Quantitative assessment of single-cell RNA-seq clustering with CONCORDEX

Kayla Jackson[1,3], A. Sina Booeshaghi[1], Ángel Gálvez-Merchán[1], Lambda Moses[1], Tara Chari[1], and Lior Pachter[1,2,*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA
[2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA
[3]Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[*]To whom correspondence should be addressed

April 2023

Abstract

Many single cell RNA-sequencing (scRNA-seq) data analysis workflows rely on methods that embed and visualize the properties of a k-nearest neighbor (kNN) graph in two-dimensions. These visualizations are typically combined with categorical labels assigned to individual data points and can support a range of analysis tasks, despite the fact that these embeddings are known to distort the local and global properties of the graph. Rather than relying on a two-dimensional visualization, we introduce a method for quantitatively assessing the concordance between a set of labels and the k-nearest neighbor graph. Our method, called CONCORDEX, computes for each node the fraction of neighbors with the same or different label, and compares the result to the mean obtained with random labeling of the graph. CONCORDEX can be used for any categorical label and can be interpreted via an intuitive heatmap visualization. We demonstrate its utility for assessment of clustering results and "integration". Since CONCORDEX can be used to directly visualize properties of a kNN graph, we also use CONCORDEX to evaluate how well two-dimensional embeddings capture the local and global structure of the underlying graph. We have made CONCORDEX available as a Python-based command line tool (`https://github.com/pachterlab/concordex`) and as a software package in Bioconductor:
`https://bioconductor.org/packages/concordexR`.

Introduction

Computational analysis of single-cell RNA-sequencing (scRNA-seq) data typically includes dimensionality reduction with linear methods such as principal component analysis (PCA) and nonlinear methods such as UMAP or t-SNE. These embeddings support a range of analysis tasks, including cluster validation ([Ran+21], [IGA22]), evaluation of modality integration ([Lue+22]) and batch correction ([Tra+20]), and inferring relationships between cell types ([CP22], [Qiu+23]).

Non-linear methods for dimensionality reduction are assumed to faithfully represent the data and are therefore used in a quantitative manner. For example, when two-dimensional embeddings suggest that nearby cells are from the same batch, it is assumed that a transformation must be applied to the data to correct for batch effect. Similarly, when embeddings are used to visualize cell types, if distinct types are distant in the two-dimensional embedding then the clustering is assumed to be valid. These interpretations rely primarily on visual inspection of the embedding in order to make conclusions about the underlying local and global structure of the data. They are not quantitative partly because they depend on methods that distort the geometry of cells ([Coo+22], [KB19], [CP22] and partly because visualization of the embeddings can be misleading [Kay23].

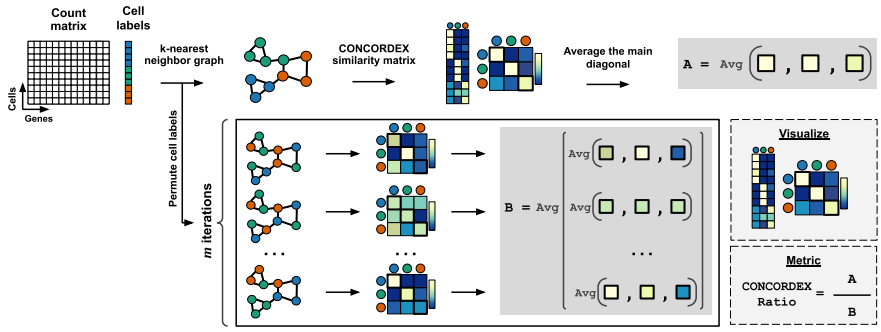
Rather than rely on visual assessment of two-dimensional embeddings that are based on kNN graphs (e.g. UMAP), we present a concordance index that we name CONCORDEX. CONCORDEX is an automated method for assessing and visualizing colorings applied to the nodes of kNN graphs, where a coloring refers to any discrete label applied to the nodes of the graph. We show that CONCORDEX can be used to analyze any graph with labeled nodes and we demonstrate its utility in revealing patterns that are challenging to estimate visually in non-linear embeddings. CONCORDEX is available as a Python command line tool (`https://github.com/pachterlab/concordex`) and as the software package concordexR in Bioconductor (`https://bioconductor.org/packages/concordexR`).

Results

Non-linear embeddings like t-SNE and UMAP attempt to place nodes of kNN graphs in two dimensions. The analyses they support are, for the most part, equivalent to asking to what extent the nodes in a kNN graph respect a given partition, or coloring of the nodes. Partitions can arise from cell-type labels, batches, samples, or other metadata.

