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

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

[1] Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

[2] Department of Biochemistry, Osaka International Cancer Institute, Osaka 541-8567, Japan

[3] Department of Clinical Bio-resource Research and Development, Graduate School of Medicine Kyoto University, Kyoto 606-8501, Japan

[4] Center for Drug Design and Research, National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka, 567-0085, Japan

[*] Correspondence: kiida@protein.osaka-u.ac.jp

Summary

Single-cell RNA sequencing (scRNA-seq) analysis has advanced our knowledge of the transcriptional states of cells. By analyzing scRNA-seq data, one can deconvolve cell states into thousands of gene expression profiles and perform cell clustering. However, conventional gene-based analyses depend on manual literature searches, which are difficult to interpret with clustering results. Here, we present an analysis pipeline named ASURAT, which simultaneously performs unsupervised clustering and biological interpretation in terms of cell type, disease, biological process, and signaling pathway activity. ASURAT uses a correlation graph-based decomposition of functional gene set to define multiple biological terms, which plays a key role in cell state inference. We apply ASURAT to several public scRNA-seq and spatial transcriptome datasets for human peripheral blood mononuclear cells, small cell lung cancer, and pancreatic ductal adenocarcinoma. Our hybrid approach using the existing methods and ASURAT identifies functional subpopulations, adding another layer of biological interpretability to conventional gene-based analyses.

Introduction

Single-cell RNA sequencing (scRNA-seq) has profoundly advanced our knowledge of cells, owing to its immense potential for discovering the transcriptional principles governing cell fates at the single-cell level (La Manno et al., 2018). scRNA-seq has been widely used to improve our understanding of individual cells (Ganesh et al., 2020), intra- and intertumoral heterogeneity (Stewart et al., 2020), intercellular interaction (Chen et al., 2020), tumorigenesis (Chen et al., 2019), drug resistance (Maynard et al., 2020; Stewart *et al.*, 2020), and the effects of viral infection on immune cell populations (Devitt et al., 2019). Various clustering methods, wherein cells are partitioned according to transcriptome-wide similarity, have been proposed (Pasquini et al., 2021) and applied to cell type annotation (Kim et al., 2020). However, interpreting single-cell data remains challenging (Andrews et al., 2021; Aran et al., 2019; Gao et al., 2019; Kiselev et al., 2019; Lahnemann et al., 2020).

Conventionally, single-cell transcriptomes are analyzed and interpreted by means of unsupervised clustering followed by manual literature searches of marker genes chosen from a large number of differentially expressed genes (DEGs) (Andrews *et al.*, 2021; Lahnemann *et al.*, 2020). Today, several computational tools for cell type inference, such as Garnett (Pliner et al., 2019), SCSA (Cao et al., 2020), and scCATCH (Shao et al., 2020), are available to assist manual annotation, as detailed in the review by Pasquini *et al*. (2021). However, this is often difficult because genes are generally associated with multiple biological functions (Cancer Genome Atlas Research et al., 2017). In cancer transcriptomics, this difficulty is exacerbated by the interdependence between disease-related biomarker genes and biological function terms; furthermore, expression levels of these marker genes can be heterogeneous depending on cancer microenvironments (Moore et al., 2019).

A possible solution is to realize cell clustering and biological interpretation at the same time. Recently, reference-based analysis has been applied in single-cell transcriptomics (Aran *et al.*, 2019; Cao *et al.*, 2020; Li et al., 2017). One such technique is reference component analysis (RCA), which is used for accurate clustering of single-cell transcriptomes along with cell-type annotation based on similarity to reference transcriptome panels (Li *et al.*, 2017). However, these methods require well-characterized transcriptomes with purified cells, which may be difficult to apply to ambiguous phenotypes. Another approach is using supervised classification (Gao *et al.*, 2019) combined with gene set enrichment analysis, incorporating biological knowledge and functions such as signaling pathway activity; hence, it may improve the interpretability over signature gene-based approaches, which place sole emphasis on individual roles of genes (Fan et al., 2016). However, we still lack a prevailing theory leveraging this information at the single-cell level.

To overcome the aforementioned limitations, a novel theoretical paradigm providing biological interpretations to computational results is needed. Thus, we propose a novel analysis pipeline named ASURAT (functional annotation-driven unsupervised clustering of single-cell transcriptomes), which simultaneously performs unsupervised cell clustering and biological interpretation in terms of cell type, disease, biological process, and signaling pathway activity. In this paper, we demonstrate the clustering performance of ASURAT using several published scRNA-seq and spatial transcriptome datasets for human peripheral blood mononuclear cells (PBMCs), small cell lung cancer (SCLC), and pancreatic ductal adenocarcinoma (PDAC). We show that ASURAT can greatly improve functional understandings of single-cell transcriptomes, adding a new layer of biological interpretability to conventional gene-based analyses.

Results

Overview of ASURAT's workflow

ASURAT was developed to simultaneously cluster and interpret single-cell transcriptomes using functional gene sets (FGSs) (**Figure 1**), which was implemented in the R programming language. FGSs are collected from knowledge-based databases (DBs) for cell type, disease, biological process, signaling pathway activity and so on (**Figure 1b**, Methods). Then, ASURAT creates multiple biological terms using FGSs (**Figure 1c**), which we later explain using a correlation graph-based decomposition. This is conceptually important in that we aim to formulate a "biological meaning" by following the method in Saussure's semiology, introducing the philosophical definition of *signe* as the "association between signifier and signified." We call such new biological terms signs (see Methods for definition). Finally, ASURAT transforms scRNA-seq data into a sign-by-sample matrix (SSM), in which rows and columns stand for signs and samples (cells), respectively (**Figure 1c**, Methods). By analyzing SSMs, individual cells can be characterized by various biological terms (**Figure 1d**).

Sign-by-sample matrix (SSM)

By preparing a normalized-and-centered read count table, a correlation matrix R among all the genes, and a set of pairs of biological description and FGS $\{(T_k, \Omega_k) | k = 1, 2, \cdots, q\}$ defined in an imported DB, we divide the correlation graph of Ω_k into several parts and obtain subsets $\Omega_k^{(j)} \subset \Omega, j = 1, 2, \ldots, m_k$. Here, a triplet $(T_k, \Omega_k^{(j)}, R)$ is termed a sign, in particular (T_k, Ω_k, R) a parent sign (see Methods for definition). In many applications, high correlations are expected to yield rich information. Hence, we divide Ω_k into the following three subsets (**Figure 2**): (i) a strongly correlated gene set (SCG), which is a set of genes with strong positive correlations with each other; (ii) a variably correlated gene set (VCG), which is a set of genes with strong negative correlations with genes in SCG; and (iii) a weakly correlated gene set (WCG), which is a set of genes with weak correlations with each other (see Methods for definition). ASURAT creates an SSM

for SCG by weighted averaging of normalized-and-centered gene set expression levels of SCGs and WCGs. Similarly, an SSM for VCG is created from VCGs and WCGs. Then, by vertically concatenating SSMs for SCGs and VCGs, we created a single SSM, which displays the degrees that cells in columns have characteristics of signs in rows, whose entries are termed sign scores.

Unsupervised clustering of SSMs

One focus of analyzing SSMs is to cluster cells and find significant signs (**Figure 1d**), where "significant" means that the sign score is specifically upregulated or downregulated at the cluster level. 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 (see Methods). To cluster cells, we use two strategies. The first is unsupervised clustering, such as Partitioning Around Medoids (PAM), hierarchical, and graph-based clustering with and without principal component analysis (PCA). The second is a method of extracting a continuous tree-like topology using diffusion map (Coifman and Lafon, 2006), followed by allocating cells to different branches of the data manifolds (Parra et al., 2019). Choosing an appropriate strategy depends on the biological context, but the latter is usually applied for developmental processes or time-course experimental data, which are often followed by pseudotime analyses.

Clustering single-cell transcriptomes of PBMCs

To validate the clustering performance of ASURAT, we analyzed two public scRNA-seq datasets (see Datasets), namely the PBMC 4k and 6k datasets, in which the cell types were inferred using computational tools based on prior assumptions (Cao *et al.*, 2020). We first excluded low-quality genes and cells and attenuated technical biases with respect to zero inflation and variation of capture efficiencies between cells using bayNorm (Tang

et al., 2020). The resulting read count tables were supplied to ASURAT and four other methods: scran (version 1.18.7) (Lun et al., 2016), Seurat (version 4.0.2) (Hao et al., 2021), Monocle 3 (version 1.0.0) (Trapnell et al., 2014), and SC3 (version 1.18.0) (Kiselev et al., 2017). To infer existing cell types and the population ratios in the PBMC 4k and 6k datasets, we implemented the existing methods using close to the default settings, performed cell clustering, and annotated each cluster by manually investigating DEGs, where we defined DEGs as genes with false discovery rates (FDRs) $< 10^{-99}$ (see Methods for statistical test). When using ASURAT, we performed unsupervised cell clustering and semi-automatic annotation based on SSMs for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (**Figure 3a**).

Among all the existing methods, Seurat and Monocle 3 could robustly reproduce most blood cell type labels, as inferred by Cao *et al*. (2020), while scran and SC3 output many unspecified cells (**Figure 3d**). We found that Seurat's pipeline, followed by manual annotations based only on a couple of DEGs, provided comparable population ratios with previous results (Cao *et al.*, 2020). However, it was quite laborious to find marker genes from numerous DEGs, which tend to increase in number of cells as well as significance levels. Based on the clustering results of Seurat, we assigned the labels (i) T cell, (ii) monocyte, (iii) B cell, and (iv) NK or NKT cell to the cells in PBMC 4k (resp. PBMC 6k) by finding marker genes from (i) 57 (114), (ii) 102 (148), (iii) 49 (33), and (iv) 32 (35) DEGs, respectively. To avoid such a time-consuming process, we implemented scCATCH (version 2.1) (Shao *et al.*, 2020), an automatic annotation tool for the clustering results of Seurat. However, population ratios inferred by Seurat with scCATCH were less consistent than those of Seurat with manual annotations (**Figure 3d**).

