Functional annotation-driven unsupervised clustering of single-cell transcriptomes

Keita Iida[1*], Jumpei Kondo[2,3], 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

ORCID (Keita Iida): https://orcid.org/0000-0002-1076-830X

ORCID (Jumpei Kondo): https://orcid.org/0000-0002-1350-0480

ORCID (Masahiro Inoue): https://orcid.org/0000-0001-7315-026X

ORCID (Mariko Okada): https://orcid.org/0000-0002-6210-8223

Keywords:

Single-cell transcriptome; biological interpretation; unsupervised clustering; small cell neuroendocrine cancer

Abstract

Single-cell RNA sequencing (scRNA-seq) analysis has significantly advanced our knowledge of functional states of cells. By analyzing scRNA-seq data, we can deconvolve individual cell states into thousands of gene expression profiles, allowing us to perform cell clustering, and identify significant genes for each cluster. However, interpreting these results remains challenging. Here, we present a novel scRNA-seq analysis pipeline named ASURAT, which simultaneously performs unsupervised cell clustering and biological interpretation in semi-automatic manner, in terms of cell type and various biological functions. We validate the reliable clustering performance of ASURAT by comparing it with existing methods, using six published scRNA-seq datasets from healthy donors and cancer patients. Furthermore, we applied ASURAT to patient-derived scRNA-seq datasets including small cell lung cancers, finding some putative cancer subpopulations showing different resistance mechanisms. ASURAT is expected to open new means of scRNA-seq analysis, focusing more on "biological meaning" than 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[1]. scRNA-seq has been widely used to improve understanding of individual cells[2], intra- and intertumoral heterogeneity[3], cell-to-cell interaction[4], tumorigenesis[5], drug resistance[3,6], and the effects of viral infection on immune cell populations[7]. Various clustering methods, wherein cells are partitioned according to transcriptome-wide similarity, have been proposed[8] and applied to cell type annotation[9]. However, interpreting single-cell data remains challenging[10-13].

Conventionally, cell types are inferred using unsupervised clustering followed by a manual literature search of differentially expressed marker genes[13]. Currently, several computational tools, such as Garnett[14] and SCSA[12], are available to assist manual annotation, as detailed in the review by Pasquini *et al.*[8]. However, this process is often difficult because marker genes are generally expressed in multiple cell types[15]. In cancer transcriptomics, this difficulty is exacerbated by the interdependence between disease-related genes and numerous biological terms; furthermore, expression levels of marker genes can be heterogeneous depending on cancer microenvironments[16].

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[10,12,17]. 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[17]. 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[11] combined with gene set enrichment analysis, incorporating biological knowledge such as pathway activity; hence, it may improve the interpretability over signature gene-based approaches, which place sole emphasis on individual roles of genes. However, we still lack a prevailing theory leveraging this information at the single-cell level.

To overcome the aforementioned limitations, a novel theoretical tool providing biological interpretations to computational results is needed. Thus, we propose a scRNA-seq analysis pipeline for simultaneous cell clustering and biological interpretation, named ASURAT. Here, "interpretation" is given by multiple biological terms such as cell type, biological process, pathway activity, chemical reaction, and various biological functions. By using ASURAT, users can create desired sets of biological terms and the corresponding spectrum matrices, which can be supplied to the subsequent unsupervised cell clusterings. In this paper, we first demonstrate the reliable clustering performance of ASURAT based on comparison with existing methods, using six published scRNA-seq datasets of healthy donors and cancer patients. Next, we applied ASURAT to single-cell lung cancer transcriptomes, which include malignant cancer types expressing neuroendocrine markers[3]. We show that ASURAT can greatly improve functional understandings of various cell types, which may contribute to clinical improvements.

Results

Overview of ASURAT

ASURAT was developed for simultaneously clustering single-cell transcriptomes and biological interpretation, which was implemented by R programming scripts (Supplementary Notes, Supplementary File 6). After inputting scRNA-seq data and knowledge-based databases (DBs), ASURAT creates lists of biological terms with respect to cell type and biological functions, which we termed signs. Then, ASURAT creates a functional spectrum matrix, termed a sign-by-sample matrix (SSM). By analyzing SSMs, users can cluster samples to aid their interpretation. We later explain the workflow (**Fig. 1**). The details of ASURAT's formulations can be found in the Methods section.

Workflow of ASURAT

In preparation, we collected DBs for Disease Ontology (DO)[18], Cell Ontology (CO)[19], Gene Ontology (GO)[20], Kyoto Encyclopedia of Genes and Genomes (KEGG)[21], and Reactome[22] using the R packages DOSE (version 3.16.0), ontoProc (version 1.12.0), clusterProfiler (version 3.18.0), KEGGREST (version 1.30.0), and reactome.db (version 1.74.0), respectively (Chapter 7, Supplementary Notes). Any DBs including corresponding tables between biological descriptions and genes can be input to ASURAT (**Fig. 1b**). Additionally, ASURAT computes a correlation matrix using Pearson or Spearman correlation coefficients from a normalized read count matrix of scRNA-seq data.

The first step is to create signs by inputting a normalized-and-centered read count matrix and knowledge-based DB. From a gene set Ω and correlation matrix R defined for each biological description T in DBs, ASURAT decomposes the correlation graph into several parts. Here, a triplet of biological description, gene subset, and correlation matrix is termed a sign, in particular (T, Ω, R) a parent sign. In many applications, high correlations are expected to have rich information. Hence, we decompose Ω into the following three categories (**Fig. 2**): (i) a strongly correlated gene set (SCG), which is a set of genes with strong positive correlations with each other; (ii) variably correlated gene set (VCG), which is a set of genes with strong negative correlations with genes in SCG; and (iii)

weakly correlated gene set (WCG), which is a set of genes with weak correlations with each other.

Next, 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 SCG and VCG, we create a single SSM. The rows and columns of an SSM stand for signs and samples (or cells), respectively, and entries stand for cell-type or functional spectra, termed as sign scores. A remarkable benefit is that users can create multiple SSMs as necessary by inputting various DB (**Fig. 1c**).

The final step is to characterize samples using SSMs to produce a conclusion. One focus of analyzing SSMs is to cluster samples and find significant signs (**Fig. 1d**), where "significant" means that the sign score is specifically upregulated or downregulated at the cluster level (cf. separation index). In ASURAT, we use two strategies: one uses unsupervised clusterings, such as Partitioning Around Medoids (PAM), hierarchical-based, and graph-based clusterings with and without principal component analysis (PCA); while the other is a method of extracting a continuous tree-like topology using diffusion map[23], followed by allocating samples to different branches of the data manifolds[24]. 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.

Comparison of performance of ASURAT with existing methods

Many unsupervised clustering methods have been proposed and their performances quantified using datasets with independently identified phenotypes. However, it remains unclear whether these methods robustly demonstrate better performance using cancer single-cell transcriptomes including ambiguous phenotypes. Conventional marker gene-based approaches may misrepresent cluster accuracy[17], and simple application of PCA may be ineffective. However, when using ASURAT, users can obtain robust and explainable clustering results, since SSMs can be created from as many DBs as needed and supplied to the subsequent unsupervised clusterings.

To validate the reliable clustering performance of ASURAT, we obtained six published scRNA-seq datasets derived from healthy donors (PBMC datasets: pbmc_4000 and pbmc_6000), cervical cancer patients (day1_norm and day7_hypo), and lung cancer patients (sc68_vehi and sc68_cisp). From all datasets, we excluded genes and cells with low qualities and attenuated technical biases with respect to zero-inflation and variation of capture efficiencies between cells using bayNorm[25]. The resulting read count tables were supplied to ASURAT and four other methods: Seurat (version 4.0.1)[26], Monocle 3 (version 0.2.3.0)[27], SC3 (version 1.18.0)[28], and PCA using prcomp() from the R stats package (version 4.0.4).

There are five blood cells in the PBMC datasets[12], which are regarded as hypothetical results. However, no consensus cell types exist, especially for cancer datasets. Hence, the clustering accuracies cannot be quantified using standard measures such as adjusted Rand index[29]. Instead, the clustering qualities were assessed using validity indices such as average silhouette width (ASW)[30], a measure of how tightly grouped cells are in clusters and the distant between clusters. To reduce computational cost, we performed two-dimensional Uniform Manifold Approximation and Projection (UMAP)[31] after the straightforward computations of Seurat, Monocle 3, PCA, and ASURAT; the resulting two-dimensional cell states were supplied to NbClust[32], and 26 validity indices were obtained (Supplementary Files). From SC3, we obtained only ASWs computed from consensus matrices and hierarchical clusterings. We hypothesized that clustering quality positively correlates with clustering accuracy, while considering that they do not guarantee interpretability. Additionally, other topology-based clustering methods were not used for computing ASWs.

For PBMC datasets with known numbers of clusters of existing cell types, we compared ASWs across all the methods within such numbers ± 1 (shaded area in **Fig. 3a**). For other datasets, we focused on the ranges of the number of clusters, wherein at least one method provides ASWs ≥ 0.6. Interestingly, the best-performing method, exhibiting the greatest ASW, was different across the datasets (**Fig. 3a**). Seurat performed best when the number of clusters $k = 4$ in pbmc_6000. Although SC3 outperformed at a different k in day7_hypo

and PBMC datasets, it could not detect >1 cluster in sc68_vehi and sc68_cisp. Compared with other methods, only the naïve usage of PCA was unremarkable across most datasets.

Notably, ASURAT outperformed existing methods at ≥1 *k* in every dataset, with one exception in sc68_cisp (**Fig. 3a**). Moreover, those ASWs were >0.5 without exception and >0.6 with only one exception (viz. sc68_cisp). The existing methods presented both strengths and weaknesses depending on the datasets. Seurat exhibited better performances with PBMC datasets, while it performed less remarkably with most cancer datasets. Although we carefully tuned Seurat's parameters by changing the normalization method, variable gene-per-cell ratio, and the number of principal components, we could not obtain well-separated clusters for day1_norm and day7_hypo (**Fig. 3b**). In contrast, Monocle 3 generally exhibited better performances on cancer datasets while performing less remarkably with PBMC datasets. We found that Monocle 3's clustering performance was unstable and strongly depended upon dimension reduction techniques.

