Accurate sub-population detection and mapping across single cell experiments with PopCorn

Yijie Wang[1], Jan Hoinka[1], and Teresa M Przytycka[1]*

National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda MD 20894, USA

* correspondence to przytyck@ncbi.nlm.nih.gov

Abstract. The identification of sub-populations of cells present in a sample, and the comparison of such sub-populations across samples are among the most frequently performed analyzes of single cell data. Current tools for these kind of data however fall short in their ability to adequately perform these tasks. We introduce a novel method, PopCorn (single cell sub-Populations Comparison), allowing for the identification of sub-populations of cells present within individual experiments while simultaneously performing sub-populations mapping across these experiments. PopCorn utilizes several novel algorithmic solutions enabling the execution of these tasks with unprecedented precision. As such, PopCorn provides a much needed tool for comparative analysis of populations of single-cells.

1 Introduction

Recent technological advances have facilitated unprecedented opportunities for studying biological systems at single-cell level resolution. For example, single-cell RNA sequencing (scRNA-seq) enables the measurement of transcriptomic information of thousands of individual cells in one experiment. Analyses of such data provide information that was not accessible using bulk sequencing which can only assess average properties of cell populations. Single cell measurements however can capture the heterogeneity of a population of cells. In particular, single cell studies allow for the identification of novel cell types, states, and dynamics [1, 2, 3, 4]. The benefits of single cell data however come at the cost of unique computational challenges [5]. These challenges emerge from the stochasticity of single cell experimental data as well as from the multitude of questions that are being addressed with this technology.

One of the most prominent uses of the scRNA-seq technology is the identification of sub-populations of cells present in a sample and comparing such sub-populations across samples [6, 7, 8, 9, 10, 11, 12, 13]. Such information is crucial for understanding the heterogeneity of cells in a sample and for comparative analyses of samples from different conditions, tissues, and species. While some information about sub-population structure can be gained from data visualization methods such as the dimensionality reduction technique tSNE [14], relaying on visualization approaches alone can be highly misleading. To address this challenge, Butler at al. developed a computational approach enabling the identification of single cell populations across data sets [15]. This method is based on the Correlated Components Analysis (CCA) followed by an alignment of the CCA basis vectors between the data sets. The method has been implemented as a part of the popular software package Seurat [15]. However, while Seurat's approach provides the first step towards addressing this important problem, it has several significant limitations. Most notably, it operates under the assumption that the input samples consist of same sub-populations of cells (albeit possibly in different proportion) and focuses on identifying the correspondence between these sub-populations. This assumption is reasonable when aiming at aligning two replicas of the same experiment or identifying conserved sub-populations across highly similar experimental data. Single cell data however is increasingly used in comparative analysis of more diverse cell populations containing unique sub-populations. To address this critical gap we developed a new approach, single cell sub-Populations Comparison (PopCorn) that allows for comparative analysis of two or more single cell populations.

There are two key ideas behind PopCorn that are fundamental for the accuracy of our approach. The first idea is to identify sub-populations of cells present within individual experiments simultaneously with performing sub-populations mapping across these experiments rather than identifying the sub-populations

first and mapping them later. This allows for integrating information across experiments thus reducing noise. The second key innovation consists of a new approach to identify sub-populations of cells within a given experiment. Unlike simple clustering approaches, PopCorn utilizes Personalized PageRank vectors [16] and a quality measure of cohesiveness of a cell population (introduced in Supplementary Materials A) to construct an auxiliary sub-population co-membership propensity graph to guide the process of identifying such sub-populations.

We tested the performance of PopCorn in two distinct settings. First, we demonstrated its potential in identifying and aligning sub-populations from single cell data from human and mouse pancreatic singe cell data [15]. Next, we applied PopCorn to the task of aligning biological replicates of mouse kidney single cell data [17]. PopCorn achieved a striking improvement over the existing tool.

Consequently, and as a result of our integrative approach, PopCorn provides novel and unmatched tool for comparative analysis of single-cells populations.

2 Method

Informally, a sub-population of cells should include cells that have a common expression pattern (consistency) which are distinct from the expression patterns of other cells (separation). However, applying this principle in the context of single cell experiments is non-trivial. Given the stochastic nature of single cell experiments, some sub-populations can be well separated in one experiment whereas the separation can be less pronounced in another - either due to technical issues or due to biological differences between the samples. In addition, the stochasticity of the experiment introduces noise to the readout of the expression level of individual genes in individual cells, which might impact the accuracy of the assessment of the similarities between cells.

The main idea of the PopCorn is based on the simultaneous identification of sub-populations of cells present within individual experiments with performing sub-populations mapping across experiments. To this purpose, PopCorn integrates two objective functions aimed at (i) ensuring a meaningful partition of cells into sub-populations within each individual experiment and (ii) ensuring the consistency of these partitions across the experiments. To jointly optimize these two objectives, we construct two weighed graphs. The first graph, also known as sub-population co-membership propensity graph , encodes the propensity of cells belonging to the same sub-population for any two cells form the same experiment (A in Fig. 1) (we use same label to denote a graph and its matrix representation). The criterion used for constructing this graph is key to the efficiency of our method and is described in the next subsection and Supplementary Materials A. The second graph, a multipartite graph, encodes pairwise similarities of cells from different experiments (B in

Fig. 1). Following the construction of these two graphs, the final step of the method consists in solving a k-partition problem that simultaneously takes into account constraints encoded by both graphs.

In the subsequent sections, we outline the main ideas behind our approach while more technical details are deferred to the supplementary materials. We additionally note that the subsequent sections assume the existence of q scRNA-seq data sets denoted by $D_1, D_2, ..., D_q$ where a data set D_i covers N_i genes over M_i cells.

