Bioinformatics, 38(18), 2022, 4330–4336
https://doi.org/10.1093/bioinformatics/btac541
Advance Access Publication Date: 4 August 2022
Original Paper

Gene expression

ASURAT: functional annotation-driven unsupervised clustering of single-cell transcriptomes

Keita Iida [1,*], Jumpei Kondo[2,3], Johannes Nicolaus Wibisana[1], Masahiro Inoue[3] and Mariko Okada[1,4]

[1]Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan, [2]Division of Health Sciences, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan, [3]Department of Clinical Bio-Resource Research and Development, Graduate School of Medicine Kyoto University, Kyoto 606-8501, Japan, and [4]Center for Drug Design and Research, National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka 567-0085, Japan

*To whom correspondence should be addressed.

Associate Editor: Valentina Boeva

Received on October 13, 2021; revised on July 4, 2022; editorial decision on July 29, 2022; accepted on August 1, 2022

Abstract

Motivation: Single-cell RNA sequencing (scRNA-seq) analysis reveals heterogeneity and dynamic cell transitions. However, conventional gene-based analyses require intensive manual curation to interpret biological implications of computational results. Hence, a theory for efficiently annotating individual cells remains warranted.

Results: We present ASURAT, a computational tool for simultaneously performing unsupervised clustering and functional annotation of disease, cell type, biological process and signaling pathway activity for single-cell transcriptomic data, using a correlation graph decomposition for genes in database-derived functional terms. We validated the usability and clustering performance of ASURAT using scRNA-seq datasets for human peripheral blood mononuclear cells, which required fewer manual curations than existing methods. Moreover, we applied ASURAT to scRNA-seq and spatial transcriptome datasets for human small cell lung cancer and pancreatic ductal adenocarcinoma, respectively, identifying previously overlooked subpopulations and differentially expressed genes. ASURAT is a powerful tool for dissecting cell subpopulations and improving biological interpretability of complex and noisy transcriptomic data.

Availability and implementation: ASURAT is published on Bioconductor (https://doi.org/10.18129/B9.bioc.ASURAT). The codes for analyzing data in this article are available at Github (https://github.com/keita-iida/ASURATBI) and figshare (https://doi.org/10.6084/m9.figshare.19200254.v4).

Contact: kiida@protein.osaka-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has deepened our knowledge of biological complexity in terms of heterogeneity and dynamic transition of cell populations, and this knowledge has immense potential for helping elucidate the regulatory principles underlying our body plans (La Manno *et al.*, 2018). scRNA-seq has been widely used to improve the molecular understanding of malignant cells in lymphoma (Zhang *et al.*, 2019), intra- and intertumoral heterogeneity in drug-treated cancer populations (Stewart *et al.*, 2020), ligand–receptor interaction in tumor immune microenvironments (Chen *et al.*, 2020) and effects of viral infection on immune cell populations (Devitt *et al.*, 2019). Various clustering methods based on gene expression similarity have been proposed (Pasquini *et al.*, 2021) and applied to annotate cell types (Kim *et al.*, 2020). However, conventional gene-based analyses require intensive manual curation to annotate clustering results; hence, efficient and unbiased interpretation of single-cell data remains challenging (Andrews *et al.*, 2021; Aran *et al.*, 2019; Kiselev *et al.*, 2019).

Conventionally, single-cell transcriptomes are analyzed and interpreted by means of unsupervised clustering followed by manual curation of marker genes selected from a large number of differentially expressed genes (DEGs) (Andrews *et al.*, 2021). Here, manual curations are based on literature searches of biological functions of DEGs. Today, several computational tools for cell type inference are available, as detailed in a review by Pasquini *et al.* (2021). However, manual curation is often difficult because a single gene is generally multifunctional and therefore associated with multiple biological function terms (Cancer Genome Atlas Research Network *et al.*, 2017). In cancer transcriptomics, this difficulty is exacerbated by

the complex interdependence between disease-related genes and their heterogeneous expressions, associated with numerous biological function terms.

A possible solution is to realize clustering and interpretation simultaneously. Recently, reference component analysis has been used for accurate clustering of single-cell transcriptomes along with unbiased cell-type annotation based on similarity to reference transcriptome panels (Li *et al.*, 2017). Yet, these methods require transcriptomic data of well-characterized reference cells as learning datasets, which might not always be available. Another approach is using functionally annotated gene sets for scoring cells, implemented in R packages including PAGODA (Fan *et al.*, 2016) and ssGSEA (Subramanian *et al.*, 2005). Given single-cell transcriptome data, these methods use statistical methods, such as principal component analysis (PCA) and gene set enrichment analysis (GSEA), for providing each cell with scores of annotations against functionally annotated gene sets, such as signaling pathway modules. Nevertheless, correlations of gene expressions are complex with positive and negative (Saxena *et al.*, 2006), strong and weak and non-linear relationships, which can be poorly captured using the existing methods (see also Section 4).

To overcome these limitations, a nonlinear framework defining biological terms in a more interpretable way is needed. Therefore, we propose the computational tool, ASURAT (functional annotation-driven unsupervised clustering of single-cell transcriptomes), which simultaneously performs unsupervised clustering and biological interpretation in terms of cell type, disease, biological process and signaling pathway, using a nonlinear correlation graph decomposition for functionally annotated gene sets. In this study, we demonstrate the clustering performance of ASURAT using scRNA-seq datasets for healthy and disease human peripheral blood mononuclear cells (PBMCs), small cell lung cancer (SCLC) and spatial transcriptome (ST) datasets for pancreatic ductal adenocarcinoma (PDAC). Our results suggest that ASURAT can greatly improve functional understanding of single-cell transcriptomes, adding a new layer of biological interpretability to conventional gene-based analyses.