To confirm whether ASURAT outperforms existing methods using other low-dimensional representation techniques, we replaced UMAP with t-distributed stochastic neighbor embedding (t-SNE)[33] and supplied the resulting two-dimensional cell states to NbClust[32]. Again, we confirmed that ASURAT generated well-separated clusters with relatively greater ASWs across datasets, while Monocle 3 broke down when used with some datasets (Supplementary Fig. S1). These results indicate that cells are better characterized in the high-dimensional sign score space than in the gene expression space.

Finally, to validate ASURAT's cell-type inference, we reanalyzed PBMC datasets using Seurat, Monocle 3, SC3, and ASURAT under almost default settings. Consequently, Seurat and Monocle 3 could reproduce most blood cell type labels (**Figs. 3c** and **d**), as inferred by Cao *et al.*[12], but a few dozen cells remained unspecified. Although SC3 provided the greatest ASWs at $k = 4$ and 6 in pbmc_4000 and pbmc_6000, respectively, it reproduced only B cell and NK or NKT cell labels. However, ASURAT identified five cell types, with none remaining unspecified (Supplementary Figs. S3 and S4). The subpopulation ratios were approximately consistent with the reported values, except for the tiny megakaryocyte subpopulation. Such a small discrepancy was unavoidable,

because Cao *et al*. used only differentially expressed genes and preselected cell types to identify the most preferable cell types. Furthermore, we reanalyzed cervical cancer datasets using ASURAT and found several putative populations of small cell neuroendocrine carcinoma and adenocarcinoma (Supplementary Figs. S5 and S6). These results demonstrate that ASURAT can perform robust, high-quality, and reliable clusterings using various single-cell transcriptomes.

Identifying chemoresistant cells in lung cancer scRNA-seq datasets

Previous work[3] indicated that small cell lung cancer (SCLC) tumors undergo a shift from chemosensitivity to chemoresistance against platinum-based therapy. However, the exact mechanism behind chemoresistance is still unclear, because transcriptional heterogeneity is often concealed in hidden biological states, which cannot be readily identified by conventional marker gene-based analyses. To investigate the cancer subtypes in the chemosensitive and chemoresistant tumors, we applied ASURAT to the scRNA-seq data of circulating tumor cell-derived xenografts from the vehicle (sc68_vehi) and cisplatin (sc68_cisp) treatment groups.

Given the normalized and centered read count matrices, we created SSMs using DO and GO DBs, and KEGG for both sc68_vehi and sc68_cisp. We then visualized the sign scores in heat maps (**Figs. 4a** and **5a**). The cells were clustered by one of the following: (i) PCA, followed by k-nearest neighbor (KNN) graph generation and Louvain algorithm using Seurat's functions[26] and (ii) diffusion map generation, followed by allocation of cells to the different branches of the data manifold using MERLoT[24]. Here, cells in sc68_vehi were clustered by (i), while those in sc68_cisp were clustered by (ii), providing the most explainable results.

We visualized the t-SNE plot of SSM using GO for sc68_vehi, wherein cell clustering labels and SCLC-related sign scores are overlaid (**Fig. 4b**). Sign IDs and the related genes are represented by, for example, DOID:5409_S (*ASCL1*, etc.) and DOID:5409_V (*MKI67*, *BIRC5*, etc.), where the suffixes "S" and "V" indicate SCG and VCG, respectively. Since *ASCL1*, *MKI67*, and *BIRC5* are important for neuronal differentiation[34], malignancy[35], and inhibition of apoptosis[36], DOID:5409_S and DOID:5409_V represent SCLC

differentiation and proliferation with cell survival, respectively. We found at least two existing subpopulations of SCLC in sc68_vehi. This was further confirmed by violin plots for the related signs (**Fig. 4c**). Remarkably, sign scores for platinum drug resistance were specifically upregulated in the group with label 2 (GO: BP). The population ratios of group 1 and 2 were 0.84 and 0.15, respectively. Consequently, we found that the SCLCs not receiving cisplatin treatment contained ≤15% putative chemoresistant cells, which was not found in the original report[3].

Likewise, we visualized the diffusion map of SSM with DO for sc68_cisp. We observed a tree-like topology in the data manifold, representing a putative cell differentiation lineage (**Fig 5b**). We defined a pseudotime $t \in [0, 1]$ (i.e., an arc-length parameter) along the branches using MERLoT[24]; a starting point $t = 0$ was set at the end of the branch with label 1. From the pseudo-time course analysis, we found at least three SCLC subpopulations (**Fig. 5c**). Strikingly, sign scores for different resistant mechanisms, such as platinum drug resistance and PD-L1 expression mediating immunosuppression, were upregulated in groups labeled 2 and 3 (DO: disease), while sign scores for intracellular protein transport with an SCLC malignancy marker CD24[37] was upregulated in the group labeled 1 (DO: disease), suggesting the recalcitrant malignancy of relapsed SCLCs against cisplatin treatments. The population ratios of groups 1, 2, and 3 were 0.39, 0.30, and 0.30, respectively. Consequently, we found 30% putative chemoresistant SCLCs and another 30% with other possible resistant cell types expressing PD-L1, while others did not exhibit these resistance mechanisms. Our results support the finding that transcriptional heterogeneity increases in chemoresistant SCLC tumors[3].

The most time-consuming step in our workflow is finalizing the set of signs by tuning ASURAT's parameters through trial and error, which is critical for downstream analyses. Here, users may face difficulty in prioritizing the importance of several signs. For sc68_cisp, we found that the sign scores for meningioma, myopathy, malignant pleural mesothelioma, and other diseases were also upregulated in the group labeled 2, but their actual relationships to the patient's disease were unknown. Nevertheless, ASURAT helped us find well-structured data manifolds and characterize cells in biologically explainable manners for cell types, biological processes, and signaling pathways.

Discussion

We developed a novel scRNA-seq analysis pipeline for simultaneous cell clustering and biological interpretation, allowing users to create systems of cell-type and functional spectra as necessary by inputting collected databases. The resulting matrices can be supplied to unsupervised clustering without gene preselection. We analyzed cancer patient- and healthy donor-derived scRNA-seq datasets: the former was to uncover the unknown characteristics of small cell neuroendocrine cancers, while the latter to confirm cell-type inference, aiming to reproduce results inferred in previous studies.

First, we demonstrated ASURAT's superiority to existing methods with respect to robust, high-quality, and reliable clustering using these datasets (**Fig. 3**). ASURAT yielded well-separated cell clusters from most transcriptomes, despite the dimension reduction processing, while other conventional methods occasionally failed, demonstrating cells were better characterized in the high-dimensional sign score space than in the gene expression space. In practice, we recommend using signature gene-based tools such as Seurat before using ASURAT to broadly understand the transcriptome. Unlike reference-based analyses[10,12,17], ASURAT does not require any bespoke reference but instead takes input from knowledge-based databases.

Next, we found the putative cancer subpopulations existing in the chemosensitive and chemoresistant tumors of SCLC. We found that sc68_vehi (vehicle treatment) contained $\leq$ 15% possible platinum-resistant cells (**Fig. 4c**), suggesting this chemoresistant mechanism latently existed before the therapy. Moreover, we found that sc68_cisp (cisplatin treatment) contained 30% platinum-resistant cells with the same ratio of cells exhibiting PD-L1 expression (**Fig. 5c**).

Notably, we demonstrated that simultaneous cell clustering and biological interpretation of single-cell transcriptomes was viable (**Fig. 1**). The formulation of correlation-based decomposition of signature gene sets was critical for ASURAT's performance (**Fig. 2**). Additionally, we searched virtually the whole parameter space to obtain the desired

interpretation results. Thus, our strategy may greatly improve functional understandings of cancer subpopulations, intracellular heterogeneity, and cellular processes.

However, some limitations are worth noting. Although small cell neuroendocrine cancers have been studied extensively for human tumors by bulk sample RNA-seq analyses[34], few publications address scRNA-seq experiments for such rare cancer subtypes. As available scRNA-seq data and knowledge-based databases expand in size and diversity, our theoretical framework for ASURAT should be generalized to prioritize biological terms more efficiently than manual screening. Furthermore, integrating systems of signs across various conditions should be addressed. One means is applying canonical correlation analysis, which has been incorporated in Seurat[26,38]. Nevertheless, extracting common systems of "biological meanings" across multiple conditions, different cell types, and possibly different species remains challenging.

We also expect ASURAT to improve scRNA-seq data-driven mathematical modeling for patient classification[39], which includes parameter estimations of dynamical systems of gene regulatory network. Since ASURAT detects significant biological functions (e.g., biological process, pathway activity, and chemical reaction) for cell clustering, one can obtain promising candidates for a core regulatory network, which may greatly reduce the numbers of parameters. Another interesting approach to this problem is implementing ASURAT to construct sign networks, which may be analyzed by nonparametric Markov random field theory[40]. We expect ASURAT to open new ways to scRNA-seq analysis from "biological meaning" perspective beyond conventional gene-based analyses.

Acknowledgements

We thank Takeya Kasukawa and Johannes Nicolaus Wibisana 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

M.O. and M.I. started the project. K.I. conceived the theory of ASURAT. K.I. developed the analysis pipeline. J.K. and M.I. prepared the cervical cancer samples and obtained the single-cell RNA sequencing data. M.I. and J.K. translated the computational results. K.I., J.K., and M.O. wrote the manuscript. M.O. supervised the work.

Conflict of interest

The authors declare no conflict of interest.

Supplementary materials

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

Fig. S1 ASURAT outperforms existing methods with respect to robust, high-quality, and reliable clusterings of various single-cell transcriptomes.

Fig. S2 Detailed workflow of **Fig. 1c** focusing on the parameter settings.

Fig. S3 Identification of the cell types in pbmc_4000 by ASURAT.

Fig. S4 Identification of the cell types in pbmc_6000 by ASURAT.

Fig. S5 Identification of the cell types and functional subpopulations in day1_norm by ASURAT.

Fig. S6 Identification of the cell types and functional subpopulations in day7_hypo by ASURAT.

Supplementary File 1 NbClust's output for 2-dim UMAP computed by Seurat across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_001_nbclust_umap_seurat.pdf).

