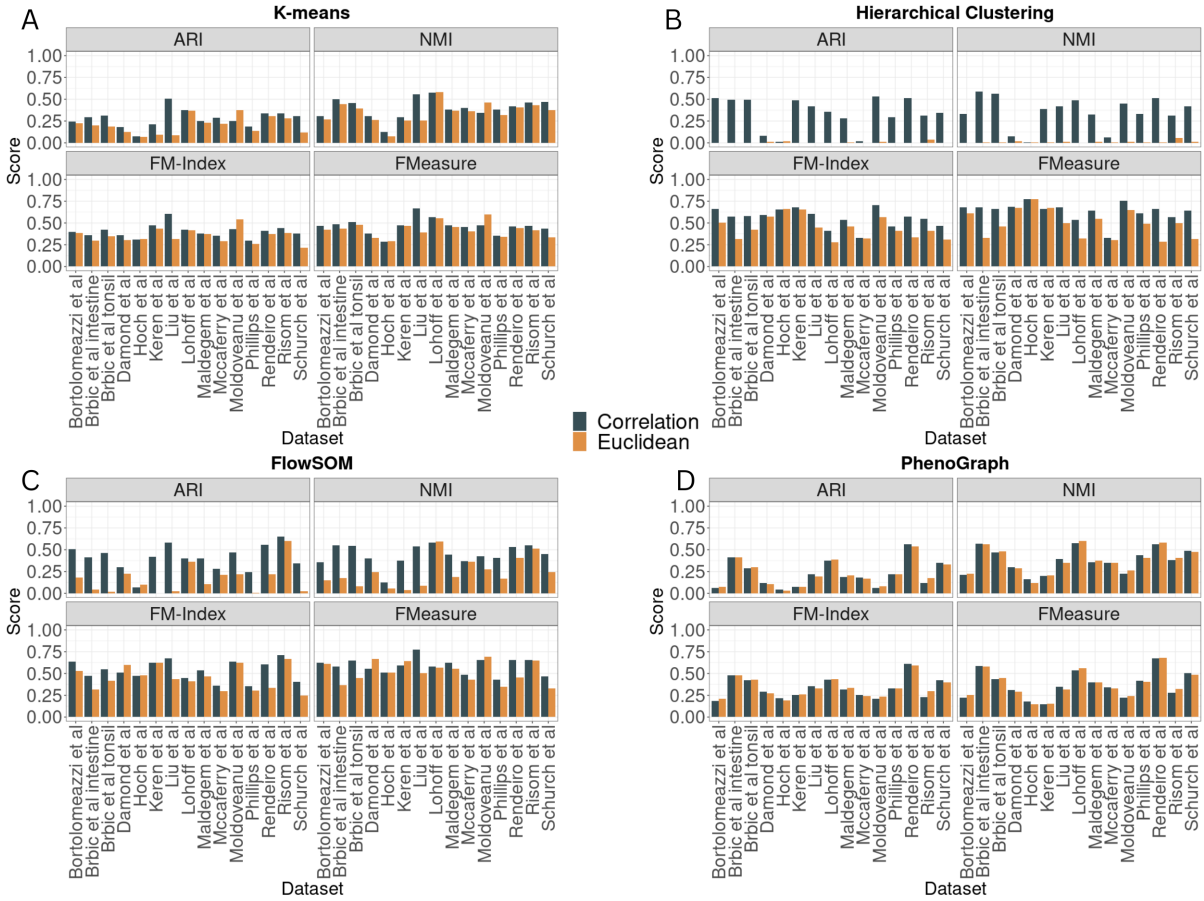
Figure 7: (A) Bar plots of cluster estimation performance for methods included in FuseSOM across all 15 datasets. The relative error metric was used to gauge performance. The Jump and Discriminant methods are the top performers. The dashed line represents a relative error of zero. In contrast, values above this line indicate an underestimation of the true number of clusters, and values below indicate an overestimation of the true number of clusters. (B) Boxplots of clustering performance based on the estimated number of clusters for methods included in FuseSOM. The Jump and Discriminant methods are top performers across all evaluation metrics.