2 Materials and methods

2.1 Overview of ASURAT workflow

ASURAT is a computational tool for simultaneously clustering and interpreting single-cell transcriptomes (Fig. 1) using functionally annotated gene sets collected from knowledge-based databases for cell type, disease, biological process and signaling pathway activity, such as Cell Ontology (CO) (Diehl *et al.*, 2016), Disease Ontology (DO) (Yu *et al.*, 2015), Gene Ontology (GO) (Yu *et al.*, 2012) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) (Fig. 1b). ASURAT creates multiple biological terms using single-cell transcriptome data and annotated gene sets (Fig. 1c). We called such biological terms 'signs'. Then, ASURAT creates sign-by-sample matrices (SSMs), in which rows and columns stand for signs and samples (cells), respectively (Fig. 1d). SSM is analogous to a read count table, where the rows represent signs with biological meaning instead of individual genes and the values contained are 'sign scores' instead of read counts. By analyzing SSMs, individual cells can be characterized by various biological terms (Fig. 1e).

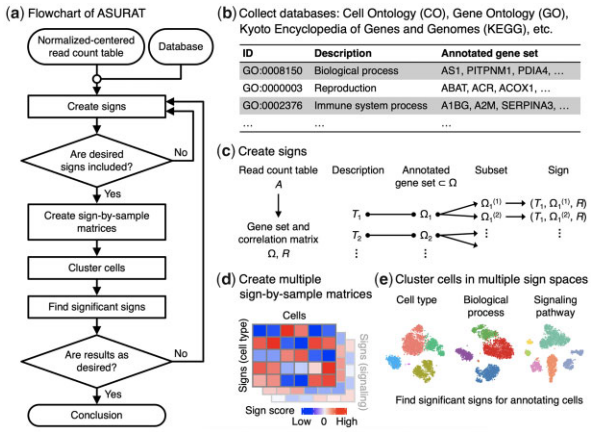

Fig. 1. Workflow of ASURAT. (**a**) Flowchart of ASURAT. (**b**) Collection of databases. (**c**) Creation of signs and (**d**) SSMs. (**e**) Analysis of SSMs to infer cell types, diseases, biological processes and signaling pathway activities

2.2 Sign

Let A be a read count table of size $p \times n$ from single-cell transcriptomic data, whose rows and columns are p genes, represented by $\Omega = \{1, 2, \cdots, p\}$, and n cells, respectively, and R a 'rapport' (e.g. correlation matrix) among Ω. Let $\mathcal{F} = \{(T_k, \Omega_k) | \Omega_k \subset \Omega,\ k = 1, 2, \cdots, q\}$ be a set of ordered pairs, where T_k and Ω_k are biological description and the annotated gene set, respectively. Consider an R-dependent representation $\Omega_k = \cup_{j=1}^{m_k} \Omega_k^{(j)}$, where m_k is an integer, for $k = 1, 2, \cdots, q$; then, the triplet $(T_k, \Omega_k^{(j)}, R)$ is termed a sign, in particular (T_k, Ω_k, R) a parent sign. Our definition is based on Saussure's semiology. According to Maruyama, the original notion of a *signe* is a segment of 'a thing of interest', created by an arbitrary decomposition based on its relations. For example, 'rainbow' is a continuum of varying light input, from which we see distinct colors of red, yellow, green and blue by our subjective decomposition based on their spectral relationships (Couper, 2015).

2.3 Correlated gene set

Let $R = (r_{i,j})$ be a correlation matrix of size $p \times p$ defined by A and a certain measure (e.g. Spearman's measure), whose diagonal elements are 1s. Let α and β be positive and negative constants satisfying $0 < \alpha \leq 1$ and $-1 \leq \beta < 0$, respectively. Let us arbitrarily fix $(T_k,\ \Omega_k) \in \mathcal{F}$ and consider the following subsets of Ω_k:

$$U_k(\alpha) = \{i \in \Omega_k | \exists j \in \Omega_k \text{ such that } r_{i,j} \geq \alpha,\ i \neq j\},$$
$$V_k(\beta) = \{i \in \Omega_k | \exists j \in \Omega_k \text{ such that } r_{i,j} \leq \beta,\ i \neq j\},$$
$$W_k(\alpha, \beta) = U_k(\alpha) \cup V_k(\beta).$$

Hereinafter we omit the arguments α and β for simplicity. Let $R_{W_k} = (\tilde{r}_{i,j})_{i,\ j \in W_k}$ be a submatrix of R. Let us identify the row vectors of R_{W_k} with points in $|W_k|$-dimensional Euclidean space and denote the set of those points as $\tilde{W}_k$. Then, let us identify all the elements in W_k with those in $\tilde{W}_k$ through the subscripts of $\tilde{r}_{i,\ j}$, $i,\ j \in W_k$. If V_k is not empty, decompose $\tilde{W}_k$ into two disjoint subsets $\Omega_k^{(s)}$ and $\Omega_k^{(v)}$ by Partitioning Around Medoids (PAM) clustering (Schubert and Rousseeuw, 2019), from which we obtain

$$W_k = \tilde{W}_k = \Omega_k^{(s)} \cup \Omega_k^{(v)}.$$

Otherwise, if V_k is empty, let $\Omega_k^{(s)} = U_k$ and $\Omega_k^{(v)} = \phi$ (empty). Thus, Ω_k is decomposed into three parts as follows:

$$\Omega_k = \Omega_k^{(s)} \cup \Omega_k^{(v)} \cup \Omega_k^{(w)}, \tag{1}$$

where $\Omega_k^{(w)} = \Omega_k - W_k$. Let $R_{\Omega_k^{(s)}} = (r_{i,\ j})_{i,\ j \in \Omega_k^{(s)}}$ and $R_{\Omega_k^{(v)}} = (r_{i,\ j})_{i,\ j \in \Omega_k^{(v)}}$ be submatrices of R and let $\mu_k^{(s)}$ (resp. $\mu_k^{(v)}$) be the mean of off-diagonal elements of $R_{\Omega_k^{(s)}}$ ($R_{\Omega_k^{(v)}}$). We assume $\mu_k^{(s)} \geq \mu_k^{(v)}$ without loss of generality. If $\mu_k^{(s)} \geq \alpha$, then $\Omega_k^{(s)}$, $\Omega_k^{(v)}$ and $\Omega_k^{(w)}$ are termed strongly, variably and weakly correlated gene sets, which are hereafter abbreviated as SCG, VCG and WCG, respectively. Otherwise, correlated gene sets cannot be defined for T_k.

