Purification of single-cell transcriptomics data with coreset selection

Róbert Pálovics [1] Tony Wyss-Coray [1 2 3] Baharan Mirzasoleiman [4]

Abstract

Despite the overall success of single-cell transcriptomics, variations in the number of cells captured from biological replicates in different regions of the embedding space of cells limit the interpretation of downstream computational analyses. Here we introduce a coreset selection based purification method to alleviate potential replicate specific biases within single-cell datasets. We first identify regions of the embedding space of cells that are not biased towards single biological replicates, and then extract a representative cell subset (coreset) covering them. We demonstrate that the extracted coresets provide a solid ground for downstream analyses. Specifically, we show that differential gene expression signatures based on purified datasets are robust against replicate specific biases across 24 different cell-type specific single-cell datasets. Furthermore, we highlight that purification can enhance supervised learning from single-cell transcriptomics data. Our results indicate substantial improvement in predictive performance (up to 0.16 gain in AUC) when testing logistic regression models on 8 cell type specific datasets across two independent cohorts.

1. Introduction

Single-cell transcriptomics is currently the best high-throughput tool to profile the state of individual cells (Svensson et al., 2018; Aldridge & Teichmann, 2020). Downstream analyses of single-cell data mainly rely on the the gene-cell count matrices that provide a summary of the experiment (Luecken & Theis, 2019). However, the number of cells captured per different biological replicates (e.g., mice), cell types (e.g., T cells), or conditions (e.g., control-disease groups) can greatly vary within these, introducing replicate specific biases (Squair et al., 2021) in the embedding space of cells, the single-cell landscape, as illustrated in Fig. 1. This imbalanced nature of the data, which is often combined with relatively low replicate numbers, makes conducting downstream analyses, such as differential gene expression (DGE), or predictive modeling difficult and unreliable (Lähnemann et al., 2020). DGE, one if not the most significant downstream task of single-cell data science is often contaminated by false discoveries due to replicate specific transcriptomic signatures (Squair et al., 2021). Additionally, the imbalanced representation of replicates across the single-cell landscape prevents the general application of predictive machine learning (ML) methods on single-cell datasets. However, the broad usage of supervised learning on massive single-cell datasets could have a tremendous impact on several predictive applications, e.g. tasks associating transcriptomics data with patient sample attributes (Phongpreecha et al., 2020).

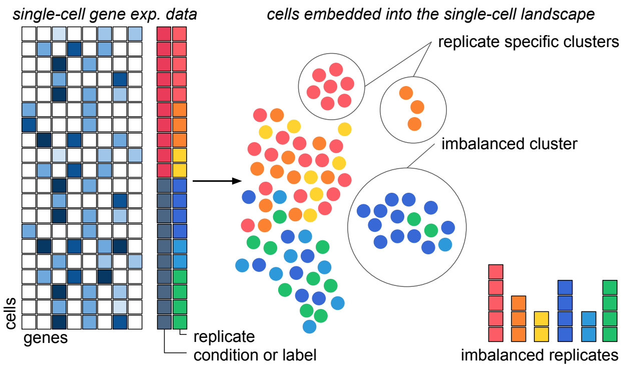
single-cell gene exp. data
cells embedded into the single-cell landscape
replicate specific clusters
imbalanced cluster
cells
genes
replicate
condition or label
imbalanced replicates

Figure 1. Single-cell landscape constructed from the gene-cell count matrix of an experiment with multiple biological replicates and two conditions. Replicate specific clusters and differences between the number of cells captured per replicates can contaminate downstream analyses.

The recently proposed batch correction and data integration (Korsunsky et al., 2019; Hie et al., 2019; Stuart et al., 2019) methods show promising results to confront technical biases within single-cell datasets. However, these cannot correct for the variability in the number of cells across the

[1] Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA [2] Paul F. Glenn Center for the Biology of Aging, Stanford University School of Medicine, Stanford, CA, USA [3] Wu Tsai Neurosciences Institute, Stanford University School of Medicine, Stanford, CA, USA [4] Department of Computer Science, University of California Los Angeles, Los Angeles, CA. Correspondence to: Robert Palovics <palovics@stanford.edu>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

single-cell landscape and most batch correction methods alter the original gene expression values. Another popular approach to correct for imbalances in the number of cells per replicate is downsampling, when an equal number of cells are sampled randomly across each replicate. Although this is a cell level approach, noise introduced by random sampling limits its applicability, and it may also miss DGE signals that are associated with meaningful cell subpopulations within the data. The 'pseudo-bulk' method (Squair et al., 2021) intends to prevent false discoveries with differential gene expression by summing single-cell profiles within cell populations. The pseudo-bulk approach has been recently shown to outperform established single-cell DGE methods (Squair et al., 2021). On the other hand, the construction of pseudo-bulk samples essentially masks and hence cancels cellular level information, dramatically shrinks sample sizes and prevent the discovery of any potentially relevant cellular subpopulations.