Tables

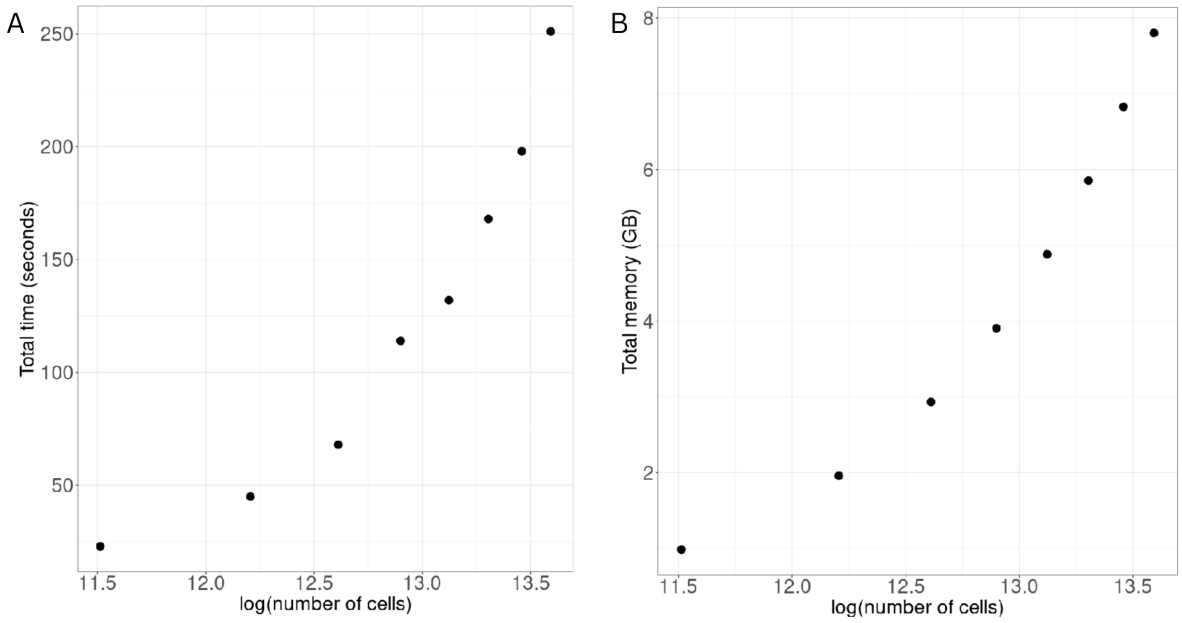
Table 1: Imaging Datasets Used

Dataset	Technology	Num Markers	Num Cells	Num Celltypes	Disease	Tissue
Shcurch et al [45]	CODEX	49	258,385	29	Colorectal Cancer	Colon
Phillips et al [34]	CODEX	52	117,170	21	Lymphoma	Skin
Brbic et al [6]	CODEX	48	248,285	21	None	Small intestine/colon
Brbic et al [6]	CODEX	44	219,926	13	Normal/Bartett's esophagus	Tonsil
Moldoveanu et al [31]	IMC	12	227,592	10	Melanoma	Skin
Van Maldegem et al [25]	IMC	17	282, 837	16	Lung Cancer	Lung
Hoch et al [15]	IMC	41	864,263	10	Melanoma	Skin
Rendeiro et al [35]	IMC	38	515,791	17	Covid-19	Lung
Damond et al [8]	IMC	36	252,059	16	Type 1 Diabetes	Pancreas
Bortolomeazzi et al [5]	IMC	30	218,615	9	Colorectal Cancer	Colon
Risom et al [37]	MIBI-TOF	22	69,672	23	Breast Cancer	Breast
Keren et al [17]	MIBI-TOF	16	201,656	6	TN Breast Cancer	Various
McCafferry et al [26]	MIBI-TOF	37	30,943	16	Tuberculosis	Various
Liu et al [22]	MIBI-TOF	12	345,490	8	Various Cancers	Various
Lohoff et al [23]	seqFISH	50	57,536	24	None	Mouse embryos