Moreover, ASURAT simultaneously performs unsupervised cell clustering and biological interpretation leveraging all defined FGSs, without relying on DEGs. We identified five cell type labels, with none remaining unspecified (**Figure 3b–d**). 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; Wagner, 2018) (**Figure 3d**). Such a small discrepancy was unavoidable, because Cao *et al*. (2020) used author-defined DEGs and preselected cell types to identify the most preferable ones. Unexpectedly, the clustering results using SSMs for GO and KEGG also showed well-separated clusters in two-dimensional Uniform Manifold Approximation and Projection (UMAP) (McInnes and Healy, 2018) spaces (Figure S1), indicating that functional states of cells are also heterogeneous with respect to biological process and signaling pathway activity. These results demonstrate that ASURAT can perform robust clustering for single-cell transcriptomes.

Clustering a single-cell transcriptome of SCLCs

Previous work reported that SCLC tumors undergo a shift from chemosensitivity to chemoresistance against platinum-based therapy (Stewart *et al.*, 2020). Stewart *et al*. (2020) analyzed scRNA-seq data obtained from circulating tumor cell-derived xenografts generated from treatment-naïve lung cancer patients, cultured with vehicle or cisplatin treatments, and reported that the gene expression profiles of the platinum-resistant tumors were more heterogeneous than those of platinum-sensitive tumors. However, the mechanism behind chemoresistance is still unclear, partly because transcriptional heterogeneity is affected by physiological states of cells such as pathological states (Stewart *et al.*, 2020), cell cycle (Dominguez et al., 2016), and metabolic processes (Jalili et al., 2021), which cannot be readily identified by conventional marker gene-based analyses alone. To better understand SCLC subtypes in chemoresistant tumors both by gene- and function-based analyses, we applied Seurat and ASURAT to the published SCLC scRNA-seq data (see Datasets) (Stewart *et al.*, 2020).

First, we investigated 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 *ASCL1* single positive subtype (Figure S2), which is consistent with the previous report (Stewart *et al.*, 2020). After quality controls, the data was normalized by bayNorm (Tang *et al.*, 2020) and the resulting read count table was supplied to the workflows of Seurat and ASURAT (see Methods). To investigate molecular subtypes and potential resistance pathways, we clustered the single-cell transcriptome into three clusters and inferred a cell cycle phase for each cell using Seurat (Hao *et al.*, 2021), as shown in UMAP spaces (**Figure 4e**). We found that the cell populations assigned to G1, S, and G2M phases are sequentially distributed in the UMAP space, indicating that the clustering results are considerably affected by the cell cycle. Then, we identified DEGs for each cluster (Group 1, 2, and 3) and performed KEGG enrichment analysis using clusterProfiler (Yu et al., 2012), but chemoresistance terms were not primarily enriched (**Figure 4f**).

Subsequently, to investigate functional heterogeneities in SCLCs, we used ASURAT while inputting a normalized-and-centered read count table and SSMs for Disease Ontology (DO), GO, and KEGG. Based on the SSM for DO, we performed a dimensionality reduction using diffusion map (Coifman and Lafon, 2006), which showed a tree-like topology. Then, we defined a pseudotime $t \in [0, 1]$ along the branches and clustered the single-cell transcriptome into three clusters using MERLoT (Parra *et al.*, 2019) (**Figure 4c**). Based on pseudotime analysis, we revealed that sign scores for platinum drug resistance (path:hsa01524_S) and PD-L1 expression-mediated immunosuppression (path:hsa05235_S) were upregulated in clusters 2 and 3, respectively. In addition, sign scores for intracellular protein transport (GO:0006886_S), with an FGS including an SCLC malignancy marker *CD24* (Kristiansen et al., 2003), was upregulated in cluster 1 (**Figure 4d**). We noticed that sign scores for hematopoietic system disease (DOID:74_S) were moderately upregulated in cluster 1 (separation index$\sim$0.38), which was supported by previous work reporting that hematopoietic cancers are similar to SCLCs in terms of gene expression profiles and drug sensitivities (Balanis et al., 2019).

Although the SCLC molecular subtypes have been extensively studied (Chen *et al.*, 2019; Ireland *et al.*, 2020; Schwendenwein et al., 2021; Wooten et al., 2019; Yatabe, 2020), data regarding the functional subtypes of *ASCL1*-positive SCLC remain limited. To identify *de novo* SCLC subtypes, future work should validate the clustering results.

Finally, we vertically concatenated all the SSMs, cell cycle phases, and expression matrices to characterize individual cells from multiple biological aspects, as shown by the heatmaps along with ASURAT's clustering result (**Figure 4a, b**). As shown, we were able to simultaneously perform unsupervised clustering and biological interpretation of single-cell transcriptomes. Moreover, our results provided new insights into SCLCs by adding a layer of DEGs using multiple Mann-Whitney *U* tests (**Figure 4a**), showing that most DEGs had been previously overlooked (Chen *et al.*, 2019; Ireland *et al.*, 2020; Schwendenwein *et al.*, 2021; Wooten *et al.*, 2019; Yatabe, 2020). Taken together, our findings provide a novel clue for the clinical improvements for relapsed SCLC tumors.

The most time-consuming step is finalizing the set of signs by tuning ASURAT's parameters to obtain information related to the purpose. Here, users may face difficulty in prioritizing the importance of several signs. In this study, we found that the sign scores for meningioma, mesothelioma, and other diseases were also upregulated in cluster 2, but their actual relationships to those diseases in patients were unknown. Nevertheless, ASURAT may help us find well-structured data manifolds and characterize cells in explainable manners.

Clustering a spatial transcriptome (ST) of PDAC tumor

To widen ASURAT's applicability to spatial transcriptome (ST) data, we analyzed the published scRNA-seq and ST datasets obtained from human PDAC patients (Moncada et al., 2020), in which they reported that various cell types, including pancreatic, immune, and cancer cells, are spatially distributed in the distinct tissue regions of the primary

PDAC tumors, and that cancer cells are accompanied by inflammatory fibroblasts, M1 macrophage, and other cell types. Since the cellular resolutions of the STs were estimated to be 20–70 cells per ST spot, which is far lower than that of scRNA-seq, computational methods have been proposed to predict cell type architectures by integrating ST and scRNA-seq datasets (Elosua-Bayes et al., 2021; Moncada *et al.*, 2020). Here, we aimed to cluster STs and add a new layer of functional annotation to PDAC tissue sections using Seurat's canonical correlation analysis (CCA)-based integration method (Butler et al., 2018) and ASURAT's unsupervised clustering. For this purpose, we obtained published ST (PDAC-A ST1) and scRNA-seq (PDAC-A inDrop from 1 to 6) datasets (Moncada *et al.*, 2020) (see Datasets).

First, we combined all the scRNA-seq datasets after confirmation of minimal batch effects (Figure S3). Then, the ST and scRNA-seq data were normalized by bayNorm (Tang *et al.*, 2020) and the resulting read count tables were supplied to Seurat (see Methods). To cluster the ST with reference to the scRNA-seq data, we performed a CCA-based data integration, followed by an unsupervised clustering of the integrated transcriptome, which is shown in the UMAP spaces and tissue image (**Figure 5a**). Unexpectedly, batch effects were not corrected between ST and scRNA-seq datasets using Seurat's CCA-based integration method; nevertheless, the inferred cancer and non-cancer regions were approximately consistent with previously annotated histological regions (Elosua-Bayes *et al.*, 2021; Moncada *et al.*, 2020), wherein several marker genes such as *REG1A*, *S100A4* and *TM4SF1*, and *CELA2A* were identified as DEGs for clusters 2, 3, and 5, respectively (FDRs$< 10^{-80}$, Mann-Whitney *U* tests).

Next, we input the ST and scRNA-seq integrated transcriptome into ASURAT's workflow (see Methods). To investigate complex PDAC tissues, we made a custom-built cell-type database (CB) by simply merging the DO and CO DBs, CellMarker (Zhang et al., 2019), MSigDB (Subramanian et al., 2005), and PanglaoDB (Franzen et al., 2019).

Based on the SSM for CB, which was computed from the integrated transcriptome, we performed a dimensionality reduction using PCA and clustered the SSM by *k*-nearest neighbor (KNN) graph generation and the Louvain algorithm, which is shown in UMAP spaces and the tissue image (**Figure 5b**). Remarkably, our hybrid approach using Seurat and ASURAT was able to remove the aforementioned batch effects and cluster the ST spots into three clusters. Our clustering result was further substantiated by cell-cycle phase inference, from which we found that G2M-phase spots were broadly distributed while G1- and S-phase spots were located in the clustered regions (**Figure 5c**).

Then, we investigated functional heterogeneities in the ST using SSMs for CB, GO, and KEGG, computed from the integrated transcriptome. We identified distinct tissue regions using multiple signs, in terms of cell type, disease, biological process, and signaling pathway activity (**Figure 5d**). In particular, we revealed that sign scores for PDAC (DOID:3498_S), which has a functional gene set including PDAC markers such as *S100P* and *MMP1*, were upregulated in the clustered ST spots approximately matching the reported cancer region (Moncada *et al.*, 2020). Similarly, the reported normal pancreatic tissue region, in which endocrine cells were activated (Moncada *et al.*, 2020), was annotated by a sign for pancreatic secretion (path:hsa04972_S). These results demonstrated how ASURAT can also be applied to ST data analyses.

Macrophages play tumor-promoting roles in PDAC, including chronic inflammation and extracellular matrix (ECM) remodeling (Poh and Ernst, 2021). Our analysis showed that the PDAC region (cluster 3) was enriched with signs for macrophages (CBID:434_S) and ECM-receptor interactions (path:hsa04512_S) (**Figure 5d**), while the gene expression of inflammatory cytokines, reported by Poh and Ernst (2021), was rarely detected because of the poor expression levels. Thus, our results partly supported the claim concerning macrophage involvement in the PDAC microenvironment (Elosua-Bayes *et al.*, 2021; Moncada *et al.*, 2020), but the chronic inflammation should be further validated.

Finally, previous studies suggested that M1 macrophages exist in stromal and cancer regions (corresponding approximately clusters 2 and 3 in **Figure 5b**, respectively) while M2 macrophages exist in duct epithelium (undetected) (Moncada *et al.*, 2020). However, we could not detect such subpopulations due to potential limitations in our hybrid approach. Seurat's CCA-based integration method requires preselections of highly variable genes from both scRNA-seq and ST datasets. Subsequently, genes are further filtered in ASURAT's workflow, in which FGSs are defined from common genes in the integrated transcriptome and imported DBs. As a result, a considerable number of genes are removed, which may result in undesirable exclusions of important biological terms. We expect that future work can solve this problem by improving CCA-based data integration methods.