CONCORDEX and CONCORDEX ratio

We use the kNN graph to assess the neighborhood composition of each cell (See Methods) and define the concordance index. Formally, let V be a set of cells,



Workflow of CONCORDEX

Figure 1: The processing pipeline to generate CONCORDEX. A kNN graph is constructed from an input count matrix and associated labels. For each cell, the longer matrix represents the fraction of neighbors that have the label indicated on the column. The aggregated similarity matrix is generated by averaging cells of the same type. CONCORDEX (A) is computed by averaging the entries on the main diagonal of the aggregated similarity matrix. We shuffle the labels on the k-nearest neighbor graph m times and recompute CONCORDEX at each iteration. The average CONCORDEX of the permuted (B) graphs is then used to compute the CONCORDEX ratio.

and consider a graph $G = (V, E)$ where E is the set of edges in a kNN graph computed according to some metric on V. Let P be a coloring of V according to elements of a set C, i.e. $P : V \to C$ where C is a finite set, with $|C| = p$. C may represent distinct clusters for the cells, batches, or any other discrete attribute associated to cells. The set $K_c = \{v | v \in V, P(u) = c\}$ is the set of cells with color $c \in C$. For each $v \in K_c$, the set $K_{v,c} = \{u | (uv) \in E, P(v) = c\}$ is the set of neighbors of v that have the same color. Similarly, the set $K_{v,c' \in C \setminus \{c\}} = \{u | (uv) \in E, P(v) = c'\}$ is the set of neighbors of v that have a different color. For a single cell, $\frac{|K_{v,c}|}{k}$ is the fraction of its neighbors with the same color. The fraction of neighbors with a different color is defined similarly.

We define a matrix with rows corresponding to cells and entries corresponding to the fraction of neighbors in each color across the columns. We aggregate the rows of according to the cells in K_c by averaging down the columns of each sub-matrix to obtain $\mathbf{S}$. The CONCORDEX between P and G is defined to be $\frac{1}{p} \sum_i \mathbf{S}_{ii}$. To assess whether the observed coloring is better than random, we shuffle the coloring P and recompute the CONCORDEX for m iterations. The ratio between the empirical CONCORDEX and average of the m permuted CONCORDEX statistics is referred to as the CONCORDEX ratio (Figure 1; Methods).

In other words, the CONCORDEX ratio measures the ratio between concordance of the partition of cells induced by P with the graph G to that of a random partition with G (Figure 1). A CONCORDEX ratio of 1 indicates

that the assignment was no better than random. A CONCORDEX ratio $k > 1$ indicates that k times as many cells assigned together than would be expected by chance. Unlike UMAP, CONCORDEX does not depend on numerous hyperparameters (only on the value of k used in construction of the kNN graph), and is deterministic.

Validation

To validate CONCORDEX, we simulated scRNA-seq datasets and compared the CONCORDEX ratio to the accuracy of cell-type assignment. (Figure 2). Specifically, we simulated scRNA-seq datasets with distinct cell types using Splatter ([ZPO17]) and then assigned cells using CellAssign ([Zha+19]), which uses marker genes for cell type assignment and has been shown to outperform similar methods [Xie+21]. The performance of CellAssign was ascertained with respect to the ground truth of the simulations. An assessment of performance would typically involve examining a UMAP constructed from the assigned cells, and an assignment would be considered to be good if cells were separated by type. Figure 2 shows that CONCORDEX is strongly correlated with the performance of CellAssign (Spearman $\rho = 0.99$, $p < 2.2 \times 10^{-16}$). We also investigated this relationship when the number of cell types varied in the simulated datasets (Supplementary Figure S1). Because there is difficulty in randomly coloring a the k-nearest neighbor graph when there are few labels (i.e. nearly all labels may appear in the cell neighborhood)([DF10], we find that the magnitude of CONCORDEX ratio varies with the number of simulated cell types, but the strength of the correlation remained the same (Supplementary Figure S1).

Uncovering Integration Effects

We evaluated the utility of CONCORDEX to assess the fidelity of dataset "integration" on data generated from in-utero and ex-utero mouse embryos at the E10.5 developmental stage ([Agu+21]. In-utero and ex-utero data appear well mixed in the UMAP visualization (Figure 3).

Since UMAP generation is often preceded by dimension reduction with PCA for scRNA-seq analysis, we first evaluated the data in the 15-dimensional PCA space. In the two-label case, a CONCORDEX near 0.5 and CONCORDEX ratio of 1 would suggest that the data are integrated. In the Log-normalized PCA space (CONCORDEX = 0.72; CONCORDEX Ratio = 1.42), most of the ex-utero cells have neighborhoods that are composed nearly entirely of cells with the same label (Figure 3a) and the in-utero cells show a similar pattern. This pattern is also apparent in the CONCORDEX matrices generated from the data in the UMAP space (CONCORDEX = 0.63; CONCORDEX Ratio = 1.25). While the CONCORDEX matrices show that in-utero and ex-utero cells in the PCA and UMAP embedded spaces are comparably separated, visual inspection of the UMAP embedding suggests that the data are well mixed. The UMAP embedding of the Scaled-Stabilized data is similarly misleading. The aggregated CONCORDEX similarity matrix for the PCA space shows that the

