Super-cells untangle large and complex single-cell transcriptome networks

Mariia Bilous[1,2,6], Loc Tran[1,2,6], Chiara Cianciaruso[3], Santiago J. Carmona[1,2], Mikael J. Pittet[3,4,5], David Gfeller[1,2,*]

[1]Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland,
[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[3]Department of Pathology and Immunology, University of Geneva, Geneva, Switzerland
[4]Department of Oncology, Geneva University Hospitals, Geneva, Switzerland
[5]Center for Systems Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA
[6]These authors contributed equally to this work

***Corresponding author:** David.Gfeller@unil.ch

Keywords: single-cell transcriptomics, computational biology, coarse-graining

Abstract

Single-cell RNA sequencing (scRNA-seq) technologies offer unique opportunities for exploring heterogeneous cell populations. However, in-depth single-cell transcriptomic characterization of complex tissues often requires profiling tens to hundreds of thousands of cells. Such large numbers of cells represent an important hurdle for downstream analyses, interpretation and visualization. Here we develop a network-based coarse-graining framework where highly similar cells are merged into super-cells. We demonstrate that super-cells not only preserve but often improve the results of downstream analyses including visualization, clustering, differential expression, cell type annotation, gene correlation, imputation, RNA velocity and data integration. By capitalizing on the redundancy inherent to scRNA-seq data, super-cells significantly facilitate and accelerate the construction and interpretation of single-cell atlases, as demonstrated by the integration of 1.46 million cells from COVID-19 patients in less than two hours on a standard desktop.

Introduction

Single-cell RNA sequencing (scRNA-seq) provides transcriptome-wide gene expression profiles at single-cell resolution. This technology has been transformative for unsupervised investigation of heterogeneous cell populations[1,2], identification of novel cell states and cell types[3], discovery of novel markers[4] and reconstruction of developmental trajectories[5,6].

The rapid developments of single-cell capture and sequencing technologies enable researchers to profile tens to hundreds of thousands of single cells[7–10]. This high throughput provides unprecedented opportunities to explore the heterogeneity of complex tissues[7,11]. For instance, even very rare cell types can be detected by unsupervised clustering[12–14]. The relationship between different clusters can then be investigated by different approaches (e.g., PAGA[15] or TooManyCells[16]), providing a high-level abstraction of the scRNA-seq data. However, the large number of cells that need to be profiled from complex biological samples in order to capture low-frequency populations is an important hurdle for downstream analyses.

Attempts to streamline the analysis of large scRNA-seq data have mainly focused on improving bioinformatics pipelines to scale with more cells[17–20] and adapt them to computational infrastructures that can cope with large memory requirements[21,22]. These developments often require the use of dedicated platforms, which are less user-friendly and not necessarily available to researchers without in-depth training in bioinformatics. As an alternative, cell subsampling approaches have been developed[23,24]. However, these approaches do not capitalize on the full information present in the initial scRNA-seq data, which is likely sub-optimal, for instance to reduce the noise inherent to scRNA-seq data due to dropout[25].

High-throughput single-cell transcriptomic profiling of biological samples typically leads to the repetitive sampling of highly similar, and possibly biologically redundant cells. Some attempts to simplify scRNA-seq data by merging such highly similar cells have been proposed (e.g., MetaCell)[26,27]. MetaCell has been mainly used for visualization and exploratory purposes[28–30]. However, it remains unclear whether the approach can be used for quantitative and robust downstream analyses, and how much biological information may be gained or lost when analyzing such simplified data[31].

Here, we developed a network-based coarse-graining framework to simplify scRNA-seq data by merging highly similar cells into “super-cells”. We demonstrate that super-cells (i) preserve the global structure of the initial data, (ii) enable efficient and robust downstream analyses, as demonstrated by the identification and validation of genes specifically expressed in tumor-infiltrating dendritic cell subtypes, (iii) significantly reduce noise from single-cell gene expression measurements and (iv) lead to ten- to hundred-fold reduction of the size of the data and the computational time and memory requirements. Overall, our results provide a robust framework for scRNA-seq data analysis at an intermediate and tunable level of resolution between single cells and clusters, and will facilitate the exploration of large scRNA-seq datasets, as demonstrated by the analysis of 1.46 million cells from COVID-19 patients.

Results

Simplifying scRNA-seq data with super-cells

To facilitate the analysis of scRNA-seq data, we developed a computational coarse-graining framework based on the idea of grouping highly similar cells into super-cells (Fig. 1a). First, scRNA-seq data are modeled as a single-cell graph with nodes representing cells and edges connecting cells with high transcriptomic similarity[15,32] (see Methods). Next, super-cells are built by merging single cells with very high internal connectivity. We used the walktrap algorithm[33], which allows users to predefine the number of super-cells. Unlike standard clustering, our aim is not to identify large cell populations that can be mapped to distinct biological cell types, but rather to merge cells that contain highly similar and likely repetitive transcriptomic information. The graining level (γ) is defined as the ratio between the number of cells and the number of super-cells. In the super-cell network, the size of a node is proportional to the number of single cells that have been merged and the weight of an edge connecting two super-cells is proportional to the total number of edges connecting cells from those two super-cells. Finally, a super-cell gene expression matrix is computed by averaging gene expression within super-cells (Fig. 1a).

To explore and benchmark this coarse-graining approach, we first applied it to three scRNA-seq datasets and compared the results of the analysis at the single-cell and super-cell levels. The first dataset is a gold standard dataset that consists of five

human adenocarcinoma cell lines (*cell_lines,* $N = 3'918$)[34] (Fig. 1b). The second one is a dataset of purified T cells from healthy donors (*Tcells,* $N = 40'560$)[10] (Fig. 1c). The third one is a dataset of tumor-infiltrating CD8 T lymphocytes (*TILs,* $N = 3'574$)[35] (Fig. 1d). For the *cell_lines* dataset, we used the cell line information to annotate single cells. For the *Tcells* and *TILs* datasets, we annotated single cells based on clustering since these cases represent more realistic data where the ground truth cell type annotation is not known a priori. In the *Tcells*, the three clusters could be mapped to CD4, naïve CD8 and cytotoxic CD8 T cells. In the *TILs,* the three clusters correspond to naïve, effector memory-like, and exhausted/progenitor-exhausted CD8 T cells. As shown in Fig. 1b-d, the super-cells at different graining levels preserve the global structure of the initial data.

Super-cells preserve clustering and differential expression analysis results

To assess whether super-cells can be used for downstream analyses, we first checked whether super-cells contain cells originating from the same cell type. To this end, we used the purity, defined as the proportion of the most abundant cell type in a super-cell. Fig. 2a shows that super-cells have purity close to 1, indicating that they consist mainly of cells of the same cell type. This high purity allowed us to annotate the super-cells according to the most abundant cell type in each super-cell and use the same coloring scheme in the visualization (Fig. 1b-d). As a negative control, random grouping of cells would result in much lower purity (Fig. 2a).

Beyond visualization, an important step in scRNA-seq data analysis is to identify distinct cell types or cell states by clustering. We used a hierarchical clustering algorithm and compared the results of clustering of super-cells to the known ground truth for *cell_lines* or to the results of single-cell clustering for *Tcells* and *TILs,* using the adjusted Rand index (ARI) (see Methods). Fig. 2b shows that the consistency of the clustering of super-cells is well preserved over different graining levels. As expected, random grouping of cells resulted in much lower ARI values. Results obtained with super-cells also showed significantly higher ARI values compared to subsampling at higher graining levels. To further investigate the consistency between the clustering at the single-cell and super-cell levels, we computed the conservation of clusters obtained with other clustering algorithms applied to the single-cell data. We observed that ARI values for clustering of the super-cells are within the same range

as ARI values for the clustering of single cells using different clustering algorithms (Fig. 2b, blue vertical line). These results indicate that super-cells can be used for clustering, and the differences with the clustering at the single-cell level are within the expected fluctuations observed with different choices of clustering algorithms.

Another important step of scRNA-seq data analysis is to identify differentially expressed genes between clusters. To explore the performance of differential expression (DE) analysis at the super-cell level, we assessed the recovery rate of differentially expressed genes found at the single-cell level using the true positive rate (TPR) (see Methods). The analysis showed that more than 75% of the differentially expressed genes observed at the single-cell level can be recovered even at relatively high graining levels (Fig. 2c), in contrast to subsampling or random grouping.

We then explored whether the original number of clusters, defined as the one that maximizes the silhouette coefficient[36], can be recovered in super-cells. Fig. 2d demonstrates that this number could be recapitulated until a graining level of at least 50 and this conservation compares favorably with subsampling and random grouping approaches. Additionally, even in cases where the predicted optimal number of clusters differs between single cells and super-cells, the clustering results in super-cells are still consistent with the reference annotation of single cells (Supplementary Fig. 1a).

As super-cells contain different numbers of single cells, they might benefit from sample-weighted algorithms. To explore the impact of the weights of super-cells, we compared the performance of sample-weighted clustering and sample-weighted DE with their unweighted versions (see Methods). Overall, Supplementary Fig. 1b-d show that, at moderate graining levels (e.g., $\gamma = 50$), both sample-weighted and unweighted algorithms perform similarly on super-cells while at higher graining levels, sample-weighted algorithms tend to display higher accuracy.