Discussion

We have developed ASURAT, a novel analysis pipeline for simultaneous cell clustering and biological interpretation using FGSs (**Figure 1**). ASURAT begins by performing a correlation graph-based decomposition of FGS to define multiple biological terms, termed signs (**Figure 2**). Importantly, our formulation of signs is derived from the philosophical concept of *signe* in Saussure's semiology, which was originally defined as the "association between signifier and signified." ASURAT then transforms scRNA-seq data into an SSM, which shows the degrees that the cells in columns have the characteristics of the signs in rows. This SSM plays a key role in characterizing individual cells from multiple aspects of their meaning. By applying ASURAT to several scRNA-seq and spatial transcriptome datasets for PBMCs, SCLC, and PDAC, we robustly reproduced the previously reported blood cell types (**Figure 3**), identified putative subtypes of chemoresistant SCLC (**Figure 4**), and identified distinct tissue regions to the PDAC tissue (**Figure 5**).

Conventionally, single-cell transcriptomes are analyzed and interpreted by means of unsupervised clustering followed by manual literature searches of marker genes chosen from a large number of DEGs, which has been a common bottleneck of signature gene-based analyses (Andrews *et al.*, 2021; Aran *et al.*, 2019; Gao *et al.*, 2019). The statistical significance of individual genes, typically defined by *p*-value or fold change, is dependent on clustering results, which are also affected by various physiological states of cells (Dominguez *et al.*, 2016; Jalili *et al.*, 2021). Here, we expect that ASURAT provides an alternative approach using FGSs and demonstrates superior performance for identifying functional subtypes even within a fairly homogeneous population such as isolated cancer cells. In practice, complemental usages of ASURAT and existing methods (Butler *et al.*, 2018; La Manno *et al.*, 2018) will provide more comprehensive understanding of single-cell and spatial transcriptomes, helping us shed 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 (Devitt *et al.*, 2019).

In omics data analyses, knowledge-based DBs are used to interpret computational results: GO, KEGG pathway, and motif enrichment analyses are often used for transcriptomic and epigenomic analyses (McLeay and Bailey, 2010; Mootha et al., 2003; Reimand et al., 2019). In contrast, we propose a unique analysis workflow, in which such DBs are used for simultaneous clustering and biological interpretation by defining signs from scRNA-seq data and FGSs. 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 database (Zhang et al., 2020). We anticipate that ASURAT will make it possible to identify various inter-sample differences among T cell receptor repertoires in terms of cellular subtype, antigen-antibody interaction, genetic and pathological backgrounds.

Finally, future challenges in data-driven mathematical analysis are worth noting. 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, which infer conditional independence structures. A non-Gaussian Markov random field theory (Morrison et al., 2017) is one of the most promising approaches to this problem, but it requires quite a large number of samples for achieving true graph edges (Morrison *et al.*, 2017). As available data expand in size and diversity, biological interpretation will become increasingly important. Hence, future work should improve methods for prioritizing biological terms more efficiently than manual screening. We hope extending ASURAT will greatly facilitate our intuitive understanding of various biological data and open new means of general functional annotation-driven data analysis.

Limitations of the study

To formulate signs, we used a correlation graph-based decomposition based on FGSs setting thresholds of positive and negative correlation coefficients (**Figure 2**), from which we obtain SCGs, VCGs, and WCGs (see Methods). Although this method is intuitive and easy to use, such three-part decomposition might be insufficient in some cases. For example, one cannot divide the FGS for DO term "lung small cell carcinoma" (DOID 5409) into more than three parts, while SCLC can be classified into at least four molecular subtypes (Schwendenwein *et al.*, 2021; Yatabe, 2020). Therefore, development of a more flexible method for dividing the correlation graphs is warranted.

Signs are derived from information in existing DBs. This inevitably introduces bias problems, such as the inherent incompleteness of the DBs and annotation bias, *viz*. some biological terms are associated with many genes, while others are associated with few (Gaudet and Dessimoz, 2017). To overcome this problem, one should monitor what signs are included during data processing (**Figure 1a**) and carefully tune the parameters to select reliable signs (Figure S4). Our R scripts help users perform this process (see Supplementary Notes).

Acknowledgments

We thank Takeya Kasukawa for comments that greatly improved the analysis pipeline. K.I. was supported by JSPS KAKENHI Grant No. 20K14361. K.I., J.K., and M.I. were supported by Shin Bunya Kaitaku Shien Program of Institute for Protein Research, Osaka University. M.O. was supported by JSPS KAKENHI Grant No. 17H06299, 17H06302, and 18H04031, and JST-Mirai program No. JPMJMI19G7. M.O. and M.I. were supported by P-CREATE, Japan Agency for Medical Research and Development.

Author contributions

Conceptualization, K.I.; Methodology, K.I.; Formal Analysis, K.I. and J.N.W.; Investigation, all the authors; Writing, all the authors; Supervision, M.O., M.I., and K.I.; Project Administration, M.O.; Funding Acquisition, M.O., M.I., J.K., and K.I.

Declaration of interests

The authors declare no competing interests.

Figures

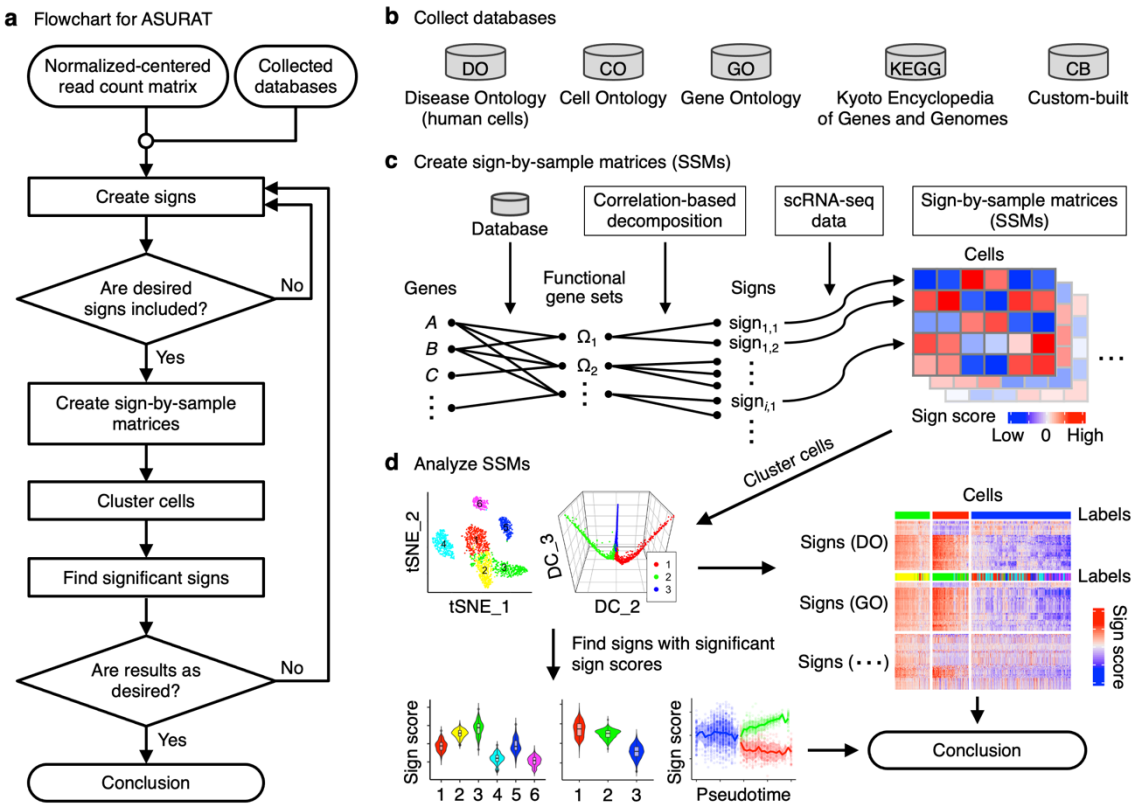


Figure 1. Workflow of ASURAT

(a) Flowchart of the procedures. (b) Collection of knowledge-based databases (DBs). (c) Creation of sign-by-sample matrices (SSMs) from normalized and centered read count matrix and the collected DBs. (d) Analysis of SSMs to infer cell types, diseases, and biological functions.

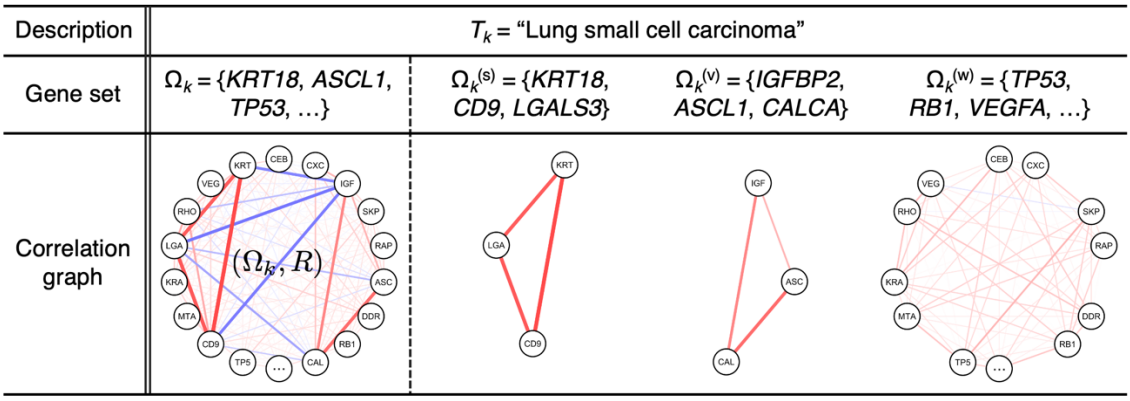


Figure 2. A representation of correlation graph-based decomposition

From single-cell RNA-seq data and a Disease Ontology (DO) term with DOID 5409, which concerns small cell lung cancer, three signs $(T_k, \Omega_k^{(j)}, R)$, $j \in \{\mathrm{s}, \mathrm{v}, \mathrm{w}\}$, 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, $\Omega_k^{(\mathrm{s})}$, $\Omega_k^{(\mathrm{v})}$, and $\Omega_k^{(\mathrm{w})}$, respectively. Red and blue edges in correlation graphs indicate positive and negative correlations, respectively; color density indicates the strength of the correlation.