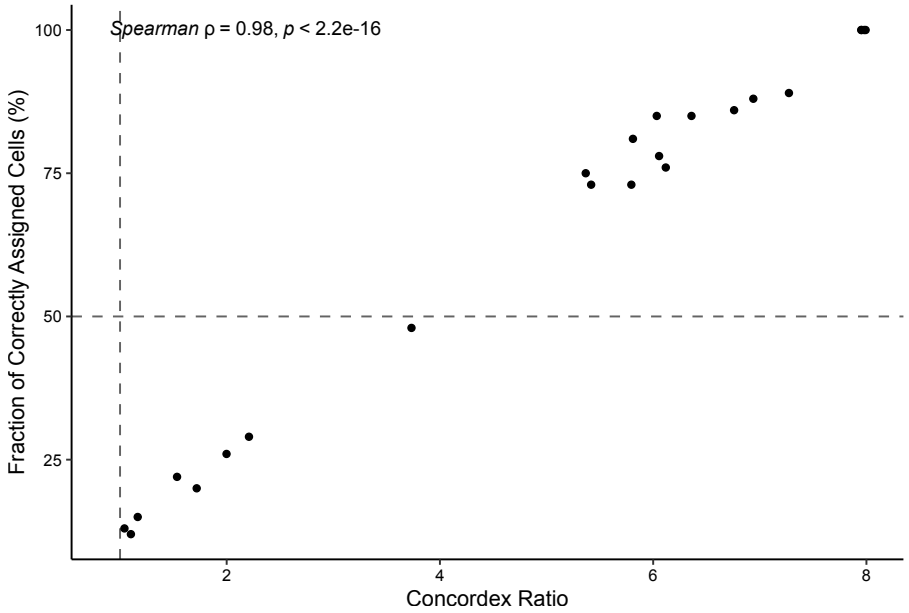


Figure 2: CONCORDEX correlates with assignment accuracy. We computed the CONCORDEX ratio for datasets simulated with Splatter. The simulated datasets contained 8 distinct cell types and 1000, 2000, 4000, or 8000 cells. The cell types were assigned by CellAssign using up to 15 marker genes.

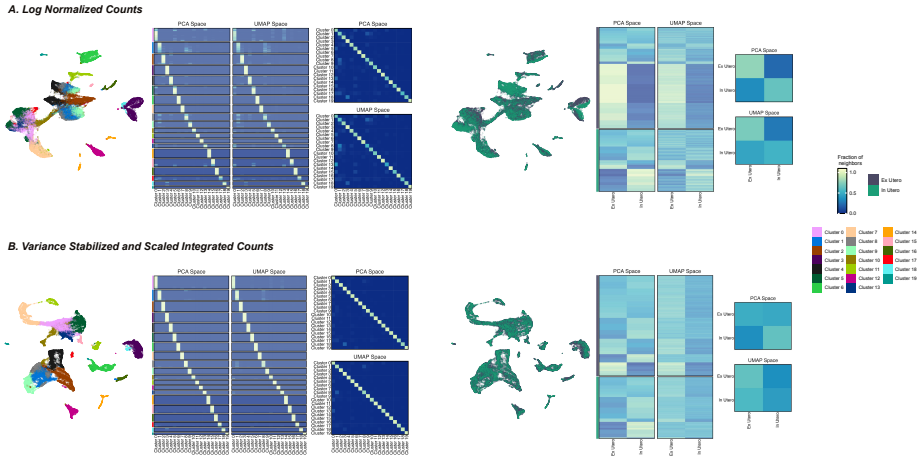
neighborhoods of the ex-utero cells are nearly balanced, but we found that this is largely due to the effects of averaging. The expanded matrix reveals that the fraction of ex-utero cells with neighborhoods that are composed mostly of cells with same label is similar to the fraction of cells with neighborhoods composed of mostly cells with the opposite label. The neighborhoods of the in-utero cells are mostly composed of other in-utero cells (Figure 3b). We also observed this pattern in the UMAP space. There, the neighborhoods of in-utero cells are biased to include ex-utero cells rather than a mix of both cell labels. The ex-utero cells become more isolated, with neighborhoods being composed of only 36% in-utero cells on average. While neither CONCORDEX nor the embeddings determine whether the mixing effects are a result of the normalization method, integration procedure, or initial reduction to a few dimensions using PCA (Figure S2), visual inspection of the UMAP embedding does not reliably reveal whether or not integration was effective. On the other hand,

CONCORDEX revealed that the combined effects of the data transformations did not fully "integrate" the datasets and that reduction to two-dimensional can suggest misleading associations between cells.