In our work, we intend to overcome the noted pitfalls of single-cell data analysis with a new computational approach. We suggest purifying the data by selecting a "representative" subset (coreset) of cells from areas of the single-cell landscape where multiple replicates are represented. Many natural notions of representativeness satisfy submodularity, an intuitive diminishing returns property: selecting new data points help less if similar data points have already been selected. Such problems can often be reduced to maximizing a submodular set function. Submodular maximization has recently achieved great success in various machine learning and data mining applications, including exemplar-based clustering, document and corpus summarization, recommender systems, etc. (El-Arini & Guestrin, 2011; Dasgupta et al., 2013; Mirzasoleiman et al., 2013).

We perform purification in two steps. First we extract cells from areas where multiple replicates are present and then we select a representative coreset of these filtered cells. Specifically, we use the well-known exemplar clustering submodular function that selects the set of most centrally located elements in the data. The selected coreset hence (1) protects downstream results from biases originating from replicate specific areas since these are foremost excluded, (2) represents *all* remaining regions of the single-cell landscape, including any interesting subpopulations, and simultaneously (3) ensures that regions are not overrepresented (4) selects centrally located cells and avoids potential outliers.

2. Method

First we perform standard prepossessing steps (e.g., quality control) of single-cell data analysis (Luecken & Theis, 2019) and then we use the log-CPM normalized raw gene-cell count matrix as the input for purification. We calculate the first M principal components (PCs) and use these to embed the cells to a lower dimensional space ("single-cell landscape"). Alternatively, one may perform additional preprocessing steps, e.g., it is possible to input batch corrected data for coreset based purification, or use cell embeddings that are different from the PCs.

Discarding replicate specific areas We identify for each cell c the k-nearest neighbors of the cell (n_c) in the PC space based on Euclidean distance. We calculate the number of those that belong to the same replicate as c: $m_c := |\{l_c = l_d, d \in n_c\}|$ where l_c is the replicate of cell c and l_d is the replicate of cell d. Cells with $0 < m_c < k$ are included for coreset selection.

Coreset selection We intend to select a set of cells, that best represent the cells we included. One approach for finding such cells is solving the exemplar clustering problem (Kaufman & Rousseeuw, 2009; Mirzasoleiman et al., 2016), that aims to minimize the sum of pairwise dissimilarities between the selected cells and the rest of the elements of the original dataset. Based on the PCs, we use Gaussian kernel to define the similarity of cells c and d as

$$s_{cd} = e^{-\frac{|p_c - p_d|^2}{2\sigma^2}}$$

where p is the embedding vector of a cell based on the PCs, and σ is the standard deviation of the PC matrix.

The set of $r|V|$ exemplar cells from the groundset V can be found as

$$S^* \in \arg\max_{\substack{S \subseteq V \\ |S| \leq r|V|}} \sum_{c \in V} \max_{d \in S} s_{cd}. \tag{1}$$

The maximization problem (1) is NP-hard. However, the above objective function is monotone and submodular, and hence a near optimal solution can be found efficiently. A set function $F : 2^V \to \mathbb{R}^+$ is submodular if $F(e|S) = F(S \cup \{e\}) - F(S) \geq F(T \cup \{e\}) - F(T)$, for any $S \subseteq T \subseteq V$ and $e \in V \setminus T$. F is *monotone* if $F(e|S) \geq 0$ for any $e \in V \setminus S$ and $S \subseteq V$. For maximizing a monotone submodular function, the greedy algorithm provides a $(1 - 1/e)$ approximation guarantee (Wolsey, 1982). The greedy algorithm starts with the empty set $S_0 = \emptyset$, and at each iteration t, it chooses an element $e \in V$ that maximizes the marginal utility $F(e|S_t) = F(S_t \cup \{e\}) - F(S_t)$. Formally, $S_t = S_{t-1} \cup \{\arg\max_{e \in V} F(e|S_{t-1})\}$. The computational complexity of the greedy algorithm is $\mathcal{O}(|V|^2 \cdot r)$. However, its complexity can be reduced to $\mathcal{O}(|V|)$ using stochastic methods (Mirzasoleiman et al., 2015), and can be further improved e.g., using lazy evaluation (Minoux, 1978) and distributed implementations (Mirzasoleiman et al., 2013).