Supplementary File 2 NbClust's output for 2-dim UMAP computed by Monocle 3 across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_002_nbclust_umap_monocle3.pdf).

Supplementary File 3 SC3's output of ASWs across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_003_average_silhouette_sc3.pdf).

Supplementary File 4 NbClust's output for 2-dim UMAP preprocessed by PCA across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_004_nbclust_umap_pca.pdf).

Supplementary File 5 NbClust's output for 2-dim UMAP computed by ASURAT across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_005_nbclust_umap_asurat.pdf).

Supplementary File 6 ASURAT's R function files (SupplementaryFile_006_R_files.zip), which is available on GitHub (https://github.com/keita-iida/ASURAT).

REFERENCES

1 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498, doi:10.1038/s41586-018-0414-6 (2018).

2 Ganesh, K. *et al.* L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat Cancer* **1**, 28–45, doi:10.1038/s43018-019-0006-x (2020).

3 Stewart, C. A. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* **1**, 423–436, doi:10.1038/s43018-019-0020-z (2020).

4 Chen, Z. *et al.* Ligand-receptor interaction atlas within and between tumor cells and T cells in lung adenocarcinoma. *Int J Biol Sci* **16**, 2205–2219, doi:10.7150/ijbs.42080 (2020).

5 Chen, H. J. *et al.* Generation of pulmonary neuroendocrine cells and SCLC-like tumors from human embryonic stem cells. *J Exp Med* **216**, 674–687, doi:10.1084/jem.20181155 (2019).

6 Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell* **182**, 1232–1251 e1222, doi:10.1016/j.cell.2020.07.017 (2020).

7 Devitt, K. *et al.* Single-cell RNA sequencing reveals cell type-specific HPV expression in hyperplastic skin lesions. *Virology* **537**, 14–19, doi:10.1016/j.virol.2019.08.007 (2019).

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

9 Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**, 2285, doi:10.1038/s41467-020-16164-1 (2020).

10 Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172, doi:10.1038/s41590-018-0276-y (2019).

11 Gao, F. *et al.* DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, 44, doi:10.1038/s41389-019-0157-8 (2019).

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

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

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

15 Cancer Genome Atlas Research, N. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378-384, doi:10.1038/nature21386 (2017).

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

17 Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708-718, doi:10.1038/ng.3818 (2017).

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

19 Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* **7**, 44, doi:10.1186/s13326-016-0088-7 (2016).

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

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

22 Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-D655, doi:10.1093/nar/gkx1132 (2018).

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

24 Parra, R. G. *et al.* Reconstructing complex lineage trees from scRNA-seq data using MERLoT. *Nucleic Acids Res* **47**, 8961-8974, doi:10.1093/nar/gkz706 (2019).

25 Tang, W. *et al.* bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics* **36**, 1174-1181, doi:10.1093/bioinformatics/btz726 (2020).

26 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Preprint at https://www.biorxiv.org/content/10.1101/2020.10.12.335331v1*, doi:10.1101/2020.10.12.335331 (2020).

27 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386, doi:10.1038/nbt.2859 (2014).

28 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486, doi:10.1038/nmeth.4236 (2017).

29 Hubert, L. & Arabie, P. Comparing partitions. *J Classif* **2**, 193-218 (1985).

30 Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **20**, 53-65, doi:10.1016/0377-0427(87)90125-7 (1987).

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

32 Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* **61**, 1-36 (2014).

33 van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J Mach Learn Res* **9**, 2579-2605 (2008).

34 Balanis, N. G. *et al.* Pan-cancer Convergence to a Small-Cell Neuroendocrine Phenotype that Shares Susceptibilities with Hematological Malignancies. *Cancer Cell* **36**, 17-34 e17, doi:10.1016/j.ccell.2019.06.005 (2019).

35 Skov, B. G., Holm, B., Erreboe, A., Skov, T. & Mellemgaard, A. ERCC1 and Ki67 in small cell lung carcinoma and other neuroendocrine tumors of the lung: distribution and impact on survival. *J Thorac Oncol* **5**, 453-459, doi:10.1097/JTO.0b013e3181ca063b (2010).

36 Belyanskaya, L. L. *et al.* Cisplatin activates Akt in small cell lung cancer cells and attenuates apoptosis by survivin upregulation. *Int J Cancer* **117**, 755-763, doi:10.1002/ijc.21242 (2005).

37 Kristiansen, G. *et al.* CD24 is an independent prognostic marker of survival in nonsmall cell lung cancer patients. *Br J Cancer* **88**, 231-236, doi:10.1038/sj.bjc.6600702 (2003).

38 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

39 Imoto, H., Zhang, S. & Okada, M. A Computational Framework for Prediction and Analysis of Cancer Signaling Dynamics from RNA Sequencing Data-Application to the ErbB Receptor Signaling Pathway. *Cancers (Basel)* **12**, doi:10.3390/cancers12102878 (2020).

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

41 Kubota, S. *et al.* Dedifferentiation of neuroendocrine carcinoma of the uterine cervix in hypoxia. *Biochem Biophys Res Commun* **524**, 398-404, doi:10.1016/j.bbrc.2020.01.024 (2020).

42 Hashimoto, S. *et al.* Comprehensive single-cell transcriptome analysis reveals heterogeneity in endometrioid adenocarcinoma tissues. *Sci Rep* **7**, 14225, doi:10.1038/s41598-017-14676-3 (2017).

43 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

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

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

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

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

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

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

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

51 Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978, doi:10.1093/bioinformatics/btq064 (2010).

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

53 Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1 30 31–31 30 33, doi:10.1002/cpbi.5 (2016).

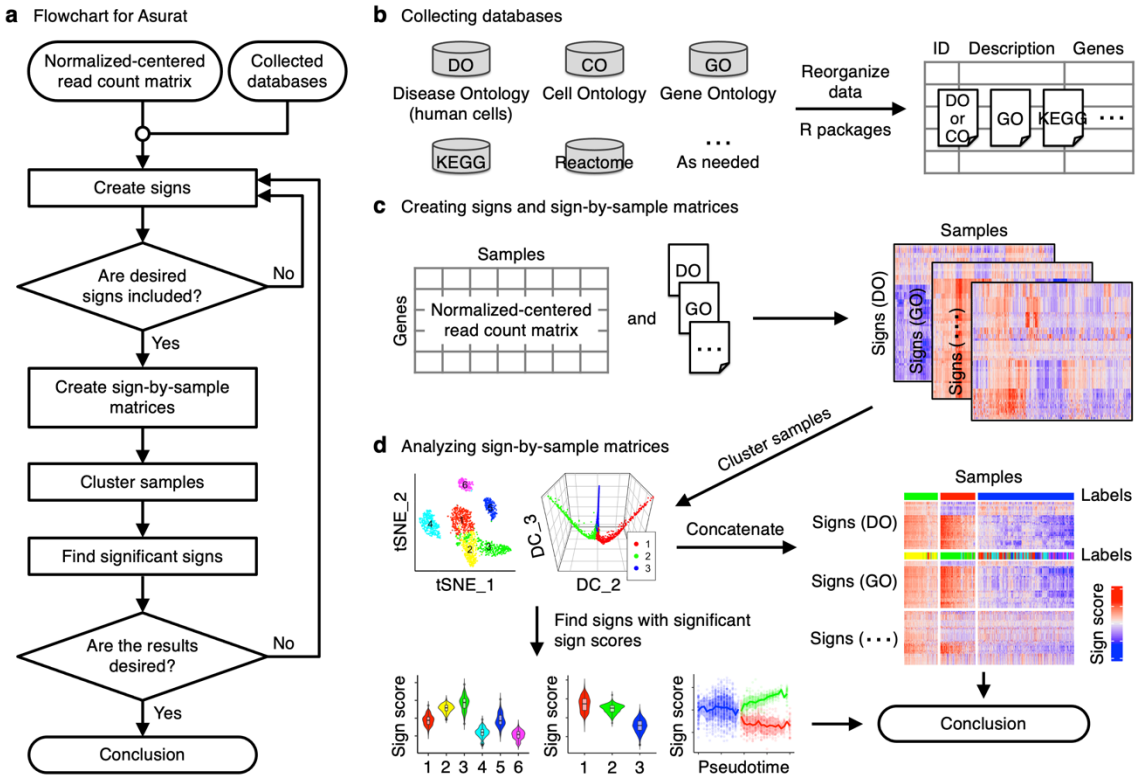

Fig. 1 Workflow of ASURAT. (**a**) Flowchart of the procedures, (**b**) collection of knowledge-based databases (DBs), (**c**) creation of sign-by-sample matrices (SSMs) from normalized and centered read count matrix and the collected DBs, and (**d**) analysis of SSMs to infer cell types and biological functions.

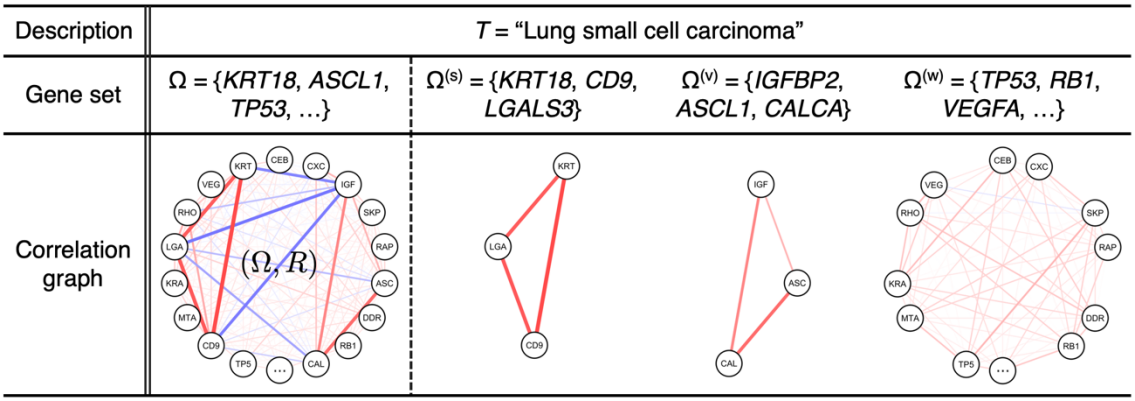

Fig. 2 An example showing decomposition of a correlation graph, which produces three signs based on a Disease Ontology (DO) term. From single-cell RNA sequencing data and a DO term with DOID 5409, which concerns small cell lung cancer, three signs $(T, \Omega^{(i)}, R)$, $i \in \{\mathrm{s}, \mathrm{v}, \mathrm{w}\}$, were produced from their parent sign (T, Ω, R) by decomposing the correlation graph (Ω, R) into strongly, variably, and weakly correlated gene sets, $\Omega^{(\mathrm{s})}$, $\Omega^{(\mathrm{v})}$, and $\Omega^{(\mathrm{w})}$, respectively. Red and blue edges in correlation graphs indicate positive and negative correlations, respectively, and color density indicates the strength of the correlation.

