Dirichlet process mixture models for single-cell RNA-seq clustering

Nigatu A. Adossa[1*], Kalle T. Rytkönen[1,2], Laura L. Elo[1,3*]

[1]Turku Bioscience Centre, University of Turku and Åbo Akademi University, FI-20520, Turku, Finland, [2]Institute of Biomedicine, Research Centre for Integrative Physiology and Pharmacology, University of Turku, FI-20014, Finland, [3]Institute of Biomedicine, University of Turku, FI-20014, Finland

[*]Corresponding authors: nigatu.a.adossa@utu.fi (NAA), laura.elo@utu.fi (LLE)

Abstract

Clustering of cells based on gene expression is one of the major steps in single-cell RNA-sequencing (scRNA-seq) data analysis. One key challenge in cluster analysis is the unknown number of clusters and, for this issue, there is still no comprehensive answer. In order to enhance the process of defining meaningful cluster resolution, we compare Bayesian Latent Dirichlet Allocation (LDA) method to its non-parametric counterpart, Hierarchical Dirichlet Process (HDP) in the context of clustering scRNA-seq data. A potential main advantage of HDP is that it does not require the number of clusters as an input parameter from the user. While LDA has been used in single-cell data analysis, it has not been compared in detail with HDP. Here, we compare the cell clustering performance of LDA and HDP using four scRNA-seq datasets (immune cells, kidney, pancreas and decidua/placenta), with a specific focus on cluster numbers. Using both intrinsic (DB-index) and extrinsic (ARI) cluster quality measures, we show that the performance of LDA and HDP is dataset dependent. We describe a case where HDP produced a more appropriate clustering compared to the best performer from a series of LDA clusterings with different numbers of clusters. However, we also observed cases where the best performing LDA cluster numbers appropriately capture the main biological features while HDP tended to inflate the number of clusters. Overall, our study highlights the importance of carefully assessing the number of clusters when analyzing scRNA-seq data.

1. Introduction

Recent advances in single-cell sequencing have enabled increased resolution of biological and medical studies of cellular functions. Single-cell RNA-sequencing (scRNA-seq) is widely used to study cellular heterogeneity in cancer, developmental biology, immunology and neurology (Tang et al., 2019). Clustering of cells based on their gene expression profiles is one of the major steps in scRNA-seq data analysis. For instance, the computational analysis of scRNA-seq data for cell-type identification has mainly relied on unsupervised clustering methods (Qi et al., 2019), such as distance-based cluster optimization, density-based clustering, or graph-based clustering methods (Petegrosso et al., 2019).

Single-cell clustering methods are mainly based on optimization of the pairwise distance between cells (Petegrosso et al., 2019; Qi et al., 2019), which is a challenging task due to the high dimensionality of the data (Remesh and Pattabiraman, 2017). The choice of the distance metric also affects the clustering result (Singh et al., 2013). Bayesian clustering, which uses sampling-based inference methods for clustering, can be utilized to address these challenges. Unlike traditional distance optimization-based techniques, the Bayesian approach uses soft cluster assignments, in which the data points are assigned to each cluster according to their probability of uncertainty, allowing a mixed cluster membership. Moreover, sampling-based Bayesian clustering methods avoid distance calculation, allowing a tractable way of dealing with high dimensional data. One such Bayesian admixture model is latent Dirichlet allocation (LDA) (Blei et al., 2003), which recently has been successfully adopted for clustering of both scRNA-seq (Dey et al., 2017; duVerle et al., 2016; Sun et al., 2018; Wang et al., 2021) and scATAC-seq (Bravo González-Blas et al., 2019) data.

One key challenge in cluster analysis is the choice of cluster resolution. This is inherently linked to one of the great advances of single-cell sequencing, which is the discovery of previously unknown cellular states or even new cell types. There are several clustering methods available for scRNA-seq data analysis with different parameters regulating the cluster resolution. For instance, Seurat 4 (Hao et al., 2021) implements the shared-nearest neighbor (SNN) graph-based clustering on PCA space with modularity optimization and a user-selected parameter regulating the cluster resolution (Butler et al., 2018). Similarly, Monocle 3 (Cao et al., 2019) implements graph-based community detection algorithms with a

user-defined input resolution parameter. However, an inappropriate choice of these parameters may impede the discovery of novel cell states or types.

To address these challenges, in this study, we investigate the utility of hierarchical Dirichlet process (HDP) (Teh et al., 2006) for clustering scRNA-seq data as a non-parametric counterpart of LDA. The HDP method has been applied, for example, to correct technical variations for scRNA-seq data (Prabhakaran et al., 2016), to segment gene regulatory networks (Wang and Wang, 2013) and to cluster bulk gene expression data (Wang and Wang, 2013). Here we apply HDP to cluster scRNA-seq data and compare its performance to LDA. We analyze in detail three publicly available scRNA-seq datasets, including artificially mixed human immune cells, and two tissue-specific subsets of kidney and pancreas cells from *Tabula Muris* (Schaum et al., 2018), with high quality cell type annotations. Additionally, we also test the scalability of the methods with a large dataset from human decidua/placenta (Vento-Tormo et al., 2018). We specifically focus on the clustering resolution necessary to capture the cellular heterogeneity using both intrinsic and extrinsic cluster quality measures.

2. Results

To study the performance of LDA and HDP clustering models in identifying the cellular heterogeneity from scRNA-seq data, we applied them to an artificial mixture of human immune cells (S1 Table), mouse kidney cells, and mouse pancreas cells (Schaum et al., 2018). For each dataset, the cluster quality was measured first intrinsically using the Davies-Bouldin index (DB-index) and secondly extrinsically using the Adjusted Rand Index (ARI) with the reference clusters from the original publications (see Materials and methods for details). In addition to DB-index, we also tested the intrinsic cluster quality with Calinski-Harabasz (CH-index) (Calinski and Harabasz, 1974), which overall gave similar results as the DB-index (Fig. S1). Finally, the clustering results of the best two k values based on the intrinsic DB-index were visualized using the UMAP plots side by side with the reference cell type annotations from the original publications. In each dataset, we ran HDP clustering with 20 repetitions and a series of LDA clusterings with an increasing number of clusters k from 2 to 20 (20 repetitions each) using the default parameters. The run time for a single analysis on a 48 core Ubuntu 16.04 EC2 cloud instance was ~2-3 minutes for LDA in the immune cells (~1000 cells), pancreas cells (~2000 cells) and kidney cells (~3000 cells), whereas the run time for HDP increased from ~6 minutes with ~1000 cells to ~15 minutes with ~2000 cells

and ~28 minutes with ~3000 cells (Table S2). LDA and HDP run times for the decidua/placenta (64,000 cells) took 1.35 hours and 4 days respectively, and this data was not used for the full comparison between LDA and HDP. The memory usage was similar between LDA and HDP (Table S2).

2.1 LDA and HDP clustering performance in human immune cells

In the intrinsic evaluation of the human immune cell data, the two lowest (best) DB-index values with LDA were obtained with k = 3 (DB = 2.3) and k = 5 (DB = 2.4) clusters (Fig 1A), whereas for HDP those were k = 7 (DB = 2.3) and k = 9 (DB = 2.4) (Fig 1B).

Fig 1. Comparison of LDA and HDP clustering performance using artificially mixed human immune cell scRNA-seq data. Intrinsic cluster quality measure defined by Davies-Bouldin index (DB-index) for (A) LDA and (B) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates the DB-index values (lower indicates better clustering). Extrinsic cluster quality measure defined by Adjusted Rand Index (ARI) for (C) LDA and (D) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates ARI (higher indicates better clustering). For (A-D) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of LDA clustering with (E) k = 3 and (G) k = 5. The UMAP plot of HDP clustering with (F) k = 7 and (H) k = 9. (I) The UMAP plot showing the reference clustering with the cell-type annotation from the original publications (S1 Table).

In the extrinsic cluster evaluation, increasing the LDA cluster number to k = 5 resulted in an increasingly better quality in terms of ARI, but larger cluster numbers did not affect the quality markedly (Fig 1C). The mean ARI values for two best DB-index informed HDP clusterings (k = 7 and k = 9) had higher ARI values (~0.6) than those of LDA (<0.5 for k = 3 and k = 5) (Fig 1D). Thus, the extrinsic quality measures were in line with the intrinsic DB-index values, suggesting that – judged by the reference clusters – HDP performed slightly better than LDA in this dataset.

We next visually inspected the best performing clusterings selected by DB-index with UMAP plots by comparing these to the reference cell type annotations (Fig 1E-I). HDP with k = 7 resolved the main reference cell types (Fig 1F), whereas LDA with k = 5 did not (Fig 1G).

Specifically, LDA with k = 5 had one cluster containing B cells, dendritic cells and lymphoblasts together, whereas HDP with k = 7 or k = 9 was able to resolve these three cell types to their own clusters. Overall, for this dataset, when comparing with the reference clusters, the DB-index informed HDP was able to predict a biologically more adequate clustering than DB-index informed LDA.

2.2 LDA and HDP clustering performance in mouse kidney cells

In the intrinsic evaluation of the mouse kidney data (Schaum et al., 2018), LDA with cluster numbers k = 6 and k = 12 showed the minimum average DB-index values of 2.2 and 2.3, respectively (Fig 2A), indicating the highest intrinsic cluster quality. The HDP clustering result partitioned the dataset into k = 11 clusters with the lowest average DB-index value of 2.6 followed by k = 17 with average DB-index value of 3.0 (Fig 2B).