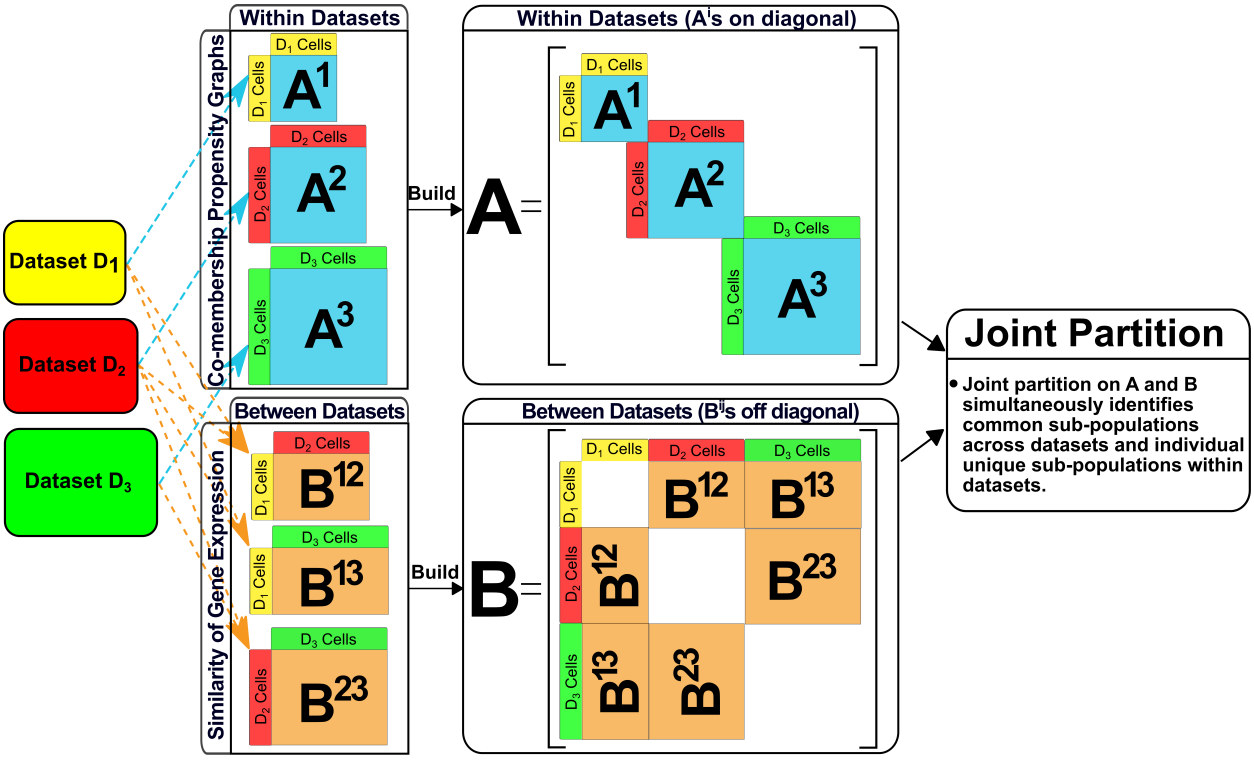


Fig. 1: The work flow of PopCorn. For each experiment i we first construct, matrix A^i representing the propensity of each pairs of cells to be in the same sub-population (this is achieved by using personalized PageRank method, see Section 2.1 and Supplementary Materials A). Matrix A summarizes all matrices A^i and is constructed by placing these on the diagonal of A. Next, we construct matrices $B^{i,j}$ which encode similarities between pairs of cells form different experiments i and j (see Section 2.2). B is constructed by placing the similarity matrices for all pairs of experiments off-diagonal as illustrated in the figure. We then perform joint partition of the graphs represented by matrices A and B by applying semi-definite programming to solve the problem.

2.1 Construction of the sub-population co-membership propensity graph

The objective of sub-population identification is to partition cells into groups while optimizing for consistency within each group and separation between the groups.

To address these challenge, we compute *sub-population co-membership propensity graph,* A^i, for every experiment D_i. A^i consists of a weighted graph with nodes corresponding to the cells from experiment D_i and edge weight represents the propensity of a given pair of cells to be in one cluster (one sub-population). To estimate such propensity, each cell "votes" which other cells should be put in the same sub-population with

itself. The voting process utilizes a personalized PageRank vector on the expression similarity graph (see (5) and Supplementary Materials A for a precise definition) for the cells in experiment D_i. For any cell, the ranking of the cells in its personalized PageRank vector utilizes a measure of expression consistency and a separation to ensure the desired properties for the sub-populations proposed (See Supplementary Materials A for more details). An edge (l,m) is included in the graph sub-population co-membership propensity graph A^i if this cells l and m obtained at least one vote to be in the same sub-population. The weight of each edge equals to the number of votes received by the given pair. Note that partitioning A^i into k sub-graphs provides a method to uncover the population structure in experiment D_i which is of an independent interest. However if performed in isolation, such sub-population assignment would not benefit from the information contained in the data from other experiments. Thus graphs A^i are used jointly with the information represented by graph B that encodes pairwise similarities between cells from different experiments as described below.

2.2 Solving joint sub-population identification and mapping problem

Solving the joint sub-population identification and mapping problem translates into grouping the cells into a set of clusters such that the resulting partition is optimized for grouping cells of similar expression pattern within data sets while ensuring cells of the same kind are also aligned across data sets.

To this end, we encode complementary information regarding the sub-population co-membership propensity and the pairwise cell-similarities from different experiments in two graphs A and B respectively.

A corresponds to the union of all graphs A^i as defined in the previous section. Let $N = \sum_i N_i$ be the overall set of cells, arranged such that cells from the same data set have consecutive indices. $A \in \mathbb{R}^{N\times N}$ is then constructed by assembling the adjacency matrices $A^i, i = 1,2,...,q$ into a block-diagonal matrix, i.e. $A^i \in \mathbb{R}^{N_i\times N_i}$ are arranged along the diagonal as shown in Fig. 1.

B in turn consists of a q-partite graph recording the similarity between cells across different experiments based on gene expression. Formally

$$B = \begin{bmatrix} O^1 & B^{12} & ... & B^{1q} \\ (B^{12})^T & O^2 & ... & B^{2q} \\ ... & ... & ... & ... \\ (B^{1q})^T & (B^{2q})^T & ... & O^q, \end{bmatrix} \tag{1}$$

where $Q^i \in \mathbb{R}^{N_i\times N_i}$ is an all zero matrix and $B^{ij} \in \mathbb{R}^{N_i\times N_j}$ measures expression consistency between cell from different experiments (see definition in (6).) Note that A, $B \in \mathbb{R}^{N\times N}$ have the same dimension.

Given these two adjacency matrices, we then compute a partition of the cells into k clusters that respects the connectivity defined by both graphs. To accomplish this, we first normalize both matrices as described in Supplementary Materials B.1 and define the normalized Laplacian matrix of A and B as L_A and L_B respectively. To encode the assignment of cells to a sub-population we define an assignment matrix $Y_{N\times k}$, where N is the total number of cells, as follows:

$$Y[j,k] = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ cell belongs to sub-population } k \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that given the normalized Laplacian matrix L_X, where X is either A or B, the problem of finding the optimal partition into k sub-populations that respects connectivity defined by matrix X is equivalent to the normalized k-cut problem [18, 19] and can be expressed as

$$\begin{aligned} \min: \quad & \text{tr}(Y^T L_X Y) \\ s.t. \quad & Y \in \mathfrak{F}_k, \end{aligned} \tag{3}$$

where $\mathfrak{F}_k = \left\{ Y : Y 1_k = 1_N, Y^T 1_N \geq c, Y_{ij} \in \{0,1\} \right\}$.

In particular, if $X = A$, this leads to clustering the scRNA-seq data sets into k different sub-populations, based solely on experiment specific features (similarity expression pattern between different cells in the same experiment) without using any information from other experiments. In contrast if $X = B$ we will find the best alignment between the cells across experiments ignoring sub-population structure withing each experiment.

Accordingly, a k-partition formulation respecting both matrices can be defined as

$$\begin{aligned} \min: \quad & \text{tr}(Y^T (L_A + \lambda L_B) Y) \\ s.t. \quad & Y \in \mathfrak{F}_k. \end{aligned} \tag{4}$$

were the parameter λ defines a scalar weight relation between the two sets of edges. To find the optimal solution we use a semi-definite programming (SDP) relaxation approach as described in the Supplementary Materials B.

2.3 Expression similarity between cells

Expression similarity is defined differently for cells form the same experiment compared to cells from separate experiments, although both share a first common step: the identification of highly variable genes

(HVGs) [20] Ω^i in each data set D^i and consequent utilization the normalized expression data $\Sigma_i \in \mathbb{R}^{|\Omega^i| \times N_i}$ of those HVGs to compute expression consistency.