Uncovering UMAP Effects on Cell-type Clustering

We further tested the utility of CONCORDEX for cluster validation on data generated from ex-utero and in-utero mouse embryos dataset ([Agu+21]). In this context, having fractions near to 1 on the main diagonal of the aggregated CONCORDEX similarity matrix implies that clusters are well isolated. We found that the CONCORDEX similarity matrix can readily visualize distinct clusters in both the PCA and UMAP (Figure 3) spaces. In the Log-normalized PCA data, CONCORDEX (CONCORDEX $= 0.83$; CONCORDEX Ratio $= 16.7$) deviates from 1 and suggests an interaction between some clusters. For example, the neighborhoods of cells in cluster 0 are composed of cells from clusters 5, 7, 10, and 13. Since the UMAP algorithm permits distortion of the intra-celltype and inter-celltype distances, we observe that CONCORDEX diminishes in the UMAP space (CONCORDEX $= 0.79$; CONCORDEX Ratio $= 15.9$). Notably, this metric can be somewhat improved by tuning the hyperparameters of UMAP algorithm (Supplementary Figure S4). Simply changing the random state improves CONCORDEX in the UMAP space (CONCORDEX $= 0.80$; CONCORDEX Ratio $= 16.03$) reflecting the arbitrary nature of otherwise equivalent embeddings. While tuning the hyper-parameters can alter how well the UMAP embedding preserves the structure of the PCA space, CONCORDEX directly visualizes the cluster relationships in the PCA-embedded data and reduces the need to embed the data in fewer dimensions.

We also observed that the UMAP embedding can be a poor representation of the k-nearest neighbor graph in general. For example, the clusters in the Scaled-stabilized data are well defined in the PCA space (CONCORDEX $=$ 0.94; CONCORDEX Ratio $= 18.7$) and UMAP space (CONCORDEX $= 0.93$; CONCORDEX Ratio $= 18.7$). Although CONCORDEX indicates that the clusters are well separated, visual inspection of the UMAP embedding suggests a close relationship between several clusters that overlap in the plot even though this structure is not apparent in the k-nearest neighbor graph generated from the PCA-embedded data. The long CONCORDEX matrix allows user to assess the degree to which clusters overlap. Stated another way, users can ask which cluster has the greatest number of cells where more than half of their neighbors have a different label. We found that the number of cells with mixed neighborhoods is often greater in the UMAP Space (Supplementary Figure S3), suggesting that clusters are more mixed in the UMAP space than they are in the PCA space. These relationships are especially difficult to estimate in the UMAP embedding alone, where overlapping clusters can appear indistinguishable. Overall, CONCORDEX provides an exact estimate for mixing between clusters and allows users to differentiate true structure in the data from overplotting, and the CONCORDEX similarity matrix is sufficient to visualize the global structure of the k-nearest neighbor graph and when combined with the



Fraction of neighbors in each cluster and growth condition

Figure 3: Left plots show the UMAP embedding of the a) Log-normalized and integrated counts or b) Variance stabilized, scaled, and integrated counts with the expanded and averaged CONCORDEX similarity matrices next to it. In the expanded CONCORDEX similarity matrices, each row is a cell and cells are grouped by their assigned cluster or batch. When the heatmaps are grouped by cluster assignment, the CONCORDEX visualizations show that almost all clusters are well separated which contradicts visual inspection of the UMAP embedding. Similarly, visual inspection of the UMAP colored by growth condition suggests that the data are effectively integrated, however, the heatmaps show that most cells have neighborhoods that are exclusive of one growth condition.

CONCORDEX measures.

Discussion

We showed that the CONCORDEX ratio is strongly correlated with accuracy of cluster assignment in a simulated dataset. In the simulation, the magnitude of the CONCORDEX ratio depended on the number of labels. This is largely due to the difficulty in randomly coloring the k-nearest neighbor graph when the number of labels is far less than the maximal degree of the graph (i.e k)([DF10]). The most extreme case occurs when there are only two labels and the distribution of labels across all neighborhoods deviates from uniformity. This increases the probability that the random labels more closely approximate the true labels, and thus the overall ratio is diminished. This pitfall reveals a general limitation of using the k-nearest neighbor graph in methods to assess cluster accuracy and suggests a need to develop metrics that are independent of the graph.

A typical use case of CONCORDEX includes assessing the existence of and relationships between predefined groups or clusters. In contrast to UMAP, the

CONCORDEX similarity matrix can be used to readily visualize distinct clusters without distorting the global relationships between them. The long matrix is especially useful for estimating the proximity of clusters and presents a more natural interpretation of the biological relationship between them. Given that UMAP plots are also used to visualize gene expression data within a cluster, we note that gene expression can be readily plotted as a heatmap grouped by predefined clusters without loss of information present in the UMAP visualization.