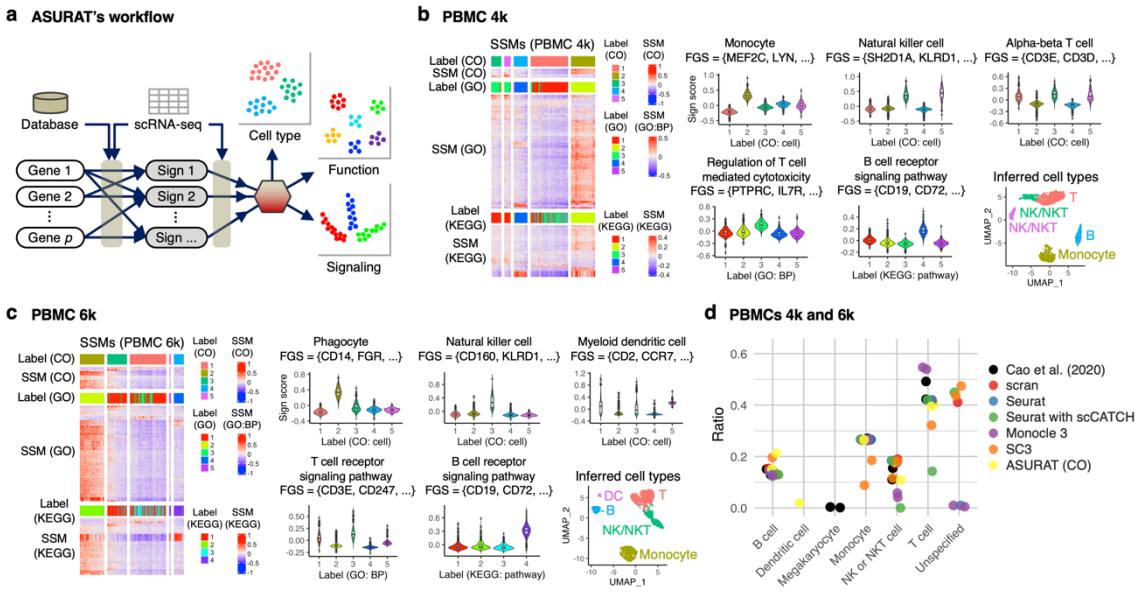
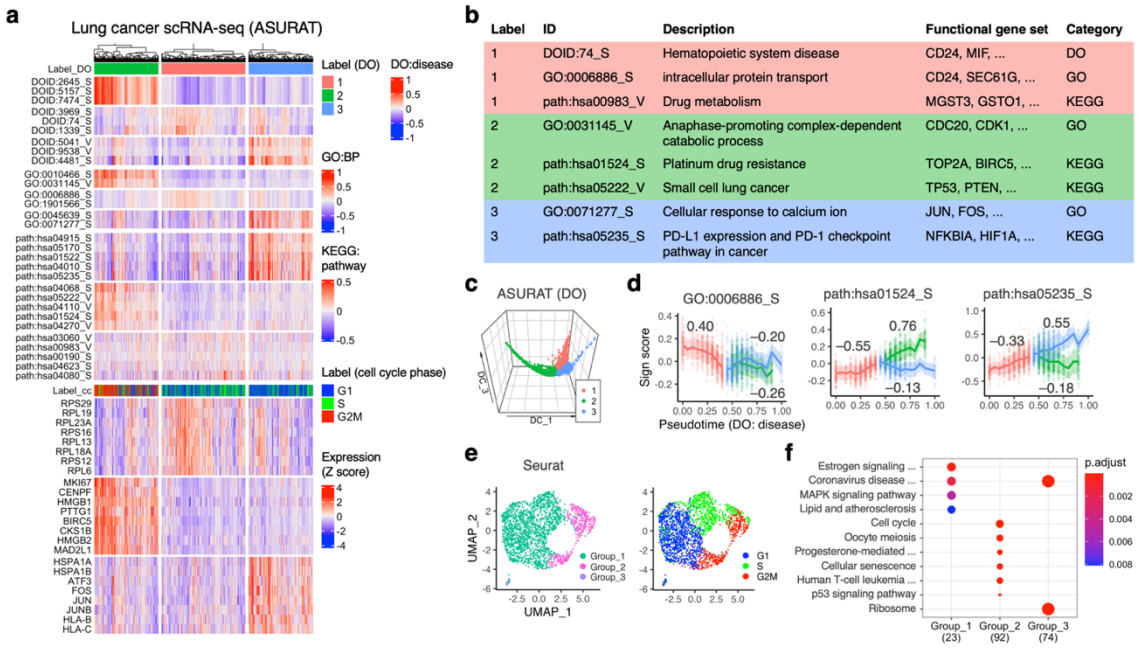


Figure 3. Clustering single-cell transcriptomes of peripheral blood mononuclear cells (PBMCs)

(a) Schematic illustration of ASURAT's workflow. (b, c) Identification of cell types from analyses of sign-by-sample matrices (SSMs) for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG). According to heatmaps and violin plots, showing representative signs and the functional gene sets (FGSs), T cell ("T"), B cell ("B"), NK or NKT cell ("NK/NKT"), monocyte, and dendritic cell ("DC") were identified as shown in Uniform Manifold Approximation and Projection (UMAP) plots. (d) Population ratios in the PBMC 4k and 6k datasets predicted by seven different methods.



Label	ID	Description	Functional gene set	Category
1	DOID:74_S	Hematopoietic system disease	CD24, MIF, ...	DO
1	GO:0006886_S	intracellular protein transport	CD24, SEC61G, ...	GO
1	path:hsa00983_V	Drug metabolism	MGST3, GSTO1, ...	KEGG
2	GO:0031145_V	Anaphase-promoting complex-dependent catabolic process	CDC20, CDK1, ...	GO
2	path:hsa01524_S	Platinum drug resistance	TOP2A, BIRC5, ...	KEGG
2	path:hsa05222_V	Small cell lung cancer	TP53, PTEN, ...	KEGG
3	GO:0071277_S	Cellular response to calcium ion	JUN, FOS, ...	GO
3	path:hsa05235_S	PD-L1 expression and PD-1 checkpoint pathway in cancer	NFKBIA, HIF1A, ...	KEGG

Figure 4. Clustering a single-cell transcriptome of small cell lung cancers

(a) Heatmaps showing (i) clustering results of ASURAT, (ii) sign scores of sign-by-sample matrices (SSMs) for Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), and (iii) scaled gene expression levels, which are concatenated vertically. Here, only the most significant signs and differentially expressed genes (DEGs) for ASURAT's clusters are shown. (b) Representative signs from (a). (c) Diffusion map of the SSM for DO, projected onto the first three coordinates. (d) Sign scores for the indicated IDs along the pseudotime, in which the standard deviations are shown by the shaded area. The value on each plot stands for the separation index for a given group versus all the others. The clustering labels are consistent with those in (a) and (b). (e) Clustering results and cell cycle phases computed by Seurat. (f) KEGG pathway enrichment analysis based on DEGs for Seurat's clusters in (e).

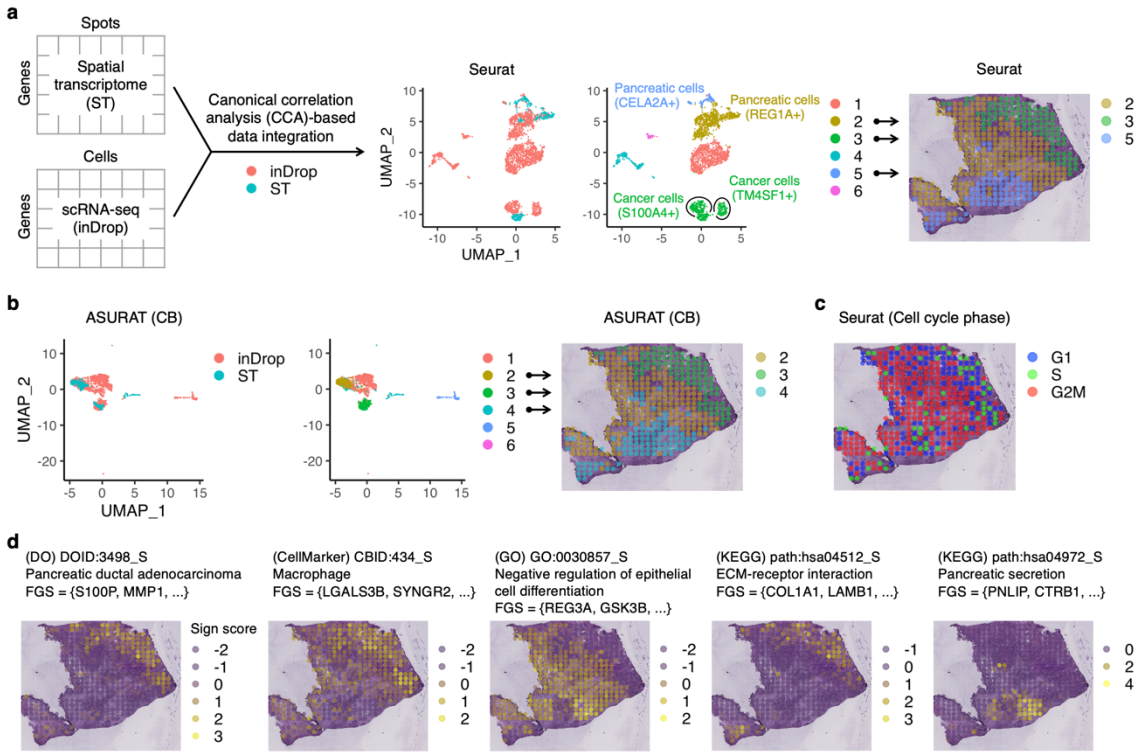


Figure 5. Clustering a spatial transcriptome (ST) of pancreatic ductal adenocarcinoma (PDAC) tumor

(a) Canonical correlation analysis (CCA)-based data integration of scRNA-seq and ST datasets using Seurat, followed by unsupervised clustering of the integrated data. Cells are manually labeled according to the indicated DEGs in Uniform Manifold Approximation and Projection (UMAP) plots. (b) ASURAT's unsupervised clustering based on a sign-by-sample matrix (SSM), derived from the integrated data, using a custom-build database (CB). Batch effects are further removed from those in CCA-based integration results. (c) Cell cycle phases inferred using the integrated data. (d) Profiles of sign scores in the PDAC tissue, predicting functional heterogeneities and distinct tissue regions.