For cells l and m belonging to the ith data set D^i, their expression consistency W^i_{lm} is computed using the cosine similarity as follows:

$$W^i_{lm} = \begin{cases} \cos(\theta_{lm}) & \cos(\theta_{lm}) = \dfrac{E_i[:,l]^T E_i[:,m]}{\|E_i[:,l]\| \, \|E_i[:,m]\|} \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

Here, $E_i \in \mathbb{R}^{R_i \times N_i}$ is derived by applying Principle Component Analysis (PCA) to Σ_i where R_i corresponds to the number of principle components.

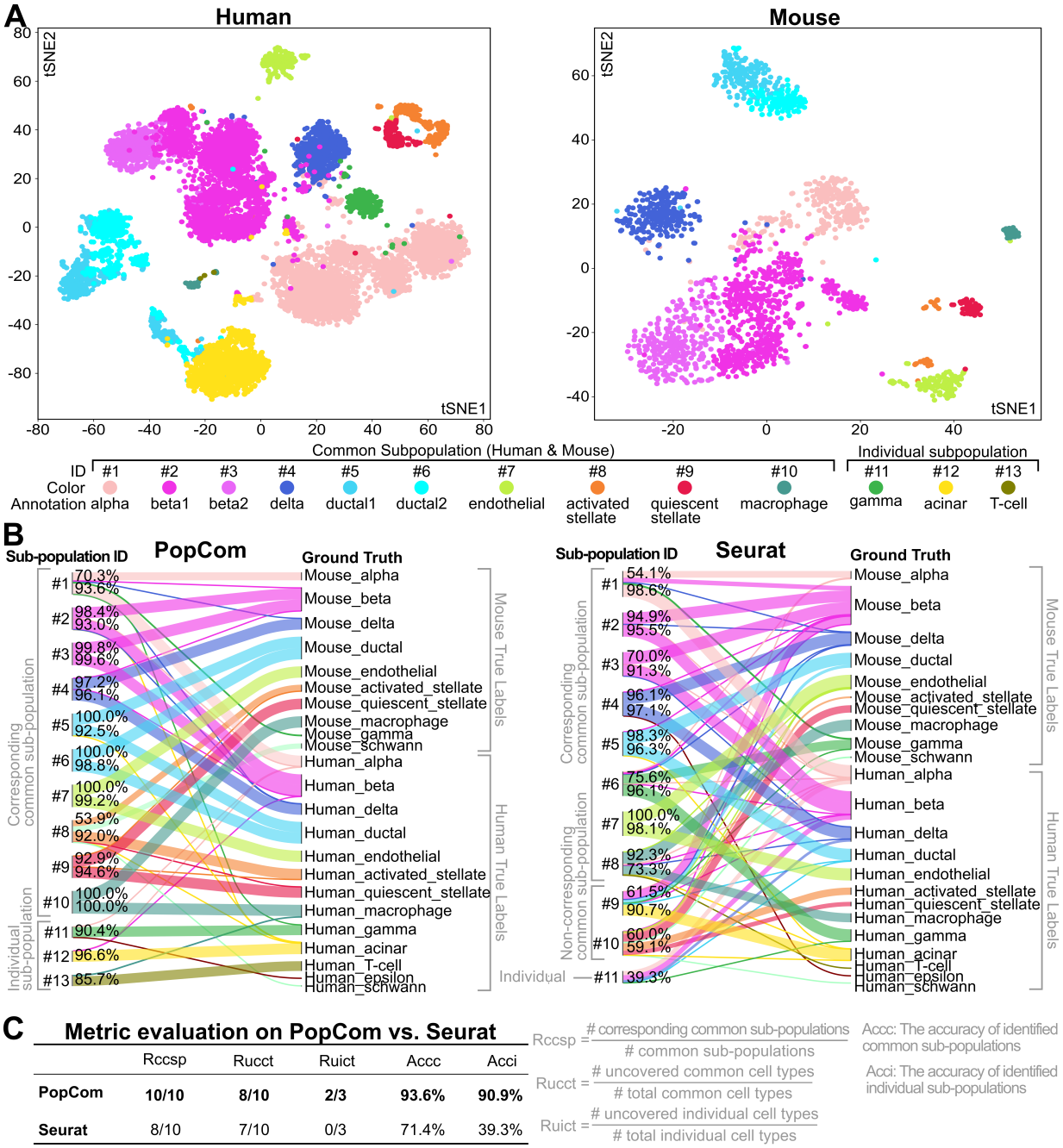
To compute expression consistency between cells across different experiments, the posibility of different scRNA-seq data sets having different sets of highly variable genes (HVGs) needs to be taken into account. To account for this variability, we compute the similarity between cells l and m across data sets using co-expressed HVGs only:

$$B^{ij}_{lm} = \begin{cases} \cos(\theta_{lm}) & \cos(\theta_{lm}) = \dfrac{\sum_\omega \Sigma_i[\omega,l]\Sigma_j[\omega,m]}{\sqrt{\sum_\omega \Sigma_i[\omega,l]^2}\sqrt{\sum_\omega \Sigma_j[\omega,m]^2}} \geq 0, \ \omega \in \Omega_i \cap \Omega_j \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

3 Results

3.1 Comparative analysis of single-cell RNA-seq experiments across species

To demonstrate the capabilities of our approach, we first applied PopCorn on two scRNA-seq data sets from different species and compare its performance to that of Seurat alignment method, a recent developed method that aims to integrates multiple single cell data sets [15]. In particular, we obtained both, human and mouse pancreatic cell transcriptomes from GEO Series accession number GSE84133. The human scRNA-seq data set contains 8,629 cells from 13 cell types whereas the mouse scRNA-seq data set includes 1,886 cells from 11 cell types, 10 of which are shared by both data sets. In addition, the human scRNA-seq data set has 3 individual cell types that do no appear in mouse scRNA-seq data set. We performed a comparative analysis on both data sets to identify *individual sub-populations* which only contain cells from a single data set, and *common sub-populations* which include cells from more than one data set. We further utilize the labels provided in [15] as our gold standard to validate the performance of the comparative analysis. The results generated by our PopCorn approach are then benchmarked against the results of the Seurat alignment



C

Metric evaluation on PopCom vs. Seurat

	Rccsp	Rucct	Ruict	Accc	Acci
PopCom	**10/10**	**8/10**	**2/3**	**93.6%**	**90.9%**
Seurat	8/10	7/10	0/3	71.4%	39.3%

$$\text{Rccsp} = \frac{\#\text{ corresponding common sub-populations}}{\#\text{ common sub-populations}}$$

$$\text{Rucct} = \frac{\#\text{ uncovered common cell types}}{\#\text{ total common cell types}}$$

$$\text{Ruict} = \frac{\#\text{ uncovered individual cell types}}{\#\text{ total individual cell types}}$$