In a typical scRNA-seq analysis workflow, the kNN graph is generated from a few principal components. Thus, the two-dimensional embedding should represent the local and global structure generated from the underlying PCA-embedded data. We found that added variance in the UMAP embedding distorts these relationships and that there is a discordance between the underlying kNN graph visualized by CONCORDEX and what can be assessed visually in the low-dimensional embedding. In the case of modality or batch integration, data can appear more mixed in the UMAP space despite contrasting evidence in the PCA space. For cluster relationships, clusters that are isolated in the PCA space can appear indistinguishable from nearby classes in the UMAP embedding. What's more, the hyper-parameters of low-dimensional representations can be adjusted to improve concordance with the PCA space or better align with user expectation. Despite this, "improved" embeddings can still fail to accurately represent the underlying structure in the PCA space. This observation reflects a significant limitation of UMAP embeddings and supports the notion that they can not be used as reliable representations of the underlying structure of the data.

We believe that CONCORDEX more accurately reflects the underlying structure of embedded scRNA-seq data. CONCORDEX can be computed on any colored kNN graph, but we showed that it has practical applications for cluster assessment and batch or integration comparison. Furthermore, CONCORDEX can be used to directly visualize and assess PCA embedded data and preclude the need to further embed data to two-dimensions.

Methods

Datasets and pre-processing

The datasets used in this study were derived from reference [Agu+21]. A detailed description of the pre-processing steps are described elsewhere ([CP22]). Briefly, two transformed count matrices were provided by the authors. The 'Log-Normalized' dataset consisted of log-transformed counts and the 'Stabilized-Scaled' dataset had been variance stabilized using Seurat and subsequently mean-centered and scaled. Both datasets were integrated using Seurat. The count matrices were mean-centered and scaled before principal component analysis (PCA). PCA analysis was performed using sklearn TruncatedSVD to 15 dimensions to agree with the analysis performed by the original authors. We used the findKNN function from the BiocNeighbors package in R to find the

exact 30 nearest neighbors for each cell in the dataset.

The UMAP algorithm was applied to the 15-dimensional PCA embeddings with default settings except where noted in Supplementary Figure S4. The subsequent visualizations were colored using labels provided by the authors.

Simulated Datasets

Simulated datasets were generated using Splatter ([ZPO17]) with default parameters. We generated datasets with 1000, 2000, 4000, or 8000 cells. For each dataset we simulated 2, 4, 6, or 8 cell types that were distinguished by 1, 3, 5, 7, 9, 11, or 13 marker genes. The probability that a simulated gene was a marker gene was fixed at 0.20.

The cell types for each simulated dataset were predicted using CellAssign ([Zha+19]). We compared the predicted cell types to the known cell types and computed to the fraction of correctly assigned cells for each simulated dataset.

CONCORDEX: k-nearest neighbor concordance index

The k-nearest neighbor graph $G = (V, E)$ can be generated from a scRNA-seq dataset where V is the set of cells and E, the set of edges, is determined according to some metric on V. The number of cells in the dataset is denoted $|V| = n$

We consider a coloring P of G according to elements of a finite set C, with $|C| = p$. The elements of C can be any discrete label, including cluster assignment, growth condition, or batch.

CONCORDEX is generated by considering the subgraph $G_c = (V_c, E_c)$ for each $c \in C$. That is, for each discrete label we consider the nodes in the graph with label c and all edges of these nodes.

We construct an $n \times p$ matrix from these subgraphs. For each $v \in V_c$ we compute the fraction of its k neighbors that have the label $c' \in C$. The rows of the matrix are colored according to coloring on V. The $p \times p$ similarity matrix, $\mathbf{S}$ is computed from the larger matrix by averaging the down the rows of the same color and so that the rows and columns are in the same order.

The CONCORDEX metric is computed by averaging the fractions on the main diagonal of $\mathbf{S}$. That is,

$$\text{CONCORDEX} = \frac{1}{p} \sum_{i=1}^{p} \mathbf{S}_{ii}$$

To compute the CONCORDEX ratio, we shuffle the coloring on V, keeping the multiplicity of each label fixed, and recompute CONCORDEX as defined above. The shuffling is carried out over m iterations and the m metrics are averaged. The ratio is defined as the observed CONCORDEX divided by the average of the m shuffled metrics.

The CONCORDEX ratio is defined as follows,

$$\text{CONCORDEX ratio} = \frac{\sum_{i=1}^{p} \mathbf{S}_{ii}}{\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{p} \mathbf{S}_{ii}^{(j)}}$$

Here, the index j keeps track of the iteration for shuffling the node colors, and the matrix $\mathbf{S}^{(j)}$ is the similarity matrix generated on iteration j.