Fig 2. Comparison of LDA and HDP clustering performance using mouse kidney cells (Schaum et al., 2018). Intrinsic cluster quality measure defined by Davies-Bouldin index (DB-index) for (A) LDA and (B) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates the DB-index values (lower indicates better clustering). Extrinsic cluster quality measure defined by Adjusted Rand Index (ARI) for (C) LDA and (D) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates ARI (higher indicates better clustering). For (A-D) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of LDA clustering with (E) k = 6 and (G) k = 12. The UMAP plot of HDP clustering with (F) k = 11 and (H) k = 17. (I) The UMAP plot showing the reference clustering with the cell-type annotation from the original publication (Schaum et al., 2018).

In the extrinsic comparison, LDA clustering with k = 6 showed an average ARI value of 0.60, which was close to the highest average ARI value of 0.61 obtained with k = 5 (Fig 2C). The HDP clustering (k = 11) had an average ARI value of 0.67 (Fig 2D), suggesting that, in this dataset, the DB-index informed HDP with a higher cluster number (k = 11) may be useful in order to achieve a more detailed cell state or subtype specific resolution than the DB-index informed LDA with k = 6.

Visual inspection with the reference cell type annotations indicated that LDA clustering with k = 6 (best by DB-index) resolved the kidney limb epithelial cells, duct epithelial cells and partitioned the kidney tubule epithelial cells into two sub-clusters (Fig 2E, 2I). However, it did not separate a cluster of immune cells (macrophages) from kidney cells, whereas LDA with k = 12 did (Fig 2G, 2I). The HDP clustering with k = 11 gave similar results as LDA with k = 12 (Fig 2F), whereas HDP with k = 17 added several apparently sporadic clusters (Fig 2H). Considering the DB-index for the selection of an approximate cluster number, the HDP k value (k = 11) had the lowest DB-index value (Fig 2B) and highest ARI value (Fig 2D), suggesting the utility of HDP in this dataset. Additionally, these results suggest that the HDP-based k value may be useful to guide the selection of the k value for LDA, when two LDA k values have similar DB-index.

2.3 LDA and HDP clustering performance in mouse pancreatic cells

We repeated the comparison of LDA and HDP using mouse pancreatic cells (Schaum et al., 2018). In the intrinsic evaluation, the LDA clustering with k = 3 showed the lowest mean DB-index value of 2.3, and with increasing k, k = 7 displayed a local minimum (DB-index = 2.5) (Fig. 3A). Based on DB-index, HDP had worse performance compared to LDA, with k = 14 showing the lowest average DB-index value of 3.4 (Fig. 3B), and k = 17 showing the second lowest average DB-index value of 3.6.

Fig 3. Comparison of LDA and HDP clustering performance using mouse pancreatic cells (Schaum et al., 2018). Intrinsic cluster quality measure defined by Davies-Bouldin index (DB-index) for (A) LDA and (B) HDP clustering. The x-axis shows the number of clusters k, and the y-axis indicates the DB-index values (lower indicates better clustering). Extrinsic cluster quality measure defined by Adjusted Rand Index (ARI) for (C) LDA and (D) HDP clustering. The x-axis shows the number of clusters k, and the y-axis indicates ARI (higher indicates better clustering). For (A-D) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of LDA clustering with (E) k = 3 and (G) k = 7. The UMAP plot of HDP clustering with (F) k = 14 and (H) k = 17. (I) The UMAP plot showing the reference clustering with the cell-type annotation from the original publication (Schaum et al., 2018).

In the extrinsic cluster quality evaluation, increasing the LDA cluster number from k = 3 to k = 7 increased ARI, but larger numbers of clusters did not affect the quality markedly, producing average ARI values in the range 0.67-0.72 (Fig 3C). Similarly, the HDP clustering with k = 14 gave an ARI value of 0.67 (Fig 3D). The visual inspection with reference annotations suggested that LDA with k = 7, but not with k = 3, was able to resolve most of the cell subtypes present in the reference (Fig 3E, 3G, 3I), whereas HDP with k = 14 and k = 17 resulted in additional cell subsets (Fig 3F, 3H).

2.4 Comparison of existing LDA clustering tools for scRNA-seq data

While the main aim of our study was to compare LDA and HDP for clustering scRNA-seq data, we also compared the Gensim implementation of LDA with two existing LDA implementations for scRNA-seq data, Celda (Wang et al., 2021) and DIMM-SC (Sun et al., 2018). Since the computational times with DIMM-SC extended to several weeks with the full datasets, we used the top 2000 most highly variable genes for this comparison (Fig. S2 and S3). Especially with Gensim LDA and Celda, the best k values defined by the lowest DB-index values were generally in line with the highest average ARI values (Fig. S2 and S3). On the other hand, while Gensim resulted in better (lower) mean DB-index values compared to the other two methods, Celda displayed higher extrinsic ARI values in the two datasets. This was also reflected in the UMAP visualization, where Celda resulted in coherent clustering of the cells (Fig. S2 and S3). Overall, the Gensim LDA and DIMM-SC showed a wider range of variability in the cluster quality values than Celda for the repeated clustering runs (Fig. S2 and S3).

2.5 Comparison of LDA, HDP and the Seurat SNN clustering

The Bayesian Dirichlet process mixture models such as LDA and HDP are different from the clustering methods used in most of the existing state-of-art single-cell clustering tools, such as the widely used Seurat tool [20]. Seurat 4 clustering uses the graph-based shared nearest-neighbor (SNN) algorithm, where the resolution parameter (r) controls the resulting number of clusters. We compared LDA and HDP with Seurat 4 [20] using the top 2000 most highly variable genes (Fig. S4 - S7). For the immune cell dataset, the Seurat clustering resulted in the best intrinsic quality (lower DB-index) when the resolution r was below 0.1, resulting in k = 5 or k = 6 (Fig. S4). It also had the highest extrinsic cluster quality defined by ARI value of

0.62, while the highest average ARI values for LDA and HDP clustering were 0.54 (with k = 8) and 0.61 (with k = 7), respectively. For the kidney, pancreas and early pregnancy datasets (Fig. S5 - S7), Seurat, LDA and HDP clustering results had relatively similar average DB-index values for the different cluster numbers and resolution parameters, but Seurat resulted in slightly better ARI values compared to HDP and LDA.

3. Discussion and conclusion

We have evaluated the clustering performance of Dirichlet process mixture models LDA and HDP on three scRNA-seq datasets using both intrinsic (Hassani and Seidl, 2017) and extrinsic (Amigó et al., 2009) cluster quality measures defined by DB-index (Davies and Bouldin, 1979) and ARI (Hubert and Arabie, 1985), respectively. For each dataset, we also selected two best cluster numbers (k) based on intrinsic DB-index for more detailed visual evaluation. The intrinsic cluster quality provides general information about how compact the data points are within the individual clusters and how well the different clusters are separated. Because intrinsic quality measures do not assess the biological relevance of the clusters, we also considered extrinsic cluster quality and using UMAPs visually compared the identified clusters to the clusters from the original publications. Overall, our study showed that the relative performance of LDA and HDP was dataset dependent and highlighted the importance of carefully assessing the number of clusters when analyzing scRNA-seq data.

The variation in DB-index and ARI values between repeated runs of LDA and HDP indicated that the clustering results varied for different runs of the same dataset. Therefore, average values over multiple runs were used to produce robust results for the comparative analysis. Further, we generally observed less variation in HDP runs compared to LDA runs, suggesting that HDP could provide more robust DB-index and ARI values.

Our comparison of LDA and HDP indicated that their performance was dataset dependent. In the immune cell dataset (Fig 1), the DB-index informed HDP resulted in a more adequate clustering than the DB-index informed LDA when evaluated by both ARI and visual inspection with the original reference annotations. This provided evidence that at least in some cases HDP is a useful addition to the previously more widely employed LDA. For the other two datasets (Fig 2 and 3), HDP did not offer a clear advantage over LDA. In the kidney data, the DB-index informed HDP performed well judged by ARI, but in the visual

inspection it did not provide conceivable advantage over the DB-index informed LDA (Fig 2). In the pancreas data, HDP suggested higher numbers of clusters than LDA, while visual inspection suggested that these may inflate the clustering (Fig 3).

For the purpose of our comparisons, the cluster annotations from the original studies were considered to provide adequate level of resolution and quality to be used as a reference in the extrinsic analysis and in the visual inspection of the best intrinsic DB-index defined cluster numbers. A more in-depth biological interrogation of the detailed clustering differences is outside of the scope of our comparison. The overall biological interpretation of the resulting cluster annotations typically demands integration with other methods, such as protein level studies and spatial analysis (Dey et al., 2017).

Recently, several single-cell specific implementations of LDA clustering have become available (Dey et al., 2017; duVerle et al., 2016; Sun et al., 2018; Wang et al., 2021), while the implementations of HDP clustering for scRNA-seq are limited. We extended our main HDP to LDA comparison to also include two scRNA-seq specific LDA implementations, Celda (Wang et al., 2021) and DIMM-SC (Sun et al., 2018). We observed that, based on intrinsic DB-index analysis, Gensim LDA performed better than Celda and DIMM-SC, whereas extrinsic ARI analysis supported the coherence of the Celda results. Celda also showed less variability between repeated runs than Gensim LDA and DIMM-SC.

The runtime and memory usage of both LDA and HDP for datasets with smaller numbers of cells (~1000, ~2000, ~3000 cells) was practical for repeated analysis runs. However, for the large dataset (~65,000 cells), the increased running time affected the practicality of their use. Additionally, the inference method used in a given LDA or HDP implementation also affects its run time. The Gensim implementations of LDA and HDP use the variational inference method (Blei and Jordan, 2006), which is easier to scale to high-dimensional data than sampling-based inference methods such as MCMC (Blei et al., 2017). The LDA tools Celda and DIMM-SC implement the expectation maximation algorithm for model parameter estimation and, in the context of this study, they appeared computationally adequate, especially, when focusing on the top 2000 most highly variable genes. Currently, BISCUIT (Prabhakaran et al., 2016), the single-cell specific implementation for HDP clustering, uses Gibb's sampling as the inference method. Gibbs sampling typically runs extensive iterations before it converges to the target posterior distribution, making it computationally expensive.