Methods

Datasets

Human peripheral blood mononuclear cells (PBMCs)

These data were obtained from peripheral blood mononuclear cells (PBMCs) of healthy donors, which include approximately 4,000 and 6,000 cells; thus, they were referred to as PBMCs 4k and 6k, respectively. The data were produced with a 10x protocol using UMIs. The single-cell transcriptome datasets were downloaded from 10x Genomics repository (https://support.10xgenomics.com/single-cell-gene-expression/datasets). The following filtered read count matrices were obtained: PBMC 4k from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k) and PBMC 6k from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k). After data quality controls, the read count tables of PBMC 4k (resp. PBMC 6k) contained 6,658 (resp. 5,169) genes and 3,815 (resp. 4,878) cells.

Human small cell lung cancer (SCLC) with cisplatin treatments

The data were obtained from circulating tumor cell-derived xenografts cultured with cisplatin treatments, which was generated from lung cancer patients (Stewart *et al.*, 2020). The data were produced with a 10x protocol using unique molecular identifiers (UMIs). The SRA files were downloaded from Gene Expression Omnibus (GEO) with the accession codes GSE138474: GSM4104164 and GSM4104165, which are referenced in Stewart et al. (2020). SRA Toolkit version 2.10.8 was used to dump the FASTQ files. Cell Ranger version 3.1.0 was used to align the FASTQ files to the GRCh38-3.0.0 human reference genome and produce the single-cell transcriptome datasets. After controlling for data quality, the read count table contained 6,581 genes and 3,923 cells.

Human pancreatic ductal adenocarcinoma (PDAC)

The single-cell RNA sequencing (scRNA-seq) and spatial transcriptome (ST) data were

obtained from PDAC patients using inDrop and ST protocols (Moncada *et al.*, 2020), respectively. The FASTQ files were downloaded from Gene Expression Omnibus (GEO) with the accession codes GSE111672: GSM3036909, GSM3036910, GSM3036911, GSM3405527, GSM3405528, GSM3405529, and GSM3405530. Mapping of raw sequencing data from inDrop and ST protocols were processed using custom pipelines from https://github.com/flo-compbio/singlecell and https://github.com/jfnavarro/st_pipeline respectively. Both pipelines used the parameters explained by Moncada et al. (2020). Prior to downstream analysis, we concatenated all the scRNA-seq datasets. After data quality controls, the read count table of the combined scRNA-seq dataset contained 5,893 genes and 2,051 cells, wherein the ST dataset contained 4,497 genes and 428 ST spots. ST data was imported and visualized using Spaniel (Queen et al., 2019).

Data preprocessing: quality control, normalization, and centering

For all the scRNA-seq datasets, the low-quality genes and cells were removed by the following three steps: (i) removing the genes for which the number of non-zero expressing cells is less than a user-defined threshold; (ii) removing the cells whose read counts, number of genes expressed with non-zero read counts, and percent of reads mapped to mitochondrial genes are within user-defined ranges; and (iii) removing the genes for which the mean of the read counts is less than a user-defined threshold. See Chapters 2 and 3 of Supplementary Notes for the details.

After applying data quality controls, the data were normalized by bayNorm (Tang *et al.*, 2020), which attenuates technical biases with respect to zero inflation and variation of capture efficiencies between cells. The resulting inferred true count matrices were supplied to a log-transformation with a pseudo-count to attenuate the impact of dispersion in the counts for highly expressed genes. Finally, subtracting the sample mean from each row vector, we obtained the normalized-and-centered read count tables. See Chapter 4 of

Supplementary Notes for the details.

Knowledge-based databases

Any knowledge-based databases (DBs) including corresponding tables between biological descriptions and genes can be used in ASURAT. Here, we collected such DBs from Disease Ontology (DO) (Yu et al., 2015), Cell Ontology (CO) (Diehl et al., 2016), Gene Ontology (GO) (Yu *et al.*, 2012), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) using R packages DOSE (version 3.16.0), ontoProc (version 1.12.0), clusterProfiler (version 3.18.0), and KEGGREST (version 1.30.0), respectively. Additionally, to investigate PDAC datasets, we made a custom-built cell-type database (CB) by simply merging DO and CO DBs, CellMarker (Zhang *et al.*, 2019), MSigDB (Subramanian *et al.*, 2005), and PanglaoDB (Franzen *et al.*, 2019). See Chapter 8 of Supplementary Notes for the details.

Definition of sign

Let A be a read count table of size $p \times n$ from transcriptomic data, whose rows and columns mean p genes and n cells, represented by $\Omega = \{1, 2, \cdots, p\}$ and $S = \{1, 2, \cdots, n\}$, respectively, and R a "relation structure" (e.g., correlation matrix) among Ω. Let $\mathcal{F} = \{(T_k, \Omega_k) | k = 1, 2, \cdots, q\}$ be a set of ordered pairs, where T_k and $\Omega_k \in 2^{\Omega}$ (2^{Ω} is the power set of Ω) are biological description and the functional gene set, respectively. Consider an R-dependent representation $\Omega_k = \bigcup_{j=1}^{m(k)} \Omega_k^{(j)}$, 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 as described in the early 20th century. According to Maruyama (2008), the original notion of a *signe* is a segment of a thing of interest, which is created by an arbitrary decomposition based on its relationships, such as what it is distinct from opposing. For example, a rainbow is a continuum of varying light input, from which we can see distinct colors of red, yellow, green, and blue by our subjective decomposition based on its spectrum relationships

(Couper, 2015).

Definition of correlated gene set

Let A be a gene-by-cell matrix of size $p \times n$ from a transcriptomic data and $R = (r_{i,j})$ a correlation matrix of size $p \times p$ defined by A and a certain measure (e.g., Pearson's measure), whose diagonal elements are 1. Let α and β be positive and negative constants satisfying $0 < \alpha \leq 1$ and $-1 \leq \beta < 0$, respectively. Let us fix a biological description T_k and the functional gene set Ω_k, where $k = 1, 2, \cdots, q$. Now, 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 us denote $\Omega_k^{(\mathrm{w})} = \Omega_k - W_k$. If V_k is not empty, represent each element of W_k as a point in the Euclidean space spanned by the row vectors of R and decompose W_k into two disjoint subsets by Partitioning Around Medoids (PAM) clustering (Schubert and Rousseeuw, 2019), that is $W_k = \Omega_k^{(\mathrm{s})} \cup \Omega_k^{(\mathrm{v})}$. Otherwise, if V_k is empty, let $\Omega_k^{(\mathrm{s})} = U_k$ and $\Omega_k^{(\mathrm{v})} = \phi$ (empty). Thus Ω_k is decomposed into three parts as follows:

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

Let $\mu_k^{(\mathrm{s})}$ (resp. $\mu_k^{(\mathrm{v})}$) be the mean of off-diagonal elements of R for $\Omega_k^{(\mathrm{s})}$ ($\Omega_k^{(\mathrm{v})}$), and assume $\mu_k^{(\mathrm{s})} \geq \mu_k^{(\mathrm{v})}$ without loss of generality. If $\mu_k^{(\mathrm{s})} \geq \alpha$, then $\Omega_k^{(\mathrm{s})}$, $\Omega_k^{(\mathrm{v})}$, and $\Omega_k^{(\mathrm{w})}$ are strongly, variably, and weakly correlated gene sets, respectively, which are abbreviated as SCG, VCG, and WCG. Otherwise, correlated gene sets cannot be defined for T_k.

For any given (T_k, Ω_k, R) the genes should strongly and positively correlate within each of $\Omega_k^{(\mathrm{s})}$ and $\Omega_k^{(\mathrm{v})}$, while they should negatively correlate between $\Omega_k^{(\mathrm{s})}$ and $\Omega_k^{(\mathrm{v})}$. Thus, we can hypothesize that SCG and VCG are predominantly associated with T_k, which may aid interpretation of biological meanings of corresponding signs. **Figure 2** shows that the

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

Though simpler methods based on decomposition of correlation graphs exist, such as one-shot PAM clustering (Schubert and Rousseeuw, 2019), tree cutting after hierarchical clustering (Murtagh and Legendre, 2014), independent component analysis (ICA)- or principal component analysis (PCA)-based methods (Hyvarinen, 1999), and several graph statistical approaches (Blondel et al., 2008; Bodenhofer et al., 2011), we found that our VCG definition is critical for providing sample clusterings in the downstream analysis. We tried replacing our decomposition method (1) with one-shot PAM clustering, but the sample clusterings frequently exhibited deteriorated performance. This occurred when both VCG and WCG (obtained from the one-shot clustering) included many weakly correlated genes, which may contribute less to the parent sign.

Definition of sign-by-sample matrix

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

$$C^{(x)} = \omega^{(x)} B^{(x)} + \left(1 - \omega^{(x)}\right) B^{(\mathrm{w})}, \tag{2}$$

where $\omega^{(x)}$, $0 \leq \omega^{(x)} \leq 1$, are weight constants. Here $C^{(\mathrm{s})}$ and $C^{(\mathrm{v})}$ are said to be sign-by-sample matrices (SSMs) for SCG and VCG, respectively, and the entry $c_{k,j}^{(x)}$ a sign score of the kth sign and jth cell (sample). Note that ensemble means of sign scores across cells (samples) are zeros because SSMs are derived from the centered gene expression matrix A.

Definition of separation index

Briefly, a separation index is a measure of significance of a given sign score for a given subpopulation. Since the row vectors of SSMs are centered (i.e., the means are zeros), wherein the degree of freedom is reduced, naïve usages of statistical tests and fold change analyses should be avoided. Nevertheless, we propose helping users to find significant signs using a nonparametric index to quantify the extent of separation between two sets of random variables. A separation index of a given random variable X takes a value from -1 to 1: the larger positive value indicates that Xs are markedly upregulated, and the probability distribution is well separated against other distributions and vice versa.