Compared to MetaCell[26], our coarse-graining approach has better performance in terms of purity, clustering consistency, identification of differentially expressed genes, and supports a broader range of graining levels γ (Supplementary Fig. 2).

Super-cells reveal genes specifically expressed in dendritic cell subtypes

To further illustrate the use of super-cells for the identification of genes expressed in specific cell types, we examined scRNA-seq data of tumor-infiltrating myeloid cells

from murine KP1.9 lung adenocarcinoma (*TIMs*, $N = 15'939$)[4]. Single-cell data were coarse-grained into super-cells (Fig. 2e, see Methods). We then performed DE analysis between conventional (cDCs) and plasmacytoid (pDCs) dendritic cells, which are known to play important roles in eliciting the immune response against cancer, at both single-cell and super-cell levels. Several genes displayed clearer DE patterns in super-cells (i.e., better ranking), including *H2-Aa* which, along with other major histocompatibility complex (MHC) class II molecules, is upregulated in cDCs[37,38]. To experimentally validate these predictions, we selected additional genes coding for trans-membrane proteins with available antibodies and that were ranked better in the super-cell DE analysis (see Methods). These included *Cd74* (upregulated in cDCs) as well as *Ly6e*, *Cd47* and *Cd44* (upregulated in pDCs) (Fig. 2f). We then performed a flow cytometry analysis of DCs of the same murine lung cancer model (Fig. 2g and Supplementary Fig. 3, see Methods). The results confirmed the DE of all the selected genes at the protein level (Fig. 2h), demonstrating that super-cells are useful for magnifying biological information in scRNA-seq data that is less detectable at the single-cell level.

Super-cells improve cell type annotation, gene correlation and imputation

We next compared marker-based cell type annotation at single-cell and super-cell levels. For this, we first annotated CD4 and CD8 T cells from the *Tcells* dataset using either single markers (Fig. 3a) or gene signatures derived from bulk RNA-seq data (Fig. 3b, see Methods) and compared this annotation with the cell type determined by protein expression during the sorting procedure (see Methods). The annotation quality, computed as the area under the ROC curve (AUC), grows with the graining level until it almost reaches saturation (AUC=1) for both CD4 and CD8 T cells. Instead, subsampling failed to improve cell type annotation. Similar results were obtained when applying the super-cell annotation approach to the *TILs* dataset (Supplementary Fig. 4). Inaccuracies in annotating cells using single markers or gene signatures are likely due to the high dropout rate[25]. In the *Tcells* dataset, less than 40% of CD4, respectively CD8 T cells, express the *CD4*, respectively *CD8A*, marker genes (Fig. 3c). By contrast, at the graining level $\gamma = 200$, 79% of CD4 super-cells and 96% of CD8 super-cells express *CD4* and *CD8A* markers, respectively.

We next investigated gene correlation, which is an important source of information to infer gene co-expression modules[39]. We first used the *TILs* dataset and explored

genes that are known to be positively correlated (e.g., *Ibgb7*, *Cxcr3*, *Ly6c*, and *Gzmk* which are markers of effector memory-like cells) or negatively correlated (e.g., *Tcf7* that is expressed in naïve and stem-like cells, versus *Pdcd1* and *Havcr2/Tim-3* that are expressed in exhausted cells). Fig. 3d shows that super-cells increase expected gene correlations and remove the noise arising from single-gene dropouts in scRNA-seq. To more systematically explore the biological relevance of correlated genes, we compared the biological relatedness of top correlated gene pairs found exclusively at the single-cell or super-cell levels within individual cell lines in the *cell_lines* dataset. As a measure of similarity between a pair of genes, we used a Gene Ontology (GO)[40] match score (see Methods). Fig. 3e shows higher values of the average GO match score of the top correlated gene pairs in the super-cells compared to the single cells until a graining level of roughly 50. These results show that pairs of correlated genes found in super-cells show higher biological relatedness compared to those found in singe cells.

Imputation methods were shown to improve signal-to-noise ratio in scRNA-seq data and lead to better correlation with bulk profiles[41]. To investigated whether super-cells can be used as an input for imputation approaches, we applied MAGIC[39] to both the single cells and the super-cells from the *cell_lines* dataset. The results show that the imputed gene expression profiles are more similar to the bulk profile when applying MAGIC on super-cells (Fig. 3f). This indicates that imputation can be applied to super-cells and leads to improved correspondence with bulk data.

Super-cells are compatible with RNA velocity

We next investigated whether super-cells can be used to study differentiation processes with RNA velocity[5]. We first considered a dataset of mouse hippocampus cells (*brain_cells,* $N = 3'396$)[42] and applied the velocyto algorithm[5] to both single cells and super-cells (see Methods). We observed consistent RNA velocity results at the super-cell (right) and the single-cell (left) levels (Fig. 3g) when plotted on a joint t-distributed stochastic neighbor embedding (tSNE). Both cases show the developmental path from neurogenic cells (light blue) to neurons (dark blue). This consistency is further confirmed with a high purity of super-cells in terms of velocity across multiple graining levels (see Methods) (Fig. 3h). To compute RNA velocity, a key parameter is the estimated equilibrium slope of each gene (i.e., linear fit between

spliced and un-spliced mRNA, referred to as "γ" in the original RNA velocity publication[5]). The number of genes for which this equilibrium slope can be estimated increases with higher graining levels for super-cells, while it stays constant for random grouping, and decreases for subsampling (Fig. 3i). Moreover, the equilibrium slopes of genes in super-cells correlate with those in single cells, and the correlation is higher than for the subsampling (Fig. 3j). The improvements of super-cells over subsampling are likely due to the regularizing and enriching effects of super-cells on spliced and un-spliced mRNA abundance (see examples in Supplementary Fig. 5a). To directly compare RNA velocity at the single-cell and super-cell levels in the tSNE plots of Fig. 3g, we computed the cosine similarity between 2D RNA velocity of each single cell and 2D RNA velocity of the super-cell it belongs to (Fig. 3k). This similarity score also suggests a high consistency of super-cell RNA velocity across multiple graining levels. In the joint tSNE used in Fig. 3g to facilitate comparison, subsampling displayed similar conservation of the 2D RNA velocity (Fig. 3k). However, when plotted separately, the differentiation process is much more visible in super-cells than after subsampling (Supplementary Fig. 5b). The same analyses were performed on a mouse pancreas scRNA-seq dataset (*pancreatic_cells,* $N = 3'696$)[43] and similar results were obtained (Supplementary Fig. 6). This demonstrates that super-cells are compatible with RNA velocity.

Super-cells facilitate data integration

ScRNA-seq atlases are built by integrating data from multiple samples. To explore the use of super-cells in scRNA-seq data integration, we analyzed a recently published dataset of 1.46 million immune cells coming from various tissues from 196 COVID-19 patients and healthy controls (*COVID-19_atlas,* $N = 1'462'702$)[44]. Analyzing these data is challenging since most existing data integration algorithms do not scale with such cell numbers on standard computational infrastructures. We applied our coarse-graining framework on each sample separately (see Methods), which led to a total of $146'304$ super-cells (Fig. 4a). We then performed data integration on super-cells with Harmony[45] (Fig. 4b). The results of integrated super-cells showed high similarity with those obtained based on single cells in the original study. In particular, the main immune cell types reported in the original study could be recapitulated. Visually, Fig. 4b shows that protocols and samples are well mixed after integration of super-cells.

This is confirmed by the improvement of the kBET acceptance rates (a quantitative measure of batch mixing)[46] both in terms of protocols (Fig. 4c, top) and samples (Fig. 4c, bottom). The clustering of integrated data has also higher consistency with the original cell type annotation (ARI=0.75, compared to ARI=0.66 in non-integrated data). To further illustrate the use of super-cells, we performed DE analysis followed by gene set enrichment analysis within monocytes and B cells from COVID-19 patients versus healthy controls, at the super-cell level (see Methods). For monocytes, the most upregulated GO term in COVID-19 patients included 'chemokine mediated signaling pathway' (GO:0070098, adjusted p-value $= 1.3 \cdot 10^{-7}$) and 'complement activation' (GO:0006956, adjusted p-value $= 4.1 \cdot 10^{-5}$). The most upregulated GO terms in B cells from COVID-19 patients included 'inflammatory response' (GO:0006954, adjusted p-value $= 5.9 \cdot 10^{-3}$), 'type I interferon signaling' (GO:0060337, adjusted p-value $= 3.7 \cdot 10^{-2}$) or 'response to viruses' (GO:0009615, adjusted p-value $= 3.7 \cdot 10^{-2}$), consistent with a humoral response to SARS-Cov2. Importantly, all these analyses could be run on a standard desktop (Macbook Pro i7 core, 16G RAM) in less than two hours.

Super-cells significantly accelerate downstream analyses and reduce memory requirements

To assess the improvement in computational efficiency obtained with super-cells, we benchmarked the time and memory needed for the downstream analyses, including visualization, clustering, DE analysis (i.e., each major cell type versus the rest) and data integration (Fig. 4d), as well as RNA velocity (Supplementary Fig. 7) (see Methods). As expected, each method and the entire pipeline run faster on super-cells and uses less memory. All analyses could be run on a standard desktop with 16G of RAM at the super-cell level, while they crashed at $\sim 200'000$ cells for DE and $\sim 480'000$ cells for the other analyses at the single-cell level (red dots in Fig. 4d). Moreover, DE analysis could not be performed at the single-cell level for datasets with more than $600'000$ cells even on a high-performance computing (HPC) platform with 512G of RAM (black dots in Fig. 4d, extrapolations in gray). This demonstrates the advantage of super-cells for exploratory analysis of large scRNA-seq datasets.