Accordingly, a single run of BISCUIT using only the top 2000 most highly variable genes took more than three days, making the current implementation impractical for extensive comparisons. Therefore, further developments HDP specific to high dimensional scRNA-seq data could enhance the current computational challenge.

We also compared the performance of LDA and HDP with the graph-based SNN clustering implemented in the widely used Seurat 4 tool as a comparator method to inspect how the LDA and HDP clustering performed when evaluated with the existing state-of-art clustering method. HDP and LDA model-based clustering in general showed comparable results both in intrinsic and extrinsic evaluation measures when compared to Seurat based clustering. However, both LDA and HDP clustering resulted in markedly higher variation in the clustering results for the repeated runs compared to Seurat (Fig. S4-S7).

Ideally, cluster analysis results from scRNA-seq data give meaningful approximations of biological cell types or states. In this regard, the nonparametric HDP clustering method, unlike the LDA, automatically generates the number of clusters without a predefined number of clusters (Limsettho et al., 2014; Teh et al., 2006). Thus, HDP avoids the additional analysis of different *k* values to select the optimal number of clusters. In addition to the direct use of HDP clusters, HDP could also be used for exploratory cluster analysis to visualize and explore the unknown cellular states from scRNA-seq data and to help guide the choice a suitable number of clusters as a starting point for more refined analysis. We observed that LDA performed more robustly in the data that had closely related cell types or states, and in these cases HDP may inflate the cluster number. On the other hand, the tendency of HDP to result in larger numbers of clusters than LDA may also open up the possibility of finding novel cell types or states, which is of high importance for both basic research as well as in the inference of disease specific conditions.

The study was limited to compare the LDA and HDP model-based clustering methods in only small to medium-sized single-cell RNA-seq data due to the very long execution time (several days) that it takes to run HDP models for large datasets. Additionally, the LDA and HDP models have multiple prior concentration parameters used as an input that can affect the clustering result. However, coherent parameter tuning for multiple parameters at the same time would have required extensive computational resources and was beyond the scope of

this manuscript. Therefore, we limited our comparisons by fixing those concentration parameters to the default values.

In conclusion, our results support the previous reports that Dirichlet process based clustering models such as LDA and HDP are useful additions for single-cell data analysis in general (Bravo González-Blas et al., 2019; Kim et al., 2019; Sun et al., 2018) and that the non-parametric HDP model is a useful addition to the previously used LDA in particular.

4. Material and methods

Sequencing data

We analyzed four publicly available scRNA-seq datasets, including artificially mixed human immune cells, tissue specific subsets of kidney and pancreas cells from *Tabula Muris* (Schaum et al., 2018) and human decidua/placenta (early pregnancy) data (Vento-Tormo et al., 2018) with high quality cell type annotations. For the first dataset, we created an artificial mixture of human immune cells from seven publicly available scRNA-seq datasets from Gene Expression Omnibus (GEO): GSE75748, GSE81861, GSE44618, GSE96562, GSE85527, GSE96564 and GSE89232 (Table S1). The pre-processed datasets provided by the authors were downloaded from GEO together with their cell-type information, which was used as a reference in our clustering analysis. For the combined analysis, we converted raw counts and FPKM (Fragments Per Kilobase Million) to TPM normalized expression values (Transcripts Per Million), similarly as previously described (Pachter et al., 2011). The final artificial mixture contained expression profiles of 1153 human immune cells across 13880 genes, including CD4+ memory cells, CD8+ memory cells, B cells, dendritic cells, fibroblasts, and lymphoblasts.

The mouse kidney and pancreas datasets were from the publicly available *Tabula Muris* study (Schaum et al., 2018). The unique molecular identifier (UMI) count matrix provided by the authors was downloaded from GEO with accession GSE109774. We selected the kidney (SMART-seq based) and pancreas (droplet-based) cells, including a total of 2782 and 1961 cells, respectively, with 23433 genes for both datasets. The pre-processed UMI count data for human early pregnancy data (droplet-based) (Vento-Tormo et al., 2018) with 64,734 cells and 31,764 genes was downloaded from ArrayExpress with the accession number of E-MTAB-

6701. We used the within cell UMI count library size normalization with scaling a factor of 10^6 (Satija et al., 2015).

LDA and HDP implementations

We used the python implementations for LDA and HDP originally designed for topic modelling from the "Gensim" package. The benefit of using Gensim was that it has both LDA and HDP implemented in a single tool to ensure direct comparability. Additionally, the variational inference-based implementation of Gensim for LDA (Hoffman et al., 2010) and HDP (Wang et al., 2011) enabled scaling to high-dimensional datasets (Rehurek and Sojka, 2010). For the analysis, we used the normalized count data rounded to their nearest integer values in a "bag-of-words" representation (Zhang et al., 2010) and default parameters (alpha=1 and eta=.01 for LDA; alpha=1, gamma=1 and eta=.01 for HDP). With LDA, the number of clusters *k* was varied from 2 to 20, whereas HDP does not have a predefined number of clusters. The soft/mixed cluster assignments were transformed to hard cluster assignments by assigning each cell to the cluster with the highest cluster membership probability. In order to have biologically interpretable clustering results, clusters with less than 15 cells were grouped as a separate single cluster. For the visualizations of the clustering results, we used the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018).

In addition to the main LDA and HDP model comparison, we also compared the Gensim LDA implementation with two existing LDA implementations for scRNA-seq data, Celda (Wang et al., 2021) and DIMM-SC (Sun et al., 2018). Since the computational times with DIMM-SC extended to several weeks with the full datasets, for this additional analysis, we used only the top 2000 most highly variable genes. Again, default parameters were used, and the number of clusters *k* was varied from 2 to 20. Similarly, we attempted to compare the Gensim HDP implementation with the existing HDP implementation for single-cell RNA-seq data, BISCUIT (Prabhakaran et al., 2016). However, with BISCUIT, since the computational time for only a single cluster analysis run for e.g. pancreatic data with the top 2000 variable genes took more than three days, we excluded it from further analysis.

Finally, we compared the Gensim implementation of LDA and HDP with Seurat 4 SNN clustering (Hao et al., 2021). We used the 2000 most highly variable genes and the default parameters. For LDA clustering, we considered the number of clusters *k* ranging from 2 to 20 with 20 replicated runs for each *k*. The HDP clustering was also replicated 20 times with default resolution parameters. In the same way, we replicated the Seurat 4 SNN clustering 20 times with random seeding for multiple different resolution parameters ranging from 0.008 to 0.6.

Measures of cluster quality

The cluster quality was assessed using both intrinsic and extrinsic cluster quality measures. The intrinsic cluster quality measures involve compactness and separation as a criterion for cluster evaluation (Hassani and Seidl, 2017), whereas the extrinsic cluster quality measures evaluate the overall clustering in comparison with a reference clustering (Amigó et al., 2009).

Davies-Bouldin index (DB-index) (Davies and Bouldin, 1979) was used as an intrinsic cluster quality metric, which uses the intra-cluster variance and inter-cluster separation to evaluate cluster quality. For a clustering result which partition data points into *k* clusters, the DB-index is given by:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \frac{D_i + D_j}{d(c_i, c_j)}$$

where D_i is the average distance between all the data points in a given cluster i to their cluster center c_i and $d(c_i,c_j)$ is the distance between the i^{th} and j^{th} cluster centers. The smaller the DB-index, the better the compactness and separation of the clusters.

Calinski-Harabasz (CH-index) (Calinski and Harabasz, 1974) was also considered as another intrinsic cluster quality metric defined by the ratio of the overall between-cluster variance to overall within-cluster variance. The larger the CH-index, the higher the cluster quality.

Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) was used as an extrinsic cluster quality measure, which extends the *Rand index (RI)* (Rand, 1971) of the similarity between two clusters to adjust for chance. Here, ARI was used as a measure of cluster accuracy by comparing the observed clustering with the reference clustering. Given a clustering result $X = \{X_1, X_2, \ldots, X_k\}$ and the reference clustering $Y = \{Y_1, Y_2, \ldots Y_l\}$, the ARI is given by:

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where a_i is the number of data points in cluster X_i, b_j is the number of data points in cluster Y_j, n_{ij} is the number of overlapping data points in clusters X_i and Y_j, and n is the total number of data points. The higher the ARI value, the higher the agreement between the clustering results, with value of 1 being the maximum.

Acknowledgements

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

Competing interests

The authors declare no competing interests.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.: 675395 and Sigrid Juselius Foundation. Our research is also supported by University of Turku Graduate School (UTUGS), Biocenter Finland, and ELIXIR Finland. Work of KTR was also supported by Eemil Aaltonen Foundation, Juhani Aho Foundation and Waldemar von Frenckell Foundation.

Authors' contributions

NAA conceived the study, conducted the analysis and wrote the manuscript; KTR supervised the work and wrote the manuscript. LLE conceived the study, supervised the work and participated in writing of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

Amigó, E., Gonzalo, J., Artiles, J. and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr. Boston.* **12**, 461–486.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.*

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., Edu, J. B., Koltcov, S., Koltsova, O., Nikolenko, S., Blei, D. M., et al. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.*

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J. and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420.

Calinski, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Commun. Stat. - Simul. Comput.* **3**, 1–27.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502.

Davies, D. L. and Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227.

Dey, K. K., Hsiao, C. J. and Stephens, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.* **13**,.

duVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H. and Tsuda, K. (2016). CellTree: An R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics* **17**, 363.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29.

Hassani, M. and Seidl, T. (2017). Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam J. Comput. Sci.* **4**, 171–183.