Figure 2 shows that the SCG and VCG include *KRT18* and *ASCL1*, which have negative and positive contributions for lung small cell carcinoma, respectively. Thus, we interpret that $(T_k,\ \Omega_k^{(s)},\ R)$ and $(T_k,\ \Omega_k^{(v)},\ R)$ for DOID 5409 relate positively

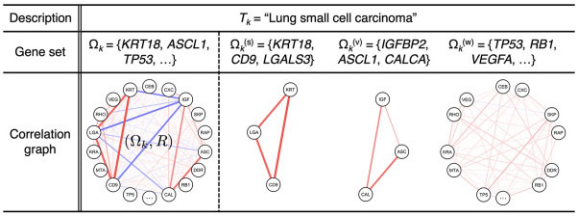

Fig. 2. Representation of correlation graph decomposition. From scRNA-seq data and a Disease Ontology term with DOID 5409, which concerns SCLC, three signs $(T_k, \Omega_k^{(j)}, R)$, $j \in \{s, v, w\}$, were produced from their parent sign (T_k, Ω_k, R) by decomposing the correlation graph (Ω_k, R) into strongly, variably and weakly correlated gene sets (e.g. positive and negative correlations are observed between *KRT18* and *CD9*, and *KRT18* and *IGFBP2*, respectively). The edge width indicates the strength of the correlation.

and negatively with this cell type, respectively. Though there exist simpler methods for decomposing graphs, such as one-shot PAM clustering, hierarchical clustering and tree cutting (Murtagh and Legendre, 2014), PCA-based methods (Hyvarinen, 1999) and several graph statistical approaches (Blondel *et al.*, 2008; Bodenhofer *et al.*, 2011), we found that the VCG definition is critical for clustering cells. In fact, we tried replacing our decomposition method (1) with one-shot PAM clustering, but the results often exhibited deteriorated performance since both VCG and WCG (obtained from the one-shot clustering) included weakly correlated genes.

2.4 Sign-by-sample matrix

Let $A = (a_{i,j})$ be a gene-by-cell matrix of size $p \times n$ from a single-cell transcriptomic data, whose entries stand for normalized-and-centered gene expression levels. For simplicity, let us assume that annotated gene sets Ω_k can be decomposed into non-empty $\Omega_k^{(s)}$, $\Omega_k^{(v)}$ and $\Omega_k^{(w)}$, for $k = 1, 2, \cdots, q$. Let $B^{(x)} = (b_{k,j}^{(x)})$, $x \in \{s, v, w\}$, be matrices of size $q \times n$, whose entries are defined as follows:

$$b_{k,j}^{(x)} = \frac{1}{|\Omega_k^{(x)}|} \sum_{i \in \Omega_k^{(x)}} a_{i,j},$$

where $|\Omega_k^{(x)}|$ stands for the number of elements in $\Omega_k^{(x)}$. Additionally, let $C^{(x)} = (c_{k,j}^{(x)})$, $x \in \{s, v\}$, be $q \times n$ matrices, whose entries are defined as follows:

$$c_{k,j}^{(x)} = \omega^{(x)} b_{k,j}^{(x)} + (1 - \omega^{(x)}) b_{k,j}^{(w)}, \quad (2)$$

where $\omega^{(x)}$, $0 \leq \omega^{(x)} \leq 1$, are weight constants. Here, $C^{(s)}$ and $C^{(v)}$ are termed SSMs for SCG and VCG, respectively, while the entry $c_{k,j}^{(x)}$ as sign score of the kth sign for jth sample (cell). Sign score profiles can be represented as points in a Euclidean space, termed sign space in this article. Notably, the ensemble means of sign scores across cells are zeros since SSMs are derived from the centered gene expression matrix A.

2.5 Significant sign

Using ASURAT, one can create multiple signs and SSMs by setting the appropriate parameters (Supplementary Note S1 and Fig. S1). Using SSMs, we can infer cell states by performing unsupervised clustering and investigating significant signs, where 'significant' means that the sign scores (2) are specifically upregulated or downregulated at the cluster level. Significant signs are analogous to DEGs. Here, naïve usages of statistical tests and fold change analyses should be avoided because the row vectors of SSMs are centered. Hence, we propose a nonparametric separation index, which quantifies the extent of separation between two sets of random variables (Supplementary Note S2 and Fig. S2).

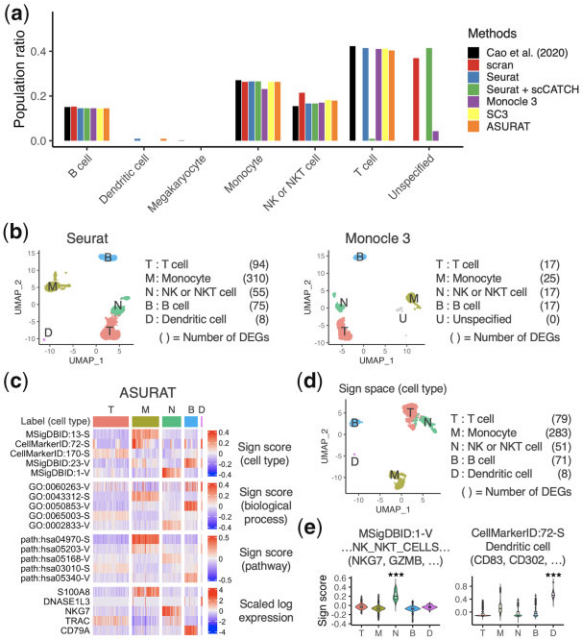

Fig. 3. Identification of cell types in peripheral blood mononuclear cell 4k single-cell transcriptome. (**a**) Population ratios predicted via seven different methods. (**b**) Uniform manifold approximation and projection (UMAP) plots, computed using Seurat and Monocle 3, for the manual investigation of DEGs based on the adjusted P-values $< 10^{-100}$. (**c**) SSMs and scaled log-transformed read count table, which are vertically concatenated and clustered based on the SSM for cell type. Only top significant signs and DEGs are shown in rows. (**d**) UMAP plot of the SSM for cell type. (**e**) Violin plots showing sign scores of significant signs, in which separation indices (I) for the clusters marked with asterisks against the others show $***I > 0.9$. Suffixes '-S' and '-V' after IDs indicate the signs are defined by strongly and variably correlated gene sets, respectively