Although super-cell construction needs to be run only once, it can be computationally demanding for large numbers of cells ($> 100'000$). To address this

issue, we included an option to perform an approximate coarse-graining, which first builds super-cells using a subset of cells and then maps the rest of the cells to the most similar super-cells (Supplementary Fig. 8a) (see Methods). Supplementary Fig. 8b,c show that the approximate coarse-graining has similar performance as the exact one. To further demonstrate the ability of the approximate coarse-graining to deal with large-scale scRNA-seq datasets, we applied it to a human pan-cancer atlas of tumor-infiltrating myeloid cells (*TIM_atlas,* $N = 108'566$)[38] (Fig. 4e). Focusing on DCs, we performed DE analysis between cDCs and pDCs for the genes tested in Fig. 2h. Our results confirm the improved DE signal in super-cells compared to single cells for *HLA-DRB1, HLA-DPA1, CD74* and *CD47* (Fig. 4f). The signal for *LY6E* and *CD44* was opposite to their expression pattern in mouse at both the single-cell and super-cell levels, suggesting that their expression in DCs subtypes may not be conserved between human and mouse.

Taking everything together, the entire analysis, including coarse-graining (both exact and approximate), dimensionality reduction, visualization, clustering, DE analysis and data integration, runs faster with super-cells than with single cells and requires less memory (Supplementary Fig. 9). These results indicate that super-cells significantly facilitate and accelerate the analysis of large scRNA-seq data.

Discussion

ScRNA-seq technologies are revolutionizing biological sciences by providing transcriptome-wide information for very large numbers of individual cells. Here we present a coarse-graining pipeline for scRNA-seq data based on super-cells. We demonstrate that super-cells preserve the structure of single-cell data at multiple graining levels and serve as a compromise structure between the single-cell level and level of clusters. The fact that super-cells are compatible with clustering, differential expression, cell type annotation, gene correlation, imputation, RNA velocity and data integration indicates that this framework can be readily applied to the vast majority of analyses performed on scRNA-seq data. Moreover, super-cells can magnify biologically relevant information, as demonstrated with the identification and validation of markers of tumor-infiltrating dendritic cell subtypes (Fig. 2e-h).

When exploring scRNA-seq data, different methods or choices of parameters are typically tested for visualization, clustering, differential expression, RNA velocity or

data integration[47,48]. Our coarse-graining framework is therefore especially appropriate for such exploratory analyses. Moreover, some analyses, like DE, are challenging to run in terms of time and memory requirements. Super-cells are powerful to simplify and accelerate these analyses.

For the construction of super-cells, we used the walktrap algorithm. Owing to its hierarchical structure, this algorithm enables users to explore different graining levels without having to recompute the super-cells for each choice of γ. This can be useful considering the heterogeneity in size and complexity of scRNA-seq datasets. The method scales as $O(N^2 \log N)$ and faster algorithms for network partition exist, such as Louvain clustering[49]. However, several of these algorithms, including MetaCell[26], do not allow users to explicitly define the graining levels. The fact that our approach leads to better preservation of downstream analyses results than MetaCell (Supplementary Fig. 2) suggests that the walktrap algorithm provides a flexible and robust way to define super-cells. Practically, we recommend using $\gamma \in [10,50]$, as this already provides a significant speed-up in the analysis of large datasets.

Super-cells are characterized by weights. We have therefore used sample-weighted versions of principal component, clustering, silhouette coefficient and DE analysis. However, we also demonstrated that unweighted algorithms for clustering and DE analysis can be used up to a reasonable graining level ($\gamma = 50$). This indicates that super-cells are compatible with a diversity of already existing methods for scRNA-seq data analysis, but further improvements are expected when explicitly considering weights at large graining levels. We therefore anticipate that our work will promote new bioinformatics developments in sample-weighted algorithms for scRNA-seq data analysis.

Beyond the coarse-graining procedure itself, our work provides a robust framework for systematic comparison of the results obtained at single-cell and super-cell resolution. This is critical to ensure that super-cells can indeed be used without losing important information from the single-cell data. As such, we believe that this framework will be useful to benchmark other coarse-graining approaches that will be or are being developed[50].

Overall, super-cells provide an optimal level of representation for scRNA-seq data between the partly redundant single-cell level and the cluster level which can mask biologically relevant intracluster heterogeneity. Our work demonstrates that super-

cells significantly reduce the size of the data and accelerate downstream analyses, while preserving the global structure of the system and the results of such downstream analyses. As the throughput of scRNA-seq technologies keeps increasing[51], we anticipate that super-cell-like approaches will play an increasingly important role in facilitating visualization, analysis, sharing and interpretation of such data.

Methods

Super-cell construction

Super-cells are built based on a log-normalized gene expression matrix. The set of features (genes) used for the construction of super-cells is defined by default as the set of the most variable genes but can also be provided by the user. Based on this set of features, a low dimensional embedding is computed using principal component analysis (PCA) (function *irlba*() from *irlba* R package[52]). The top principal components (top 10 by default in this work) are used to build a single-cell network with a k-nearest neighbors (k-NN, $k = 5$ for all studied datasets) algorithm (*nn2()* function from *RANN* R package[53]): each cell (node) is connected to k most similar cells based on the Euclidean distance. To construct super-cells, we apply the *walktrap* clustering algorithm (available as *clust_walktrap()* function from *igraph* R package[54]) to the singe-cell network. Super-cells are constructed by merging single cell (nodes) at a user-specified graining level (γ). The graining level is defined as the ratio between the number of single cells (N_c) and the number of super-cells (N_{SC}). The size of a super-cell is defined as the total number of cells it contains. The weight of an edge connecting two super-cells is computed as the total number of edges connecting cells of those super-cells. A gene expression profile of super-cells is computed by averaging gene expression within super-cells.

This coarse-graining method is implemented in an R package (SuperCell) that can be used for the all steps of the downstream analyses and is compatible with other state-of-the-art tools for scRNA-seq data analyses, including Seurat[18] and SingleCellExperiment[17].

ScRNA-seq and bulk RNA-seq datasets

Datasets analyzed in this study (Supplementary Table 1) are available in GEO under accession numbers GSE118767 (scRNA-seq of *cell_lines*), GSE116390 (scRNA-seq

of *TILs*), GSE127465 (scRNA-seq of *TIMs*), GSE95315 (scRNA-seq of *brain_cells*), GSM3852755 (scRNA-seq of *pancreatic_cells*), GSE158055 (scRNA-seq of *COVID-19_atlas*), GSE154763 (scRNA-seq of *TIM_atlas,* for the full list of integrated datasets, see Supplementary Table 2), GSE60424 (bulk RNA-seq of sorted CD4/CD8 T cells), GSE86337 (bulk RNA-seq of 5 cancer cell lines). ScRNA-seq data of purified T cells (*Tcells*) are available from 10xGenomics database.

Downstream analyses of scRNA-seq datasets

The top 1000 and 500 variable genes were used to perform the PCA for *cell_lines* and *Tcells* datasets, respectively. For *TILs*, we used the same gene set as in the original study[35]. Single-cell and super-cell data were clustered with hierarchical clustering. The ground truth cell type assignment was set to the cell line annotation for *cell_lines* and to the results of the single-cell clustering for *Tcells* and *TILs*. Based on the ground-truth cell type assignment, the reference set of differentially expressed genes was identified as a union of significantly (i.e., two-tailed t-test adjusted p-value < 0.05) upregulated (i.e., logFC > 0.5 for *cell_lines* and *TILs* or logFC > 0.25 for *Tcells*) genes in one of cell types.

Clustering consistency

To assess the consistency of the clustering of super-cells, Hubert and Arabie's adjusted Rand index (ARI) was computed. As a reference clustering, the ground truth assignment of cells was used for *cell_lines* and clustering at the single-cell level for *Tcells* and *TILs*. ARI is computed between reference clustering and clustering of super-cells using the same number of clusters.

Cluster-specific differentially expressed genes recovery rate

To assess the ability of super-cells to recover cluster-specific differentially expressed genes, the TPR was used. The reference set of differentially expressed genes (M) was computed as described above in the downstream analyses section. Then, a set of differentially expressed genes at the super-cell level ($\widetilde{M}$) was computed as the top $n = |M|$ significantly (i.e., two-tailed t-test adjusted p-value < 0.05) upregulated genes ranked by logFC. The cluster-specific differentially expressed genes recovery rate was computed as $TPR = \frac{|M \cap \widetilde{M}|}{|M|}$.

Weights of super-cells in downstream analyses

The downstream analyses of super-cell data are very similar to downstream analyses of scRNA-seq with one exception that super-cells have different sizes that have to be considered. For this, we have adapted the pipeline of scRNA-seq data analyses in a way that considers the weights of super-cells.