Data Availability

Data and code to reproduce the figures in this manuscript are available at the following Github repository: `https://github.com/pachterlab/JBMMCP_2023/`

References

[DF10] Martin Dyer and Alan Frieze. "Randomly coloring random graphs". en. In: *Random Structures & Algorithms* 36.3 (2010). _eprint: https://onlinelibrary.wiley.com/doi/pd pp. 251–272. ISSN: 1098-2418. DOI: 10.1002/rsa.20286. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20286 (visited on 05/22/2023).

[ZPO17] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Splatter: simulation of single-cell RNA sequencing data". In: *Genome Biology* 18.1 (Sept. 2017), p. 174. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1305-0. URL: https://doi.org/10.1186/s13059-017-1305-0 (visited on 04/20/2023).

[KB19] Dmitry Kobak and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics". en. In: *Nature Communications* 10.1 (Nov. 2019). Number: 1 Publisher: Nature Publishing Group, p. 5416. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13056-x. URL: https://www.nature.com/articles/s41467-019-13056-x (visited on 05/01/2023).

[Zha+19] Allen W. Zhang et al. "Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling". en. In: *Nature Methods* 16.10 (Oct. 2019). Number: 10 Publisher: Nature Publishing Group, pp. 1007–1015. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0529-1. URL: https://www.nature.com/articles/s41592-019-0529-1 (visited on 04/20/2023).

[Tra+20] Hoa Thi Nhu Tran et al. "A benchmark of batch-effect correction methods for single-cell RNA sequencing data". In: *Genome Biology* 21.1 (Jan. 2020), p. 12. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1850-9. URL: https://doi.org/10.1186/s13059-019-1850-9 (visited on 05/01/2023).

[Agu+21] Alejandro Aguilera-Castrejon et al. "Ex utero mouse embryogenesis from pre-gastrulation to late organogenesis". en. In: *Nature* 593.7857 (May 2021). Number: 7857 Publisher: Nature Publishing Group, pp. 119–124. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03416-3. URL: https://www.nature.com/articles/s41586-021-03416-3 (visited on 04/19/2023).

[Ran+21] Bobby Ranjan et al. "scConsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data". In: *BMC Bioinformatics* 22.1 (Apr. 2021), p. 186. ISSN: 1471-2105. DOI: 10.1186/s12859-021-04028-4. URL: https://doi.org/10.1186/s12859-021-04028-4 (visited on 05/22/2023).

[Xie+21] Bingbing Xie et al. "Automatic cell type identification methods for single-cell RNA sequencing". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 5874–5887. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2021.10.027. URL:

https://www.sciencedirect.com/science/article/pii/S2001037021004499.

[CP22] Tara Chari and Lior Pachter. "The Specious Art of Single-Cell Genomics". In: *bioRxiv* (2022). Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2022/12/22/2021.08.25.457696.full.pdf. DOI: 10.1101/2021.08.25.457696. URL: https://www.biorxiv.org/content/early/2022/12/22/2021.08.25.457696.

[Coo+22] Shamus M. Cooley et al. *A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-seq data*. en. Pages: 689851 Section: New Results. Jan. 2022. DOI: 10.1101/689851. URL: https://www.biorxiv.org/content/10.1101/689851v6 (visited on 05/01/2023).