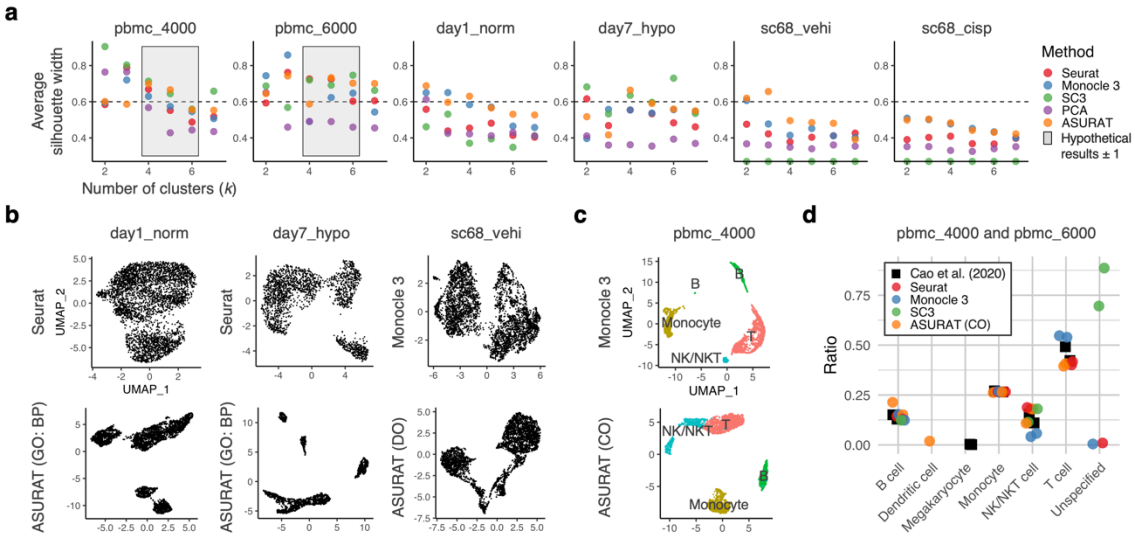

Fig. 3 ASURAT outperforms existing methods for robust, high-quality, and reliable clustering of various single-cell transcriptomes. (**a**) Average silhouette widths (ASWs) versus the number of clusters (*k*), computed by two-dimensional Uniform Manifold Approximation and Projection (UMAP) and k-means clustering for Seurat, Monocle 3, PCA, and ASURAT, while they were computed by consensus matrix-based hierarchical clustering for SC3. The dashed line on the graph represents ASW = 0.6 and the shaded area the hypothetical result. (**b**) Comparison of UMAP plots between different methods using various datasets. The input databases for ASURAT are indicated in parentheses. (**c**) Visualizations of the cell types on UMAP plots for pbmc_4000, which was reanalyzed using the inherent algorithms of Monocle 3 and ASURAT. (**d**) Population ratios in the peripheral blood mononuclear cell (PBMC) datasets, predicted by five different methods.

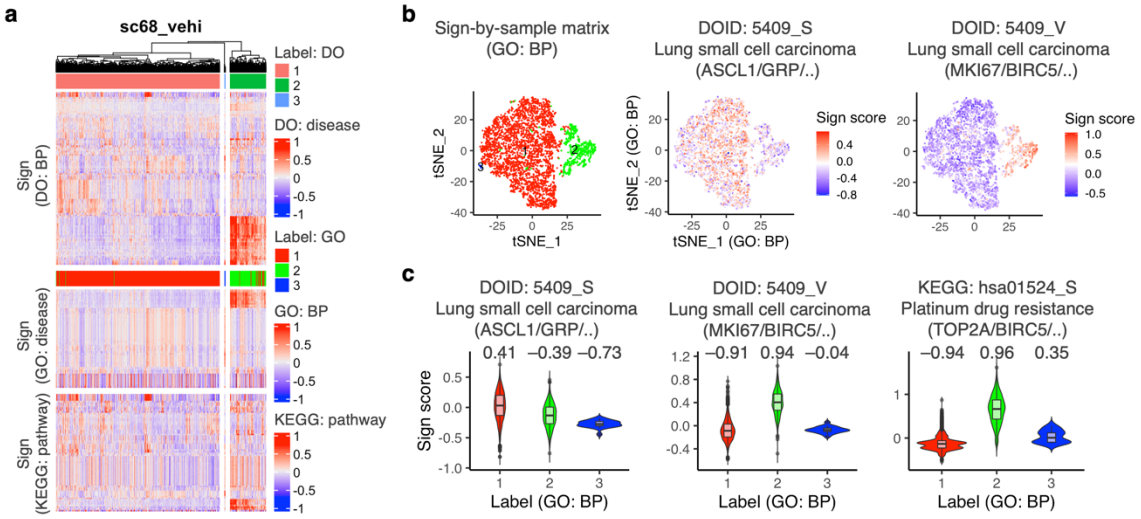

Fig. 4 Identification of the putative cell types in sc68_vehi by ASURAT. (**a**) Heat maps showing the sign scores of sign-by-sample matrices (SSMs) for Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), which are concatenated vertically. (**b**) The t-distributed stochastic neighbor embedding (t-SNE) plots of the SSM for GO, showing cell clustering and sign scores for the indicated sign IDs. (**c**) Violin plots showing the distributions of sign scores for the indicated sign IDs. Each plot represents the separation index for the given group versus all other cells.

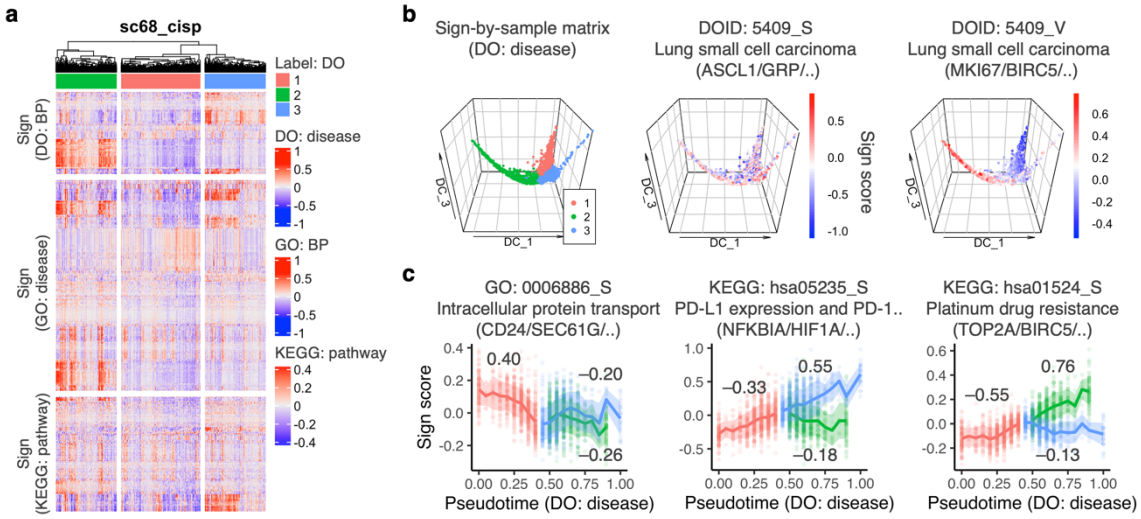

Fig. 5 Identification of the putative cell types in sc68_cisp by ASURAT. (**a**) Heat maps showing the sign scores of sign-by-sample matrices (SSMs) for Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), which are concatenated vertically. (**b**) Diffusion map of the SSM for DO projected onto the first three coordinates, showing cell clustering and sign scores for the indicated sign IDs. (**c**) Sign scores for the indicated sign IDs plotted along the pseudotime, with standard deviations shown as the shaded area. Each plot represents the separation index for the given group versus all other cells.

Methods

Datasets and data processing

Human lung cancer datasets

These data were obtained from circulating tumor cell-derived xenografts cultured with vehicle (symbolized by sc68_vehi) and cisplatin (sc68_cisp) treatments, which were generated from lung cancer patients[3]. The data were produced with the 10x protocol using unique molecular identifiers (UMIs) (https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v2-chemistry). The SRA files were downloaded from Gene Expression Omnibus (GEO) with the accession code GSE138474: GSM4104164 and GSM4104165, which are referenced in Stewart et al[3]. 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 quality controls, the read count matrices of sc68_vehi (resp. sc68_cisp) contained 6581 (resp. 6347) genes and 3923 (resp. 2285) cells.

Human cervical cancer datasets

These data were obtained from cancer tissue originated spheroids (CTOS line cerv21) including small cell neuroendocrine carcinoma, cultured for 1 d under normoxic conditions (symbolized by day1_norm) and 7 d hypoxic conditions (day7_hypo), which were generated from cervical cancer patients[41]. The data were produced by the Nx1-seq protocol using UMIs. The FASTQ files were downloaded from the DNA Data Bank of Japan (DDBJ) with accession codes DRA007915: DRX155817 and DRX155818. The Nx1-seq data were aligned and annotated as described previously[42]. Briefly, the barcode sequences were extracted from the read 1 FASTQ files. The read 2 FASTQ files, which included each cell mRNA, were directly aligned to Refseq transcript sequences (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot) using bowtie 2.2.6[43]. The aligned reads were linked to their paired extracted barcode sequences. By counting mapped reads per barcode, the gene count data in individual cells were obtained. After quality controls, the read count matrices of day1_norm (resp. day7_hypo) contained 5272 (resp. 6213) genes and 3663 (resp. 1947) cells.