Hoffman, M. D., Blei, D. M. and Bach, F. (2010). Online learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, .

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classif.* **2**, 193–218.

Kim, H. J., Yardımcı, G. G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J., et al. (2019). Capturing cell type-specific chromatin structural patterns by applying topic modeling to single-cell Hi-C data. *bioRxiv*.

Limsettho, N., Hata, H. and Matsumoto, K. I. (2014). Comparing hierarchical dirichlet process with latent dirichlet allocation in bug report multiclass classification. In *2014 IEEE/ACIS 15th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2014 - Proceedings*, p. Institute of Electrical and Electronics Engineers Inc.

McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861.

Pachter, L., Loir P., Pachter, L. and Loir P. (2011). Models for transcript quantification from RNA-Seq. *arXiv*.

Petegrosso, R., Li, Z. and Kuang, R. (2019). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinform.* **21**, 1209–1223.

Prabhakaran, S., Azizi, E., Carr, A. and Pe'er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *33rd International Conference on Machine Learning, ICML 2016*, pp. 1691–1715. International Machine Learning Society (IMLS).

Qi, R., Ma, A., Ma, Q. and Zou, Q. (2019). Clustering and classification methods for single-cell RNA-sequencing data. *Brief. Bioinform.* **21**, 1196–1208.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proc. Lr. 2010 Work. New Challenges NLP Fram.* 45–50.

Remesh, R. and Pattabiraman, V. (2017). A survey on the cures for the curse of dimensionality in big data. *Asian J. Pharm. Clin. Res.* **10**, 355–360.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502.

Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M. B., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372.

Singh, A., Yadav, A. and Rana, A. (2013). K-means with Three different Distance Metrics. *Int. J. Comput. Appl.* **67**, 13–17.

Sun, Z., Wang, T., Deng, K., Wang, X. F., Lafyatis, R., Ding, Y., Hu, M. and Chen, W. (2018). DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* **34**, 139–146.

Tang, X., Huang, Y., Lei, J., Luo, H. and Zhu, X. (2019). The single-cell sequencing: New developments and medical applications. *Cell Biosci.* **9**,.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* **101**, 1566–1581.

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J. E., Stephenson, E., Polański, K., Goncalves, A., et al. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*.

Wang, L. and Wang, X. (2013). Hierarchical Dirichlet process model for gene expression clustering Computational methods for biomarker discovery and systems biology research. *Eurasip J. Bioinforma. Syst. Biol.* **2013**,.

Wang, C., Paisley, J. and Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *Journal of Machine Learning Research*, pp. 752–760.

Wang, Z., Yang, S., Koga, Y., Corbett, S. E., Johnson, W. E., Yajima, M. and Campbell, J. D. (2021). Celda: A Bayesian model to perform co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data. *bioRxiv* 2020.11.16.373274.

Zhang, Y., Jin, R. and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. Cybern.* **1**, 43–52.

Figures

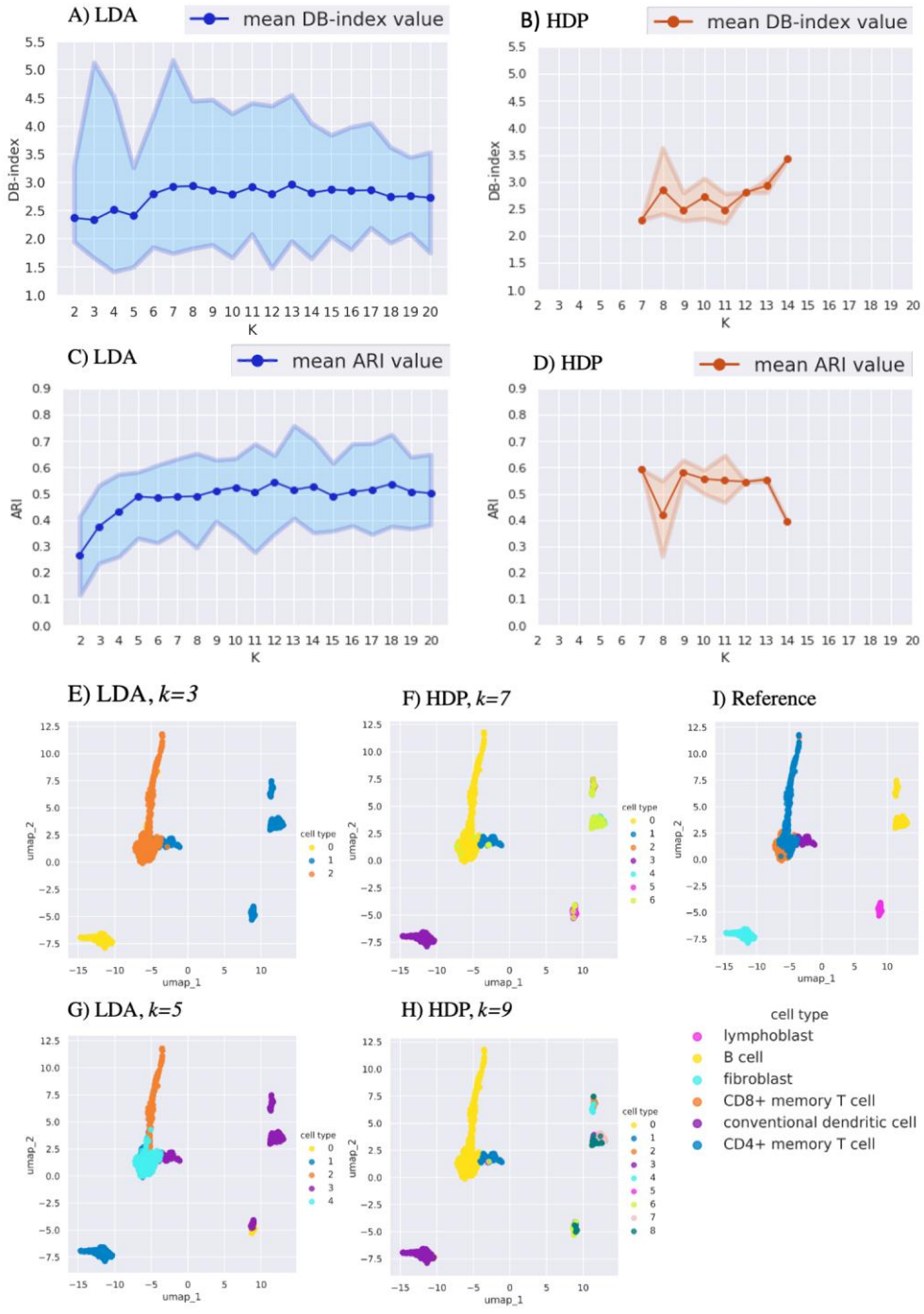

Figure 1. Comparison of LDA and HDP clustering performance using artificially mixed human immune cell scRNA-seq data. Intrinsic cluster quality measure defined by Davies-Bouldin index (DB-index) for (A) LDA and (B) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates the DB-index values (lower indicates better clustering). Extrinsic cluster quality measure defined by Adjusted Rand Index (ARI) for (C) LDA and (D) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates ARI (higher indicates better clustering). For (A-D) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of LDA clustering with (E) *k*=3 and (G) *k*=5. The UMAP plot of HDP clustering with (F) *k*=7 and (H) *k*=9. (I) The UMAP plot showing the reference clustering with the cell-type annotation from the original publications (Table S1).