Accc: The accuracy of identified common sub-populations

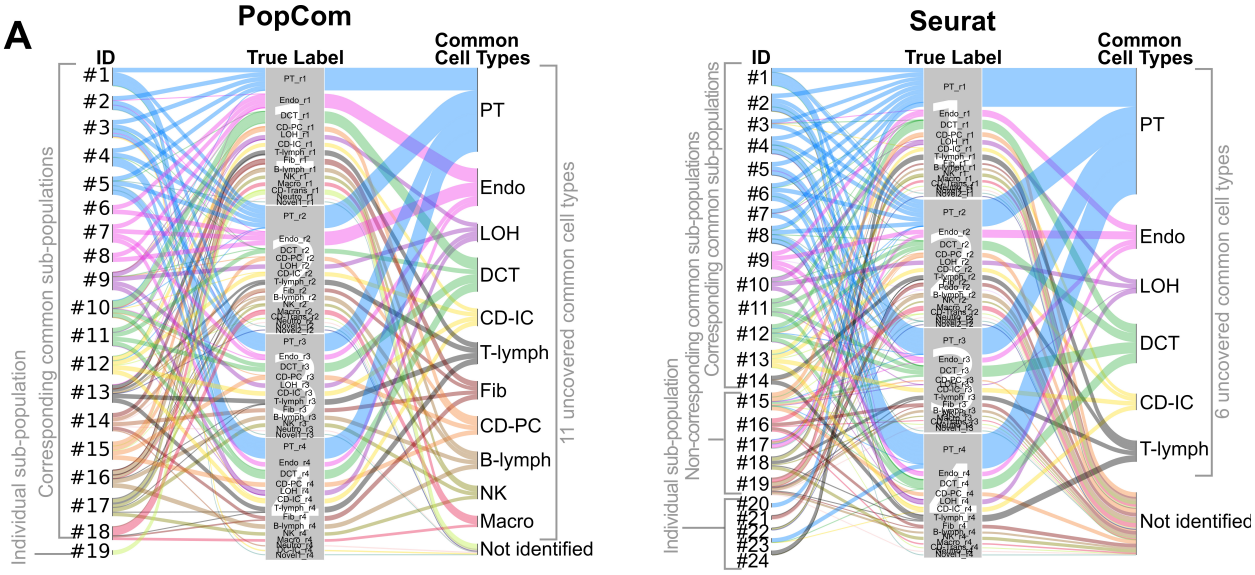
Acci: The accuracy of identified individual sub-populations

Fig. 2: (A) Two t-SNE plots for human and mouse scRNA-seq data sets, respectively. Different colors indicate different cell annotations, which can be determined via true labels (Supplementary Materials C). Cells of the same color denote a sub-population identified by PopCorn. (B) Sankey diagrams of the resulting mapping between identified sub-populations to ground truth labels of both PopCorn and Seurat. The width of the flow bar is proportional to the Acc_{ssp} score (the accuracy of a *split sub-population*) as defined in Supplementary Materials C.2. The Acc_{ssp} scores for the majority of the cell types in each identified sub-population are given at the beginning of the flow bar. As shown, all *common sub-populations* identified by PopCorn are *corresponding common sub-populations*. In addition, PopCorn identifies three *individual sub-populations* that are annotated to acinar, gamma, and T-cell, and acinar and T-cell are known to be unique cell populations in the human experiments. In contrast, sub-population #9 of Seurat incorrectly assigned acinar cells in human and beta cells in mouse into a single population; sub-population #10 of Seurat designated both activated_stellate and quiescent_stellate cells in human and mouse into one population but failed to separate them. In addition, Seurat failed to identify any individual cell populations. (C) Comparison of PopCorn and Seurat on various metric scores that are defined in Supplementary Materials C.

method. For a detailed description of the parameter selection regarding both methods, we refer the reader to Supplementary Materials D.1.

In order to ensure an impartial comparison, we define several metric scores which evaluate the results of the competing methods from distinct, yet comlementary contexts. The first metric, R_{ccsp}, aims at measuring how accurate a method can group cells of the same cell type across data sets together and is defined as the ratio of the number of identified *corresponding common sub-populations* to the total number of identified *common sub-populations*. Next, we evaluate how many common cell types across data sets and individual cell types can be recovered by a method via R_{ucct} (the ratio of uncovered common cell types to the total number of common cell types) and R_{uict} (the ratio of uncovered individual cell types to the total number of individual cell types). Last but not least, we utilize *Accc* and *Acci* to quantify the purity of the identified *common sub-populations* and *individual sub-populations*. Purity in this context refers to the percentage of the majority cell population in an identified sub-population. The detailed definitions of all metric scores can be found in Supplementary Materials C.

Fig. 2 A illustrates two t-distributed stochastic neighbor embedding (tSNE) plots that summarize the results generated by PopCorn for the two benchmark data sets. Although the layouts of the cell sub-populations in human and mouse are distinctive in appearance, the figure illustrates PopCorn's ability to identify *common sub-population* (sub-population #1 to #10) and *individual sub-populations* (sub-population #11 to #13). Fig. 2 B visualizes the content of each identified sub-population discovered by PopCorn and the Seurat alignment method. The right figure indicates that for sub-population # 9 and sub-population # 10, Seurat assigned cells of different types in human and mouse into one group. Specifically, sub-population # 9 classified human acinar cells and mouse beta cells as one common sub-population. Human acinar cells however are a unique sub-population that does not have a correspondence population in mouse experiments. In contrast, our PopCorn method successfully identified the human acinar cells as a unique sub-population (sub-population # 12 of our in the left figure of Fig. 2). In addition, we observed that sub-population # 10 of Seurat contains both mouse activated_stellate and quiescent_stellate cells and human activated_stellate and quiescent_stellate cells, but fails to designate them into separate sub-populations. PopCorn however correctly assigned mouse activated_stellate and human activated_stellate cells into one sub-population (sub-population # 9 of our PopCorn), and mouse quiescent_stellate and human quiescent_stellate cells into another sub-population (sub-population # 10 of our PopCorn). Last but not least, Seurat fails to identify any known *individual sub-populations*. However, PopCorn identifies 2 known *individual sub-populations* (acinar and T-cell cells). Fig. 2 C compares the results of PopCorn and Seurat on the metrics as defined above and shows that PopCorn outperforms Seurat on all metrics scores.



B

Metric evaluation on PopCom vs. Seurat

	Rccsp	Rucct	Accc
PopCom	**18/18**	**11/16**	**92.4%**
Seurat	14/19	6/16	76.8%

$$\text{Rccsp} = \frac{\#\text{ corresponding common sub-populations}}{\#\text{ common sub-populations}}$$

$$\text{Rucct} = \frac{\#\text{ uncovered common cell types}}{\#\text{ total common cell types}}$$

Accc: The accuracy of identified common sub-populations

Fig. 3: (A) Sankey diagrams of the results mapping between identified sub-populations to ground turth labels of both PopCorn and Seurat. The numbers in the background of the center gray area correspond to the respective replica indexes. The width of the flow bar is proportional to the Acc_{ssp} score (accuracy of a *split sub-population*) defined in Supplementary Materials C.2. The Acc_{ssp} scores are given in Supplementary Materials D.2. As shown, all *common sub-populations* identified by PopCorn are *corresponding common sub-populations*. In contrast, 5 out of 19 common sub-populations identified by Seurat are *non-corresponding common sub-populations*, which assinged cell types of different kinds in individual data sets together. Furthermore, PopCorn uncovered 11 common cell types whereas Seurat uncovered only 6. (B) Comparison between PopCorn and Seurat on mouse kidney single cell data sets. PopCorn outperforms Seurat on all metrics as described in Supplementary Materials C.