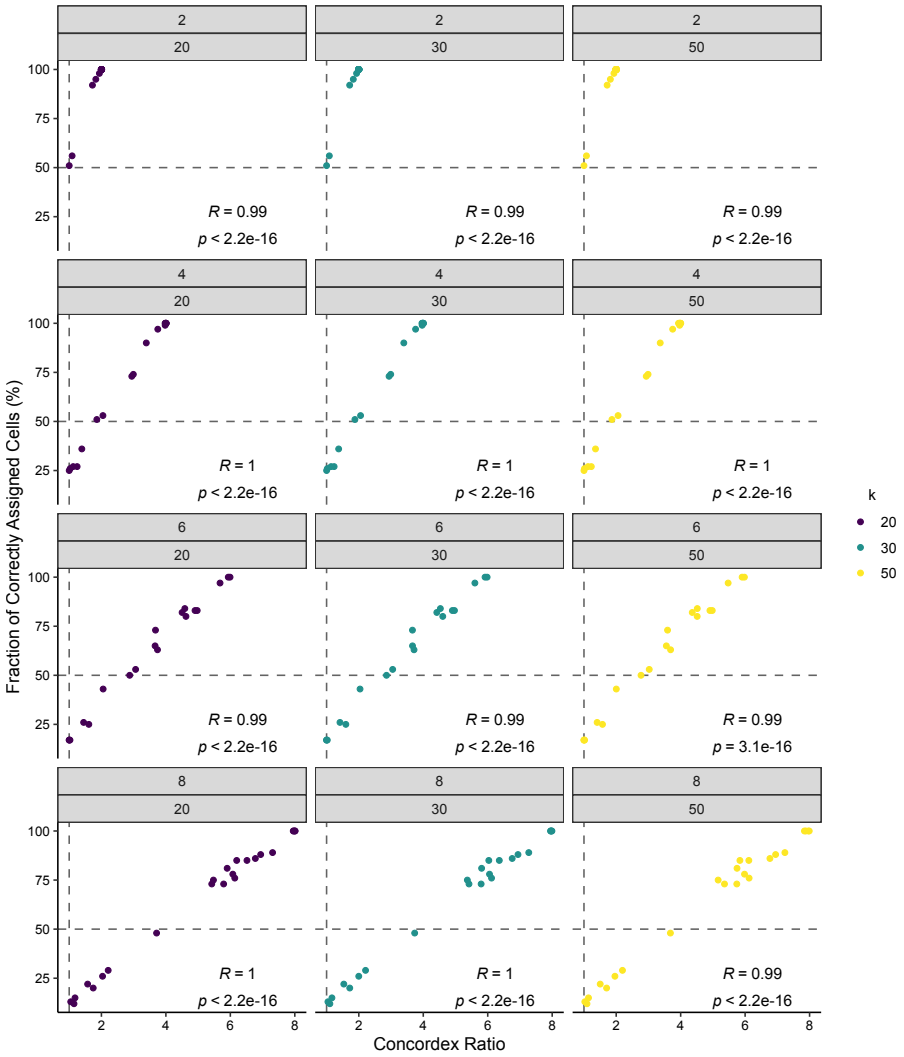
[IGA22] Aleksandr Ianevski, Anil K. Giri, and Tero Aittokallio. "Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data". en. In: *Nature Communications* 13.1 (Mar. 2022). Number: 1 Publisher: Nature Publishing Group, p. 1246. ISSN: 2041-1723. DOI: 10.1038/s41467-022-28803-w. URL: https://www.nature.com/articles/s41467-022-28803-w (visited on 05/22/2023).

[Lue+22] Malte D. Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". en. In: *Nature Methods* 19.1 (Jan. 2022). Number: 1 Publisher: Nature Publishing Group, pp. 41–50. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01336-8. URL: https://www.nature.com/articles/s41592-021-01336-8 (visited on 05/22/2023).

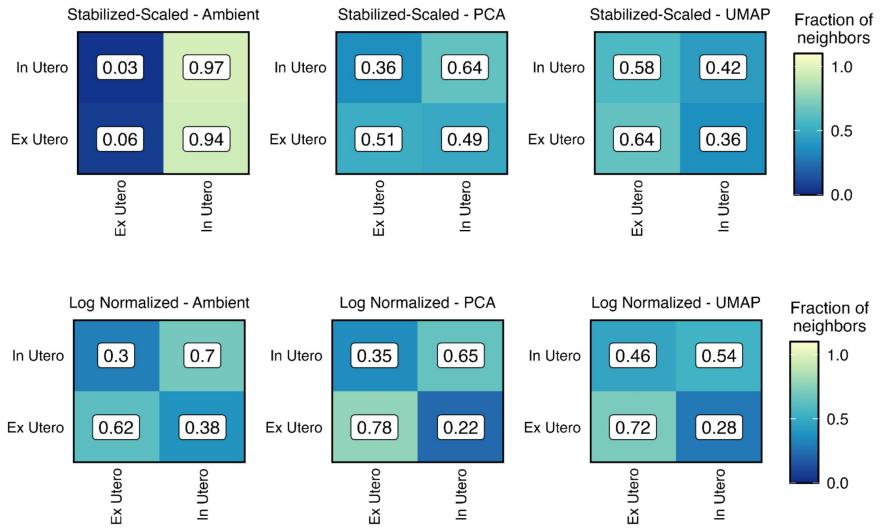
[Kay23] Matthew Kay. *ggblend: Blending and Compositing Algebra for 'ggplot2'*. R package version 0.1.0. 2023. URL: https://CRAN.R-project.org/package=ggblend.

[Qiu+23] Chengxiang Qiu et al. *A single-cell transcriptional timelapse of mouse embryonic development, from gastrula to pup*. en. Pages: 2023.04.05.535726 Section: New Results. Apr. 2023. DOI: 10.1101/2023.04.05.535726. URL: https://www.biorxiv.org/content/10.1101/2023.04.05.535726v1 (visited on 05/01/2023).