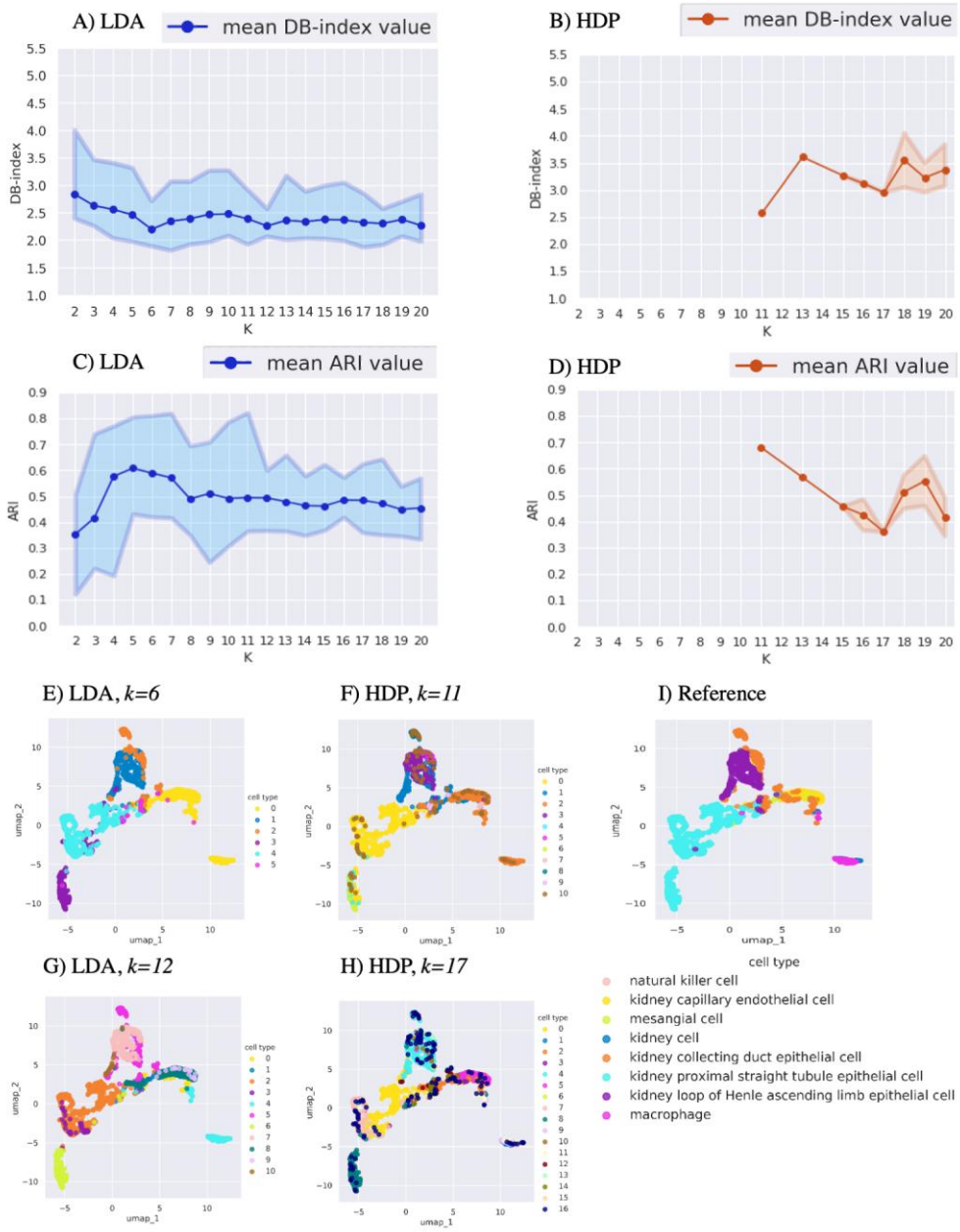

Figure 2. Comparison of LDA and HDP clustering performance using mouse kidney cells (Schaum et al., 2018). Intrinsic cluster quality measure defined by Davies-Bouldin index (DB-index) for (A) LDA and (B) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates the DB-index values (lower indicates better clustering). Extrinsic cluster quality measure defined by Adjusted Rand Index (ARI) for (C) LDA and (D) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates ARI (higher indicates better clustering). For (A-D) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of LDA clustering with (E) *k*=6 and (G) *k*=12. The UMAP plot of HDP clustering with (F) *k*=11 and (H) *k*=17. (I) The UMAP plot showing the reference clustering with the cell-type annotation from the original publication (Schaum et al., 2018).

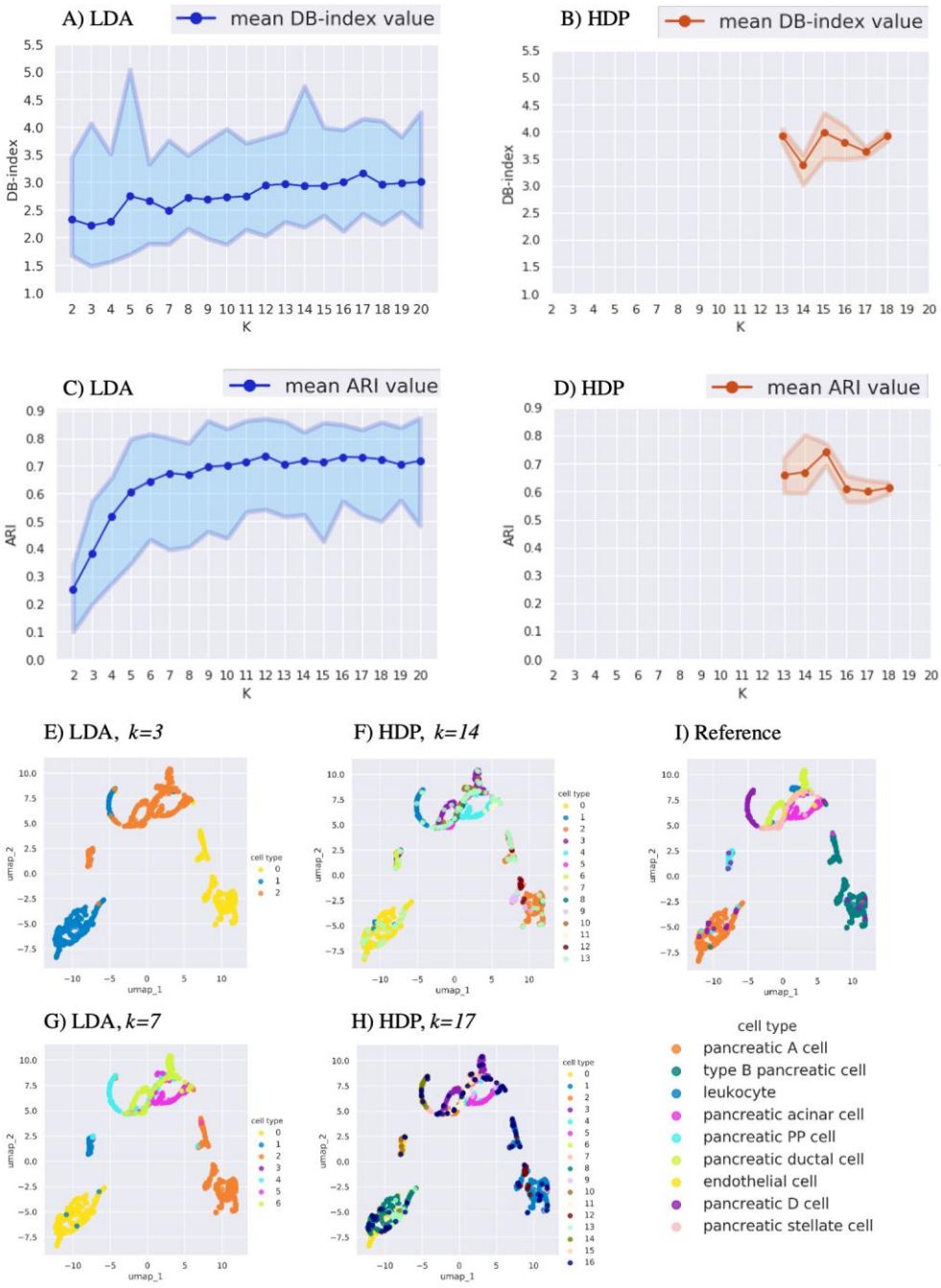

Figure 3. Comparison of LDA and HDP clustering performance using mouse pancreatic cells (Schaum et al., 2018). Intrinsic cluster quality measure defined by Davies-Bouldin index (DB-index) for (A) LDA and (B) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates the DB-index values (lower indicates better clustering). Extrinsic cluster quality measure defined by Adjusted Rand Index (ARI) for (C) LDA and (D) HDP clustering. The *x*-axis shows the number of clusters *k*, and the *y*-axis indicates ARI (higher indicates better clustering). For (A-D) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of LDA clustering with (E) *k*=3 and (G) *k*=7. The UMAP plot of HDP clustering with (F) *k*=14 and (H) *k*=17. (I) The UMAP plot showing the reference clustering with the cell-type annotation from the original publication (Schaum et al., 2018).

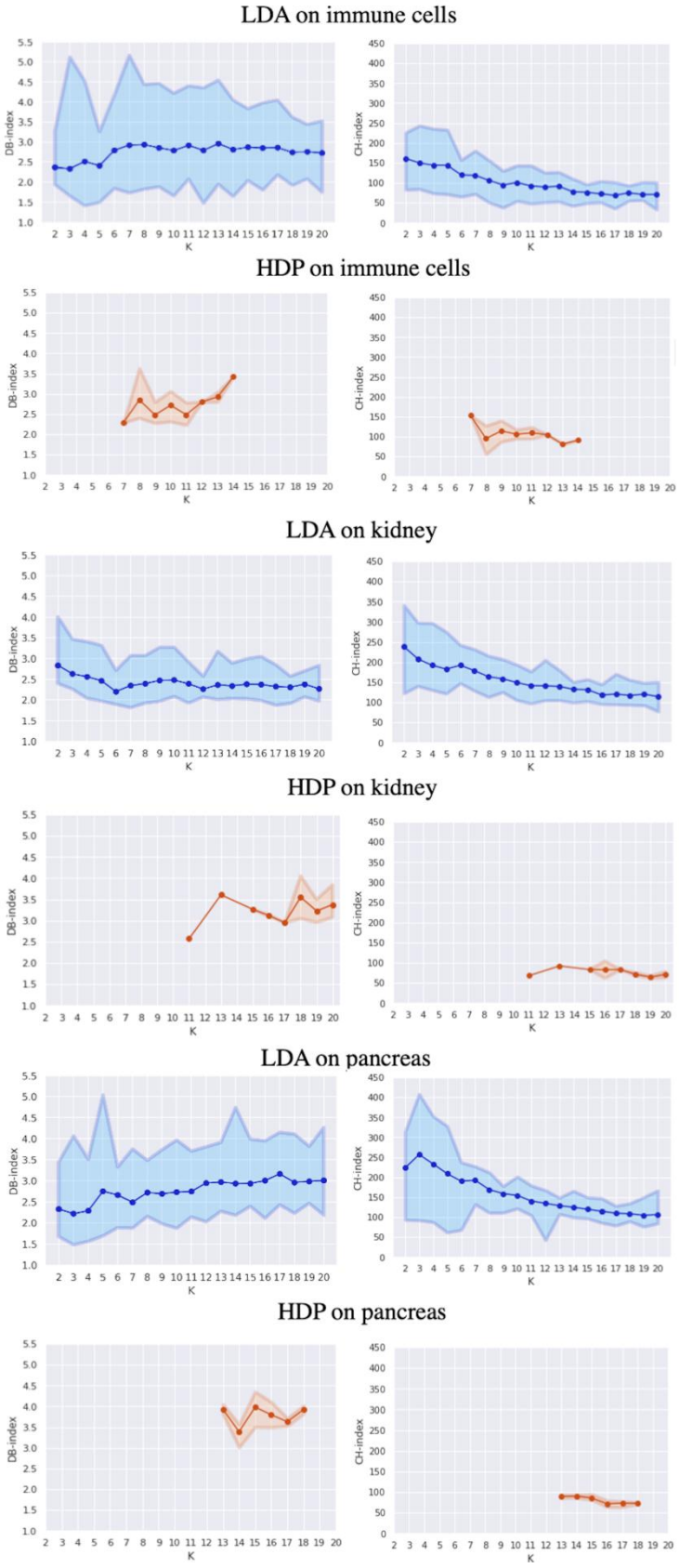

Fig. S1. Intrinsic cluster quality measures defined by DB-index and CH-index for LDA and HDP clustering results on the artificially mixed human immune cell data and the mouse kidney and pancreatic cell datasets.