3 Results

3.1 Clustering single-cell transcriptomes of PBMCs from healthy donors

To validate the clustering and interpretation performances of ASURAT in comparison with existing methods, we analyzed two public scRNA-seq datasets, namely PBMCs 4k and 6k (Supplementary Note S3), in which the cell types were inferred via computational tools based on prior assumptions (Cao *et al.*, 2020). We first excluded low-quality genes and cells (Supplementary Note S4); the resulting read count tables were supplied to ASURAT and four other methods: scran (Lun *et al.*, 2016), Seurat (Hao *et al.*, 2021), Monocle 3 (Trapnell *et al.*, 2014) and SC3 (Kiselev *et al.*, 2017). To infer the cell types, we used existing methods, performed unsupervised clustering of cells, and annotated each cluster by manually investigating DEGs based on the adjusted P-values $< 10^{-100}$ or false discovery rates $< 10^{-100}$ (Supplementary Note S5). Using ASURAT, we created SSMs from CO, Molecular Signatures Database (MSigDB) (Subramanian *et al.*, 2005) and CellMarker databases (Zhang *et al.*, 2019) for cell type, GO database for biological process and KEGG for signaling pathway. Then, cells were clustered via k-nearest neighbor graph generation and the Louvain algorithm (Hao *et al.*, 2021) based on the SSM for cell type and annotated by significant signs and DEGs.

Among all methods used, Seurat, Monocle 3 and ASURAT could robustly reproduce most blood cell types (Fig. 3a and Supplementary Fig. S5), as inferred by Cao *et al.* (2020). We found that manual annotations provided comparable population ratios with previous results (Cao *et al.*, 2020). Yet, it was quite laborious to manually investigate marker genes from numerous DEGs (Fig. 3b). To avoid such laborious process, we applied scCATCH

(Shao *et al.*, 2020) to automatically annotate the clustering results of Seurat. Nevertheless, population ratios inferred by scCATCH were less consistent than those by manual annotation (Fig. 3a and Supplementary Fig. S5). We also used ssGSEA (Subramanian *et al.*, 2005) for providing each cell with enrichment scores against 'cell type signature gene sets' defined in MSigDB without relying on DEGs (Supplementary Note S5). Although ssGSEA seemed to have imperfect clustering performance (Supplementary Figs S3e and S4e), the resulting scores were approximately consistent with Seurat annotations with a few exceptions (Supplementary Fig. S6).

Using ASURAT, we identified six cell types, including dendritic cell and megakaryocyte (Fig. 3a and Supplementary Fig. S5), by primarily investigating significant signs for cell type and secondarily the other signs and DEGs (Fig. 3c–e and Supplementary Fig. S7). The population ratios were approximately consistent with the reported values (Cao *et al.*, 2020), except for the small dendritic cell population possibly included in PBMCs (Villani *et al.*, 2017). Such a small discrepancy was unavoidable, since Cao *et al.* (2020) used author-defined DEGs and preselected cell types to identify the most preferable ones. Our results were robust against variations of the number of cells by randomly downsampling the cells from PBMC 4k data and analyzing these data using Seurat and ASURAT with almost the same parameters (Supplementary Note S6). These results demonstrate that ASURAT can perform robust clustering for single-cell transcriptomes.

3.2 Clustering single-cell transcriptomes of PBMCs from control and sepsis donors

ASURAT uses database-derived biological terms for clustering single-cell transcriptomes, which inevitably introduces annotation bias; some biological terms are associated with many genes, while others are associated with few (Gaudet and Dessimoz, 2017). Hence, it is important to validate the clustering performance in terms of cell state granularity. Here, we analyzed scRNA-seq datasets of PBMC published by the clinical cohort study for bacterial sepsis (Reyes *et al.*, 2020) (Supplementary Note S3), in which 65 subjects with different health conditions were included, total of 106 545 $CD45^+$ cells and 19 806 $LIN^-CD14^-HLA\text{-}DR^+$ dendritic cells were profiled, 16 immune-cell states were defined and the immune signatures of sepsis against bacterial infection were studied.

After excluding low-quality genes and cells (Supplementary Note S4), we inferred the cell types, using scran, Seurat, Monocle 3 and ASURAT with almost the same parameters as in the analyses of PBMCs 4k and 6k (Supplementary Note S7). Among all the methods, ASURAT could reproduce most immune cell types (Supplementary Fig. S12), which was consistent with the previous report (Reyes *et al.*, 2020). ASURAT identified 11 clusters by performing an unsupervised clustering of the SSM for cell type (Fig. 4a). Among these, we identified three monocyte subpopulations (Fig. 4a–c and Supplementary Fig. S11): M1 and M2, opposite subtypes with decreased and increased functions of fatty acid degradation; M3, similar to T cells, are characterized by increased function of cell adhesion, containing CD2, CD8A and CD58 (Fig. 4c).

We investigated the differences in cell state composition across each subject type and found that the fractions of total monocytes in subjects with infections (Leuk-UTI, Int-URO, URO, Bac-SEP, ICU-SEP and ICU-NoSEP) are larger than those in healthy controls (Control) (Fig. 4d), which is consistent with the reported result (Reyes *et al.*, 2020). Moreover, the fractions of M2 in subjects with organ dysfunction (Int-URO, URO, Bac-SEP and ICU-SEP) are smaller than those in subjects without organ dysfunction (Control and Leuk-UTI) except for severely ill patients without infection (ICU-NoSEP). Our results suggest that sepsis is associated with impairments of lipid metabolism in monocytes, supported by a previous proteomic study (Sharma *et al.*, 2019). These results demonstrate that ASURAT can cluster cells in fine-grained manners and identify functional subtypes.