Let us consider a vector $\boldsymbol{a}$ of size n, i.e., the number of samples, whose elements stand for the sign scores, and assume that the elements are sorted in ascending order. For simplicity suppose that the samples are classified into two clusters labeled 0 and 1. Let $\boldsymbol{v}$ be a vector of the labels corresponding to $\boldsymbol{a}$, and $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ be vectors having the same elements with $\boldsymbol{v}$ but the elements are sorted in lexicographic orders in forward and backward directions, respectively. Then we define separation index as follows:

$$I(\boldsymbol{v}) = 1 - \frac{2d(\boldsymbol{v}, \boldsymbol{w}_0)}{d(\boldsymbol{v}, \boldsymbol{w}_0) + d(\boldsymbol{v}, \boldsymbol{w}_1)}, \tag{3}$$

where $d(\boldsymbol{v}, \boldsymbol{w}_i)$ is an edit distance (or Levenshtein distance (Lowrance and Wagner, 1975)) with only adjacent swapping permitted. For example, if $\boldsymbol{v} = (1, 0, 0, 1, 1)$, then $\boldsymbol{w}_0 = (0, 0, 1, 1, 1)$ and $\boldsymbol{w}_1 = (1, 1, 1, 0, 0)$. From (3) one can calculate $d(\boldsymbol{v}, \boldsymbol{w}_0) = 2$ and $d(\boldsymbol{v}, \boldsymbol{w}_1) = 4$, and thus $I(\boldsymbol{v}) = 1/3$. As another example, if $\boldsymbol{v} = (0, 1, 1, 0, 0)$, then $I(\boldsymbol{v}) = -1/3$. From this example, one can see that the positive and negative values of I

mean that the given sign has positive and negative contributions for cluster "1," respectively.

Parameter settings of ASURAT

To obtain desired results, it is critical to tune ASURAT's parameters for creating SSMs. Depending on the DBs, there are six to nine parameters for creating SSMs, but many of them have been preset to unbiased and sensible default values (Figure S4). We found that our default settings worked well in our scRNA-seq analyses, but the three parameters should be tuned by users, as described below.

As formulated in (1), positive and negative constants α and β from thresholds of correlation coefficients are required for decomposing correlation graphs and creating signs (see **Figure 2** for the demonstration). In addition, unreliable signs are discarded with user-defined criteria, which were preset as follows: the sum of the number of genes in SCG and VCG is less than n_{min} or the number of genes in WCG is less than $n_{\mathrm{min}}^{(\mathrm{w})}$ (the default value is 2). Furthermore, users can remove redundant signs with similar biological meanings if information contents (ICs) (Yu et al., 2010) are defined.

Analysis of scRNA-seq datasets of PBMCs 4k and 6k

To compare the cell-type inference abilities of existing methods and ASURAT, we prepared two scRNA-seq datasets, namely PBMCs 4k and 6k (see Datasets). Subsequently, data quality controls and normalization by bayNorm were carefully performed for each dataset. See Chapters 2–4 of Supplementary Notes for the details.

Using scran (version 1.18.7) (Lun *et al.*, 2016), we normalized the data using scran functions quickCluster(), computeSumFactors(), and logNormCounts(), selected highly variable genes using modelGeneVar() and getTopHVGs() based on a variance modeling with a gene-per-cell ratio of 0.2 (as suggested in a previous work (Cruz and Wishart,

2007)), and set the principal components using denoisePCA(). Cells were clustered using buildSNNGraph() and cluster_louvain(). Then, candidates of differentially expressed genes (DEGs) were detected using pairwiseTTests() and combineMarkers(), and DEGs were defined as genes with false discovery rates (FDRs)$< 10^{-99}$ (T tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer et al., 2016) as follows: B cells (resp. marker genes *CD79A*, *MS4A1*, *IGHM*), monocytes (*S100A8*, *LYZ*, *CD14*), NK or NKT cells (*NKG7*, *GZMA*, *FGFBP2*), and T cells (*MAL*). See Chapter 13 of Supplementary Notes for the details.

Using Seurat (version 4.0.2) (Hao *et al.*, 2021), we normalized the data using Seurat's function NormalizeData() with a log normalization (default), selected highly variable genes using FindVariableFeatures() based on a variance-stabilizing transformation with a gene-per-cell ratio of 0.2 (as suggested in previous work (Cruz and Wishart, 2007)), scaled and centered gene expression levels, and performed PCA. The principal components that explain 90% of the total variability were used for the computations of FindNeighbors(). Cells were clustered using FindClusters(). Then, candidates of DEGs were detected using FindAllMarkers() and DEGs were defined as genes with false discovery rates (FDRs)$< 10^{-99}$ (Mann-Whitney *U* tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer *et al.*, 2016) as follows: T cells (resp. marker genes *TRAC*, *CD3D*, *IL32*, *TCF7*, *CD27*), monocytes (*S100A8*, *LYZ*, *CD14*), B cells (*CD79A*, *MS4A1*, *IGHM*, *VPREB3*, *BANK1*), and NK or NKT cells (*CD3D*, *NKG7*, *GZMA*, *FGFBP2*). Additionally, to automatically annotate the clustering results, we used the R function findmarkergenes() in the scCATCH (version 2.1) package (Shao *et al.*, 2020), which identified monocytes, B cells, and T cells. See Chapter 14 of Supplementary Notes for the details.

Using Monocle 3 (version 1.0.0) (Trapnell *et al.*, 2014), we ran Monocle 3 function preprocess_cds() under the default settings, in which data were normalized by a log transform with a pseudo-count of 1, scaled and centered in gene expression levels, and were subjected PCA with the dimensionality of the reduced space set to 50. Cells were clustered by cluster_cells() using Uniform Manifold Approximation and Projection (UMAP) (McInnes and Healy, 2018). Then, candidates of DEGs were detected using top_markers() and DEGs were defined as genes with false discovery rates (FDRs)$< 10^{-99}$ (Monocle's marker significance tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer *et al.*, 2016) as follows: T cells (resp. marker genes *CD3D*, *TCF7*, *CD3E*, *IL32*), monocytes (*S100A8*, *LYZ*, *CD14*), B cells (*CD79A*, *CD79B*, *BANK1*, *MS4A1*), and NK or NKT cells (*GNLY*, *NKG7*, *GZMA*). See Chapter 15 of Supplementary Notes for the details.

Using SC3 (version 1.18.0) (Kiselev *et al.*, 2017), we performed SC3 function runPCA() inputting log-normalized read count tables with a pseudo-count of 1. Cells were clustered using sc3(), and reasonable numbers of clusters were manually determined by sc3_plot_markers(). Then, candidates of DEGs were detected using get_marker_genes() and DEGs were defined as genes with false discovery rates (FDRs)$< 10^{-99}$ (Kruskal-Wallis tests). According to the DEGs, we identified several different cell types by manually searching for marker genes in GeneCards version 5.2 (Stelzer *et al.*, 2016) as follows: NK or NKT cells (resp. marker genes *GZMA*, *GZMB*, *GZMH*, *GZMK*, *GNLY*), T cells (*TRGC2*, *TCL1A*), monocytes (*GSN*, *LILRB4*, *S100A8*, *CD14*, *S100A12*), and B cells (*CD79A*, *CD79B*, *MS4A1*, *SPI1*, *LYN*). See Chapter 16 of Supplementary Notes for the details.

Using ASURAT, we created SSMs using DBs of CO, GO, and KEGG. After dimensionality reduction by PCA, cells were clustered by *k*-nearest neighbor (KNN)

graph generation and Louvain algorithm using Seurat functions FindNeighbors() and FindClusters() (Hao *et al.*, 2021). Subsequently, separation indices (SIs) were computed for all the signs for a given cluster versus all the others, then cell types were identified by manually selecting significant signs with the larger values of SIs> 0.5 (**Figure 3**). See Chapter 17 of Supplementary Notes for the details.

Analysis of an SCLC scRNA-seq dataset

For the analysis of an SCLC scRNA-seq dataset, Seurat's workflow started by normalizing data using the Seurat function NormalizeData() with a log normalization (default). Then, highly variable genes were selected by FindVariableFeatures() based on a variance stabilizing transformation with a gene-per-cell ratio of 0.2 (as suggested in previous work (Cruz and Wishart, 2007)). Then, data were scaled and centered by ScaleData(), and PCA was applied by RunPCA() with highly variable genes. Subsequently, a KNN graph was generated by FindNeighbors(), with the principal components that explain 90% of the total variability, and cells were clustered by FindClusters() with a Louvain algorithm. Additionally, cell cycle phases were inferred by CellCycleScoring() with cell cycle-related genes defined in the Seurat package. Finally, KEGG enrichment analysis was done by compareCluster() in clusterProfiler package (Yu *et al.*, 2012). See Chapter 14 of Supplementary Notes for the details.

ASURAT's workflow started with the collection of DO, GO, and KEGG databases. First, we excluded functional gene sets including too few or too many genes. Next, we created multiple signs using a correlation graph-based decomposition. Then, we removed redundant signs with similar biological meanings using doSim() in DOSE package (Yu *et al.*, 2015). Then, we created SSMs for DO, GO, and KEGG. Based on the SSM for DO, we performed a dimensionality reduction using the diffusion map and clustered cells using MERLoT (Parra *et al.*, 2019). Finally, we vertically concatenated all the SSMs, cell cycle phases inferred by Seurat, and expression matrix for characterizing individual cells

from multiple biological aspects. The DEGs were identified using FindAllMarkers() in Seurat package. See Chapters 9–12 of Supplementary Notes for the details.

Code availability

An open-source implementation of ASURAT is available on GitHub (https://github.com/keita-iida/ASURAT) under the GPLv3 license. All the input and output files used in the present paper and user-friendly documentation written in R bookdown can be downloaded from the above URL.

Supplementary materials

Notes Clear documentation (R bookdown files) showing the commands and outputs for all the analyses in the present paper, as well as an introduction to ASURAT, which is also available on GitHub (https://github.com/keita-iida/ASURAT_0.0.0.9001).

Supplementary Files ASURAT's R function files, which are available on GitHub (https://github.com/keita-iida/ASURAT_0.0.0.9001).