3.2 Comparative analysis of single-cell RNA-seq data sets with multiple replicas

Next, we tested the robustness of PopCorn to perform under the presence of multiple replicas, a crucial biological imperative to monitor the quality and repeatability of an experiment. Computational approaches processing the resulting data are expected to be resilient against reasonable technological and biological variations and to produce consistent results across replicas. To test this, we applied PopCorn and Seurat to mouse kidney scRNA-seq data recently published in [17] and performed comparative analysis on the four replicas. The replicas, identified by GEO association numbers GSM2871706, GSM2871707, GSM2871708, and GSM2871709, contained 2,943 cells, 5,060 cells, 1,383 cells, and 2,704 cells, respectively. Moreover, the four replicas have 16 cell types in common and no distinctive individual cell types. Finally, and analogous to the previous analysis procedure, the cell labels provided in [17] were used as ground truth to evaluate the performance of each method. The parameter selection of both methods can be found in Supplementary Materials D.1.

Fig. 3 summarizes the findings of applying the above described test scenario on PopCorn and Seurat and shows that PopCorn outperforms Seurat on all metric scores. As shown in Fig. 3 A, PopCorn successfully identifies 18 common sub-populations, all of which are *corresponding common sub-populations*. In contrast, Seurat identifies 19 *common sub-populations*, 5 of which are *non-corresponding common sub-populations* erroneously grouping together cells of different types in different replicas. Specifically, sub-population #15 of Seurat incorrectly assigns DCT cells in replica 3 into a group of CD-PC cells from replica 1, 2, and 4 and sub-population #16 groups Macro cells in replica 2 with Fib cells in replicas 3 and 4. In addition, sub-population #17 contains B-lymph cells in replica 1 and Endo cells in replica 2 while sub-population #18 falsely assigns T-lymph cells in replica 3 and NK cells in replicas 1 and 4 together, and sub-population #19 mistakenly groups Macro cells in replica 1 together with B-lymph cells in replicas 2, 3, and 4. It is also worth noting that out of the 16 common cell types as stated in [17], PopCorn identified a total of 11 as evidenced by the number of unique annotations for *common sub-populations*. In contrast, Seurat was only able to uncover 6 common cell types (Fig. 3 A, right). This prompted us to further investigate the properties of the remaining common cell types that were not uncovered by PopCorn. Interestingly, we found these cell types, Podo, CD-Trans, Novel1, Neutro and Novel2, to have particularly small cell populations of 10, 56, 42, 25, and 10 cells respectively corresponding to 0.1, 0.5, 0.4, 0.2, and 0.1 percent of all cells in 4 replicas.

4 Conclusions

We developed PopCorn, a new method for the identification of sub-populations of cells present within individual single cell experiments and mapping of these sub-populations across the experiments. In contrast to alternative approaches PopCorn performs these two tasks simultaneously by optimizing a function that combines both objectives. When applied to complex biological data the results produced by our approach are of unprecedented quality, robustness, and reproducibility across replicas that was not available in previous method. Several innovations developed in this work contributed to this success. First, incorporating the above mentioned tasks into a single problem statement was crucial for integrating the signal form different experiments while identifying sub-populations within each experiment. Next, the sub-population co-membership propensity graph introduced here to guide sub-population identification in individual experiments significantly aids the reliable identification of groups of similar cells that are well separated from the remaining populations. Taken together, these two ideas enable highly accurate identification of sub-populations and superior alignment of cells across populations.

With these qualities, PopCorn has great potential to become a fundamental tool in the analysis of single cell data.

5 Availability

A preliminary reference implementation of PopCorn is available upon request.

6 Acknowledgements

This research was supported by the Intramural Research Programs of the National Library of Medicine.

References

[1] E. Shapiro, T. Biezuner, and S. Linnarsson. "Single-cell sequencing-based technologies will revolutionize whole-organism science". In: *Nat. Rev. Genet.* 14.9 (Sept. 2013), pp. 618–630.

[2] H. Wu and B. D. Humphreys. "The promise of single-cell RNA sequencing for kidney disease investigation". In: *Kidney Int.* 92.6 (Dec. 2017), pp. 1334–1342.

[3] R. Bacher and C. Kendziorski. "Design and computational analysis of single-cell RNA-sequencing experiments". In: *Genome Biol.* 17 (Apr. 2016), p. 63.

[4] R. Rostom, V. Svensson, S. A. Teichmann, and G. Kar. "Computational approaches for interpreting scRNA-seq data". In: *FEBS Lett.* 591.15 (Aug. 2017), pp. 2213–2225.

[5] O. Stegle, S. A. Teichmann, and J. C. Marioni. "Computational and analytical challenges in single-cell transcriptomics". In: *Nat. Rev. Genet.* 16.3 (Mar. 2015), pp. 133–145.

[6] C. Mayer, C. Hafemeister, R. C. Bandler, R. Machold, R. Batista Brito, X. Jaglin, K. Allaway, A. Butler, G. Fishell, and R. Satija. "Developmental diversification of cortical inhibitory interneurons". In: *Nature* 555.7697 (Mar. 2018), pp. 457–462.

[7] S. Sun, T. Babola, G. Pregernig, K. S. So, M. Nguyen, S. M. Su, A. T. Palermo, D. E. Bergles, J. C. Burns, and U. Muller. "Hair Cell Mechanotransduction Regulates Spontaneous Activity and Spiral Ganglion Subtype Specification in the Auditory System". In: *Cell* 174.5 (Aug. 2018), pp. 1247–1263.

[8] B. R. Shrestha, C. Chia, L. Wu, S. G. Kujawa, M. C. Liberman, and L. V. Goodrich. "Sensory Neuron Diversity in the Inner Ear Is Shaped by Activity". In: *Cell* 174.5 (Aug. 2018), pp. 1229–1246.

[9] K. G. Paulson et al. "Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA". In: *Nat Commun* 9.1 (Sept. 2018), p. 3868.

[10] L. Duan et al. "PDGFRβ Cells Rapidly Relay Inflammatory Signal from the Circulatory System to Neurons via Chemokine CCL2". In: *Neuron* 100.1 (Oct. 2018), pp. 183–200.

[11] M. Verma, Y. Asakura, B. S. R. Murakonda, T. Pengo, C. Latroche, B. Chazaud, L. K. McLoon, and A. Asakura. "Muscle Satellite Cell Cross-Talk with a Vascular Niche Maintains Quiescence via VEGF and Notch Signaling". In: *Cell Stem Cell* 23.4 (Oct. 2018), pp. 530–543.

[12] J. Ordovas-Montanes et al. "Allergic inflammatory memory in human respiratory epithelial progenitor cells". In: *Nature* 560.7720 (Aug. 2018), pp. 649–654.

[13] L. E. Byrnes, D. M. Wong, M. Subramaniam, N. P. Meyer, C. L. Gilchrist, S. M. Knox, A. D. Tward, C. J. Ye, and J. B. Sneddon. "Lineage dynamics of murine pancreatic development at single-cell resolution". In: *Nat Commun* 9.1 (Sept. 2018), p. 3922.

[14] G. Dimitriadis, J. P. Neto, and A. R. Kampff. "t-SNE Visualization of Large-Scale Neural Recordings". In: *Neural Comput* 30.7 (July 2018), pp. 1750–1774.

[15] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nat. Biotechnol.* 36.5 (June 2018), pp. 411–420.