We identify exemplars for each condition C (e.g., control and treatment) separately in order to ensure that representative cells are extracted from all conditions. We construct groundsets one-by-one for each condition C from the cells that belong to C and apply the greedy implementation based exemplar clustering to select r_C fraction of those cells.

3. Experiments and results

We validate the proposed method in two experimental settings. First, we investigate if purification protects downstream DGE results against replicate specific biases, by performing a synthetic data augmentation based experiment. Second, we examine if purification can improve supervised learning and can result in more generalizable models. In particular, we base our experiments on two independent single-cell datasets on aging and investigate how purification impacts the accuracy of single-cell age prediction.

3.1. Datasets

The datasets used throughout the experiments are ageing focused single-cell atlases and include cells from young and aged mice. We use the annotated, log-CPM normalized count data from both experiments. Whenever PC calculation is required, we select the first $M := 20$ number of PCs.

Tabula Muris Senis (TMS). We use the SmartSeq-2 based data on male young (3-month-old) and aged (18/24-month-old) mice of TMS (Consortium et al., 2020). The data contains in total 52,553 cells across a large number of cell types from 20 tissues. We select cell types that have at least 2 replicates with a minimum of 20 cells both in the control (young) and treatment (aged) groups for the experiments of Section 3.2. This selection criteria results in 24 cell types in total (see Table 2 of the Appendix).

Murine aging cell atlas (Calico). The dataset of Kimmel et al. (Kimmel et al., 2019) includes droplet based data in total on 60,092 cells from 3 tissues (kidney, lung, spleen) of young (7/8-months-old) and aged (22/23-months-old) mice. We base our classification based validation in Section 3.3 on the 8 cell types that are present both in TMS and Calico.

3.2. Differential gene expression based evaluation

First we assess if the proposed method is protective against replicate specific biases contaminating DGE results. In order to ensure that the data we use is imbalanced and biased, we construct a synthetic, data augmentation based analysis. We first identify a few replicate specific outlier cells, and augment the data based on these. The overrepresented outliers introduce a controlled replicate specific bias in the data. Next, we purify the *augmented data* and perform DGE on the original data (D_0), the augmented data (D_a), and the purified dataset (D_p) respectively. We expect more similar results between D_p and D_0 than between D_a and D_0. For each cell type, we apply the following six steps (see Fig. 4 of the Appendix): (1) First, we calculate the ($M := 20$) PCs based on the log-CPM normalized expression profiles of the cells. (2) To identify a potential outlier group within the dataset, we calculate the distance of every cell from its closest neighbor belonging to a *different replicate* within the PCA space. (3) We select the cell c with highest such distance as well as any neighboring cells d from the same replicate ($l_c = l_d, c \in n_c$). (4) Next, we use SMOTE (Chawla et al., 2002) to augment the data based on the selected "outlier" cells and hence introduce replicate specific bias in the data. (5) We conduct Wilcoxon-Mann-Whitney based DGE on the original data (D_0), the augmented data (D_a) and the purified data (D_p). (6) Finally, we calculate Spearman correlation (S) based on the $-log_{10}$(p-values) obtained with DGE. We compute and compare $S(D_0, D_a)$ and ($S(D_0, D_p)$).

In case of each cell type selected from TMS (Consortium et al., 2020) data we increase the original dataset in size ($1.05X - 1.5X$) and compare $S(D_0, D_a)$ and $S(D_0, D_p)$. We set the number of PCs to $M := 20$, the number of neighbors to $k := 10$ across all cell types and select $r = 0.9$ fraction of the cells when performing coreset selection. Our results are summarized in Figure 2 where each point represents a cell type and the diagonal line indicates equal performance on the augmented and purified datasets. Our results indicate that purification indeed leads to higher similarities (each point is above the diagonal), even when the least, 5% outliers are added to the original dataset. In contrast, DGE performed on the augmented dataset becomes dissimilar to the DGE performed on the original data.