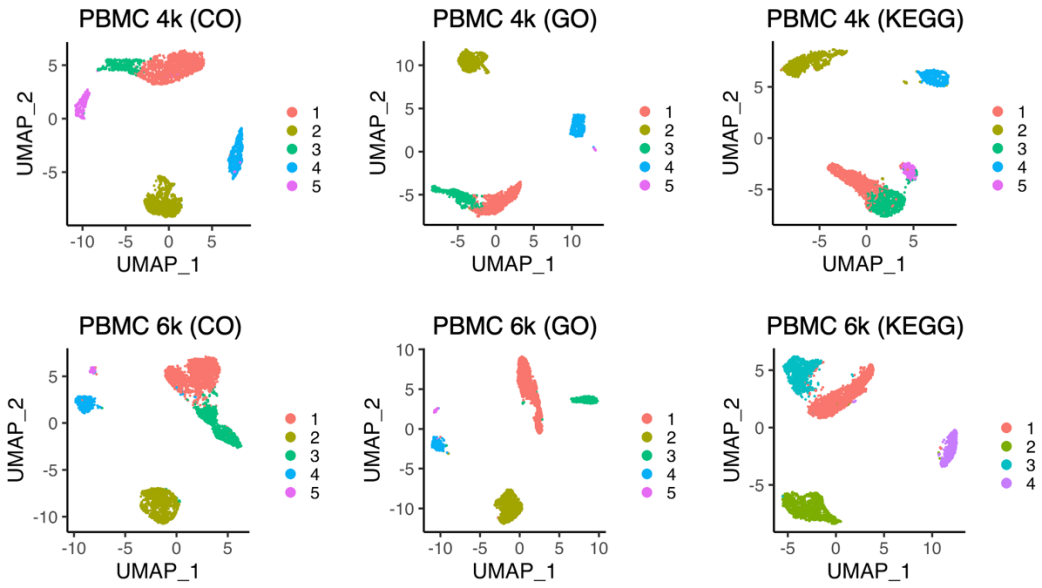


Figure S1. Clustering single-cell transcriptomes of peripheral blood mononuclear cells (PBMCs) 4k and 6k using ASURAT

Clustering single-cell transcriptomes of peripheral blood mononuclear cells (PBMCs) 4k and 6k using ASURAT.

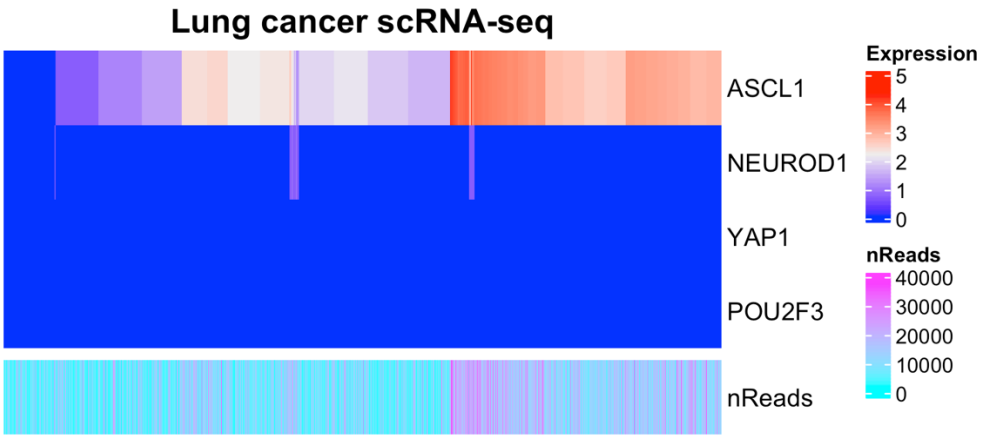


Figure S2. Heatmaps of expression levels of known small cell lung cancer marker genes

Heatmaps of expression levels of known small cell lung cancer marker genes.

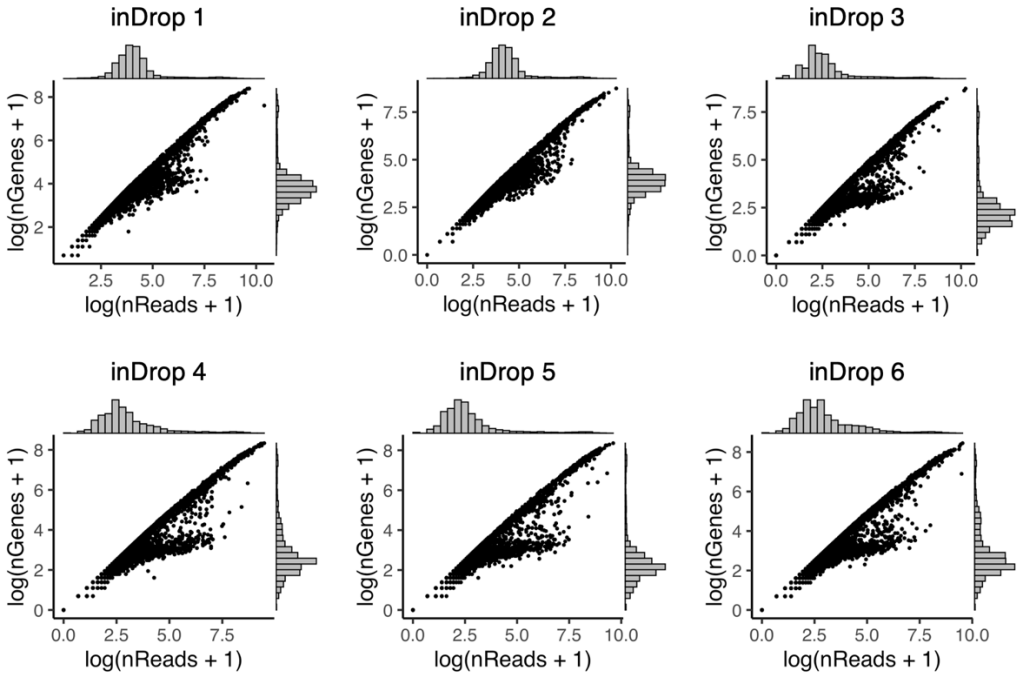


Figure S3. Data qualities across all the cells in single-cell RNA-seq datasets PDAC-A inDrop from 1 to 6

Data qualities across all the cells in single-cell RNA-seq datasets PDAC-A inDrop from 1 to 6.

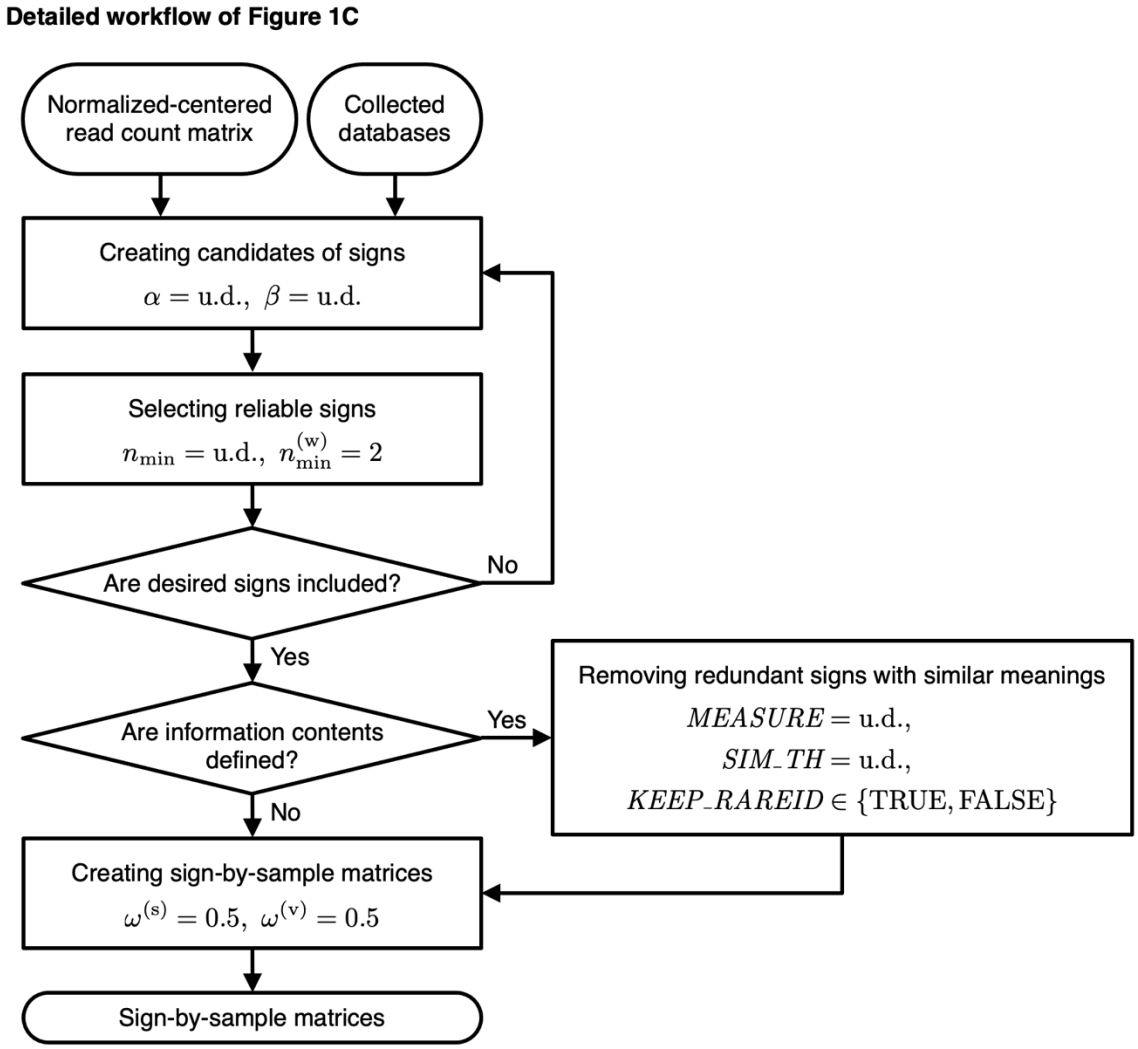


Figure S4. Detailed workflow of Figure 1c focusing on the parameter settings

Detailed workflow of Figure 1c focusing on the parameter settings.

REFERENCES

Andrews, T.S., Kiselev, V.Y., McCarthy, D., and Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc *16*, 1-9. 10.1038/s41596-020-00409-w.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol *20*, 163-172. 10.1038/s41590-018-0276-y.

Balanis, N.G., Sheu, K.M., Esedebe, F.N., Patel, S.J., Smith, B.A., Park, J.W., Alhani, S., Gomperts, B.N., Huang, J., Witte, O.N., and Graeber, T.G. (2019). Pan-cancer Convergence to a Small-Cell Neuroendocrine Phenotype that Shares Susceptibilities with Hematological Malignancies. Cancer Cell *36*, 17-34 e17. 10.1016/j.ccell.2019.06.005.

Blondel, V.D., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J Stat Mech-Theory E, P10008.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. Bioinformatics *27*, 2463-2464. 10.1093/bioinformatics/btr406.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411-420. 10.1038/nbt.4096.