[16] Reid Andersen, Fan Chung, and Kevin Lang. "Local Graph Partitioning using PageRank Vectors". In: FOCS, 2006, pp. 475–486.

[17] J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch, and K. Susztak. "Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease". In: *Science* 360.6390 (May 2018), pp. 758–763.

[18] J. Shi and J. Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (Aug. 2000), pp. 888–905.

[19] Eric P. Xing and Michael I. Jordan. *On Semidefinite Relaxation for Normalized k-cut and Connections to Spectral Clustering*. Tech. rep. UCB/CSD-03-1265. EECS Department, University of California, Berkeley, June 2003.

[20] S. H. Yip, P. C. Sham, and J. Wang. "Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data". In: *Brief. Bioinformatics* (Feb. 2018).

Supplementary Materials

A Construction of co-membership propensity graphs

In this section, we introduce how we construct a co-membership propensity graph based on a similarity matrix. Let us assuming we have a similarity matrix W that is derived by using (5). W can be viewed as the weighted adjacency matrix for a weighted un-directed graph $G(V,E)$. W_{ij} encodes the similarity between node (cell) i and node (cell) j.

For a given node $v \in V$, the group H_v^* of nodes that tend to be in the same partition with v can be found by using the personalized PageRank vector of cell v [16]. The personalized PageRank vector $p(\alpha, v)$ of v on G is the stationary distribution of the random walk on G, in which at every step, the random walker has the probability of α to restart the random walk at v and otherwise performs a lazy random walk. Mathematically, $p(\alpha, v)$ is the unique solution to

$$p(\alpha, v) = \alpha e_v + (1-\alpha)\, p(\alpha, v) H \tag{7}$$

where $\alpha \in [0,1]$ is the "teleportation" constant, e_v is the indicator vector of v (where $e_s = 1$ if $s = v$ and $e_s = 0$ if $s \neq v$) and $H = \frac{1}{2}\left(I + D^{-1}W\right)$ is the underlying probability transition matrix of the lazy random walk. D is a diagonal matrix with the weight sum of each node $d(v) = \sum_m W_{vm}, \forall v \in V$ on its diagonal $D_{vv} = d(v)$. We apply the modified local algorithm in [16] to efficiently approximate $\hat{p} = p(\alpha, v)$. The algorithm in [16] is for unweighted graphs and we extend the algorithm for weighted graph as shown in **ApproximatePageRank_weight**(v, α, ε), which unities **Push**$_u$(p, r).

Push$_u$(p, r)

1. Let $p' = p$ and $r' = r$, except for the following changes:
   (a) $p'(u) = p(u) + \alpha r(u)$
   (b) $r'(u) = \frac{(1-\alpha)}{2} r(u)$
   (c) For each v such that $(u,v) \in E$: $r'(v) = r(v) + \frac{(1-\alpha)}{2} \frac{W_{uv}}{d(u)} r(u)$
2. Return (p', r')

Then we sort the nodes based on $\hat{p}$ and attain an ordered set $H_v = \{h_1, h_2, ..., h_n\}, h_i \in V$, whose elements satisfy $\hat{p}(h_i) \geq \hat{p}(h_{i+1})$. It is easy to verify that when $\alpha > 0.5$, so that node v is always on top of the list H_v, meaning $h_1 = v$. Therefore, we use $\alpha = 0.8$ through out the paper to make sure that v is in H_v^*. Based on the ordering in H_v, we generate a collection of sets $S_j = \{h_1, h_2, ..., h_j\}$ for $j \in \{0, 1, ..., |H_v|\}$, which we call

ApproximatePageRank_weight(v, α, ε)

1. Let $p = 0$ and $r = e_v$.
2. While $\max_{u,h \in V} \frac{r(u)W_{uh}}{d(u)} \geq \varepsilon$:
   (a) Choose any node u where $\frac{r(u)W_{uh}}{d(u)} \geq \varepsilon$
   (b) Apply $\mathbf{Push}_u$ at node u, and update p and r
3. Return p with $\max_{u,h \in V} \frac{r(u)W_{uh}}{d(u)} < \varepsilon$

sweep sets [16]. We let

$$H_v^* = \max_{j \in [1, |H_v|]} \frac{\Psi(S_j)}{\Phi(S_j)} \tag{8}$$

be the node set including v that has propensity to be well-separated (characterized by $\Phi(\cdot)$) and densely connected (characterize by $\Psi(\cdot)$). $\Psi(S)$ is the weighted density between nodes in set S defining as

$$\Psi(S) = \frac{\sum_{lm} W_{lm}}{|S|^2}, v_l, v_m \in S. \tag{9}$$

And $\Phi(S)$ is the conductance of set S that characterizes separation of S.

$$\Phi(S) = \frac{\partial(S)}{\min(\mathrm{vol}(S), \mathrm{vol}(V^i) - \mathrm{vol}(S))}, \tag{10}$$

where $\partial(S) = \sum_{ij} W_{ij}, i \in S, j \notin S$ and $\mathrm{vol}(S) = \sum_{x \in S} d(x)$. H_v^* can be represented by an indicator vector $e_{H_v^*} = \sum_{s \in H_v^*} e_s$. After obtain H_v^* for every $v \in V^i$, we can compute the adjacency matrix of the co-membership propensity graph for G as follows

$$A = \sum_{v \in V} e_{H_v^*} e_{H_v^*}^T. \tag{11}$$

Based on the method introduced above, we can compute A^i for each single cell data set D^i.

B Optimization

B.1 Normalized Laplacian matrix

For adjacency matrix A, the normalized Laplcian matrix of A is defined as $L_A = I - D_A^{-1/2} A D_A - 1/2$, where I is an identify matrix and D_A is the weighted degree matrix with $D_{ii} = \sum_j A_{ij}$ on its diagonal. Similarly, the the normalized Laplcian matrix of B is $L_B = I - D_B^{-1/2} B D_B - 1/2$.

B.2 SDP relaxation of (4)

The joint partition problem (4) is a NP-hard problem. In order to efficiently obtain promising results, we relax the problem into SDP relaxation. Based on previous results [19], we know the following problems (P1) and (P2) can be relaxed to (SDP1) and (SDP2).

$$(P1)\quad \begin{aligned} \min: \quad & \mathrm{tr}(Y^T L_A Y) \\ s.t. \quad & Y \in \mathfrak{F}_k. \end{aligned} \qquad (P2)\quad \begin{aligned} \min: \quad & \mathrm{tr}(Y^T L_B Y) \\ s.t. \quad & Y \in \mathfrak{F}_k. \end{aligned}$$

$$(SDP1)\quad \begin{aligned} \min: \quad & \mathrm{tr}(L_A Z_A) \\ s.t. \quad & Z_A \mathrm{diag}(D_B^{-1/2}) = \mathrm{diag}(D_B^{-1/2}) \\ & \mathrm{tr}(Z_A) = k \\ & Z_A = Z_A^T, Z_A \succeq 0 \\ & (Z_A)_{ij} \geq 0, \forall i, j \end{aligned} \qquad (SDP2)\quad \begin{aligned} \min: \quad & \mathrm{tr}(L_B Z_B) \\ s.t. \quad & Z_B \mathrm{diag}(D_B^{-1/2}) = \mathrm{diag}(D_B^{-1/2}) \\ & \mathrm{tr}(Z_B) = k \\ & Z_B = Z_B^T, Z_B \succeq 0 \\ & (Z_B)_{ij} \geq 0, \forall i, j \end{aligned}$$