Human peripheral blood mononuclear cell datasets

These datasets were obtained from peripheral blood mononuclear cells (PBMCs) of healthy donors, which include approximately 4000 (symbolized by pbmc_4000) and 6000 (pbmc_6000) cells. 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: 4000 PBMCs from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k) and 6000 PBMCs from a healthy donor (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k). After quality controls, the read count matrices of pbmc_4000 (resp. pbmc_6000) contained 6658 (resp. 5169) genes and 3815 (resp. 4878) cells.

Data preprocessing: quality control, normalization, and centering

For all the single-cell RNA sequencing (scRNA-seq) data, the genes and cells with low qualities 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 read counts is less than a user-defined threshold (Chapter 3, Supplementary Notes).

After quality controls, the data were normalized by bayNorm[25], 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 matrices (Chapter 4, Supplementary Notes).

Definition of sign

Let T be a biological description, Ω a variable (e.g., gene) set defined for T, and R a relation structure (e.g., correlation matrix) among Ω. Assume that Ω can be represented by a union of its subsets based on R, that is $\Omega = \bigcup_{i=1}^{n} \Omega^{(i)}$. Then, the triplet $(T, \Omega^{(i)}, R)$ is termed a sign, in particular (T, Ω, R) a parent sign.

Definition of correlated gene set

Let $A = (a_{i,j})$ be a gene-by-sample matrix of size $p \times n$ from transcriptome data, whose entries stand for normalized and centered gene expression levels, and $A = (r_{i,j})$ a correlation matrix of size $p \times p$ defined by A and a certain 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, and let us fix a biological description T_k and the associated gene set $\Omega_k = \{1, 2, \cdots, m_k\}$, where $k = 1, 2, \cdots, K$ for some K. 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 $\Omega_k^{(\mathrm{w})} = \Omega_k \setminus W_k$, where "\" means set difference. 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[44], 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:

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

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 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. **Fig. 2** shows that $\Omega^{(\mathrm{s})}$ and $\Omega^{(\mathrm{v})}$ include *KRT18* and *ASCL1*, which respectively have negative and positive contributions for lung small cell carcinoma. Thus, we interpret that $(T, \Omega^{(\mathrm{s})}, R)$ and $(T, \Omega^{(\mathrm{v})}, R)$ 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[44], tree cutting after hierarchical clustering[45], independent component analysis (ICA)- or principal component analysis (PCA)-based methods[46], and several graph statistical approaches[47,48], we found 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 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-sample matrix of size $p \times n$ from a transcriptomic data, whose entries stand for normalized and centered gene expression levels, and $G = \{1, 2, \cdots, p\}$ a set representing p genes. Assume that we have q biological descriptions and the associated gene sets, denoted by T_k and Ω_k, $k = 1, 2, \cdots, q$, respectively. Let us assume that Ω_k can be decomposed into non-empty $\Omega_k^{(\mathrm{s})}$, $\Omega_k^{(\mathrm{v})}$, and $\Omega_k^{(\mathrm{w})}$ for any k. 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 $c_{k,j}^{(x)}$ as a sign score of the kth sign and jth sample (**Fig. 1c**). Note that ensemble means of sign scores

across 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 groups labeled 0 and 1. Let $\boldsymbol{v}$ be a vector of the labels corresponding to $\boldsymbol{a}$, and $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ 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[49]) 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 group "1," respectively.

Drawbacks

Signs are derived from information in existing databases (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 with few[50].

To overcome this problem, one should monitor what signs are included during data processing (**Fig. 1a**) and carefully tune the parameters to select reliable signs (Supplementary Fig. S2). Our R programming scripts help users perform this process (Supplementary Notes).

Parameter setting

To obtain explainable results of cell clustering in the downstream analysis of ASURAT, it is critical to tune the parameters in the sign creation step (Supplementary Fig. S2). There are six to nine parameters for creating SSMs depending on the database used but many of them have been preset to unbiased and sensible default values. 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 **Fig. 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 $n_{\min}^{(\mathrm{w})}$ (the default value is 2). Furthermore, users can remove redundant signs with similar biological meanings if information contents (ICs)[51] are defined.

Comparison of clustering validity indices of ASURAT with existing methods

To benchmark the clustering qualities of existing methods and ASURAT, we prepared six cancer patient- and healthy donor-derived single-cell RNA-seq datasets. Subsequently, careful quality control and normalization by bayNorm were performed for each dataset. However, 22 additional non-negligible outliers were detected for sc68_vehi by ASURAT, which led to a substantial average silhouette width (ASW) (much greater than 0.9). Hence, those cells were removed from sc68_vehi and the resulting read count table containing 6581 genes and 3901 cells was obtained (Chapter 14.2, Supplementary Notes). Note that such additional preprocessing was undertaken only for the comparison of ASWs.

Using Seurat version 4.0.1[26], we normalized the data by log transform with a pseudo-

count of 1 (default), selected variable genes based on variance stabilizing transformation with a gene-per-cell ratio of 0.2 (as suggested in previous work[52]), 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 Uniform Manifold Approximation and Projection (UMAP)[31] and t-distributed stochastic neighbor embedding (t-SNE)[33], and the resulting two-dimensional cell states were supplied to NbClust[32] (Chapter 14.3.1, Supplementary Notes).

Using Monocle 3 version 0.2.3.0[27], we ran R function preprocess_cds() in the Monocle 3 package using the default settings, in which data were normalized by log transform with a pseudo-count of 1, scaled and centered in gene expression levels, and performed PCA with a dimensionality of the reduced space of 50. The results were used for the computations of UMAP and t-SNE, and resulting two-dimensional cell states were supplied to NbClust (Chapter 14.3.2, Supplementary Notes).

Using SC3 version 1.18.0[28], we normalized the data by log transform with a pseudo-count of 1 (default), performed PCA, and ran R function sc3() in the SC3 package, with the arguments ks = 2:7 and biology = TRUE. This function automatically computed a consensus matrix for each number of clusters and output the ASW based on the hierarchical clustering of the consensus matrix (Chapter 14.3.3, Supplementary Notes). However, sc3() stopped processing and reported errors for sc68_vehi and sc68_cisp irrespective of the arguments.

Using PCA-based clustering, we normalized the data by log transform with a pseudo-count of 1 and ran prcomp() in R stat package. The principal components that explain 90% of the total variability were used for the computations of UMAP and t-SNE, and the resulting two-dimensional cell states were supplied to NbClust (Chapter 14.3.4, Supplementary Notes).

Databases were downloaded in December 2020 and verified for human and mouse scRNA-seq datasets. Using ASURAT, we normalized the data by log transform with a pseudo-count of 1, scaled and centered gene expressions, and created SSMs based on

Disease Ontology (DO) for sc68_vehi and sc68_cisp, Gene Ontology (GO) for day1_norm, day7_hypo, and pbmc_6000, and Cell Ontology (CO) for pbmc_4000. These SSMs were used for the computations of UMAP and t-SNE without preprocessing by PCA, and the resulting two-dimensional cell states were supplied to NbClust (Chapter 14.3.5, Supplementary Notes).

Cell-type inference of PBMC datasets by existing methods and ASURAT

To benchmark the abilities of cell-type inference of existing methods and ASURAT, we prepared the normalized read count tables of pbmc_4000 and pbmc_6000 in the same manner described in the previous section. Using R functions FindClusters() and FindAllMarkers() in Seurat, cluster_cells() and top_markers() in Monocle 3, and sc3_plot_markers() in SC3 packages, we identified several different cell types by manually searching marker genes in GeneCards version 5.2[53] (Chapter 14.4, Supplementary Notes). Seurat identified T cells (resp. marker genes *CD3D*, *CD3E*, *IL32*, *TRAC*), monocytes (*S100A8*, *LYZ*, *CD14*), B cells (*CD79A*, *MS4A1*, *IGHM*, *VPREB3*, *BANK1*), and NK/NKT cells (*NKG7*, *CD160*, *KLRF1*, *GZMA*, *GZMB*, *FGFBP2*, *GNLY*), Monocle 3 identified T cells (*CD3D*, *CD3E*, *CD27*, *IL32*, *TRAC*, *TCF7*), monocytes (*S100A8*, *LYZ*, *CD14*), B cells (*CD79A*, *CD79B*, *MS4A1*, *IGHM*, *VPREB3*, *BANK1*), and NK/NKT cells (*NKG7*, *GNLY*, *CD160*, *GZMA*, *FGFBP2*), and SC3 identified B cells (*CD79A*, *MS4A1*) and NK/NKT cells (*TPD52L2*, *GZMA*, *GZMB*, *GZMH*, *GZMK*).

Using ASURAT, we created SSMs based on CO, GO, and Kyoto Encyclopedia of Genes and Genomes (KEGG), clustered the cells by k-nearest neighbor (KNN) graph generation and Louvain algorithm using Seurat's functions[26] after dimension reduction by PCA, analyzed the separation index (3) of each sign score for each cluster, found the signs upregulated in specific clusters, and inferred the cell types (Supplementary Figs. S3 and S4; Chapter 14.4.4, Supplementary Notes): T cells (respectively marker genes *CD3D*, *CD3E*, *CD247*, *PTPRC*, *IL7R*, etc.), monocytes (*MEF2C*, *LYN*, *CCL3*, *CD14*, *FGR*, etc.), B cells (*CD19*, *CD72*, *CD79B*, *BTK*, *DAPP1*, etc.), NK/NKT cells (*SH2D1A*, *KLRD1*, *NCR3*, *GZMB*, *CD160*, *FGR*, *ITGB2*, *FCGR3A*, etc.), and dendritic cells (*HLA-DOB*, *CCR7*, *CD2*, *FCGR2B*, *BLK*, etc.).