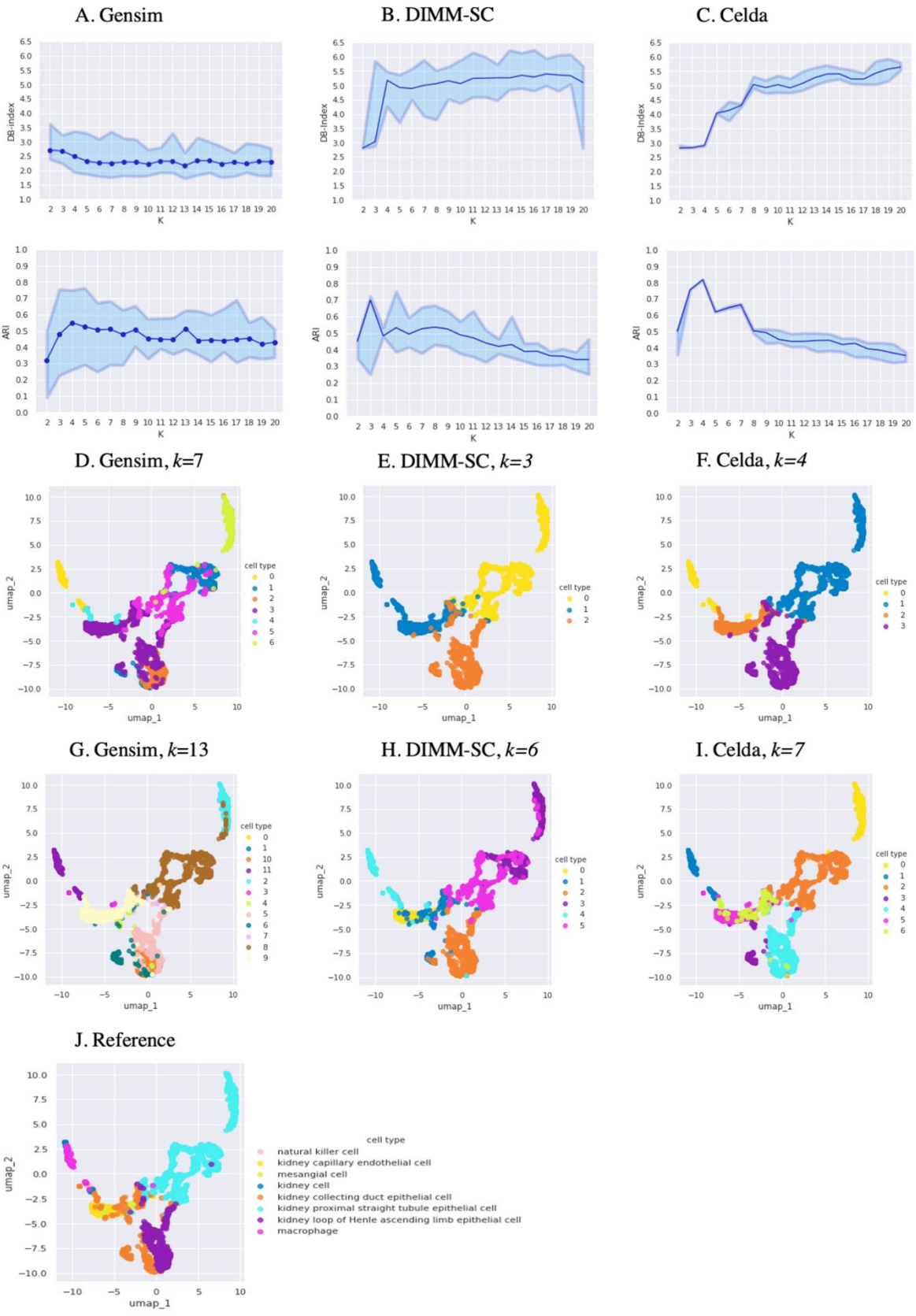

Fig. S2. Comparison of single-cell specific LDA clustering tools in the mouse kidney dataset: (A) Gensim LDA, (B) DIMM-SC, and (C) Celda. The intrinsic cluster quality measure was defined by Davies-Bouldin index (DB-index) and the extrinsic cluster quality measure by Adjusted Rand Index (ARI). The x-axis shows the number of clusters (*k = 2-20)*, and the y-axis indicates the DB-index values (lower indicates better clustering) and ARI values (higher indicates better clustering). For (A-C) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of Gensim LDA clustering with (D) *k* = 7 and (G) *k* = 13. The UMAP plot of DIMM-SC clustering with (E) *k* = 3 and (H) *k* = 6. The UMAP plot of Celda clustering with (F) *k* = 4 and (I) *k* = 7. (J) The UMAP plot showing the reference clustering with the cell-type annotation from the original publications. The top 2000 most highly variable genes were used as input for the runs.

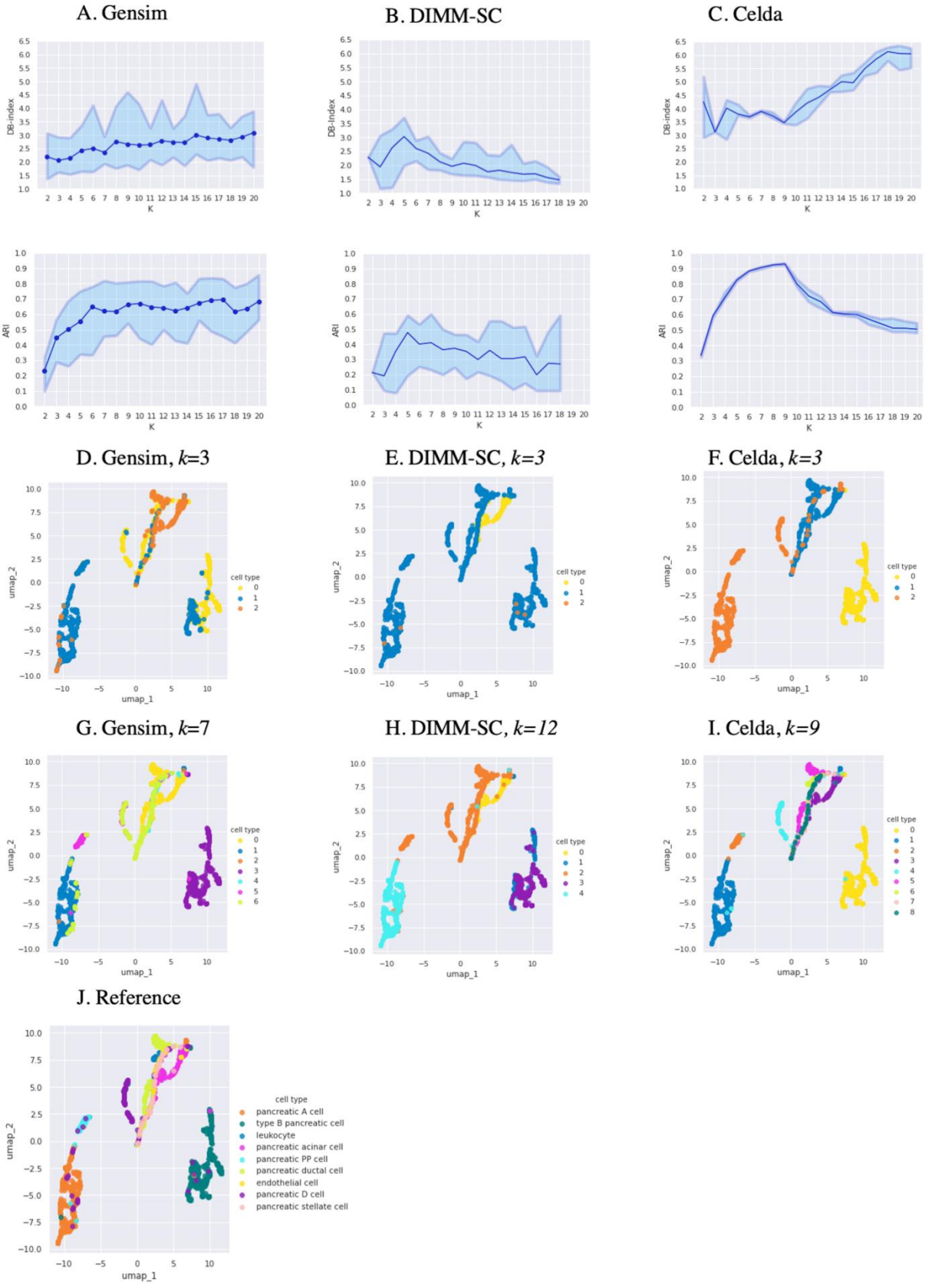

Fig. S3. Comparison of single-cell specific LDA clustering tools in the mouse pancreas dataset: (A) Gensim LDA, (B) DIMM-SC and (C) Celda. The intrinsic cluster quality measure was defined by Davies-Bouldin index (DB-index) and the extrinsic cluster quality measure by Adjusted Rand Index (ARI). The x-axis shows the number of clusters (*k=2-20)*, and the y-axis indicates the DB-index values (lower indicates better clustering) and ARI values (higher indicates better clustering). For (A-C) each run was repeated 20 times and the top, middle and bottom lines show the maximum, mean and minimum quality values, respectively. The UMAP plot of Gensim LDA clustering with (D) k = 3 and (G) k = 7. The UMAP plot of DIMM-SC clustering with (E) k = 3 and (H) k = 12 (clusters with less than 15 cells are collapsed). The UMAP plot of Celda clustering with (F) k = 3 and (I) k = 9. (J) The UMAP plot showing the reference clustering with the cell-type annotation from the original publications. The top 2000 most highly variable genes were used as input for the runs.

