Single-cell transcriptomics data purification with coreset selection

Róbert Pálovics[1], Tony Wyss-Coray[1], Baharan Mirzasoleiman[2]

1. Wu Tsai Neurosciences Institute, Stanford University School of Medicine, Stanford, CA, USA

2. Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA

Motivation

- The number of cells captured per biological replicates can vary across the "*single-cell landscape*" (**Fig.1**) introducing replicate specific biases that
  - result in false discoveries contaminating differential gene expression (DGE) results
  - prevent the broad usage of predictive machine learning (ML) methods on single-cell data (e.g., age or clinical data prediction)
- We propose a coreset selection based purification method to alleviate potential replicate specific biases within single-cell datasets

Method

Input and preprocessing

- Annotated log-CPM normalized gene-cell count matrix with metadata (replicate info, ...)
- First *M* principal components (PCs)

Step I: Discard replicate specific areas (Fig. 2)

- Calculate for each cell *c* the *k*-nearest neighbors of the cell (n_c) in the PC space based on Euclidean distances
- Include *c* if (replicate of c: l_c):

$$0 < |\{d : l_c = l_d, d \in n_c\}| < k$$

Step II: Coreset selection (Fig. 3)

- We intend to select a set of cells best representing the included ones from Step I
- Define the similarity of cells *c* and *d* as

$$s_{cd} = \exp\{-|p_c - p_d|^2/(2\sigma^2)\}$$

where σ is the standard deviation of the PC matrix and *p* is a PC

- Objective: select a set of $r|V|$ cells,

$$S^* \in \arg\max_{\substack{S \subseteq V \\ |S| \leq r|V|}} F(S);\ F(S) := \sum_{c \in V} \max_{d \in S} s_{cd}$$

- Greedy solution: start with an empty set and at each iteration *t* choose a cell *e* that maximizes the marginal utility

$$F(e|S_t) = F(S_t \cup \{e\}) - F(S_t)$$

- We identify coresets for each condition (e.g., control and treatment) separately