In those SDP relaxations $Z_A = Y_A Y_A^T$ and $Z_B = Y_B Y_B^T$, where $Y_A = D_A^{-1} Y_A ((Y_A)^T D_A Y_A)^{-1/2}$ and $Y_B = D_B^{-1} Y_B ((Y_B)^T D_B Y_B)^{-1/2}$. Comparing (SDP1) and (SDP2), we find that they have exact the same constraints except the first ones. We therefore relax the first constraint of (SDP1) and let $Z = Z_A = Z_B$ and combine (SDP1) and (SDP2) to obtain our final SDP formulation as follow.

$$(SDP)\quad \begin{aligned} \min: \quad & \mathrm{tr}((L_A + \lambda L_{\bar{B}})Z) \\ s.t. \quad & Z\mathrm{diag}(D_B^{-1/2}) = \mathrm{diag}(D_B^{-1/2}) \\ & \mathrm{tr}(Z) = k \\ & Z = Z^T, Z \succeq 0 \\ & Z_{ij} \geq 0, \forall i, j, \end{aligned} \tag{14}$$

Problem (14) can be solved by well-established toolbox and we use cvxpy to solve the relaxed SDP problem.

B.3 Rounding

In general, due to relaxation, the optimal solution of (SDP) is not feasible for (4). Therefore, we need to recover a closest feasible solution to the original problem (4). We treat row i of Z, the optimal solution of (SDP), as the feature of cell i. Then we apply k-means to Z 100 times and pick one k-means solution which yields the minimum objective function value of (4).

B.4 Preclustering

Since our objective function unitizes a full $N \times N$ matrix $L_A + \lambda L_{\bar{B}}$, the memory usage of the algorithm may be prohibitive for increasing number of data sets. This motivates us to use *supercells*, which can be obtained by over-segmentation of the data sets (in our case, hierarchical clustering is used but any other superpixel identification methods in the image processing filed can be applied). Using N_S supercells is equivalent to constraint Z to be block-constant and thus reduces the size of the SDP to a problem of $N_S \times N_S$.

C Evaluation metric

C.1 Terminology

Given q data sets $\{D^1, D^2, ..., D^q\}$, we use $C = \{C^1, C^2, ..., C^q\}$ to present a sub-population that is identified by a method, where C^i is a *split sub-population* that only contains cells from data set D^i. We call a sub-population a *individual sub-population* once C only contains cells from one data set ($C^i \neq \emptyset$ and $C^j = \emptyset, \forall j \neq i$). When C contains cells from more than just one data set, we call C a *common sub-population*.

C.2 Annotation of a *split sub-population*

For a non-empty *split sub-population* $C^i = \{C^i_1, C^i_2, ...\}$ in C, we can annotate each cell C^i_j in C^i by its ground truth label T^i_j. We use $T^i = \{T^i_1, T^i_2, ...\}$ to present the label set of C^i, where T^i_j is the ground truth label for C^i_j. Assuming we have L ground truth label $T = \{T_1, T_2, ..., T_L\}$. Then the annotation I_{C^i} of the *split sub-population* C^i can be found by

$$I_{C^i} = \arg\max_{T_l \in T} \frac{1}{|C^i|} \sum_j \mathbf{1}(T^i_j \in T_l), \tag{15}$$

where $\mathbf{1}(\cdot)$ is an indicator function. $\mathbf{1}(T^i_j \in T_l) = 1$ when $T^i_j \in T_l$, and $\mathbf{1}(T^i_j \in T_l) = 0$ otherwise.

Furthermore, we can define the accuracy Acc_{ssp} of the *split sub-population* as follow.

$$Acc_{ssp}(C^i) = \frac{1}{|C^i|} \sum_j \mathbf{1}(T^i_j = I_{C^i}), \tag{16}$$

C.3 Annotation of a *common sub-population*

After finding the annotation I_{C^i} for each non-empty *split sub-population* C^i in C, we can check whether they have the same annotations. If all the non-empty *split sub-populations* in C are annotated to the some annotation, we consider the *common sub-population* as a *corresponding common sub-population* and use

$I_C = I_{C^i}$ to denote the annotation for the *corresponding common sub-population*, where i is the index of a non-empty *split sub-population* in C. Otherwise, if the non-empty *split sub-populations* are annotated to different annotations, we consider the *common sub-population* as a *non-corresponding common sub-population* and set $I_C = NA$.

C.4 The ratio R_{ccsp} of the *corresponding common sub-population*

R_{ccsp} is the ratio of the number of the *corresponding common sub-populations* that identified by a method to the number of all identified *common sub-populations*. For example, if all *common sub-populations* are *corresponding common sub-populations*, $R_{ccsp} = 100\%$; If all *common sub-populations* are *non-corresponding common sub-populations*, $R_{ccsp} = 0\%$.

Let $\mathfrak{C} = \{\mathfrak{C}_I, \mathfrak{C}_c\}$ be the output of a method, where $\mathfrak{C}_I$ is the set of all *individual sub-populations* and $\mathfrak{C}_c$ is the set of all *common sub-populations*. The ratio of the *corresponding common sub-population* R_c is defined as follow:

$$R_{ccsp} = \frac{\sum_{C \in \mathfrak{C}_c} \mathbf{1}(I_C \neq NA)}{|\mathfrak{C}_c|} \tag{17}$$

C.5 The ratio R_{ucct} of uncoveblack common cell types

R_{ucct} measures the percentage of the common cell types that can be uncoveblack. Let T_c be the set of ground truth labels that share by multiple data sets. We know $T_c \in T, T = \{T_1, T_2, ..., T_L\}$ and then R_{ucct} can be computed as

$$R_{ucct} = \frac{\sum_{T_l \in T_c} \mathbf{1}(T_l \in I_{\mathfrak{C}_c})}{|T_c|}, \tag{18}$$

where $I_{\mathfrak{C}_c} = \{I_C | C \in \mathfrak{C}_c\}$.

C.6 The ratio R_{uict} of uncoveblack individual cell types

R_{uuct} measures the percentage of the individual cell types that can be uncoveblack. Let T_I be the set of ground truth labels that only appear in one data set. We know $T_I \in T, T = \{T_1, T_2, ..., T_L\}$ and then R_{uict} can be computed as

$$R_{uict} = \frac{\sum_{T_l \in T_I} \mathbf{1}(T_l \in I_{\mathfrak{C}_I})}{|T_I|}, \tag{19}$$

where $I_{\mathfrak{C}_I} == \{I_C | C \in \mathfrak{C}_I\}$.