3.3. Classification based evaluation

Next we investigate if purification can enhance the generalizability of classification models. We use 8 cell-type specific datasets on aging that can be found in two independent cohorts (TMS and Calico). We train logistic regression classifiers with L_1 regularization ($\alpha : 0.02 - 2$) to estimate for each cell the binary age label of its source organism (young vs. aged). (1) We train our models on the Calico data which has significantly more cells than TMS. To optimize the hyperparameters with cross-validation, we set aside cells from one young and one aged replicate in each round for validation set and use the rest of the cells for training. Our aim here is to construct a realistic training scenario where the validation set is independent from the training set. (2) We repeat the same training procedure but purify the training set before model fitting. (3) We test the predictive performance of the models on the independent cohort of TMS and measure the Area Under the ROC curve (AUC). (4) As an additional baseline we also measure the performance of a similarly optimized classifier that uses data that contains equal number of cells from each replicate.

Our results are summarized in Table 1. Purification im-

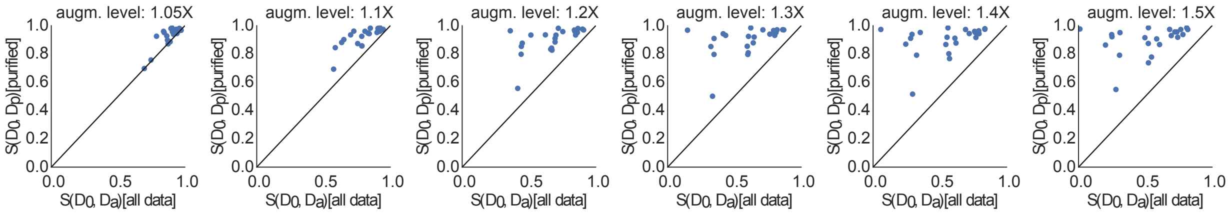

Figure 2. Spearman correlation of cell type specific DGE results, each point represents a cell type in TMS. Correlation between the augmented and the original data based DGE is indicated on the x-axis, and correlation between the purified and the original data is shown on the y-axis. Diagonal lines represent equal similarities, points above the diagonal indicate improvement with purification. From left to right, results at increasing augmentation levels are shown (1.05,1.1X, 1.2X, 1.3X, 1.4X, 1.5X).

Table 1. Cell-type specific classification results on aged vs. young cells. Classifiers are trained on the Calico dataset (Kimmel et al., 2019), evaluation is based on the cohort of TMS (Consortium et al., 2020). Columns from left to right: hyperparameters used indicating the number of PCs M, number of neighbors k, fraction of the young cells selected r_y, fraction of aged cells selected r_a, regularization coef. of the baseline model α_{ori}, regularization coef. of the purified model α_{pure}; AUC measured on TMS in case of models trained on the original data AUC_{ori}, the purified data AUC_{pure} and the balanced subsampled data AUC_{bal}; difference of AUC_{pure} and AUC_{all}.

ORGAN, CELL TYPE	M	k	r_y	r_a	α_{ORI}	α_{PURE}	AUC_{ORI}	AUC_{PURE}	AUC_{BAL}	$AUC_{\text{PURE}} - AUC_{\text{ORI}}$
KIDNEY, ENDO. CELL	20	20	0.4	0.2	1.00	0.06	0.661	**0.767**	0.726	0.106
KIDNEY, EPITH. CELL	20	10	1.0	0.2	1.00	0.80	0.564	**0.638**	0.522	0.074
KIDNEY, MONOCYTE	20	20	0.4	0.8	0.40	0.20	0.526	**0.680**	0.388	0.154
LUNG, ENDO. CELL	20	20	0.8	1.0	0.10	0.08	0.841	**0.885**	0.820	0.044
LUNG, MONOCYTE	20	15	0.8	1.0	0.60	0.80	0.822	**0.826**	0.810	0.004
LUNG, T CELL	20	10	0.6	0.2	0.10	0.01	0.801	**0.902**	0.769	0.101
SPLEEN, B CELL	20	15	1.0	1.0	0.08	0.04	0.841	**0.862**	0.816	0.021
SPLEEN, T CELL	20	20	1.0	0.8	0.02	0.04	0.886	**0.906**	0.883	0.02

proves the predictive performance in each cell type, especially in case of cell types where training on the whole original datasets results in fairly low performance, i.e., in cell types from the kidney. Substantial improvements can be observed for cell types where training on the whole data results in good performance as well. For example, there is an additional 0.1 gain in AUC in case of the lung T cells (purification is visualized in Fig. 3).