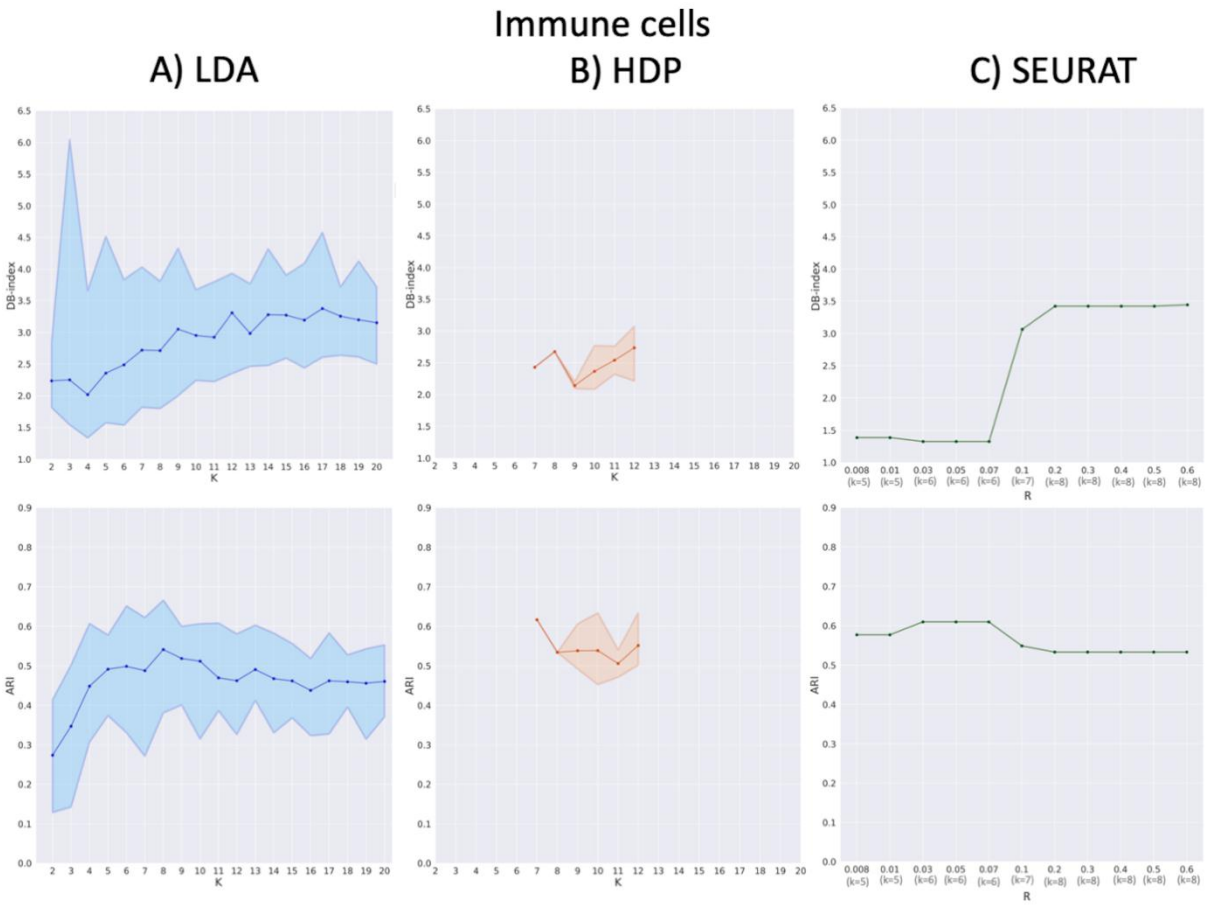

Fig. S4. Comparison of LDA, HDP and Seurat clustering based on intrinsic (Davies-Bouldin index, DB-index) and extrinsic (Adjusted Rand Index, ARI) cluster quality measures on immune cell data: (A) LDA, (B) HDP, and (C) Seurat SNN. The *x*-axis shows the number of clusters (*k=2-20)* for LDA and HDP and the resolution parameter *r* (from 0.008 to 0.6) for Seurat SNN. For Seurat, the average cluster numbers for the given resolution parameters are shown in brackets. The *y*-axis shows the maximum, mean and minimum values for DB-index (lower indicates better clustering) and ARI values (higher indicates better clustering) across 20 repeated runs.

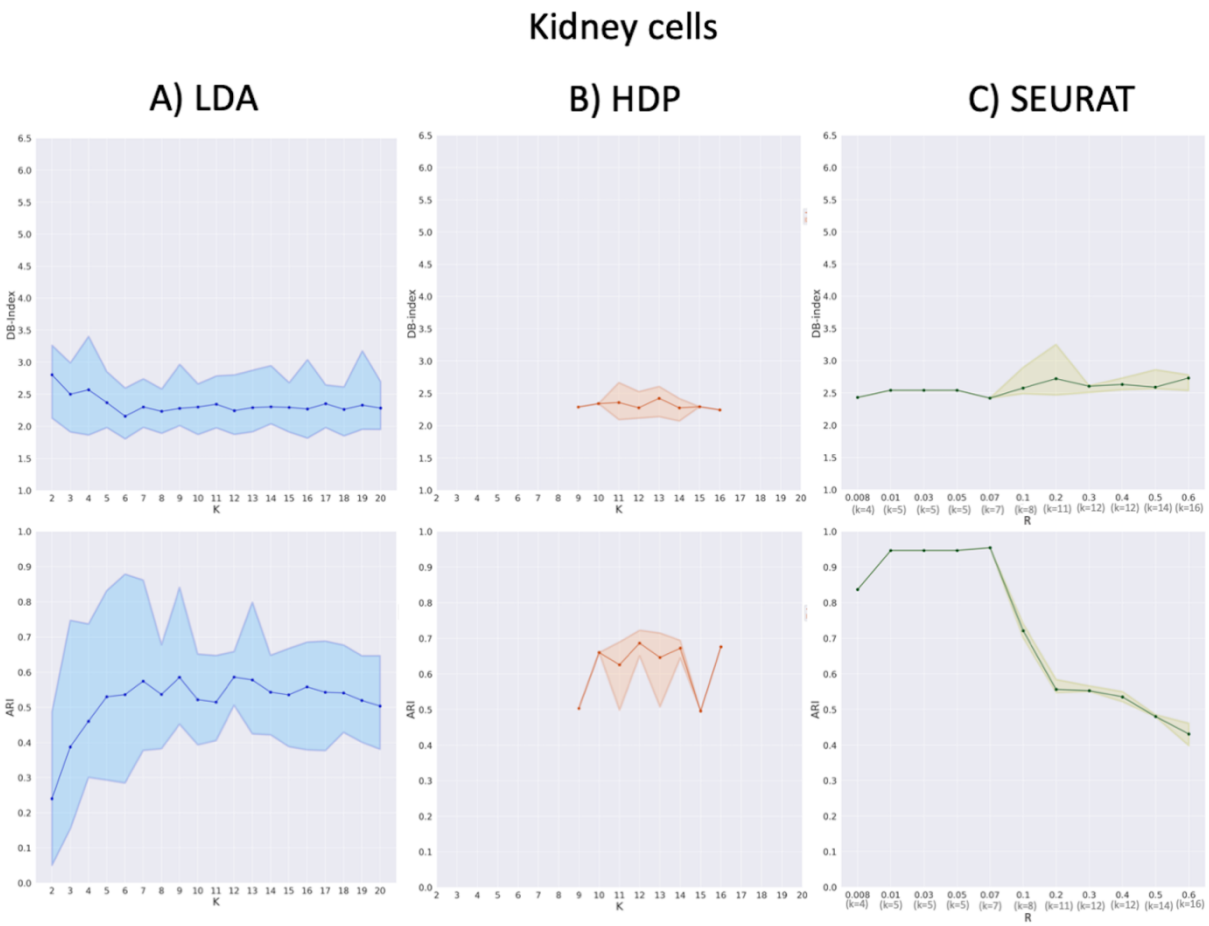

Fig. S5. Comparison of LDA, HDP and Seurat clustering based on intrinsic (Davies-Bouldin index, DB-index) and extrinsic (Adjusted Rand Index, ARI) cluster quality measures on mouse kidney cell data: (A) LDA, (B) HDP, and (C) Seurat SNN. The *x*-axis shows the number of clusters (k=2-20) for LDA and HDP and the resolution parameter *r* (from 0.008 to 0.6) for Seurat SNN. For Seurat, the average cluster numbers for the given resolution parameters are shown in brackets. The *y*-axis shows the maximum, mean and minimum values for DB-index (lower indicates better clustering) and ARI values (higher indicates better clustering) across 20 repeated runs.