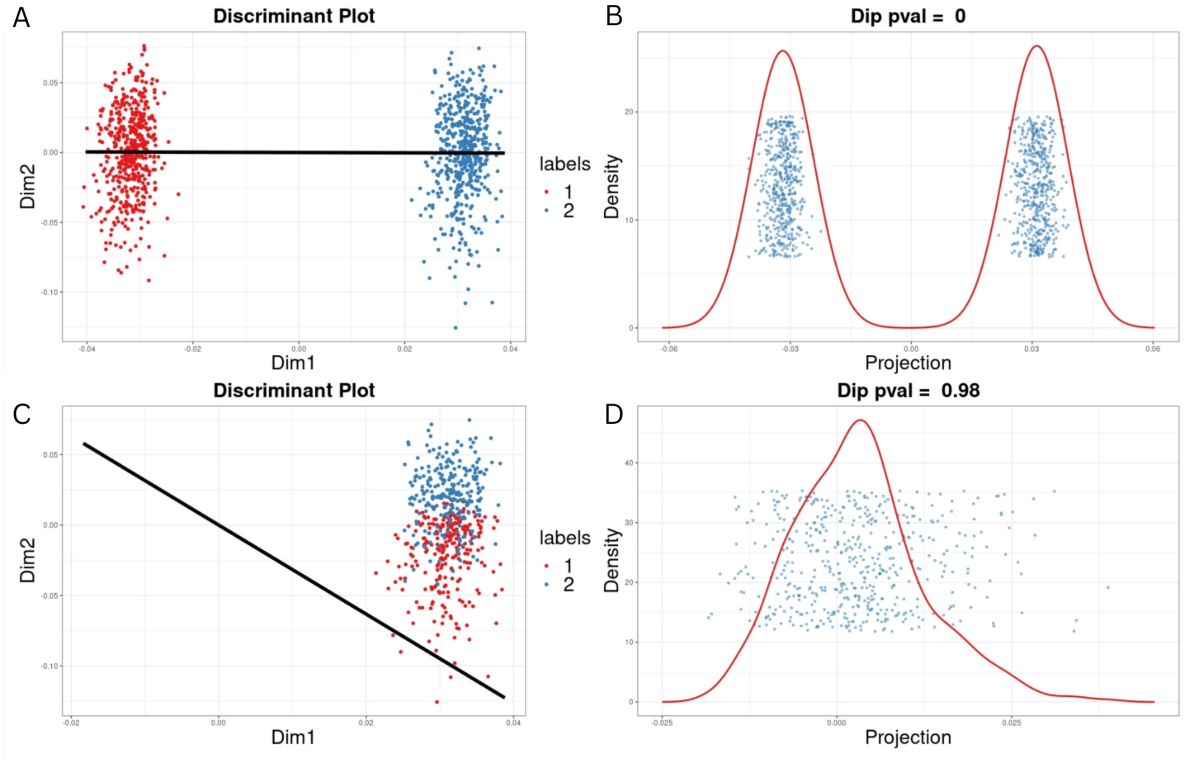
Supplementary Figures



Supplementary Figure S1: (A) Bar plots of K-means clustering of a subset of 20K cells for 15 datasets using Pearson correlation and Euclidean distance. K-means with Pearson correlation, on average, performs better than Euclidean for most of the datasets across all evaluation metrics. (B) Bar plots of Hierarchical clustering of a subset of 20K cells for 15 datasets using Pearson correlation and Euclidean distance. On average, hierarchical clustering with Pearson correlation performs better than Euclidean for most datasets across all evaluation metrics. C) Bar plots of FlowSOM clustering of a subset of 20K cells for 15 datasets using Pearson correlation and Euclidean distance. FlowSOM clustering with Pearson correlation, on average, performs better than Euclidean for most datasets across all evaluation metrics. D) Bar plots of PhenoGraph clustering of a subset of 20K cells for 15 datasets using Pearson correlation and Euclidean distance. There is no evidence of performance differences between Pearson and Euclidean.



Supplementary Figure S2: (A) Scatter plot of total memory usage for FuseSOM across an increasing number of cells. (B) Scatter plot of total running time for FuseSOM across an increasing number of cells.



Supplementary Figure S3: An example of discriminant coordinates cluster number estimation. (A) and (C) Shows two classes separated by a discriminant line in black. (B) and (D) Shows the Dip test applied to the projection of the two classes onto this discriminant line. Note that for (B), a significant p-value is obtained, reflecting the class separation. For (D), the opposite is observed. P-values $< 2e^{-16}$ were rounded to 0.