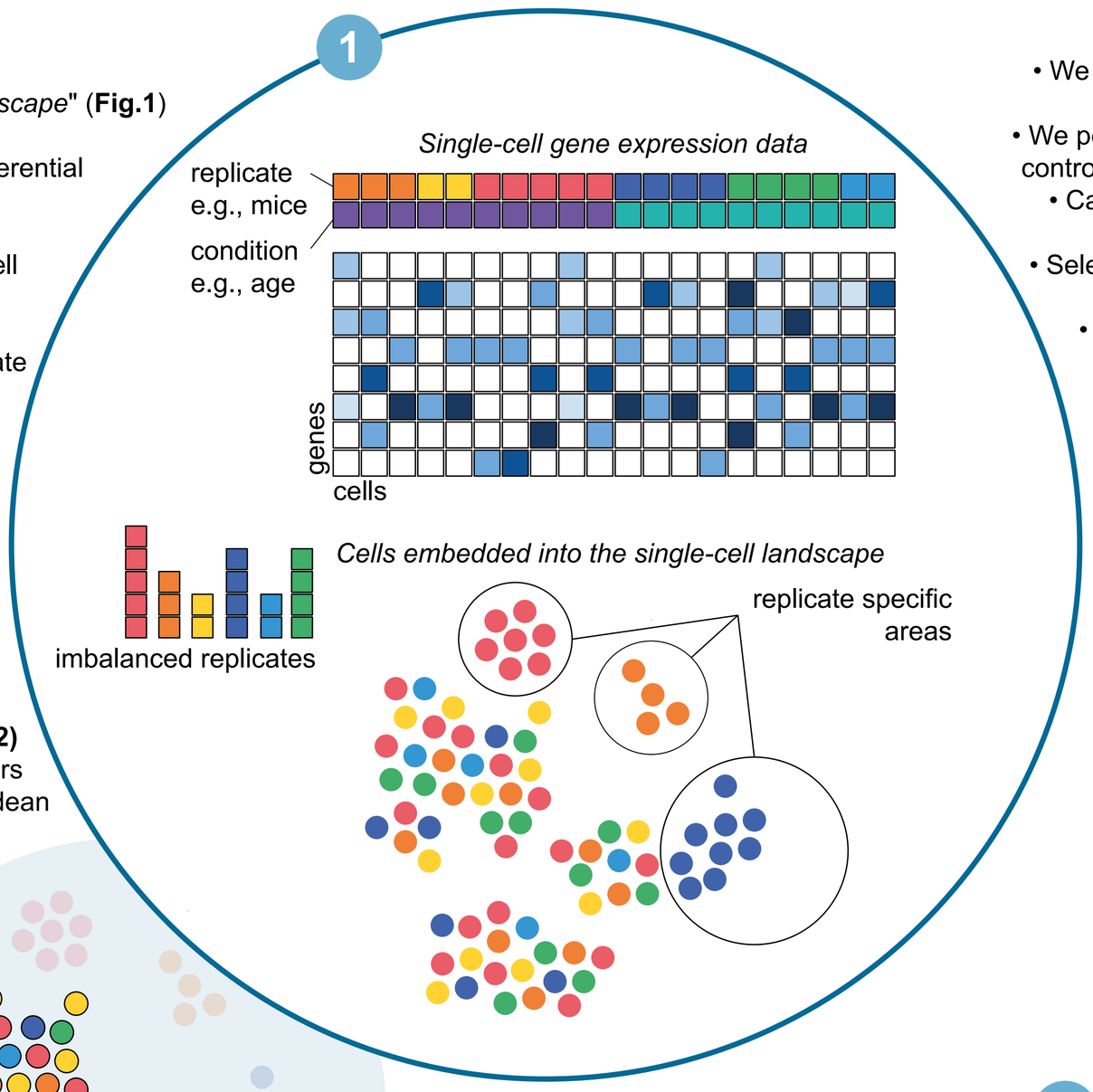
Data

Tabula Muris Senis (TMS)

- SmartSeq-2 data on male *young* (3 mo.) and *aged* (18/24 mo.) mice from 20 tissues
- Select cell types that have at least 2 replicates with a minimum of 20 cells both in the control (*young*) and treatment (*aged*) groups: *24 cell types in total*
- Number of cells range between 100-10,000 per cell type

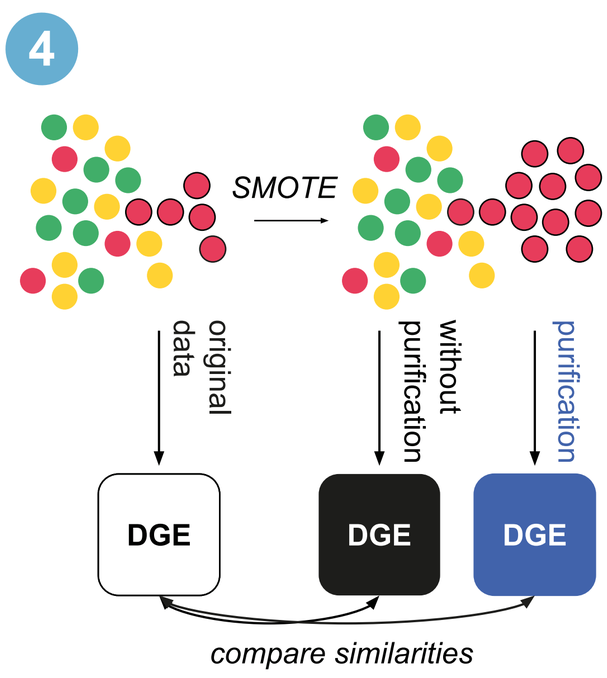
Murine aging cell atlas (Calico)

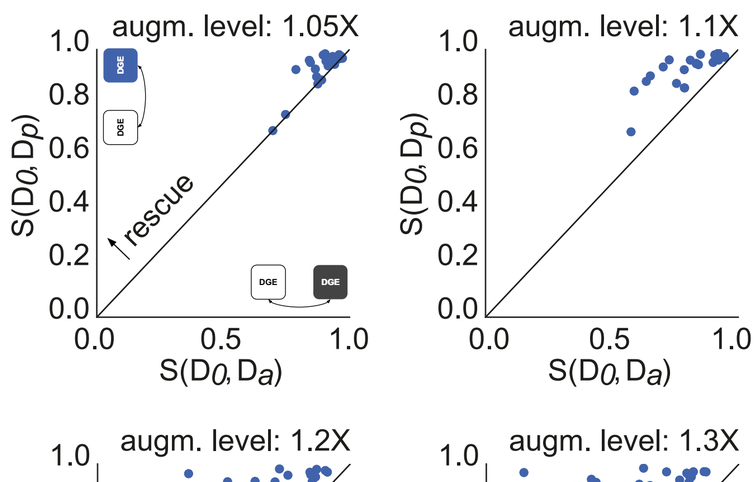
- Droplet based data from 3 tissues (kidney, lung, spleen) of *young* (7/8 mo.) and *aged* (22/23 mo.) mice
- *8 cell types* in total found both in TMS and Calico
- Number of cells range between 500-20,000 per cell type