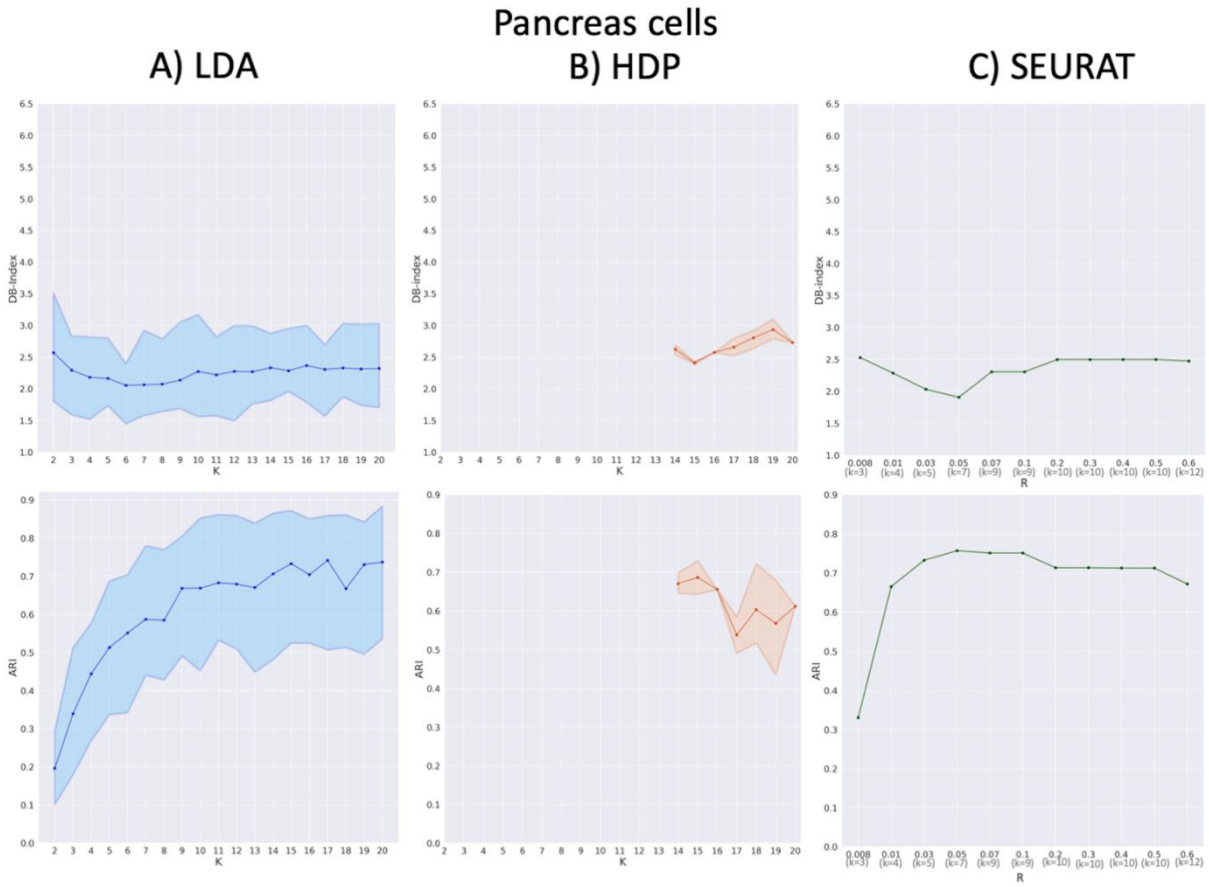

Fig. S6. Comparison of LDA, HDP and Seurat clustering based on intrinsic (Davies-Bouldin index, DB-index) and extrinsic (Adjusted Rand Index, ARI) cluster quality measures on mouse pancreas cell data: (A) LDA, (B) HDP, and (C) Seurat SNN. The *x*-axis shows the number of clusters (k=2-20) for LDA and HDP and the resolution parameter *r* (from 0.008 to 0.6) for Seurat SNN. For Seurat, the average cluster numbers for the given resolution parameters are shown in brackets. The *y*-axis shows the maximum, mean and minimum values for DB-index (lower indicates better clustering) and ARI values (higher indicates better clustering) across 20 repeated runs.

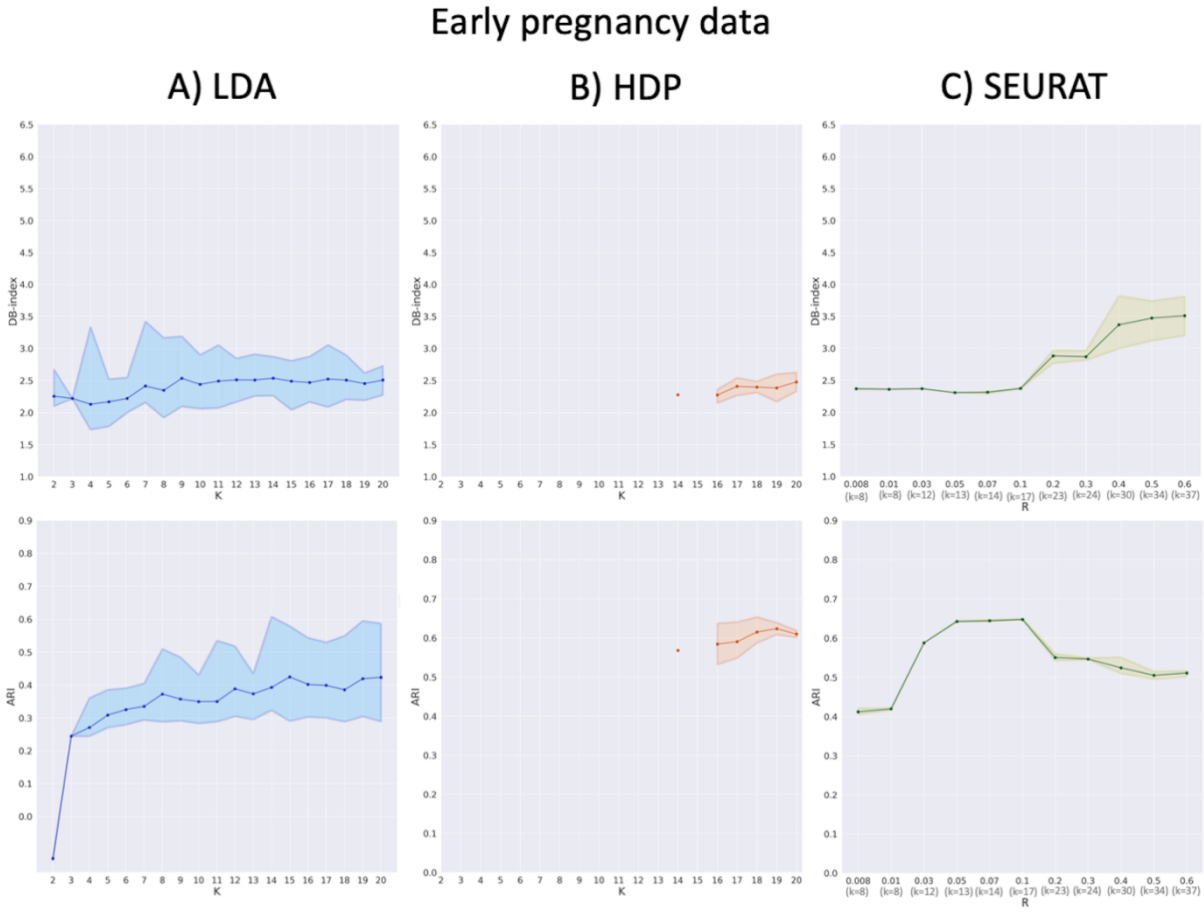

Fig. S7. Comparison of LDA, HDP and Seurat clustering based on intrinsic (Davies-Bouldin index, DB-index) and extrinsic (Adjusted Rand Index, ARI) cluster quality measures on human early pregnancy data: (A) LDA, (B) HDP, and (C) Seurat SNN. The *x*-axis shows the number of clusters (k=2-20) for LDA and HDP and the resolution parameter *r* (from 0.008 to 0.6) for Seurat SNN. For Seurat, the average cluster numbers for the given resolution parameters are shown in brackets. The *y*-axis shows the maximum, mean and minimum values for DB-index (lower indicates better clustering) and ARI values (higher indicates better clustering) across 20 repeated runs.

Table S1. Artificial mixture of human immune cells.

Selected cells by Cell-type (n of cells)	GEO accession (n of cells)	Library preparation	Sequencing platform	Downloaded data format
Fibroblast (159)	GSE75748 (1810)	Fluidigm C1	Illumina HiSeq 2500	TPM
Lymphoblast (59)	GSE81861 (1220)	Fluidigm C1	Illumina HiSeq 2000	RPKM
B-cell (174)	GSE44618 (62)	SMART-seq 1	Illumina HiSeq 2000	RPKM
	GSE81861(1220)	Fluidigm C1	Illumina HiSeq 2000	RPKM
CD4+ memory T cell (393)	GSE96562 (149)	SMART-Seq 1	Illumina HiScanSQ	Raw count data
	GSE96568 (246)	SMART-Seq 1	Illumina HiSeq 2500	Raw count data
CD8+ memory T cell (263)	GSE85527 (219)	Nextera XT DNA Library Preparation Kit (Illumina)	Illumina HiSeq 2500	Raw count data
	GSE96564 (45)	SMART-Seq 1	Illumina HiSeq 2500	Raw count data
Conventional dendritic cell (105)	GSE89232 (957)	SMART-Seq 2	Illumina HiSeq 2500	TPM

Table S2. Running time and memory usage for Gensim LDA and HDP clustering.

	Artificially mixed immune dataset		**Pancreas, Tabula muris**		**Kidney, Tabula muris**		**Decidua/placenta**	
	# genes	# cells	# genes	# cells	# genes	# cells	# genes	# cells
	13,000	1,153	23,000	1,961	23,000	2,782	23,000	64,734
LDA	1.7 min/ 2.6 GB		2.8 min/ 4.2 GB		2.3 min/ 6.0 GB		1.35 hrs/208 GB	
HDP	5.7 min/ 2.7 GB		15.2 min/ 4.3 GB		28.1 min/ 6.1 GB		4 days/ 208 GB	