Cell-type inference of cervical cancer datasets by ASURAT

To validate ASURAT's reliable cell-type inference, the normalized read count tables of day1_norm and day7_hypo were prepared in the same manner as described in the previous section. Previous work studying human cervical cancers using CTOS methods indicated that some small cell neuroendocrine carcinomas (SCNCs) exhibited combined phenotypes with other non-SCNC cells[41]. Additionally, hypoxia drove divergent differentiation of SCNCs, but detailed molecular information remained to be elucidated. Using ASURAT, we created SSMs based on DO, GO, and KEGG, and clustered the cells by one of the following: (i) PCA, followed by KNN graph generation and Louvain algorithm using Seurat's functions[26] and (ii) diffusion map generation, followed by allocation of cells to the different branches of the data manifold by using MERLoT[24]. Here, cells in day1_norm were clustered by (i), while those in day7_hypo were clustered by (i) and (ii) for SSM using DO and GO, respectively (Supplementary Figs. S5 and S6).

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 information

Functional annotation-driven unsupervised clustering of single-cell transcriptomes

Keita Iida[1*], Jumpei Kondo[2,3], 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

ORCID (Keita Iida): https://orcid.org/0000-0002-1076-830X

ORCID (Jumpei Kondo): https://orcid.org/0000-0002-1350-0480

ORCID (Masahiro Inoue): https://orcid.org/0000-0001-7315-026X

ORCID (Mariko Okada): https://orcid.org/0000-0002-6210-8223

Keywords:

Single-cell transcriptome; biological interpretation; unsupervised clustering; small cell neuroendocrine cancer

Supplementary Notes

Supplementary Notes are written in separate files, which are structured as follows:

Supplementary Files

Supplementary Files are prepared in separate files, which are structured as follows:

Supplementary File 1.

NbClust's output for 2-dim UMAP computed by Seurat across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_001_nbclust_umap_seurat.pdf).

Supplementary File 2.

NbClust's output for 2-dim UMAP computed by Monocle 3 across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_002_nbclust_umap_monocle3.pdf).

Supplementary File 3.

SC3's output of ASWs across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_003_average_silhouette_sc3.pdf).

Supplementary File 4.

NbClust's output for 2-dim UMAP preprocessed by PCA across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_004_nbclust_umap_pca.pdf).

Supplementary File 5.

NbClust's output for 2-dim UMAP computed by ASURAT across six cancer patient- and healthy donor-derived scRNA-seq datasets (SupplementaryFile_005_nbclust_umap_asurat.pdf).

Supplementary File 6.

ASURAT's R function files (SupplementaryFile_006_R_files.zip).

Supplementary Figures

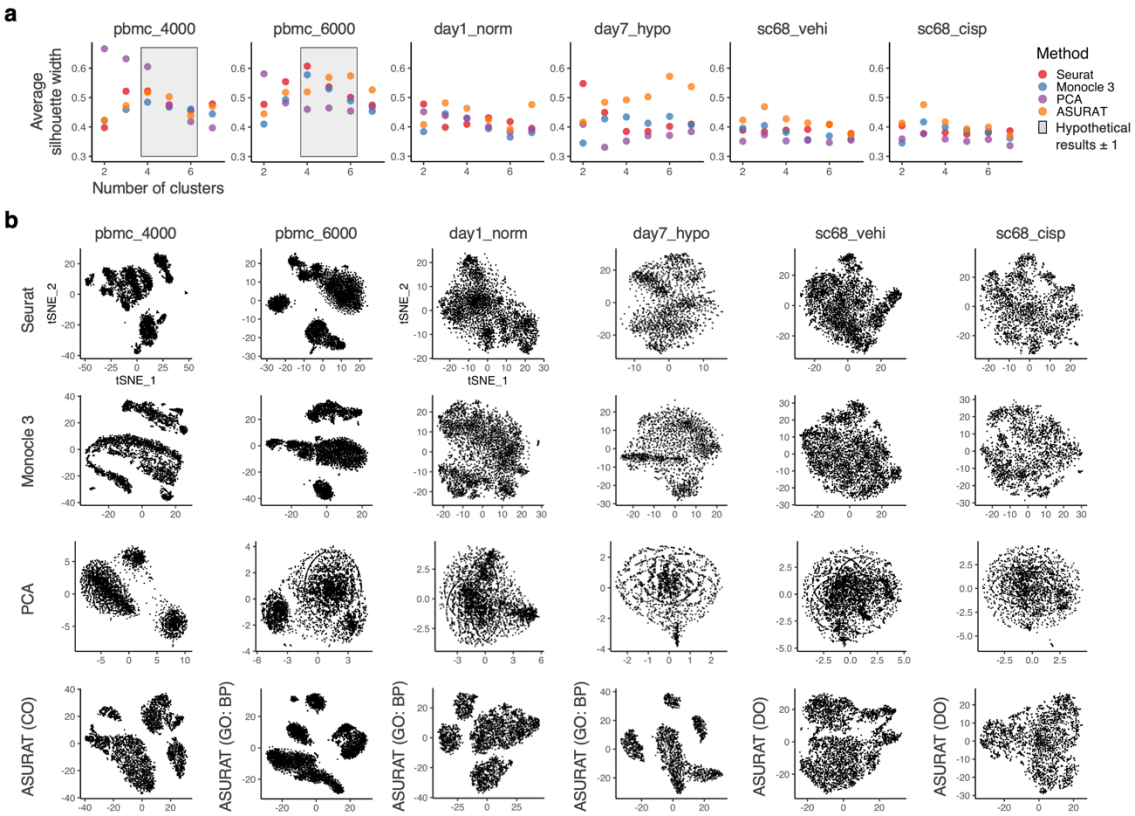

Fig. S1 ASURAT outperforms existing methods with respect to producing robust, high-quality, and reliable clusterings of various single-cell transcriptomes. (**a**) Average silhouette widths (ASWs) versus the number of clusters, computed by two-dimensional t-distributed stochastic neighbor embedding (t-SNE) and k-means clustering for Seurat, Monocle 3, PCA, and ASURAT. (**b**) Comparison of t-SNE plots between different methods using various datasets. The input databases for ASURAT are indicated in parentheses.

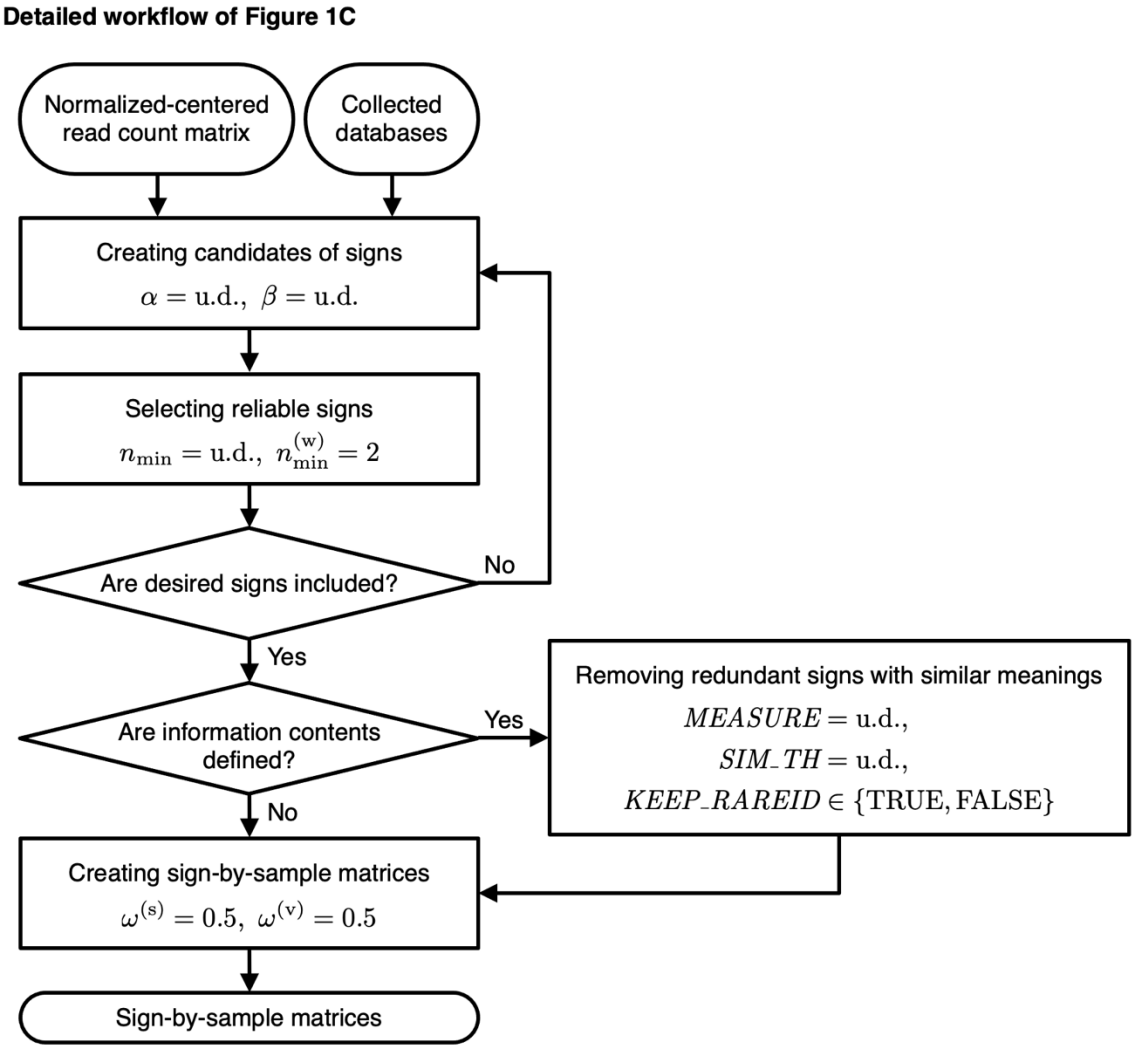

Fig. S2 Detailed workflow of **Fig. 1c** focusing on the parameter settings. The indicated values are preset as default in ASURAT, while "u.d." stands for the value or argument that users must define. Here, α and β are positive and negative threshold values of correlation coefficients, $n_{\min}$ and $n_{\min}^{(\text{w})}$ positive integers for selecting reliable signs, *MEASURE* the name of information content (IC)-based method defining semantic similarities, *SIM_TH* a threshold value used to regard two biological terms as similar, *KEEP_RAREID* determines whether the signs with larger ICs are kept or not (if TRUE, the signs with larger ICs are kept), and $\omega^{(\text{s})}$ and $\omega^{(\text{v})}$ weight constants are used to define SSMs.