For the scaling of the gene expression matrix before PCA, a weighted version of a scale function was applied (*wt.scale()* function from *corpcor* R package[55]). Next, a singular value decomposition (SVD) was computed (function *irlba*() from *irlba* R package[52]) for the sample-weighted gene expression matrix $\left(X_{for_SVD} = \frac{1}{N.c}(X^T W X) = U\Sigma V^T\right)$, where X is a scaled and transposed (cells as rows) gene expression matrix at a super-cell level, W is a diagonal matrix with sizes of super-cells on its diagonal, Σ are the singular vectors of X_{for_SVD}, U and V are left- and right-singular vectors of X_{for_SVD}. Then, the principal component embedding of super-cells (X_{PCA}) was computed as $X_{PCA} = XV$.

For the sample-weighted clustering, a sample-weighted version of the hierarchical clustering (*hclust* function with the parameter *members* set to the super-cell size vector) was used.

The computation of the silhouette coefficient was also adjusted for considering the weights of super-cells. If $a(i)$ is the mean distance between a super-cell i belonging to a cluster C_m and all other super-cells from the same cluster, then $a(i) = \frac{1}{|C_m| - size(i)} \sum_{j \in C_m, j \neq i} d(i,j) \cdot size(j)$, where $d(i,j)$ is a distance between super-cells i and j, $size(i)$ is the size of super-cell i, and $|C_m|$ is the size of cluster C_m. If $b(i)$ is the minimal average distance from a super-cell i to any super-cell from a different cluster, then $b(i) = \min_{k \neq m} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \cdot size(j)$. The silhouette of a super-cell i is $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$ if $|C_m| > 1$ and $s(i) = 0$ if $|C_m| = 1$. The overall silhouette value S is a sample-weighted mean of $s(i)$, $S = \frac{1}{N.c} \sum s(i) \cdot size(i)$.

DE analysis was performed with a sample-weighted version of the two-tailed t-test (*wtd.t.test()* function from *weights* R package[56]) with logFC computed considering super-cell size.

Benchmarking with MetaCell

MetaCells for multiple graining levels were constructed using the MetaCell package[26] and the vignette instructions provided at the authors' GitHub (https://github.com/tanaylab/metacell/tree/master/vignettes). Since MetaCell does not explicitly allow users to tune the graining level, we had to vary the parameter of the minimal MetaCell size (*min_mc_size*). This parameter was set to the size of the smallest super-cell at the corresponding graining level. This explains why the range of graining level values for MetaCell is narrow in Supplementary Fig. 2. As a gene expression matrix of MetaCell data, we used the output field of MetaCell named *mc_fp* (MetaCell footprint).

Marker genes in the mouse *TIMs* dataset

Data pre-processing and cell type annotation in the *TIMs* dataset at the super-cell level was performed the same way as the single-cell analyses in the original study[4]. In brief, super-cells were build based on the first 60 principal components computed on highly variable genes (i.e., genes with the variance larger than a mode variance), Super-cells were annotated to a particular cell type with a Bayesian classifier developed in Zillionis et al.[4] applied to the super-cell count matrix. To plot super-cells, we used SPRING layout[32] from the original study averaged within each super-cell.

For the annotated *TIMs* dataset, DE analysis between cDCs and pDCs was performed at the single-cell and super-cell levels. Among the top 150 differentially expressed genes at the super-cell level, we selected those that were detected in both subtypes (i.e., in at least 95% of cDC and pDC super-cells) and those that were ranked higher (by p-value significance and then by logFC for identical p-values) at the super-cell level ($rank_{single-cell} - rank_{super-cell} \geq 3$). Among them, genes coding for trans-membrane proteins with available antibodies were tested for their protein expression using flow cytometry (see Supplementary Tables 3 and 4 for the full list of genes matching these criteria).

Mouse tumor model

Murine KP1.9 lung adenocarcinoma tumor cells were cultured in Iscove's DMEM media (Corning) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Cells were injected into 8 weeks old C57BL/6J male mice

(Charles River) intravenously (2.5x105 cells in 100 µl PBS) to develop orthotopic tumors in the lung. The tumor cell line was derived from lung tumor nodules of a C57BL/6 KrasLSL-G12D/WT;p53Flox/Flox (KP) mouse and was kindly provided by Dr. Zippelius (University Hospital Basel, Switzerland). Mice were analyzed for tumor phenotypes 4 weeks post-cancer cell injection[4,57]. All animals were housed at the Agora In Vivo Center (AIVC) in Lausanne. Experiments were performed following protocols approved by the Veterinary Authorities of the Canton Vaud according to Swiss law (animal license VD3612).

Flow cytometry analysis of mouse lung tumors

Single cell suspensions were obtained from lung tumor tissue of C57BL/6J male mice after transcardial PBS perfusion. Small tissue pieces were generated from perfused lungs using scissors and digested in DMEM containing 2.5% FBS, 25% Accutase (Sigma), 0.5 mg/ml collagenase type IV (Worthington Biochemical Corporation), 0.5 mg/ml hyaluronidase (Sigma) and 5 Units/ml DNAse I (Sigma) for 20 min at 37 C while shaking (800 rpm). Digested lung tissue was gently meshed through 70 µm cell strainers using a plunger. After red blood cells lysis (BD Pharm Lyse), single cell suspensions were incubated with Fc blocker (1:100, BioLegend) in staining buffer (PBS with 2% FBS and 1mM EDTA), followed by cell viability staining with Live/Dead Fixable Zombie UV (1:1000, BioLegend) in PBS. Cells were then stained with fluorochrome-conjugated antibodies (see Supplementary Table 5) for 30 min at 4 C in staining buffer, prior to fixation with IC Fixation buffer (Invitrogen) for 30 min. Samples were acquired using an LSRFortessa apparatus (BD Biosciences), and data were analyzed with FlowJo (v10.7.1). Cells were gated based on their size and granularity (FSC-A vs SSC-A), followed by doublet and dead- cell exclusion. Conventional DCs (cDC) and pDCs were defined as CD45+F4/80–CD11c+MHCII+SiglecH– and CD45+F4/80– CD11c+MHCII+SiglecH+B220+, respectively.

Cell type annotation with single genes and gene signature recovered from bulk RNA-seq

Cell type annotation in *Tcells* was performed with either a single marker or a gene signature recovered from bulk RNA (see below). A score was computed as the sum of max-normalized expression of signature genes. AUC was computed using the score

as a predictor and CD4/CD8 sorting information as a label with *prediction()* function from *ROCR* R package[58].

To compute CD4/CD8 gene signature from bulk RNA-seq, a dataset of bulk RNA-seq of sorted immune cells from donors were downloaded from GEO under accession number GSE60424[59]. DE analysis of 20 sorted CD4 and 20 sorted CD8 T cells populations from 10 donors was performed using *edgeR* R package[60]. The disease status was used as a covariate to eliminate disease bias. Genes with $FDR < 0.01$ and $|logFC| > 0.5$ were considered to be true markers ($N = 550$). Only 174 of those markers were found among genes in the *Tcells* dataset. The top 5 and top 50 markers were used as a gene signature of CD4 and CD8 T cell types.

The cell type annotation in *TILs* (Supplementary Fig. 4) was performed similarly to the one in *Tcells*, but instead of using the bulk signature, a set of top marker genes of each cluster found at the single-cell level was used. Due to the lack of sorting information in this dataset, the label (ground truth cell type annotation) was set to the single-cell clustering results.

GO match score of top correlated genes

For each cell line from the *cell_lines* dataset, genes that were expressed in more than 50% of cells of this cell line were kept. The gene pairwise Pearson correlation was computed using a gene expression matrix at the single-cell and at the super-cell levels. Then, the top $1'000$ highly correlated gene pairs at each level were selected. For each gene pair, a GO match score was computed as a Jaccard coefficient of their GO id sets. We then computed an average GO match score for gene pairs found exclusively in super-cells and normalized it by an average GO match score of gene pairs found exclusively at the single-cell level ($\gamma = 1$).

Imputation in super-cells

The imputation in single-cell and super-cell data were performed using MAGIC[39] with the default configuration. Bulk RNA-seq data of human adenocarcinoma used for comparison with the imputed profiles were extracted from GEO under accession number GSE86337. Only the common genes between the bulk data and the single cell data (*cell_lines*) were kept for the imputation. For each cell line, the imputed profiles of the super-cells and single cells were compared with the corresponding bulk

profile using the rank-based Spearman correlation. For the comparison, a pseudo-bulk profile is computed as an average scRNA-seq gene expression within a particular cell line.

RNA velocity in super-cells

Super-cells were built in the same way described in the super-cell construction section starting from the total counts (sum of spliced and un-spliced counts). The spliced and un-spliced counts for super-cells were computed by taking their mean value within each super-cell. The RNA velocity of the single cells and super-cells was computed independently using *relative.velocity.estimate()* function from the *Velocyto.R* package[5] with fit.quantile = 0.02.

The purity of super-cells in terms of velocity was computed based on the cosine similarity of RNA velocities of single cells belonging to one super-cell. The purity is determined as a maximum of a median of the cosine between RNA velocity vectors of cells based on the formula:

$$velocity_purity(C) = max_{x \in C} median_{y \in C} cos\left(\vec{v}(x), \vec{v}(y)\right),$$

where $\vec{v}(x)$ is a velocity of a cell x belonging to a super-cell C.