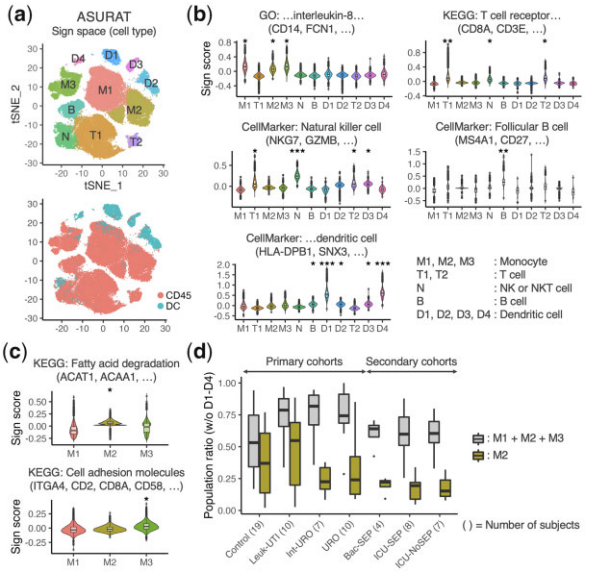

Fig. 4. Identification of cell states in peripheral blood mononuclear cell single-cell transcriptomes from control and sepsis donors. (**a**) t-distributed stochastic neighbor embedding (t-SNE) plots of the SSM for cell type, showing (top) the clustering result and (bottom) reported labels for $CD45^+$ cells and dendritic cells (DCs) by Reyes *et al.* (2020). (**b** and **c**) Violin plots showing sign scores of significant signs, in which separation indices (I) for the clusters marked with asterisks against the others show $***I > 0.9$, $**I > 0.6$ and $*I > 0.4$. (**d**) Population ratios of total monocytes (M1, M2 and M3) and subcluster M2 to all cells except for the inferred dendritic cells across each subject type. Control, uninfected and healthy control; Leuk-UTI, subjects with urinary tract infection (UTI) with leukocytosis but no organ dysfunction; Int-URO and URO, subjects with UTI with mild (or transient) and clear (or persistent) organ dysfunction, respectively; Bac-SEP, bacteremic subjects with sepsis in hospital wards; ICU-SEP and ICU-NoSEP, bacteremic subjects admitted to the intensive care unit with and without sepsis, respectively

3.3 Clustering a single-cell transcriptome of SCLC

SCLC tumors undergo a transition from chemosensitivity to chemoresistance states against platinum-based therapy through changes in transcriptional heterogeneity (Stewart *et al.*, 2020). The ability of SCLC to change phenotype in response to environmental cues involves multiple physiological states of cells, such as pathological states, cell cycle phases (Dominguez *et al.*, 2016) and metabolic processes (Jalili *et al.*, 2021). However, functional states cannot be readily identified using conventional gene-based analysis alone, and hence the mechanism behind chemoresistance remains unclear. To better understand cancer subtypes in chemoresistant tumors, we applied Seurat and ASURAT to the SCLC scRNA-seq data with cisplatin treatments, obtained from circulating tumor cell-derived xenografts generated from treatment-naïve lung cancer patients (Stewart *et al.*, 2020) (Supplementary Note S3).

First, we examined the expression levels of known SCLC marker genes (Ireland *et al.*, 2020), namely *ASCL1*, *NEUROD1*, *YAP1* and *POU2F3*, and confirmed that almost all of the cells are of the *ASCL1* single-positive subtype (Supplementary Fig. S13), which is consistent with the previous report (Stewart *et al.*, 2020). Then, we excluded low-quality genes and cells (Supplementary Note S4). To investigate molecular subtypes and potential resistance pathways, we performed unsupervised clustering of cells, using Seurat (Supplementary Note S8). We found that the populations assigned to G1, S and G2M phases are sequentially distributed in the uniform manifold approximation and projection (UMAP) space (Fig. 5a), while few clusters were observed when we regressed out the cell cycle effects (Supplementary Fig. S14), indicating that cell cycle signals are informative in SCLC heterogeneity. Next, we performed KEGG pathway enrichment analysis based on the DEGs with adjusted P-values $<10^{-2}$, the approximate highest threshold where meaningful enrichments of KEGG terms were obtained, but

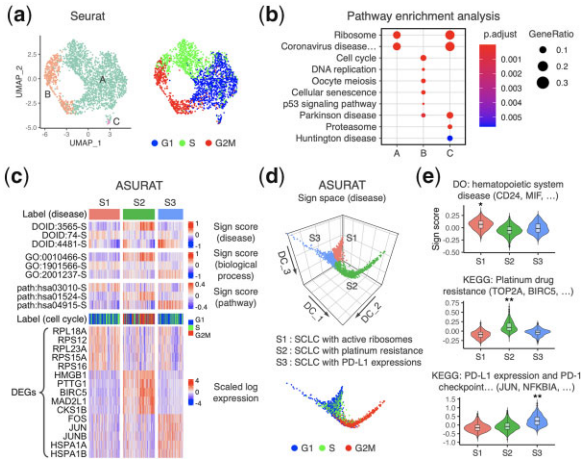

Fig. 5. Clustering result of a single-cell transcriptome of SCLCs. (**a**) Uniform manifold approximation and projection (UMAP) plots, showing (left) the clustering result and (right) cell cycle phases computed using Seurat. (**b**) Pathway enrichment analysis for the clustering result in (a), based on the DEGs with adjusted *P*-values $< 10^{-2}$, in which top five enriched terms are shown. (**c**) SSMs and scaled log-transformed read count table, which are vertically concatenated and clustered based on the SSM for disease. Only top significant signs and DEGs are shown in rows. (**d**) Diffusion map plots of the SSM for disease, showing (top) the clustering result and (bottom) cell cycle phases. (**e**) Violin plots showing sign scores of significant signs, in which separation indices (*I*) for the clusters marked with asterisks against the others show $**I > 0.6$ and $*I > 0.4$

chemoresistance-related terms were not primarily enriched (Fig. 5b, Supplementary Note S8).