Cancer Genome Atlas Research, N., Albert Einstein College of, M., Analytical Biological, S., Barretos Cancer, H., Baylor College of, M., Beckman Research Institute of City of, H., Buck Institute for Research on, A., Canada's Michael Smith Genome Sciences, C., Harvard Medical, S., Helen, F.G.C.C., et al. (2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378-384. 10.1038/nature21386.

Cao, Y., Wang, X., and Peng, G. (2020). SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. Front Genet *11*, 490. 10.3389/fgene.2020.00490.

Chen, H.J., Poran, A., Unni, A.M., Huang, S.X., Elemento, O., Snoeck, H.W., and Varmus, H. (2019). Generation of pulmonary neuroendocrine cells and SCLC-like tumors from human embryonic stem cells. J Exp Med *216*, 674-687. 10.1084/jem.20181155.

Chen, Z., Yang, X., Bi, G., Liang, J., Hu, Z., Zhao, M., Li, M., Lu, T., Zheng, Y., Sui, Q., 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. 10.7150/ijbs.42080.

Coifman, R.R., and Lafon, S. (2006). Diffusion maps. Appl Comput Harmon A *21*, 5-30. 10.1016/j.acha.2006.04.006.

Couper, P. (2015). A Student's Introduction to Geographical Thought: Theories, Philosophies, Methodologies 10.4135/9781473910775.

Cruz, J.A., and Wishart, D.S. (2007). Applications of machine learning in cancer prediction and prognosis. Cancer Inform *2*, 59-77.

De Simone, M., Rossetti, G., and Pagani, M. (2018). Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges. Front Immunol *9*, 1638. 10.3389/fimmu.2018.01638.

Devitt, K., Hanson, S.J., Tuong, Z.K., McMeniman, E., Soyer, H.P., Frazer, I.H., and Lukowski, S.W. (2019). Single-cell RNA sequencing reveals cell type-specific HPV expression in hyperplastic skin lesions. Virology *537*, 14-19. 10.1016/j.virol.2019.08.007.

Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., et al. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics *7*, 44. 10.1186/s13326-016-0088-7.

Dominguez, D., Tsai, Y.H., Gomez, N., Jha, D.K., Davis, I., and Wang, Z. (2016). A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. Cell Res *26*, 946-962. 10.1038/cr.2016.84.

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res *49*, e50. 10.1093/nar/gkab043.

Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J.B., Zhang, K., Chun, J., and Kharchenko, P.V. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods *13*, 241-244. 10.1038/nmeth.3734.

Franzen, O., Gan, L.M., and Bjorkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford) *2019*. 10.1093/database/baz046.

Ganesh, K., Basnet, H., Kaygusuz, Y., Laughney, A.M., He, L., Sharma, R., O'Rourke, K.P., Reuter, V.P., Huang, Y.H., Turkekul, M., et al. (2020). L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. Nat Cancer *1*, 28-45. 10.1038/s43018-019-0006-x.

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., and Wang, X. (2019). DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis *8*, 44. 10.1038/s41389-019-0157-8.

Gaudet, P., and Dessimoz, C. (2017). Gene Ontology: Pitfalls, Biases, and Remedies. Methods Mol Biol *1446*, 189-205. 10.1007/978-1-4939-3743-1_14.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw *10*, 626-634. 10.1109/72.761722.

Ireland, A.S., Micinski, A.M., Kastner, D.W., Guo, B., Wait, S.J., Spainhower, K.B., Conley, C.C., Chen, O.S., Guthrie, M.R., Soltero, D., et al. (2020). MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate. Cancer Cell *38*, 60-78 e12. 10.1016/j.ccell.2020.05.001.

Jalili, M., Scharm, M., Wolkenhauer, O., Damaghi, M., and Salehzadeh-Yazdi, A. (2021). Exploring the Metabolic Heterogeneity of Cancers: A Benchmark Study of Context-Specific Models. J Pers Med *11*. 10.3390/jpm11060496.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30. 10.1093/nar/28.1.27.

Kim, N., Kim, H.K., Lee, K., Hong, Y., Cho, J.H., Choi, J.W., Lee, J.I., Suh, Y.L., Ku, B.M., Eum, H.H., et al. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun *11*, 2285. 10.1038/s41467-020-16164-1.

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet *20*, 273-282. 10.1038/s41576-018-0088-9.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat Methods *14*, 483-486. 10.1038/nmeth.4236.

Kristiansen, G., Schluns, K., Yongwei, Y., Denkert, C., Dietel, M., and Petersen, I. (2003). CD24 is an independent prognostic marker of survival in nonsmall cell lung cancer patients. Br J Cancer *88*, 231-236. 10.1038/sj.bjc.6600702.

Kubota, S., Tanaka, M., Endo, H., Ito, Y., Onuma, K., Ueda, Y., Kamiura, S., Yoshino, K., Kimura, T., Kondo, J., and Inoue, M. (2020). Dedifferentiation of neuroendocrine carcinoma of the uterine cervix in hypoxia. Biochem Biophys Res Commun *524*, 398-404. 10.1016/j.bbrc.2020.01.024.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lonnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494-498. 10.1038/s41586-018-0414-6.

Lahnemann, D., Koster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand

challenges in single-cell data science. Genome Biol *21*, 31. 10.1186/s13059-020-1926-6.

Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon, L.K., Tan, W.S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet *49*, 708-718. 10.1038/ng.3818.

Lowrance, R., and Wagner, R.A. (1975). An extension of the string-to-string correction problem. J Assoc Comput Mach *22*. 10.1145/321879.321880.

Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol *17*, 75. 10.1186/s13059-016-0947-7.

Maynard, A., McCoach, C.E., Rotow, J.K., Harris, L., Haderk, F., Kerr, D.L., Yu, E.A., Schenk, E.L., Tan, W., Zee, A., et al. (2020). Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. Cell *182*, 1232-1251 e1222. 10.1016/j.cell.2020.07.017.

McInnes, L., and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426.

McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165. 10.1186/1471-2105-11-165.

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M., and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol *38*, 333-342. 10.1038/s41587-019-0392-8.

Moore, D., Simoes, R.M., Dehmer, M., and Emmert-Streib, F. (2019). Prostate Cancer Gene Regulatory Network Inferred from RNA-Seq Data. Curr Genomics *20*, 38-48. 10.2174/1389202919666181107122005.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet *34*, 267-273. 10.1038/ng1180.

Morrison, R.E., Baptista, R., and Marzouk, Y. (2017). Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting. Adv Neur In *30*.

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif *31*, 274-295. 10.1007/s00357-014-9161-z.

Parra, R.G., Papadopoulos, N., Ahumada-Arranz, L., Kholtei, J.E., Mottelson, N., Horokhovsky, Y., Treutlein, B., and Soeding, J. (2019). Reconstructing complex lineage trees from scRNA-seq data using MERLoT. Nucleic Acids Res *47*, 8961-8974. 10.1093/nar/gkz706.

Pasquini, G., Rojo Arias, J.E., Schafer, P., and Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. Comput Struct Biotechnol J *19*, 961-969. 10.1016/j.csbj.2021.01.015.

Pliner, H.A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. Nat Methods *16*, 983-986. 10.1038/s41592-019-0535-3.

Poh, A.R., and Ernst, M. (2021). Tumor-Associated Macrophages in Pancreatic Ductal Adenocarcinoma: Therapeutic Opportunities and Clinical Challenges. Cancers (Basel) *13*. 10.3390/cancers13122860.

Queen, R., Cheung, K., Lisgo, S., Coxhead, J., and Cockell, S. (2019). Spaniel: analysis and interactive sharing of Spatial Transcriptomics data. bioRxiv. 10.1101/619197.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc *14*, 482-517. 10.1038/s41596-018-0103-9.

Rempala, G.A., Seweryn, M., and Ignatowicz, L. (2011). Model for comparative analysis of antigen receptor repertoires. J Theor Biol *269*, 1–15. 10.1016/j.jtbi.2010.10.001.

Schubert, E., and Rousseeuw, P.J. (2019). Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. SISAP 2020, 171–187. 10.1007/978-3-030-32047-8_16.

Schwendenwein, A., Megyesfalvi, Z., Barany, N., Valko, Z., Bugyik, E., Lang, C., Ferencz, B., Paku, S., Lantos, A., Fillinger, J., et al. (2021). Molecular profiles of small cell lung cancer subtypes: therapeutic implications. Mol Ther Oncolytics *20*, 470–483. 10.1016/j.omto.2021.02.004.

Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., and Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience *23*, 100882. 10.1016/j.isci.2020.100882.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics *54*, 1 30 31–31 30 33. 10.1002/cpbi.5.

Stewart, C.A., Gay, C.M., Xi, Y., Sivajothi, S., Sivakamasundari, V., Fujimoto, J., Bolisetty, M., Hartsfield, P.M., Balasubramaniyan, V., Chalishazar, M.D., 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. 10.1038/s43018-019-0020-z.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545–15550. 10.1073/pnas.0506580102.

Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S., and Shahrezaei, V. (2020). bayNorm: Bayesian gene expression recovery, imputation and normalization

for single-cell RNA-sequencing data. Bioinformatics *36*, 1174-1181. 10.1093/bioinformatics/btz726.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol *32*, 381-386. 10.1038/nbt.2859.

Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science *356*. 10.1126/science.aah4573.

Wagner, F.Y., I. (2018). Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. bioRxiv. doi.org/10.1101/456129.

Wooten, D.J., Groves, S.M., Tyson, D.R., Liu, Q., Lim, J.S., Albert, R., Lopez, C.F., Sage, J., and Quaranta, V. (2019). Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. PLoS Comput Biol *15*, e1007343. 10.1371/journal.pcbi.1007343.

Yatabe, Y. (2020). Reassessing the SCLC Subtypes. J Thorac Oncol *15*, 1819-1822. 10.1016/j.jtho.2020.10.011.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics *26*, 976-978. 10.1093/bioinformatics/btq064.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284-287. 10.1089/omi.2011.0118.

Yu, G., Wang, L.G., Yan, G.R., and He, Q.Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics *31*, 608-609. 10.1093/bioinformatics/btu684.

Zhang, W., Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., et al. (2020). PIRD: Pan Immune Repertoire Database. Bioinformatics *36*, 897–903. 10.1093/bioinformatics/btz614.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res *47*, D721–D728. 10.1093/nar/gky900.