To obtain comparable 2D velocities vectors between single cells and super-cells, a joint tSNE of single-cell and super-cell velocities was computed using *tSNE.velocity.plot()* function with the parameters nPCs = 10 and perplexity = 50. We then computed the velocity similarity score as a median of the cosine similarities between the velocity of each single cell and the velocity of corresponding super-cell, i.e.,

$$similarity = median\ cos(\vec{v}(x), \vec{V}(C))$$

where $\vec{v}(x)$ is the velocity of cell x, and $\vec{V}(C)$ is the velocity of a supercell C (that contains x) in the joint tSNE. We computed the similarity between single-cell RNA velocity and that of random grouping in the same way. For subsampling method, to make a reasonable comparison, we define the similarity as follows:

$$similarity = median\ cos\left(\vec{v}(x), \vec{v}'(x')\right)$$

where $\vec{v}(x)$ is the velocity of cell x and $\vec{v}'(x')$ is the velocity of subsampled cell x' that is nearest to x (based on the Euclidean distance).

A key parameter estimated with RNA velocity is the equilibrium slope of each gene, which corresponds to the ratio between spliced and un-spliced mRNA that

represents the equilibrium between mRNA production and degradation. This equilibrium slope can be computed for a limited number of genes (Fig. 3i). To compare the consistency of RNA velocity results in super-cells and single cells, the Pearson correlation between equilibrium slope values was computed for the genes for which this equilibrium slope can be computed in all the methods (super-cell, subsampling and random grouping) across all graining levels.

Data integration of the *COVID-19_atlas* dataset

For each sample of the *COVID-19_atlas* dataset, super-cells were first constructed by specifying the graining levels to ensure more than 20 super-cells are obtained in each sample:

$$\gamma = min\left(\left[\frac{n}{20}\right], 10\right)$$

where [] denotes the integer part and n denotes the sample size. We then used the *SelectIntegrationFeatures()* function from *Seurat* package[18] to get $2'000$ genes for the integration. We next merged the super-cell samples into a single object of $146'304$ super-cells and performed a PCA reduction and Harmony integration[45] of all the samples.

To evaluate the batch effects in terms of protocols and samples before and after integration, kBET acceptance rate[46] was computed for non-integrated and integrated datasets for the cell types present in all samples in a sufficient number (i.e., B cells, monocytes, CD4 T cells and CD8 T cells). UMAP coordinates were used for this estimation.

To compute kBET acceptance rates for protocols, we randomly subsampled m super-cells from 10x5' protocol where m is number of super-cells in 10x3' protocol. We then applied the function *kBET()* on those subsampled batches for $k \in \{10, 20, 30, 40, 50\}$ to obtain the rejection rate $r(k)$. The kBET acceptance rate is then determined as $1 - r(k)$.

The kBET acceptance rates for samples were computed in the same way. In the subsampling step, 15 super-cells from each sample were subsampled where possible.

DE analysis within either monocytes or B cells from COVID-19 patients versus healthy controls was performed at the super-cell level with Seurat using only PBMC samples. Gene set enrichment was applied to the top 20 differentially expressed genes.

Approximate coarse-graining for large datasets

For the datasets with size $> 100'000$, we offer the option to use an approximate coarse-graining, which repeats all the steps of the super-cells construction but uses a set of subsampled cells, whose number can be specified by the user, to build an initial subsampled super-cell structure. Then, the remaining cells are mapped to the most similar super-cell based on the Euclidean distance (Supplementary Fig. 8a).

Integration of the human *TIM_atlas* dataset

Publicly available and well annotated datasets generated or reported in ref[38] and used to build the atlas of tumor infiltrating myeloid cells (*TIM_atlas*, $N = 108'566$) are listed in Supplementary Table 2. Integration of the listed datasets was performed as described in the original study[38]. Briefly, we performed two rounds of integration using scanorama package[61]. In the first round, 10x datasets were integrated and corrected. In the second round, corrected 10x data were integrated with all the other datasets.

All the downstream analyses, including super-cell construction, cell type annotation and the DE analysis were performed on the corrected gene expression data. Cell type annotation is available for the dataset generated in the original study[38]. Cells from the remaining datasets integrated into the atlas were annotated using a logistic regression, trained on the first 100 principal components of annotated data. Super-cells were annotated based on the most abundant cell type in each super-cell. The DE analysis between pDC and cDCs was performed at the single-cell and super-cell levels the same way as in the *TIMs* dataset.

Computational time and memory allocation

Assessment of the computational time and memory allocation for the data integration pipeline was performed on the *COVID-19_atlas* dataset. We benchmarked each step of the analyses separately, iteratively increasing the number of single cells by adding new samples. The benchmarking of the visualization includes dimensionality reduction (PCA) and UMAP. The benchmarking of the clustering includes PCA, graph construction and Seurat clustering. The benchmarking of the DE analysis consists of Seurat DE analysis with the t-test for 12 clusters (defined as the main cell types annotated in the original study) and $2'000$ genes. Since DE analysis

could not be performed for datasets with more than $600'000$ cells, the computational time and memory allocation were extrapolated using linear model (gray lines in Fig. 4d and Supplementary Fig. 9a). The benchmarking of the data integration consists of PCA and Harmony integration. The benchmarking of the 'Pipeline' consists of all the steps, including PCA, UMAP, graph construction, clustering, DE analysis and Harmony integration. The benchmarking of the coarse-graining followed by the data integration downstream analyses (Supplementary Fig. 9a) includes coarse-graining and all the steps of the 'Pipeline'. The following functions of the Seurat package[18] were used: *RunPCA()*, *RunUMAP()*, *RunHarmony()*, *FindNeighbors()* and *FindClusters()*, *FindAllMarkers()*. The benchmarking was performed on one node of an HPC cluster with 512G of RAM. To test the limits reached on standard desktop, the benchmarking was performed on an Intel core i7 with 16G of RAM.

To test computational time and memory allocation for the entire standard scRNA-seq data analyses pipeline for one sample, including exact or approximate coarse-graining followed by dimensionality reduction (PCA), clustering (Seurat) and DE analysis (t-test), scRNA-seq datasets of a different number of cells were generated by subsampling a larger dataset (GSE136831)[62]. DE analysis was performed for 3 clusters and $10'000$ genes. The benchmarking was performed on an Intel core i7 with 16G of RAM.

To assess computational time and memory allocation, the function *mark()* from R package *bench*[63] was used. The total computation time and memory allocation for the entire pipeline were computed as a sum of the computational times of each step and a max value of the memory allocation of each step, respectively.

Data availability

ScRNA-seq datasets analyzed in this study are public and listed in Supplementary Tables 1 and 2.

Code availability

The code to build super-cells is available as *SuperCell* R package (https://github.com/GfellerLab/SuperCell).

References

1. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* (2018) doi:10.1016/j.cell.2018.05.060.
2. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
3. Park, J.-E. *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
4. Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* (2019) doi:10.1016/j.immuni.2019.03.009.
5. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
6. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979 (2017).
7. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
8. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).
9. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
10. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* (2017) doi:10.1038/ncomms14049.
11. Regev, A. *et al. The Human Cell Atlas*. http://biorxiv.org/lookup/doi/10.1101/121202 (2017) doi:10.1101/121202.
12. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
13. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* **17**, 144 (2016).

14. Johansen, N. & Quon, G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.* **20**, 166 (2019).
15. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
16. Schwartz, G. W. *et al.* TooManyCells identifies and visualizes relationships of single-cell clades. *Nat. Methods* **17**, 405–413 (2020).
17. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4096.
19. Do, V. H., Rojas Ringeling, F. & Canzar, S. Linear-time cluster ensembles of large-scale single-cell RNA-seq and multimodal data. *Genome Res.* **31**, 677–688 (2021).
20. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
21. Patterson, J. Rapids - the platform inside and out. (2019).
22. Xiong, J., Gong, F., Wan, L. & Ma, L. NeuralEE: A GPU-Accelerated Elastic Embedding Dimensionality Reduction Method for Visualizing Large-Scale scRNA-Seq Data. *Front. Genet.* **11**, 786 (2020).
23. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Syst.* **8**, 483-493.e7 (2019).
24. Ren, X., Zheng, L. & Zhang, Z. SSCC: A Novel Computational Framework for Rapid and Accurate Clustering Large-scale Single Cell RNA-seq Data. *Genomics Proteomics Bioinformatics* **17**, 201–210 (2019).

25. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).

26. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).

27. Iacono, G. *et al.* bigSCale: an analytical framework for big-scale single-cell data. *Genome Res.* **28**, 878–890 (2018).

28. Bost, P. *et al.* Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients. *Cell* **181**, 1475-1488.e12 (2020).

29. Cohen, Y. C. *et al.* Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing. *Nat. Med.* (2021) doi:10.1038/s41591-021-01232-w.

30. Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A. & Tanay, A. Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. *Nat. Genet.* **52**, 709–718 (2020).

31. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).

32. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinforma. Oxf. Engl.* **34**, 1246–1248 (2018).

33. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* (2006) doi:10.7155/jgaa.00124.

34. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* (2019) doi:10.1038/s41592-019-0425-8.

35. Carmona, S. J., Siddiqui, I., Bilous, M., Held, W. & Gfeller, D. Deciphering the transcriptomic landscape of tumor-infiltrating CD8 lymphocytes in B16 melanoma tumors with single-cell RNA-Seq. *OncoImmunology* **9**, (2020).

36. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* (1987) doi:10.1016/0377-0427(87)90125-7.

37. Gerhard, G. M., Bill, R., Messemaker, M., Klein, A. M. & Pittet, M. J. Tumor-infiltrating dendritic cell states are conserved across solid human cancers. *J. Exp. Med.* **218**, e20200264 (2021).

38. Cheng, S. *et al.* A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* **184**, 792-809.e23 (2021).

39. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).

40. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

41. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).

42. Hochgerner, H., Zeisel, A., Lönnerberg, P. & Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21**, 290–299 (2018).

43. Bastidas-Ponce, A. *et al.* Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**, dev173849 (2019).

44. Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895-1913.e19 (2021).

45. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

46. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).

47. Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).

48. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

49. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* (2008) doi:10.1088/1742-5468/2008/10/P10008.

50. Kuchroo, M. *et al. Multiscale PHATE Exploration of SARS-CoV-2 Data Reveals Multimodal Signatures of Disease*. http://biorxiv.org/lookup/doi/10.1101/2020.11.15.383661 (2020) doi:10.1101/2020.11.15.383661.

51. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* (2021) doi:10.1038/s41592-021-01153-z.

52. Baglama, J., Reichel, L. & Lewis, B. W. *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*. (2019).

53. Arya, S., Mount, D., Kemp, S. E. & Jefferis, G. *RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric*. (2019).

54. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* (2006).

55. Schafer, J. *et al. corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. (2017).

56. Pasek, J. *weights: Weighting and Weighted Statistics*. (2020).

57. Pfirschke, C. *et al.* Immunogenic Chemotherapy Sensitizes Tumors to Checkpoint Blockade Therapy. *Immunity* **44**, 343–354 (2016).

58. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. Package 'ROCR'. (2015).

59. Linsley, P. S., Speake, C., Whalen, E. & Chaussabel, D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PloS One* **9**, e109760 (2014).

60. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp616.

61. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

62. Adams, T. S. *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).

63. Hester, J. *bench: High Precision Timing of R Expressions*. (2020).

Figure 1

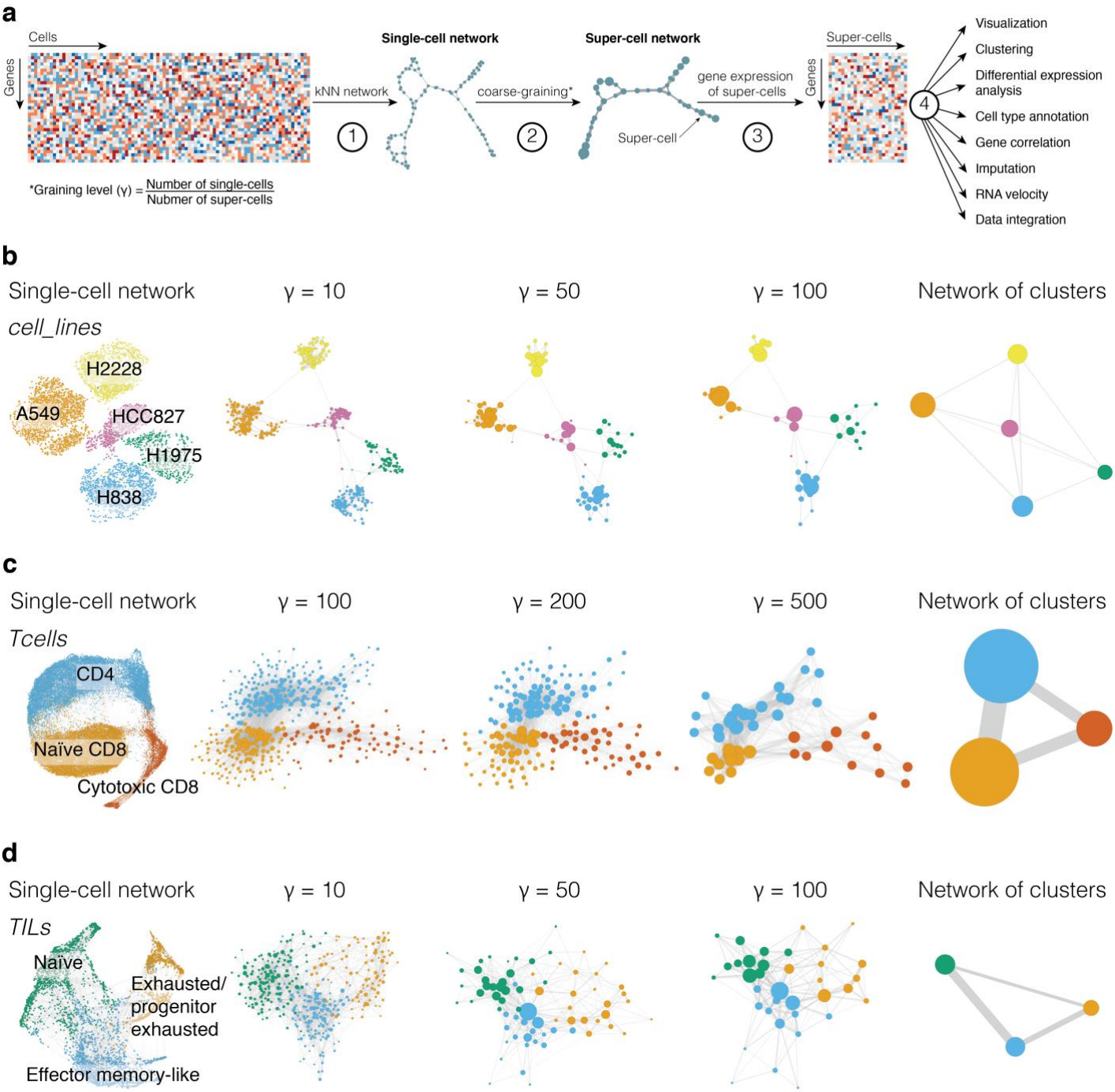
a
Cells
Genes
kNN network
Single-cell network
coarse-graining*
Super-cell network
Super-cell
gene expression of super-cells
Super-cells
Genes
Visualization
Clustering
Differential expression analysis
Cell type annotation
Gene correlation
Imputation
RNA velocity
Data integration
*Graining level (γ) = Number of single-cells / Nubmer of super-cells
b
Single-cell network
γ = 10
γ = 50
γ = 100
Network of clusters
cell_lines
H2228
A549
HCC827
H1975
H838
c
Single-cell network
γ = 100
γ = 200
γ = 500
Network of clusters
Tcells
CD4
Naïve CD8
Cytotoxic CD8
d
Single-cell network
γ = 10
γ = 50
γ = 100
Network of clusters
TILs
Naïve
Exhausted/ progenitor exhausted
Effector memory-like

Figure 1. Simplifying single-cell RNA-seq data with super-cells.

a, Overview of the coarse-graining pipeline, including the following steps. (1) A single-cell network is constructed from the single-cell gene expression matrix using k-nearest neighbors (kNN) algorithm. (2) A super-cell network is constructed by grouping similar cells into super-cells at a user-defined graining level (γ). (3) A gene expression matrix of super-cells is computed by averaging gene expression within each super-cell. (4) The super-cell gene expression matrix can be used for visualization and downstream analyses such as clustering, differential expression, cell type annotation, gene correlation, imputation, RNA velocity and data integration. **b**-**d**, Examples of super-cell networks at several graining levels. For comparison, the network of clusters is shown on the right. Colors indicate the initial cell type annotation and super-cells are colored according to the majority of cells in each super-cell. (**b**) Five cancer cell lines (*cell_lines*, $N = 3'918$) shown with different colors. (**c**) T cells sorted from PBMC (*Tcells*, $N = 40'560$) with CD4 in blue, naïve CD8 in orange and cytotoxic CD8 in red. (**d**) Tumor-infiltrating CD8 T lymphocytes (*TILs*, $N = 3'574$) with naïve cells in green, effector memory-like cells in blue and exhausted/progenitor exhausted cells in orange.