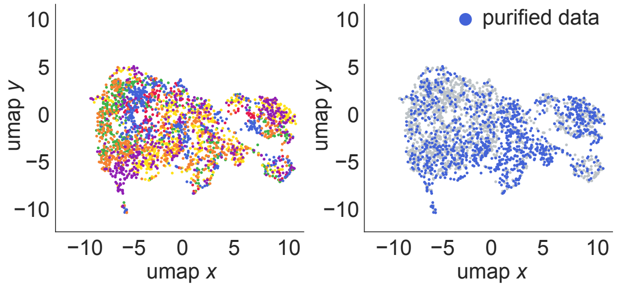

Figure 3. UMAP (McInnes et al., 2018) visualization of lung T cells of the Murine aging cell atlas (Kimmel et al., 2019). **left**: cells colored by biological replicate **right**: purified data is highlighted in blue, the rest of the cells are colored in grey.

4. Discussion and Conclusion

Here we introduced a coreset selection based method to purify single-cell datasets. We have shown that the suggested algorithm can aid downstream analyses, in particular differential gene expression, as it is protective against replicate specific biases. Additionally, we found that it leads to single-cell age classifier models that have substantially higher performance when validated on an independent cohort without the need of any integration between the two datasets. Crucially, coreset selection does not alter the gene-cell count matrix in any way. Consequently, it is possible to apply the proposed computational tool together with any downstream application including differential gene expression, trajectory analysis, cell annotation, or to couple it with any data correction method. In our future work we intend to investigate the hyperparameter fine tuning of purification.

Although here we focused on purification, coreset selection has the potential of summarizing massive single-cell datasets. The ever increasing number of cells captured in transcriptomic experiments often makes it difficult to reuse these data in secondary or follow-up analyses by different research groups. Coreset selection provides an opportunity to shrink transcriptomic data by representative summaries that include the same information as the original count matrices, but have less noise and are much easier to handle in any potential future analyses. We hope that coreset selection based purification will emerge as a best practice that extends the current single-cell analysis pipelines.

Acknowledgements

We thank our reviewers for their time and comments and the members of the laboratory of T.W.-C. for their feedback. We also gratefully acknowledge the support of Michael J. Fox Foundation (MJFF-021418).

References

Aldridge, S. and Teichmann, S. A. Single cell transcriptomics comes of age. *Nature Communications*, 11(1):1–4, 2020.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Consortium, T. M. et al. A single cell transcriptomic atlas characterizes aging tissues in the mouse. *Nature*, 583 (7817):590, 2020.

Dasgupta, A., Kumar, R., and Ravi, S. Summarization through submodularity and dispersion. 2013.

El-Arini, K. and Guestrin, C. Beyond keyword search: discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 439–447, 2011.

Hie, B., Bryson, B., and Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691, 2019.

Kaufman, L. and Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

Kimmel, J. C., Penland, L., Rubinstein, N. D., Hendrickson, D. G., Kelley, D. R., and Rosenthal, A. Z. Murine single-cell rna-seq reveals cell-identity-and tissue-specific trajectories of aging. *Genome research*, 29(12):2088–2103, 2019.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12): 1289–1296, 2019.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Luecken, M. D. and Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Minoux, M. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques*, pp. 234–243. Springer, 1978.

Mirzasoleiman, B., Karbasi, A., Sarkar, R., and Krause, A. Distributed submodular maximization: Identifying representative elements in massive data. In *Advances in Neural Information Processing Systems*, pp. 2049–2057, 2013.

Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., and Krause, A. Lazier than lazy greedy. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Mirzasoleiman, B., Karbasi, A., Sarkar, R., and Krause, A. Distributed submodular maximization. *The Journal of Machine Learning Research*, 17(1):8330–8373, 2016.