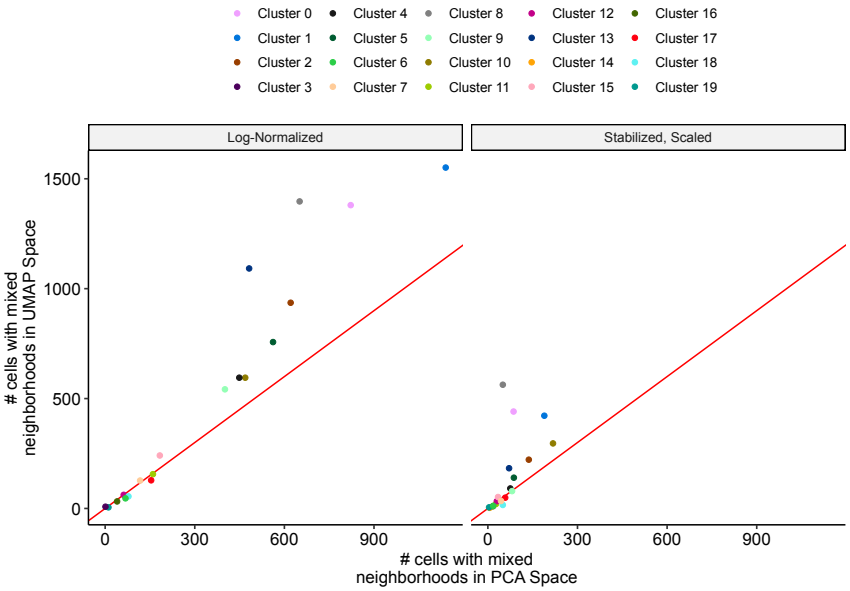
Supplementary Figures



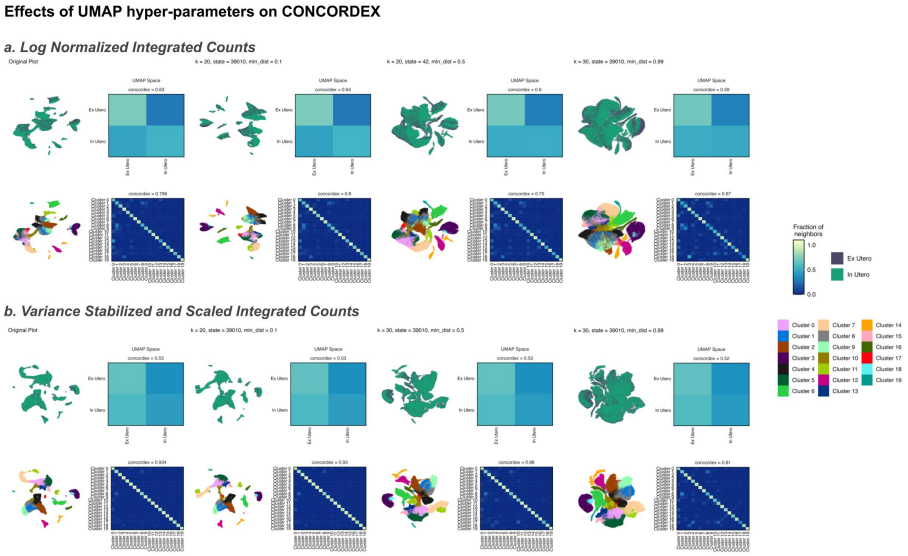
Supplementary Figure S1: Relationship between CONCORDEX ratio, assignment accuracy, number of groups, and number of neighbors. Plots show the correlation between CONCORDEX ratio and fraction of correctly assigned cell types in a simulated dataset. Each row represents the number of simulated cell types (upper label) and each column represents the number of neighbors (i.e. k) for in construction of the k-nearest neighbor graph used to compute CONCORDEX.



Supplementary Figure S2: Effect of dimension reduction on CONCORDEX. Panels show the effect of dimension reduction by PCA and UMAP on neighborhood similarity on Variance stabilized (upper) and Log Normalized (lower) datasets. Here 'ambient' refers to the centered and scaled data after initial transformation (i.e. log normalizaton or variance stabilization)



Supplementary Figure S3: Comparison of neighborhood mixing in embedded data. Cells with majority mixed neighborhoods were identified for each cluster. A cell was called mixed if more than half of its nearest neighbors belong to a different cluster. The red line is the line y=x.



Supplementary Figure S4: UMAP hyperparameters were adjusted as indicated by the figure titles. For integration results (a,b, upper panels) all UMAP tests suggest successful integration by visual inspection. The neighborhoods of the cells become artifically distorted in the log-normalized data simply by changing the random state, minimum distance, or number of neighbors. Notably, in the variance-stabilized data, choice of hyper-parameters has little effect on neighborhood composition with respect to growth condition. The user is left to choose the plot that matches their preferred aesthetics. For cluster validation (a,b, lower panels) changing hyper-parameters results in UMAP embeddings with broadly similar features (e.g. isolated clusters remain so in distinct embeddings), yet CONCORDEX can vary considerably.