Figure 2

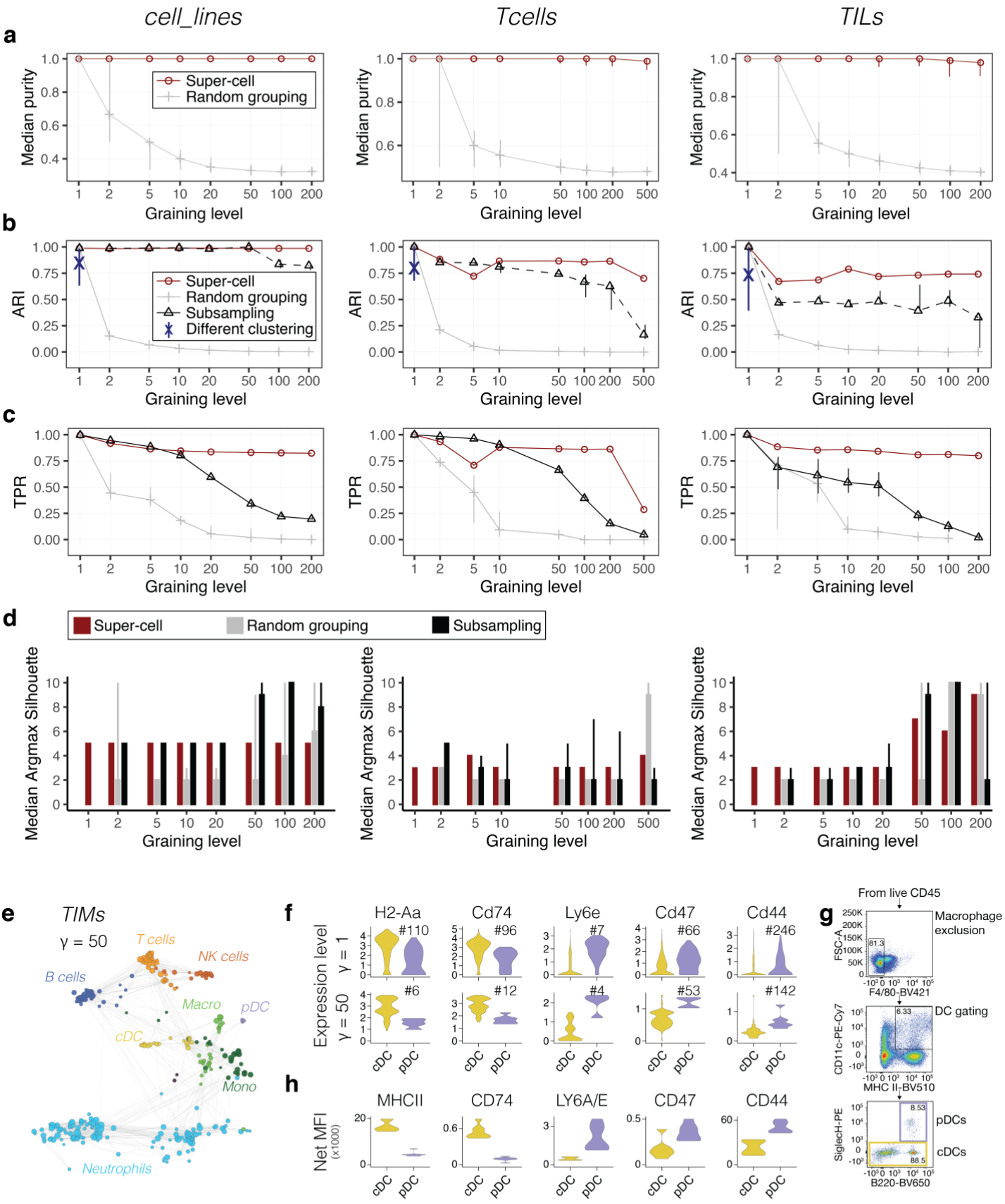
cell_lines
Tcells
TILs
a
Median purity
Super-cell
Random grouping
Graining level
b
ARI
Subsampling
Different clustering
c
TPR
d
Median Argmax Silhouette
e
TIMs
γ = 50
T cells
NK cells
B cells
Macro
pDC
cDC
Mono
Neutrophils
f
H2-Aa
Cd74
Ly6e
Cd47
Cd44
Expression level
γ = 1
γ = 50
cDC
pDC
h
MHCII
CD74
LY6A/E
CD47
CD44
Net MFI
(x1000)
g
From live CD45
Macrophage exclusion
FSC-A
F4/80-BV421
DC gating
CD11c-PE-Cy7
MHC II-BV510
SiglecH-PE
B220-BV650
pDCs
cDCs

Figure 2. Super-cells preserve clustering and differential expression results, and reveal genes specifically expressed in dendritic cell subtypes.

a, Median purity of super-cells at different graining levels for the three datasets shown in Fig. 1b-d (left *cell_lines*, middle *Tcells*, right *TILs*). For comparison, the purity after random grouping of cells is shown in gray. **b**, Consistency between the clustering obtained in super-cells and the reference annotation of single cells, computed with the adjusted Rand index (ARI). The blue line illustrates the range of ARIs when different clustering algorithms are applied at the single-cell level with "X" corresponding to the median value. **c**, Proportion of the cluster-specific differentially expressed genes found at the single-cell level and recovered by DE analysis at the super-cell level. **d**, Optimal number of clusters based on the maximum silhouette coefficient. For the subsampling (black) and random grouping (gray), the center of the error bars denotes the median, and the extrema denotes the 1st and 3rd quartiles (obtained with different random seeds). **e**, 2D visualization of the *TIMs* dataset ($N = 15'939$) at the super-cell level ($\gamma = 50$) with colors indicating cell type annotation. **f**, Expression of genes in single cells (top) and super-cells (bottom) that are more differentially expressed (i.e., better ranking) between cDCs and pDCs at the super-cell level. The number following the '#' sign indicates the ranking of each gene among the top differentially expressed ones. **g**, Flow cytometry analysis of DCs from murine KP1.9 lung adenocarcinoma ($N = 7$). **h**, Median fluorescence intensity of proteins coded by the genes from **f**. All comparison shown in **f** and **h** pass statistical significance based on two-tailed unpaired Student's t-test (p-values < 0.05).

Figure 3

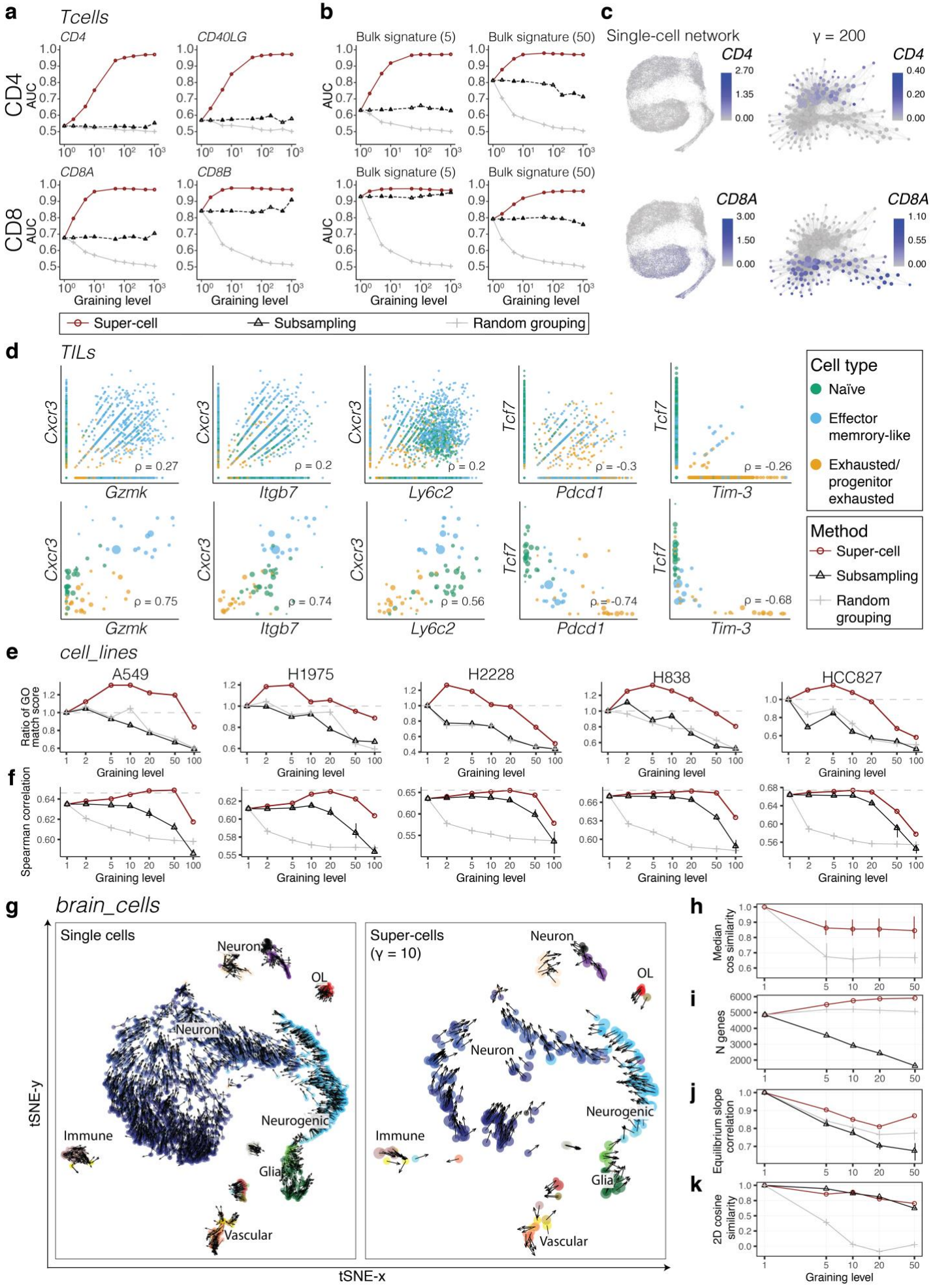
a
Tcells
CD4
CD40LG
CD8A
CD8B
CD4
CD8
AUC
Graining level
b
Bulk signature (5)
Bulk signature (50)
c
Single-cell network
γ = 200
CD4
CD8A
Super-cell
Subsampling
Random grouping
d
TILs
Cxcr3
Tcf7
Gzmk
Itgb7
Ly6c2
Pdcd1
Tim-3
ρ = 0.27
ρ = 0.2
ρ = 0.2
ρ = -0.3
ρ = -0.26
ρ = 0.75
ρ = 0.74
ρ = 0.56
ρ = -0.74
ρ = -0.68
Cell type
Naïve
Effector memrory-like
Exhausted/ progenitor exhausted
Method
Super-cell
Subsampling
Random grouping
e
cell_lines
A549
H1975
H2228
H838
HCC827
Ratio of GO match score
Graining level
f
Spearman correlation
g
brain_cells
Single cells
Super-cells (γ = 10)
Neuron
OL
Immune
Neurogenic
Glia
Vascular
tSNE-y
tSNE-x
h
Median cos similarity
i
N genes
j
Equilibrium slope correlation
k
2D cosine similarity
Graining level