Phongpreecha, T., Fernandez, R., Mrdjen, D., Culos, A., Gajera, C. R., Wawro, A. M., Stanley, N., Gaudilliere, B., Poston, K. L., Aghaeepour, N., et al. Single-cell peripheral immunoprofiling of alzheimer's and parkinson's diseases. *Science advances*, 6(48):eabd5575, 2020.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. Confronting false discoveries in single-cell differential expression. *Nature communications*, 12(1):1–15, 2021.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

Wolsey, L. A. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4): 385–393, 1982.

A. Concept of the synthetic data augmentation based experiment and summary of data sets used

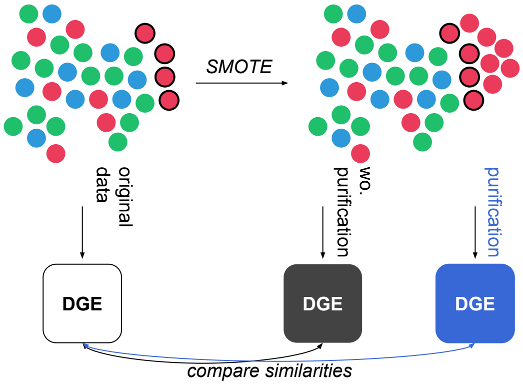

Figure 4. Concept of the synthetic data augmentation based experiment. DGE results based on the augmented data, with and without purification, are compared to the DGE results calculated from the original dataset.

Table 2. Summary of the datasets used. From left to right, for each cell type we list the number of cells, number of young cells, number of aged cells, number of replicates, number of young replicates, and the number of aged replicates. Rows highlighted in grey are only used for testing classifiers in Section 3.3. Abbreviations: CSC: crypt stem cell, endo. cell: endothelial cell, epid. cell: epidermal cell, epith. cell: epithelial cell, HSC: hematopoietic stem cell, macro.: macrophages, oligo.: oligodendrocytes, SMC: smooth muscle cell.

organ, cell type	# cells	# young cells	# aged cells	# repl.	# young repl.	# aged repl.
Tabula Muris Senis (TMS.)						
brain, endo. cell	819	300	519	6	3	3
brain, microglia	8,328	2,154	6,174	9	3	6
brain, oligo.	1,254	962	292	7	4	3
fat GAT, macro.	342	166	176	8	3	5
fat GAT, stromal cell	860	308	552	8	3	5
fat MAT, B cell	225	66	159	7	2	5
fat MAT, stromal cell	817	256	561	7	2	5
fat SCAT, macro.	414	134	280	8	3	5
fat SCAT, stromal cell	779	279	500	8	3	5
heart, fibroblast	1,482	529	953	10	4	6
heart, monocyte	716	214	502	8	3	5
intestine, CSC	580	100	480	7	2	5
intestine, secretory cell	964	186	778	8	2	6
kidney, monocyte	78	17	61	4	1	3
kidney, endo. cell	159	41	118	7	2	5
kidney, epi. cell	241	50	191	9	3	6
liver, hepatocyte	774	193	581	6	2	4
lung, endo. cell	155	45	110	7	2	5
lung, monocyte	263	88	175	9	3	6
lung, T-cell	117	12	105	5	1	4
marrow, granulocyte	2,488	534	1,954	8	2	6
marrow, HSC	1,997	943	1,054	8	3	5
pancreas, beta cell	743	226	517	5	2	3
pancreas, acinar cell	236	53	183	5	2	3
pancreas, alpha cell	298	182	116	5	2	3
skin, basal cell	495	204	291	7	3	4
skin, keratinocyte	915	447	468	8	3	5
spleen, B cell	1,550	352	1,198	8	2	6
spleen, T cell	351	88	263	8	2	6
thymus, thymocyte	1,090	418	672	8	3	5
Murine aging cell atlas (Calico)						
kidney, endo. cell	2,243	1,661	582	7	4	3
kidney, epith. cell	1,175	777	398	7	4	3
kidney, monocyte	571	256	315	7	4	3
lung, endo. cell	5,644	3,267	2,377	6	3	3
lung, monocyte	3,053	1,010	2,043	6	3	3
lung, T cell	2,541	1,107	1,434	6	3	3
spleen, B cell	17,829	8,669	9,160	7	4	3
spleen, T cell	7,138	3,523	3,615	7	4	3