C.7 The accuracy *Accc* of identified *common sub-populations*

Here we use Acc_{csp} to evaluate the purity of a *common sub-population*. Let C denote a *common sub-population* and $\bar{C} = \{\bar{C}^1, \bar{C}^2, ...\}$ be the *common sub-population* where $\bar{C}^i \neq \emptyset$. The accuracy Acc_{csp} of the *common sub-population* of C can be computed as

$$Acc_{csp}(C) = \begin{cases} \frac{\sum_i Acc_{ssp}(\bar{C}_c^i)}{|\bar{C}_c|} & \text{if } C \text{ is a corresponding common sub-population} \\ 0 & \text{if } C \text{ is a non-corresponding common sub-population.} \end{cases} \tag{20}$$

For the set $\mathfrak{C}$ of all *common sub-populations*, the accuracy of $\mathfrak{C}$ is

$$Accc(\mathfrak{C}) = \frac{\sum_{C \in \mathfrak{C}} Acc_{csp}(C)}{|\mathfrak{C}|}. \tag{21}$$

C.8 The accuracy *Acci* of identified *individual sub-populations*

The Acc_{isp} score for an *individual sub-population* C_I is

$$Acc_{isp}(C_I) = Acc_{ssp}(C_I). \tag{22}$$

For the set $\mathfrak{C}_I$ of all *individual sub-populations*, the accuracy of $\mathfrak{C}_I$ is

$$Acci(\mathfrak{C}_I) = \frac{\sum_{C \in \mathfrak{C}_I} Acc_{isp}(C)}{|\mathfrak{C}_I|}. \tag{23}$$

D Implementation details

D.1 Parameter settings

For human and mouse pancreatic single cell data sets, we set λ and k (the number of sub-populations), the only two parameters of PopCorn, to $\lambda = 1$ and $k = 10, 11, 12, 13, 14, 15$, respectively. We find that when we set $k = 14$, our PopCorn only generates 13 clusters, therefore, we use the results of $k = 13$ as our final results to compare with Seurat. For Seurat, we use the results they reported in [15] to compare with our PopCorn results because we use exact the same data sets.

For mouse kidney single cell data sets, we set $\lambda = 3$ and $k = 19$, cause when we set $k = 20$, PopCorn only generates 19 sub-populations. For Seurat, we use different number of HVGs (500, 1000, 1500) for the alignment method. And for the sub-population identification method, Seurat use the modularity maximiza-

tion method and we set the resolution parameters to 0.8, 1.0, and 1.2. We try all combinations of the above parameter settings and choose a result that generates the maximum modularity score for comparison.

D.2 Details results for mouse kidney data sets

Here we provide detailed information for Fig. 3.

A **PopCom Detailed Results**

	D1			D2			D3			D4			
	# cells	annotation	Acc_{ssp}	# cells	annotation	Acc_{ssp}	# cells	annotation	Acc_{ssp}	# cells	annotation	Acc_{ssp}	$Accc_{csp}$
#1	298	PT	100%	748	PT	98%	316	PT	99.37%	587	PT	96.42%	98.45%
#2	208	PT	95.65%	284	PT	98.24%				47	PT	95.74%	96.54%
#3	348	PT	98.28%	685	PT	97.08%	23	PT	91.30%	16	PT	62.50%	87.29%
#4	297	PT	98.99%	1122	PT	98.13%	186	PT	98.39%	347	PT	98.85%	98.59%
#5	329	PT	97.26%	729	PT	97.67%	63	PT	98.41%	117	PT	98.29%	97.91%
#6	23	Endo	100%	25	Endo	100%							100.00%
#7	70	Endo	100%	88	Endo	100%	68	Endo	100%	119	Endo	100%	100.00%
#8	48	Endo	100%	55	Endo	100%							100.00%
#9	217	LOH	95.39%	168	LOH	95.83%	140	LOH	99.29%	228	LOH	98.68%	97.30%
#10	256	DCT	82.42%	300	DCT	74%	76	DCT	93.42%	164	DCT	93.29%	85.78%
#11	363	DCT	98.07%	326	DCT	98.47%	129	DCT	98.45%	284	DCT	98.94%	98.48%
#12	98	CD-IC	85.71%	147	CD-IC	90.48%	111	CD-IC	100%	220	CD-IC	99.55%	93.94%
#13	102	T-lymph	88.24%	118	T-lymph	59.32%	140	T-lymph	99.29%	355	T-lymph	98.59%	86.36%
#14	29	Fib	93.10%	65	Fib	81.54%	16	Fib	87.50%	25	Fib	72%	83.54%
#15	88	CD-PC	97.73%	82	CD-PC	91.46%	12	CD-PC	100%	36	CD-PC	100%	97.30%
#16	50	B-lymph	88%	24	B-lymph	70.83%	12	B-lymph	91.67%	19	B-lymph	100%	87.63%
#17	23	NK	73.91%	40	NK	57.50%	60	NK	85%	109	NK	75.23%	72.91%
#18	75	Macro	84%	45	Macro	100%				17	Macro	58.82%	80.94%
#19	19	CD-Trans	100%										

B **Seurat Detailed Results**

	D1			D2			D3			D4			
	# cells	annotation	Acc_{ssp}	# cells	annotation	Acc_{ssp}	# cells	annotation	Acc_{ssp}	# cells	annotation	Acc_{ssp}	$Accc_{csp}$
#1	146	PT	96.58%	32	PT	100%	133	PT	99.25%	215	PT	97.21%	98.26%
#2	27	PT	100%	637	PT	99.37%	84	PT	98.81%	85	PT	92.94%	97.78%
#3	18	PT	83.30%	314	PT	99.36%							91.33%
#4	302	PT	98.01%	890	PT	97.08%	24	PT	95.83%	23	PT	91.30%	95.56%
#5	300	PT	99.67%				106	PT	97.17%	82	PT	98.78%	98.54%
#6	165	PT	98.18%	847	PT	99.31%	79	Endo	98.73%	54	PT	100%	99.06%
#7	73	PT	94.52%							394	PT	98.22%	96.37%
#8	459	PT	97.60%	816	PT	97.06%	41	PT	95.12%	35	PT	91.43%	95.30%
#9	123	Endo	100%	160	Endo	98.75%	70	Endo	100%	129	Endo	100%	99.69%
#10	209	LOH	98.56%	168	LOH	97.02%	135	LOH	100%	227	LOH	99.12%	98.68%
#11	386	DCT	98.45%	345	DCT	99.71%	135	DCT	97.04%	270	DCT	99.25%	98.61%
#12	182	DCT	98.89%	210	DCT	95.24%	41	DCT	97.56%	167	DCT	97.00%	97.17%
#13	111	CD-IC	73.87%	152	CD-IC	87.50%	113	CD-IC	99.15%	228	CD-IC	98.25%	94.97%
#14				58	T-lymph	96.55%				341	T-lymph	98.83%	97.69%
#15	119	CD-PC	100%	123	CD-PC	95.12%	48	DCT	47.92%	44	CD-PC	86.36%	0.00%
#16				45	Macro	100%	21	Fib	76.19%	30	Fib	67.67%	0.00%
#17	93	B-lymph	49.46%	34	Endo	79.41%							0.00%
#18	110	Macro	56.36%	59	B-lymph	94.92%	21	B-lymph	57.14%	41	B-lymph	46.34%	0.00%
#19	20	NK	75%				191	T-lymph	72.25%	111	NK	72.07%	0.00%
#20							126	PT	98.41%				98.41%
#21							69	Fib	79.71%				79.71%
#22							66	T-lymph	50%				50.00%
#23										223	PT	95.96%	95.96%
#24				95	T-lymph	96.84%							

Fig. 4: (A) Table for the detailed results for PopCorn on mouse kidney 4 replicas. (B) Table for the detailed results for Seurat on mouse kidney 4 replicas. Gray shade shows the *non-corresponding common sub-populations*.