Figure 3. Super-cells improve cell type annotation, gene correlation, imputation and RNA velocity.

a-**b**, AUC of recovery of CD4 (top) and CD8 (bottom) T cells from the *Tcells* dataset using single markers (**a**) or signatures defined form bulk (**b**) consisting of the top 5 or top 50 genes. **c**, Expression of *CD4* (top) and *CD8A* (bottom) in T cells from the *Tcells* dataset at the single-cell and the super-cell ($\gamma = 200$) levels. **d**, Gene correlation at the single-cell (top) and the super-cell ($\gamma = 50$) (bottom) levels for selected gene pairs in the *TILs* dataset, with the corresponding sample-weighted Pearson correlation (ρ). **e**, Comparison of the GO similarity of super-cell top correlated genes and single-cell top correlated genes identified for individual cell lines from the *cell_lines* dataset. The y-axis shows the ratio between mean GO match scores of the top correlated genes at the super-cell and the single-cell levels. **f**, Mean Spearman correlation between bulk and MAGIC-imputed data in each cell line of the *cell_lines* dataset. The dashed lines show the correlation between the pseudo-bulk (i.e., averaged gene expression within a cell line) and bulk gene expression. **g**, Joint tSNE visualization of RNA velocity for the *brain_cells* dataset ($N = 3'396$) for single cells (left) and super-cells ($\gamma = 10$) (right) colored by cell type annotation. **h**, Velocity purity in super-cells (computed as a cosine similarity of single-cell velocities within each super-cell). **i**, Number of genes with valid estimated equilibrium slope values. **j**, Pearson correlation of gene equilibrium slope values obtained in single-cell and super-cell RNA velocities. **k**, Cosine similarity between 2D single-cell and super-cell RNA velocities shown in **g**. For the subsampling and random grouping, the center of the error bars denotes the median, and the extrema denotes the 1st and 3rd quartiles (obtained with different random seeds).

Figure 4

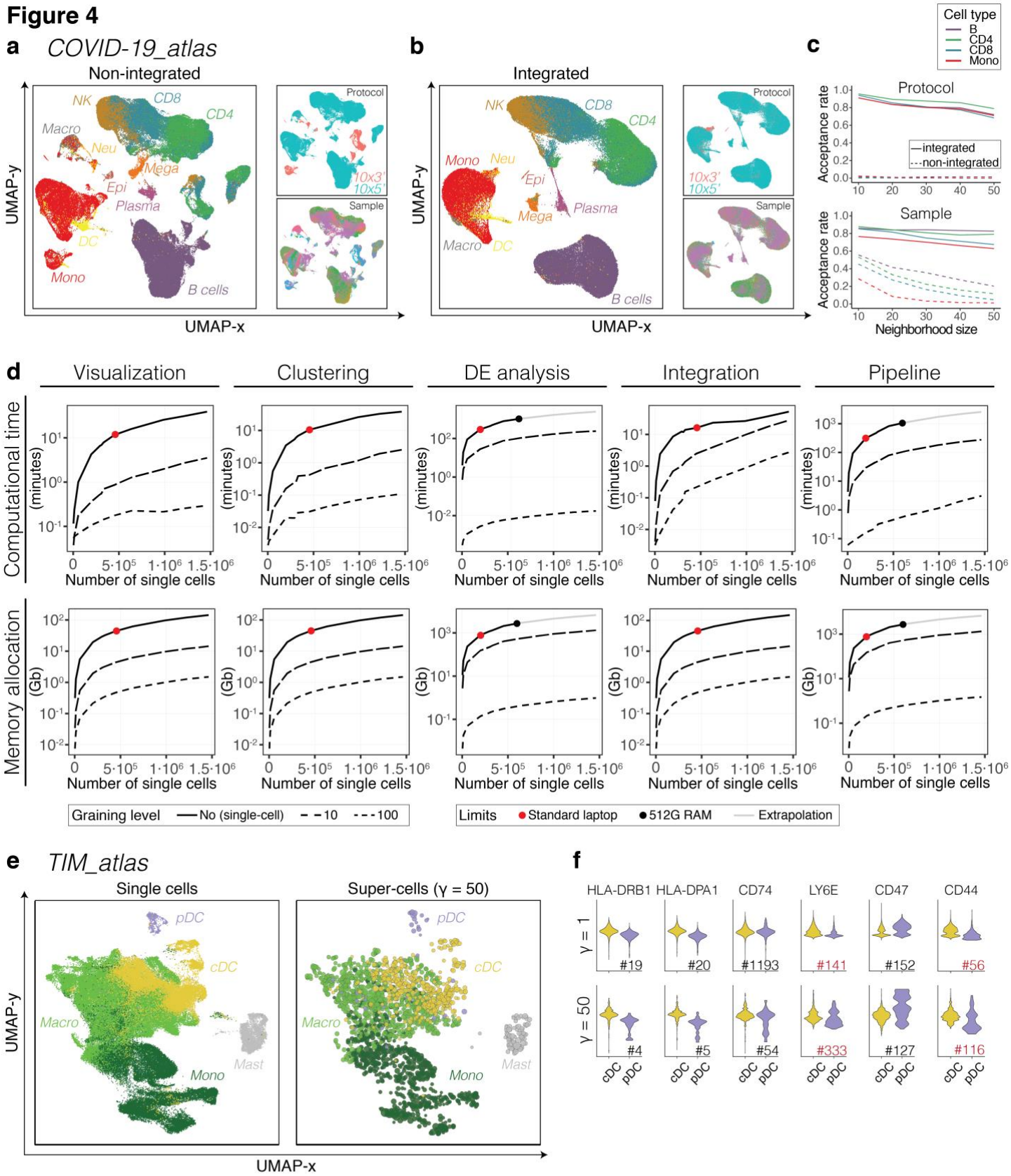
a
COVID-19_atlas
Non-integrated
NK
CD8
CD4
Macro
Neu
Mega
Epi
Plasma
DC
Mono
B cells
Protocol
10x3'
10x5'
Sample
UMAP-y
UMAP-x
b
Integrated
c
Cell type
B
CD4
CD8
Mono
Acceptance rate
integrated
non-integrated
Neighborhood size
d
Visualization
Clustering
DE analysis
Integration
Pipeline
Computational time
Memory allocation
(minutes)
(Gb)
Number of single cells
Graining level
No (single-cell)
10
100
Limits
Standard laptop
512G RAM
Extrapolation
e
TIM_atlas
Single cells
Super-cells (γ = 50)
pDC
cDC
Macro
Mono
Mast
f
HLA-DRB1
HLA-DPA1
CD74
LY6E
CD47
CD44
γ = 1
γ = 50
#19
#20
#1193
#141
#152
#56
#4
#5
#54
#333
#127
#116
cDC
pDC

Figure 4. Super-cells facilitate data integration and accelerate downstream analyses.

a-**b**, UMAP visualization of the non-integrated (**a**) and Harmony-integrated (**b**) *COVID-19_atlas* dataset ($N = 1'462'702$) at the super-cell ($\gamma = 10$) level. Super-cells are colored according to the cell type annotation, protocol or sample. **c**, Batch effect level in terms of protocol (top) and sample (bottom) in the non-integrated and Harmony-integrated *COVID-19_atlas* dataset, computed as the kBET acceptance rate for the four most frequent cell types. **d**, Computational time (top) and memory allocation (bottom) for the visualization (UMAP), clustering (Seurat), DE analysis (t-test, each cell type versus the rest), data integration (Harmony) and all steps together ('Pipeline') for the super-cells (dashed lines) and single cells (solid line). Red dots show the limits reached on standard desktops (16G of RAM). Black dots correspond to the limits reached on a machine with 512G RAM (linear extrapolations shown in gray). **e**, UMAP visualization of the *TIM_atlas* dataset ($N = 108'566$) at the single-cell (left) and the super-cell ($\gamma = 50$) (right) levels computed with the approximate coarse-graining. Cells are colored according to the cell type annotation. **f**, Relative (z-score) expression of genes experimentally studied in Fig. 2h at the single-cell (top) and super-cell ($\gamma = 50$) (bottom) levels. The number following the '#' sign indicates the ranking of each gene among the top differentially expressed ones. All comparison shown pass statistical significance based on two-tailed unpaired Student's t-test (p-values < 0.05) except for *CD74* at the single-cell level (p-value = 1). Ranks for genes showing a different behavior both at single-cell and super-cell levels between mouse and human are shown in red.