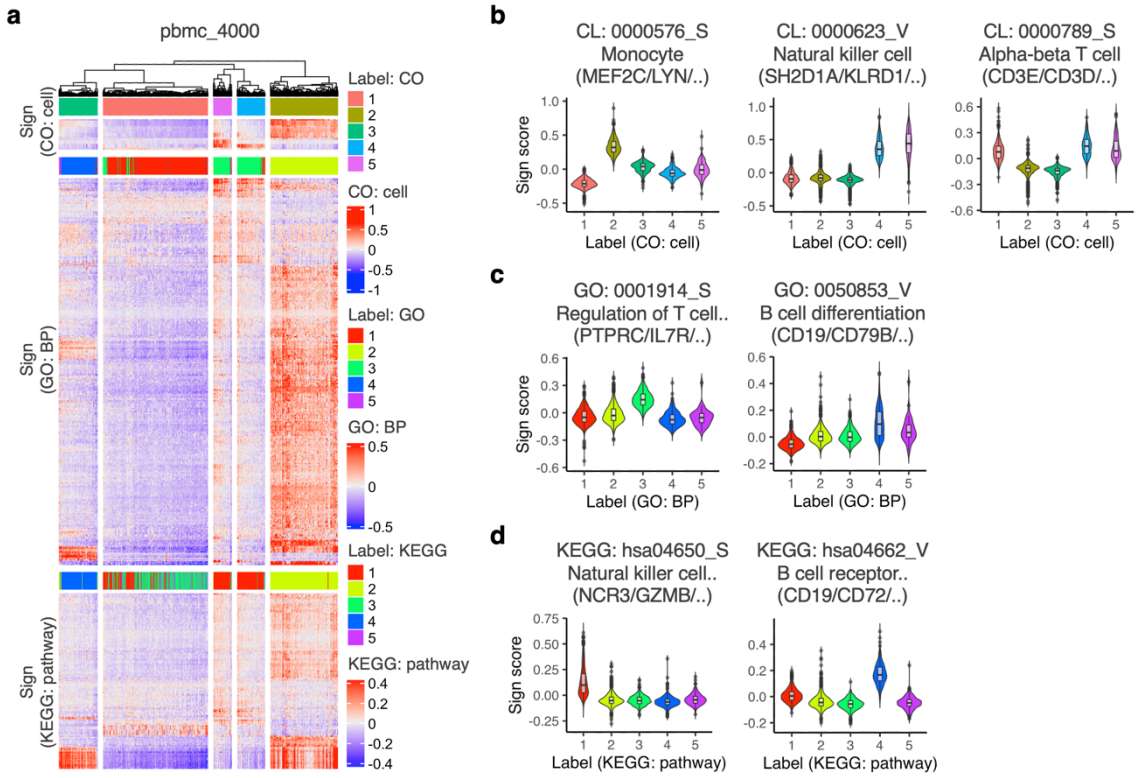

Fig. S3 Identification of the cell types in pbmc_4000 by ASURAT. (**a**) Heat maps showing the sign scores of sign-by-sample matrices (SSMs) for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), which are concatenated vertically. The cells are clustered by k-nearest neighbor (KNN) graph generation and Louvain algorithm by using Seurat's functions in the R package after dimension reduction by principal component analysis. (**b**)-(**d**) Violin plots showing the distributions of sign scores for the indicated sign IDs. The cell type labels were inferred by CO as follows: T cell (label 1), monocyte (label 2), B cell (label 3), and NK/NKT cell (label 4 and 5).

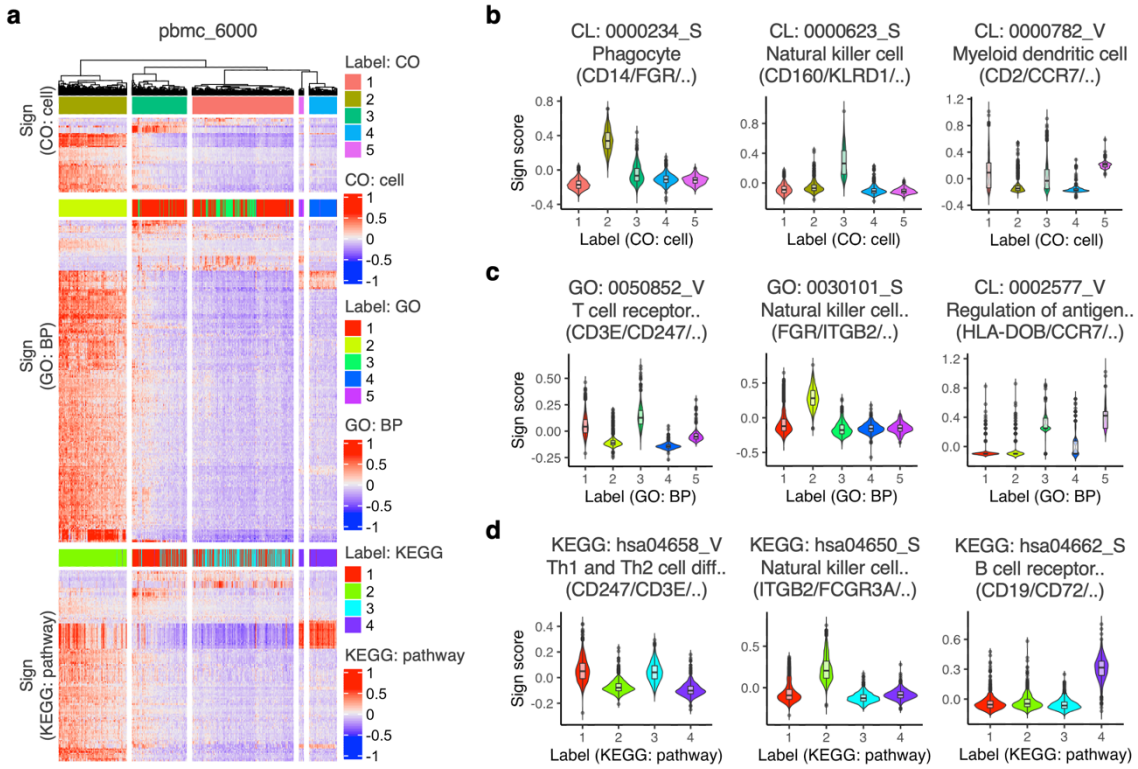

Fig. S4 Identification of the cell types in pbmc_6000 by ASURAT. (**a**) Heat maps showing the sign scores of sign-by-sample matrices (SSMs) for Cell Ontology (CO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), which are concatenated vertically. The cells are clustered by k-nearest neighbor (KNN) graph generation and Louvain algorithm by using Seurat's functions in the R package after dimension reduction by principal component analysis. (**b**)-(**d**) Violin plots showing the distributions of sign scores for the indicated sign IDs. The cell type labels were inferred by CO as follows: T cell (label 1), monocyte (label 2), NK/NKT cell (label 3), B cell (label 4), and dendritic cell (label 5).

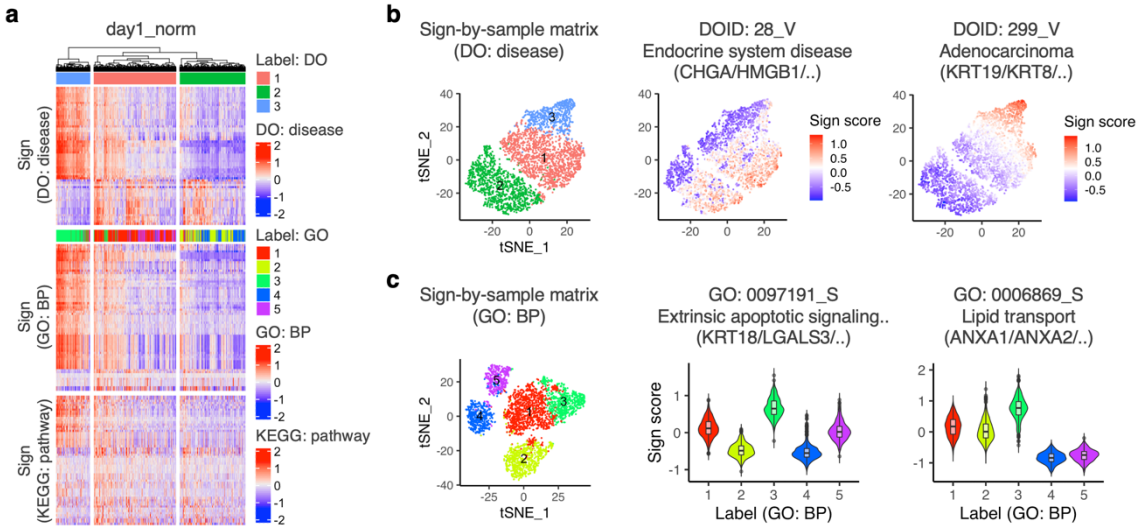

Fig. S5 Identification of putative the cell types and functional subpopulations in day1_norm by ASURAT. (**a**) Heat maps showing the sign scores of sign-by-sample matrices (SSMs) for Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), which are concatenated vertically. The cells were clustered by k-nearest neighbor (KNN) graph generation and Louvain algorithm by using Seurat's functions in the R package after the dimension reduction by principal component analysis. (**b**) t-SNE plots of the SSM for DO, showing the cell clustering and sign scores for the indicated sign IDs. (**c**) t-SNE plots of the SSM for GO and violin plots showing the distributions of sign scores for the indicated sign IDs.