Subsequently, to investigate functional subtypes of SCLCs, we used ASURAT to create SSMs from DO, GO and KEGG databases for disease, biological process and signaling pathway, respectively (Fig. 5c). We performed a dimensionality reduction using a diffusion map (Coifman and Lafon, 2006), followed by unsupervised clustering of cells using MERLoT (Parra *et al.*, 2019) based on the SSM for disease (Fig. 5d, Supplementary Note S8). Investigating significant signs and DEGs based on the separation indices $I > 0.4$ and adjusted *P*-values $<10^{-70}$, respectively, we found three SCLC subpopulations (Fig. 5c–e): S1, cancer cells characterized by significant signs for hematopoietic system disease and DEGs for ribosomal protein (*RPS12*, *RPL18A*, etc.); S2, characterized by significant signs for platinum drug resistance and DEGs for cancer-related genes (*HMGB1*, *PTTG1*, etc.); and S3, characterized by significant signs for programmed death ligand 1 (PD-L1) expression-mediated immunosuppression and DEGs for proto-oncogenes (*FOS*, *JUN*, etc.). Although SCLC molecular subtypes have been extensively studied (Balanis *et al.*, 2019; Chen *et al.*, 2019; Ireland *et al.*, 2020; Schwendenwein *et al.*, 2021), these functional subpopulations have been previously overlooked. Identifying *de novo* SCLC subtypes by future work will validate our clustering results. ASURAT provides a novel clue for the clinical improvements for relapsed SCLC tumors.

3.4 Clustering an ST of PDAC

Recent studies of spatially resolved transcriptomic profiling for human PDAC tumors have uncovered that cancer and non-cancer cells are spatially distributed in the distinct tissue regions of primary tumors, and that PDAC cells are accompanied by inflammatory fibroblasts and immune cells (Elosua-Bayes *et al.*, 2021; Moncada *et al.*, 2020). In an original study (Moncada *et al.*, 2020), the cellular resolutions of the STs were estimated at 20–70 cells per ST spot, far lower than those of scRNA-seq. Thus, computational methods have been proposed to predict existing cell types by integrating ST and scRNA-seq datasets (Elosua-Bayes *et al.*, 2021; Moncada *et al.*, 2020). Here, we applied Seurat and ASURAT to the published PDAC ST and scRNA-seq data (Moncada *et al.*, 2020) (Supplementary Note S3), aiming to compare the clustering results of ASURAT with those of existing methods.

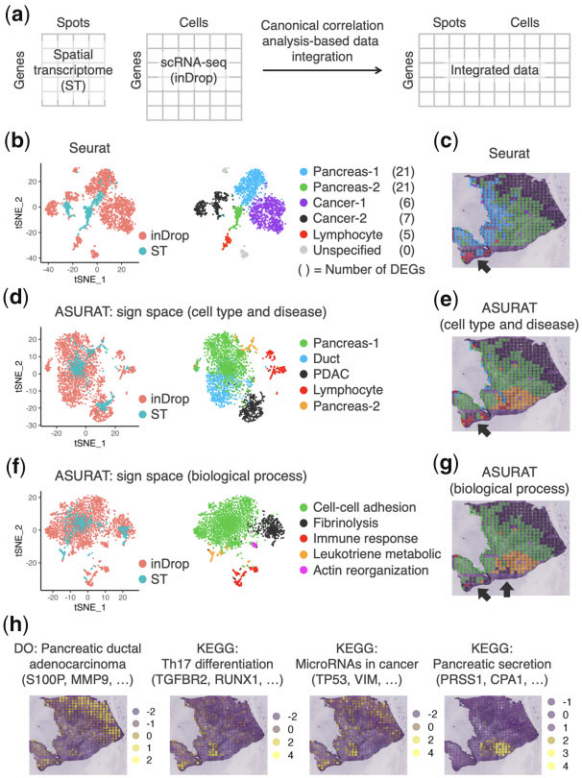

Fig. 6. Clustering results of a ST of PDAC. (**a**) Canonical correlation analysis-based data integration of ST and scRNA-seq data using Seurat functions. (**b–g**) t-distributed stochastic neighbor embedding (t-SNE) plots and clustering results shown in the PDAC tissue, based on the indicated methods: (b), (d) and (f), showing (left) the labels for ST and scRNA-seq data and (right) annotation results by manual investigation based on the DEGs with adjusted *P*-values $< 10^{-100}$ and significant signs; (c), (e) and (g), showing the clustering results, in which labels are the same with those in (b), (d) and (f), respectively. The black colored spots pointed by the arrows indicate the spots newly predicted as atypical region, which might be a normal pancreas involved in cancer. (**h**) Profiles of sign scores in the PDAC tissue, predicting cancer, inflammation and pancreas spots

First, we combined all scRNA-seq datasets after confirming minimal batch effects (Supplementary Fig. S15). We excluded low-quality genes and cells from the ST and scRNA-seq data (Supplementary Note S4). To cluster the ST with reference to the combined scRNA-seq data, we integrated these two data, using a canonical correlation analysis-based data integration method (Butler *et al.*, 2018) (Fig. 6a). Then, we used Seurat to perform unsupervised clustering for the integrated data. Unexpectedly, batch effects were not corrected between ST and scRNA-seq datasets after data integration (Fig. 6b); nevertheless, the inferred cancer and non-cancer regions were approximately consistent with the previously annotated histological regions (Moncada *et al.*, 2020) (Fig. 6c), wherein several marker genes such as *S100P* and *FSCN1* were identified as DEGs for the putative cancer cluster (Supplementary Note S9).

Next, to investigate complex cell state composition of the PDAC tissue, we used ASURAT to create SSMs from DO, CO, MSigDB and CellMarker databases for cell type and disease, GO database for biological process and KEGG for signaling pathway. Then, we performed unsupervised clustering of the integrated data based on the SSMs for cell type and disease, and biological process (Supplementary Note S9). Remarkably, ASURAT could remove the aforementioned batch effects (Fig. 6d and f) and identify cancer and non-cancer populations (Supplementary Figs S16 and S17). Moreover, we noticed that the spots grouped in the same cluster with PDAC cells are dispersed within the normal pancreas region, which might be a normal pancreas involved in cancer (Fig. 6e and g).

To further investigate the cell states in these spots, we profiled all sign scores for 4282 signs across the tissue (Supplementary Fig.