Results I

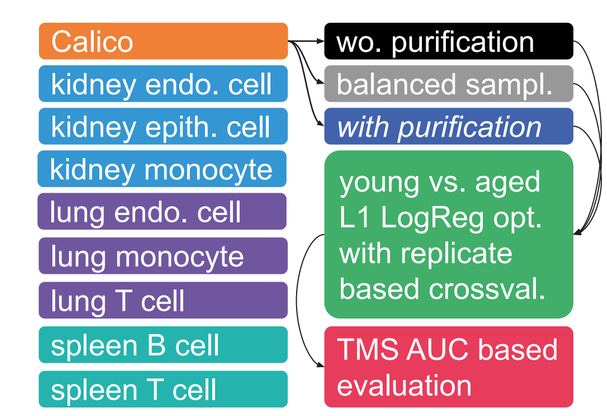
- We found that purification protects against replicate specific biases that contaminate DGE results
- We performed a synthetic, data augmentation based, controlled experiment in 24 cell types of TMS (**Fig. 4**):
  - Calculate the distance of each cell from its closest neighbor from a different replicate
  - Select the cell with highest such distance as well as any neighboring cells from the same replicate
  - Create controlled bias: use SMOTE to augment the data based on the selected "outlier" cells
  - Perform DGE (Mann-Whitney *U*) between *young* and *aged* cells on the original data (D_0), the augm. data (D_a) and the purified data (D_p)
  - Purif. parameters: $k = 10$; $M = 20$; $r = 0.9$ $|V|$
  - Calculate Spearman correlations based on the obtained p-values: $S(D_0, D_a)$; $S(D_0, D_p)$
- Results indicate that $S(D_0, D_a) < S(D_0, D_p)$, i.e, purification rescues the augm. data (**Fig. 5**)





Results II

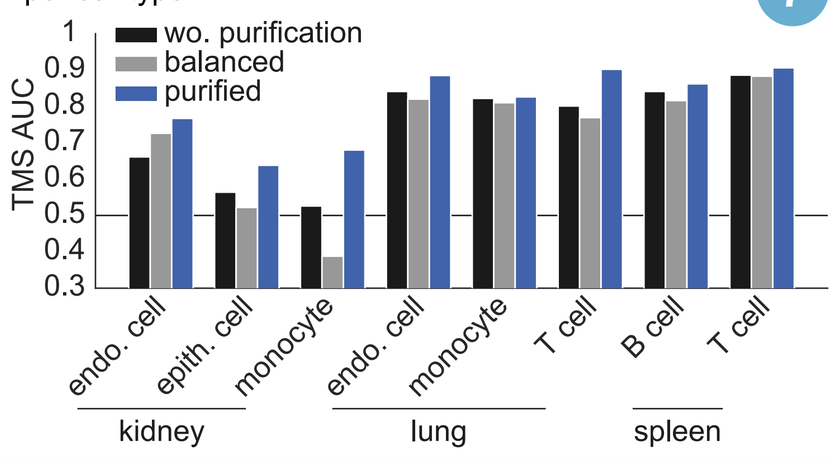
- Purification improved the prediction of age at the cell level, experiment is shown in **Fig. 6**:
- Train cell type specific LogReg classifiers with L1 reg. ($\alpha : 0.02 - 2$) for *young* vs. *aged* on Calico with replicate based cross-validation
- Repeat the same training procedure but purify the training set before model fitting
  - Balanced sampling: select equal number of cells per replicate uniformly at random
  - Purification results in the highest AUC scores measured on TMS (**Fig. 7**)
  - **Fig. 8**: Purification of the lung T cells

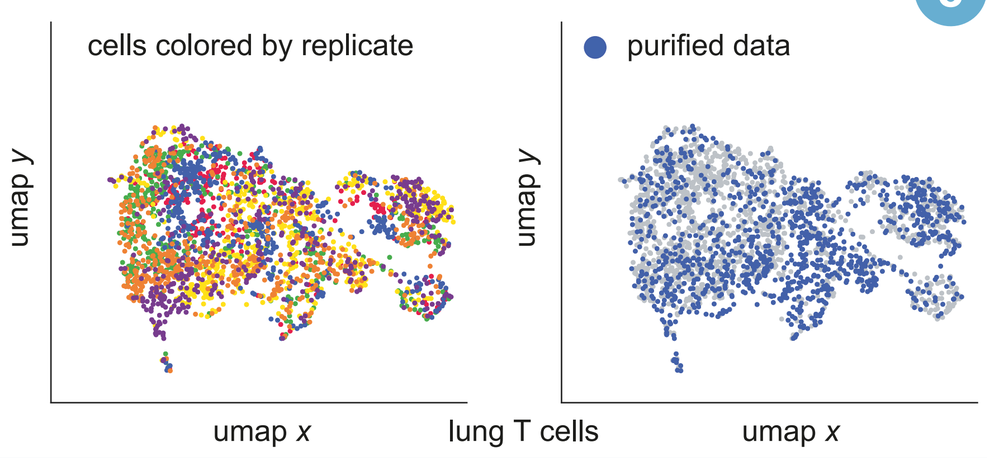


Conclusion

We introduced a coreset selection based method to purify single-cell data. Purification is protective against replicate specific biases and aids downstream analyses, in particular differential gene expression.

Additionally, it substantially improves the predictive performance of supervised models trained on single-cell data. Purification leads to more generalizable cell level aging classifier models indicated by the higher predictive performance when validated on multiple cell types of an independent cohort.





UCLA

Stanford University