S18). The sign scores for PDAC were increased in the ST spots approximately matching the reported PDAC region (Moncada *et al.*, 2020), while those for micro RNAs in cancer were increased both in the previously annotated PDAC spots and the newly predicted atypical spots (Fig. 6h). These newly predicted spots were also annotated by a sign for Th17 cell differentiation, suggesting tumor-associated inflammation or antitumor immunity through intercellular communications between Th17 and cancer cells (Muller-Hubenthal *et al.*, 2009), which remains to be elucidated in PDAC (Liu *et al.*, 2019). In more than 90% of PDAC cases, *KRAS* is mutated at the G domain of the 12th residue (Ischenko *et al.*, 2021; Luchini *et al.*, 2020). Hence, we speculated that it is possible to validate our clustering results of cancer and non-cancer spots by comparing the frequencies of *KRAS* mutations using ST data. Unfortunately, we were unable to detect any read mapped to the specific reported region, possibly owing to the shallow read depth and inherent 3′ bias present in the data. Simultaneous genetic and transcriptional profiling may address this problem in the future (Lee *et al.*, 2020).

4 Discussion

We developed ASURAT, an original computational tool for simultaneous cell clustering and biological interpretation using database-derived functional terms. ASURAT performs correlation graph decompositions of functionally annotated gene sets to define multiple biological terms, termed signs. The notions of SCG and VCG are critical for capturing complex correlation structures compared with existing methods such as PAGODA2 (Fan *et al.*, 2016) and ssGSEA (Subramanian *et al.*, 2005) (Supplementary Note S10 and Fig. S19). ASURAT then transforms scRNA-seq data into SSMs, whose rows and columns represent signs and samples (cells), respectively. This SSM plays a key role in characterizing individual cells by various biological terms. Applying ASURAT to several single-cell and ST datasets for PBMCs, SCLC and PDAC, we robustly reproduced the previously reported blood cell types and identified putative subtypes of chemoresistant SCLC and distinct regions within the PDAC tissue.

ASURAT uses database-derived biological terms for clustering single-cell transcriptomes, which inevitably introduces annotation bias (Supplementary Note S11); some biological terms are associated with many genes, whereas others are associated with only a few genes (Gaudet and Dessimoz, 2017). Moreover, in some cases, there might be no functional category for a cell type of interest. Nevertheless, users still have a means to manually characterize cells using combinations of significant signs in different functional categories as well as DEGs. Automatic curation will be a potential addition to ASURAT in the future.

Conventionally, single-cell transcriptomes are analyzed and interpreted by means of unsupervised clustering followed by manual curation of marker genes selected from DEGs, which has been a common bottleneck of gene-based analyses (Andrews *et al.*, 2021; Aran *et al.*, 2019). The statistical significance of individual genes, typically defined by *P*-value or fold change, is dependent on clustering results. ASURAT can provide an alternative approach and demonstrates superior performance for identifying functional subtypes even within a fairly homogeneous population such as isolated cancer cells. In practice, complementing ASURAT with existing methods (Butler *et al.*, 2018; La Manno *et al.*, 2018) will provide a more comprehensive understanding of single-cell and STs, shedding light on putative transdifferentiation of neuroendocrine cancers (Balanis *et al.*, 2019; Kubota *et al.*, 2020), intercellular communication in tumor immune microenvironments (Maynard *et al.*, 2020) and virus infection on immune cell populations.

In omics data analyses, knowledge databases are used to interpret computational results: pathway and motif enrichment analyses are often used for transcriptomic and epigenomic analyses (McLeay and Bailey, 2010; Reimand *et al.*, 2019). In contrast, we propose a unique computational workflow, in which such databases are used for simultaneous clustering and biological interpretation by defining signs. This framework is potentially applicable to any multivariate data with variables linked with annotation information. We can also find such datasets in studies of T cell receptor sequencing (De Simone *et al.*, 2018; Rempala *et al.*, 2011) along with a pan-immune repertoire (Zhang *et al.*, 2020). We anticipate that ASURAT will allow for the identification of various inter-sample differences among T cell receptor repertoires in terms of cellular subtype, antigen–antibody interaction, genetic and pathological backgrounds.

Since ASURAT can create multivariate data (i.e. SSMs) from multiple signs, ranging from cell types to biological functions, it will be valuable to consider graphical models of signs, from which we may infer conditional independence structures. A non-Gaussian Markov random field theory is one of the most promising approaches to address this problem, although requires a large number of samples for achieving true graph edges (Morrison *et al.*, 2017). As the increase in size and diversity of the available data, biological interpretation will become increasingly important. Hence, future work should improve methods for prioritizing biological terms more efficiently than manual screening. We believe that ASURAT will greatly expand our understanding of various biological data and open new means of general functional annotation-driven data analysis.

Data availability

The PBMCs datasets from healthy donors are available in the 10x Genomics repository at https://support.10xgenomics.com/single-cell-gene-expression/datasets: '4k PBMCs from a Healthy Donor' and '6k PBMCs from a Healthy Donor'. The PBMCs datasets from control and sepsis donors are available in the Broad Institute Single Cell Portal at https://singlecell.broadinstitute.org/single_cell: SCP548, which are referenced in Reyes *et al.* (2020). The SCLC and PDAC datasets are available in Gene Expression Omnibus with accession codes GSM4104164 and GSM3036909, GSM3036910, GSM3036911, GSM3405527, GSM3405528, GSM3405529 and GSM3405530, which are referenced in Stewart *et al.* (2020) and Moncada *et al.* (2020), respectively. All the data analyzed in this article are available at Github (https://github.com/keita-iida/ASURATBI) and figshare (https://doi.org/10.6084/m9.figshare.19200254.v4).

Acknowledgements

The authors thank Takeya Kasukawa for the comments that improved the analysis pipeline.

Funding

K.I. was supported by the JSPS KAKENHI [20K14361]. J.N.W. was supported by the Honjo International Scholarship Foundation. K.I., J.K. and M.I. were supported by the Shin Bunya Kaitaku Shien Program of Institute for Protein Research, Osaka University. M.O. was supported by the JSPS KAKENHI [17H06299, 17H06302 and 18H04031; JST-Mirai program number JPMJMI19G7; and JST CREST program number JPMJCR21N3]. M.O. and M.I. were supported by the P-CREATE, Japan Agency for Medical Research and Development. M.O. and K.I. were supported by the JST Moonshot R&D [JPMJMS2021].

Conflict of Interest: none declared.

References

Andrews,T.S. *et al.* (2021) Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.*, **16**, 1–9.

Aran,D. *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172.

Balanis,N.G. *et al.* (2019) Pan-cancer convergence to a small-cell neuroendocrine phenotype that shares susceptibilities with hematological malignancies. *Cancer Cell*, **36**, 17–34.e7.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.

Bodenhofer,U. *et al.* (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics*, **27**, 2463–2464.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
Cancer Genome Atlas Research Network. *et al.* (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**, 378–384.
Cao,Y. *et al.* (2020) SCSA: a cell type annotation tool for single-cell RNA-seq data. *Front. Genet.*, **11**, 490.
Chen,H.J. *et al.* (2019) Generation of pulmonary neuroendocrine cells and SCLC-like tumors from human embryonic stem cells. *J. Exp. Med.*, **216**, 674–687.
Chen,Z. *et al.* (2020) Ligand-receptor interaction atlas within and between tumor cells and T cells in lung adenocarcinoma. *Int. J. Biol. Sci.*, **16**, 2205–2219.
Coifman,R.R. and Lafon,S. (2006) Diffusion maps. *Appl. Comput. Harmon. Anal.*, **21**, 5–30.
Couper,P. (2015) *A Student's Introduction to Geographical Thought: Theories, Philosophies, Methodologies*. SAGE Publications, Inc., Los Angeles.
De Simone,M. *et al.* (2018) Single cell T cell receptor sequencing: techniques and future challenges. *Front. Immunol.*, **9**, 1638.
Devitt,K. *et al.* (2019) Single-cell RNA sequencing reveals cell type-specific HPV expression in hyperplastic skin lesions. *Virology*, **537**, 14–19.
Diehl,A.D. *et al.* (2016) The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*, **7**, 44.
Dominguez,D. *et al.* (2016) A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell Res.*, **26**, 946–962.
Elosua-Bayes,M. *et al.* (2021) SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.*, **49**, e50.
Fan,J. *et al.* (2016) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods*, **13**, 241–244.
Gaudet,P. and Dessimoz,C. (2017) Gene ontology: pitfalls, biases, and remedies. *Methods Mol. Biol.*, **1446**, 189–205.
Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587. e3529.
Hyvarinen,A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, **10**, 626–634.
Ireland,A.S. *et al.* (2020) MYC drives temporal evolution of small cell lung cancer subtypes by reprogramming neuroendocrine fate. *Cancer Cell*, **38**, 60–78.e12.
Ischenko,I. *et al.* (2021) KRAS drives immune evasion in a genetic model of pancreatic cancer. *Nat. Commun.*, **12**, 1482.
Jalili,M. *et al.* (2021) Exploring the metabolic heterogeneity of cancers: a benchmark study of context-specific models. *J. Pers. Med.*, **11**, 496.
Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kim,N. *et al.* (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.*, **11**, 2285.
Kiselev,V.Y. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
Kiselev,V.Y. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.
Kubota,S. *et al.* (2020) Dedifferentiation of neuroendocrine carcinoma of the uterine cervix in hypoxia. *Biochem. Biophys. Res. Commun.*, **524**, 398–404.
La Manno,G. *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.
Lee,J. *et al.* (2020) Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.*, **52**, 1428–1442.
Li,H. *et al.* (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**, 708–718.
Liu,X. *et al.* (2019) The reciprocal regulation between host tissue and immune cells in pancreatic ductal adenocarcinoma: new insights and therapeutic implications. *Mol. Cancer*, **18**, 184.
Luchini,C. *et al.* (2020) KRAS wild-type pancreatic ductal adenocarcinoma: molecular pathology and therapeutic opportunities. *J. Exp. Clin. Cancer Res.*, **39**, 227.
Lun,A.T. *et al.* (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.
Maynard,A. *et al.* (2020) Therapy-Induced evolution of human lung cancer revealed by Single-Cell RNA sequencing. *Cell*, **182**, 1232–1251.e22.
McLeay,R.C. and Bailey,T.L. (2010) Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.
Moncada,R. *et al.* (2020) Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.*, **38**, 333–342.
Morrison,R.E. *et al.* (2017) Beyond normality: learning sparse probabilistic graphical models in the non-Gaussian setting. In: *Advances in Neural Information Processing Systems*, Vol. **30**, pp. 2356–2366.
Muller-Hubenthal,B. *et al.* (2009) Tumour biology: tumour-associated inflammation versus antitumor immunity. *Anticancer Res.*, **29**, 4795–4805.
Murtagh,F. and Legendre,P. (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classif.*, **31**, 274–295.
Parra,R.G. *et al.* (2019) Reconstructing complex lineage trees from scRNA-seq data using MERLoT. *Nucleic Acids Res.*, **47**, 8961–8974.
Pasquini,G. *et al.* (2021) Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.*, **19**, 961–969.
Reimand,J. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.*, **14**, 482–517.
Rempala,G.A. *et al.* (2011) Model for comparative analysis of antigen receptor repertoires. *J. Theor. Biol.*, **269**, 1–15.
Reyes,M. *et al.* (2020) An immune-cell signature of bacterial sepsis. *Nat. Med.*, **26**, 333–340.
Saxena,V. *et al.* (2006) Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res.*, **34**, e151.
Schubert,E. and Rousseeuw,P.J. (2019) Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In: *International Conference on Similarity Search and Applications 2019*, Vol. 11807, pp. 171–187.
Schwendenwein,A. *et al.* (2021) Molecular profiles of small cell lung cancer subtypes: therapeutic implications. *Mol. Ther. Oncolytics*, **20**, 470–483.
Shao,X. *et al.* (2020) scCATCH: automatic annotation on cell types of clusters from Single-Cell RNA sequencing data. *iScience*, **23**, 100882.
Sharma,N.K. *et al.* (2019) Lipid metabolism impairment in patients with sepsis secondary to hospital acquired pneumonia, a proteomic analysis. *Clin. Proteomics*, **16**, 29.
Stewart,C.A. *et al.* (2020) Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat. Cancer*, **1**, 423–436.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
Villani,A.C. *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.
Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
Yu,G. *et al.* (2015) DOSE: an R/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.
Zhang,A.W. *et al.* (2019) Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods*, **16**, 1007–1015.
Zhang,W. *et al.* (2020) PIRD: pan immune repertoire database. *Bioinformatics*, **36**, 897–903.
Zhang,X. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.