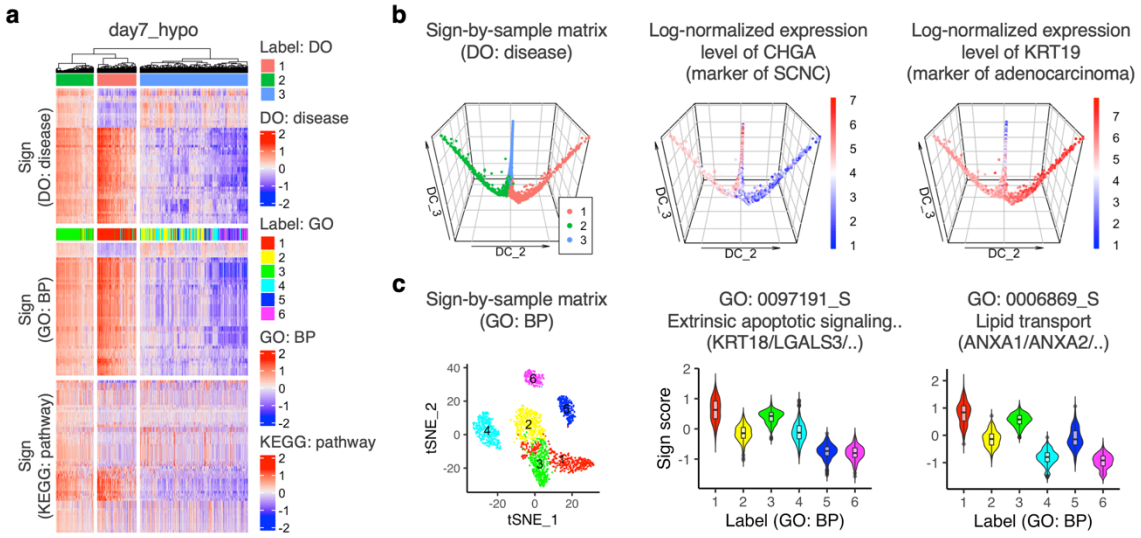

Fig. S6 Identification of putative the cell types and functional subpopulations in day7_hypo by ASURAT. (**a**) Heat maps showing the sign scores of sign-by-sample matrices (SSMs) for Disease Ontology (DO), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG), which are concatenated vertically. The cells were clustered by (i) k-nearest neighbor (KNN) graph generation and Louvain algorithm by using Seurat's functions in the R package after the dimension reduction by principal component analysis for the SSM for DO, and (ii) diffusion map, followed by allocations of samples to the different branches of the data manifold by using MERLoT for the SSM for GO. (**b**) t-SNE plots of the SSM for DO, showing cell clustering and sign scores for the indicated sign IDs. (**c**) t-SNE plots of the SSM for GO and violin plots showing the distributions of sign scores for the indicated sign IDs.

Supplementary File 1

day1_norm

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	11.6055	6545.1842	2023.4373	5.4614	5654.5078	280077356.1	53547181.54	16886.2314	3.6435	2.7878	0.3819	0.6484	0.558	1.8001	-985.8611	-0.4441	0.3388	8443.1157	0.743	1.6248	0.3171	0.0049	0	0.6522	1.9797	0.7596
3	0.4597	6091.4104	1623.8283	8.8946	9037.0043	250278321.8	45443496.74	10875.3937	6.7871	4.3286	0.3534	0.9787	0.4387	1.8808	-842.9433	-0.4678	0.4473	3625.1312	0.6383	0.6811	0.7635	0.0016	0	0.9217	1.593	0.5844
4	1.3173	6402.1796	1221.3225	14.2116	11568.5766	222921082.4	12660140.15	7533.1632	9.7563	6.2491	0.3072	0.853	0.4542	1.0742	-100.4768	-0.069	0.4311	1883.2908	0.6059	0.5865	1.0128	0.0071	0	0.9242	1.315	0.5518
5	2.6163	6707.8493	621.9483	18.1842	13617.9894	199061667.3	8364841.041	5647.9555	14.3587	8.335	0.3134	0.7553	0.4813	1.4563	-343.7384	-0.3128	0.3991	1129.5911	0.5826	0.9744	1.1576	0.0117	0	0.9166	1.1299	0.345
6	0.442	6401.3157	865.061	15.6998	14893.8983	202336934.9	8129189.005	4827.2129	17.4713	9.7521	0.3063	0.9206	0.4139	0.8646	145.9518	0.1563	0.3718	804.5355	0.5428	0.6824	1.3718	0.0058	0	1.2961	1.0465	0.3248
7	1.7717	6738.6193	385.6434	19.3124	16350.9804	185017760.9	4563635.803	3903.7769	22.3108	12.059	0.3407	0.9216	0.4041	2.3202	-583.7913	-0.5677	0.35	557.6824	0.5112	0.9908	1.5608	0.0063	0	1.2814	0.9499	0.3567
Number_clusters	2	7	5	7	3	5	4	3	7	5	6	2	2	2	2	2	3	3	2	2	2	5	0	2	0	6
Value_Index	11.6055	6738.6193	599.3742	19.3124	3382.4965	27134682.54	32783356.58	2668.6073	4.8395	-0.6687	0.3063	0.6484	0.558	1.8001	-985.8611	-0.4441	0.4473	4817.9845	0.743	1.6248	0.3171	0.0117	0	0.6522	0	0.3248

day7_hypo

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	6.4904	4485.0373	705.8133	14.2897	3810.1019	136617237	37655607.8	12065.5426	6.0542	3.3059	0.3518	0.6571	0.6159	0.4811	1278.1255	1.0777	0.4418	6032.7713	0.803	2.9526	0.3331	0.012	0	0.6154	2.2981	0.5371
3	0.2122	3407.4511	2609.9035	9.8341	5751.9381	113382322	34732299.3	8852.9361	16.4202	4.5056	0.3068	0.8016	0.4681	1.353	-201.4274	-0.2603	0.4018	2950.9787	0.6529	0.2325	0.6807	0.007	0	0.8767	1.8962	0.494
4	8.2066	6188.1834	732.518	41.0003	8222.7284	56662087.7	3896360.67	3779.199	20.2684	10.5546	0.3246	0.6714	0.5545	1.7711	-409.2586	-0.4344	0.4606	944.7998	0.6906	1.1801	0.6966	0.0114	0	0.6913	1.2678	0.2584
5	4.4803	6570.6092	336.7232	43.8981	9499.1546	45961603.4	2086330.1	2744.5092	25.2207	14.5337	0.3375	0.7172	0.5333	1.7803	-379.1373	-0.4374	0.4254	548.9018	0.6213	1.346	0.8826	0.0077	0	0.8935	1.088	0.2403
6	0.7615	6232.0433	391.4234	41.1042	10273.8336	44458938.8	2054009.88	2338.9575	30.5107	17.0537	0.3247	0.8357	0.4834	1.0048	-3.0731	-0.0047	0.3924	389.8262	0.585	1.2115	0.9965	0.0192	0	1.1138	0.9963	0.2056
7	0.9348	6302.6562	208.2151	41.518	10899.1159	43891282.4	1352772.52	1946.4376	37.7813	20.4928	0.2926	0.8721	0.461	0.9893	5.16	0.0108	0.3649	278.0625	0.5367	1.5481	1.17	0.0085	0	1.1686	0.9082	0.1972
Number_clusters	4	5	3	5	4	4	4	4	3	4	7	2	2	3	3	2	4	3	2	2	2	6	0	2	0	7
Value_Index	8.2066	6570.6092	1904.0902	43.8981	2470.7903	46019750	30835938.6	4039.0472	10.366	-2.0699	0.2926	0.6571	0.6159	1.353	-201.4274	1.0777	0.4606	3081.7926	0.803	2.9526	0.3331	0.0192	0	0.6154	0	0.1972

sc68_vehi

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	5.2345	5244.2173	1762.9078	-5.9883	4971.1443	459442647.7	112489668.6	21823.1138	2.4766	2.345	0.3387	0.8596	0.4756	1.4411	-1016.1964	-0.3059	0.5049	10911.5569	0.6082	0.6605	0.4953	0.0038	0	0.7295	2.1548	1.1806
3	0.5806	4687.9294	1028.6	-9.667	7927.4464	484497589.9	39205675.89	15028.2068	5.4607	3.4053	0.3501	0.8773	0.4221	0.5653	1218.0018	0.7685	0.4488	5009.4023	0.6212	1.0573	0.7098	0.0032	0	0.7759	1.812	0.8493
4	0.5539	4291.7488	1479.9549	-13.1145	9583.195	563425158.6	33330825.62	11890.5431	6.4413	4.3039	0.35	1.0234	0.378	1.4604	-488.3405	-0.315	0.4236	2972.6358	0.5761	0.5835	1.0186	0.0044	0	0.9437	1.619	0.7589
5	3.4838	4809.9666	922.7938	-4.8492	12436.9093	423599137.9	21211381.96	8617.7859	9.3672	5.9384	0.3621	0.9497	0.4035	0.8854	164.7498	0.1293	0.4048	1723.5572	0.5432	0.558	1.3534	0.0064	0	0.9563	1.3896	0.7975
6	0.4852	4942.6814	956.2717	-2.48	13788.7834	431335050.6	13257256.95	6967.4892	12.2007	7.3449	0.3314	0.8453	0.4119	2.2238	-704.9684	-0.5495	0.3741	1161.2482	0.5154	0.3536	1.592	0.0043	0	0.9628	1.2346	0.4804
7	1.8043	5288.164	729.4613	2.2512	15613.9299	367719423.8	5175289.287	5594.0735	16.4309	9.1482	0.322	0.7822	0.4261	1.5664	-415.4762	-0.3611	0.3524	799.1534	0.5045	0.5313	1.6937	0.006	0	0.9902	1.1046	0.4214
Number_clusters	2	7	3	7	3	5	3	3	7	5	7	7	2	2	2	2	2	3	3	1	2	5	0	2	0	7
Value_Index	5.2345	5288.164	734.3077	2.2512	2956.3022	147561933.5	73283992.79	3657.2433	4.2302	-0.2279	0.322	0.7822	0.4756	1.4411	-1016.1964	-0.3059	0.5049	5902.1546	0.6212		0.4953	0.0064	0	0.7295	0	0.4214

sc68_cisp

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	2.5107	1663.6929	1337.0659	-3.9078	2919.5304	180061066.6	46145839.95	14975.011	2.5378	1.7287	0.3483	1.1647	0.3903	0.6765	741.1774	0.4778	0.3244	7487.5055	0.5391	0.3958	0.5655	0.0135	1e-04	0.8517	2.3195	1.3046
3	0.6875	1986.6895	1036.3124	-8.449	4559.9734	197612510.7	17134497.42	9444.0131	3.5427	2.7412	0.3307	0.9148	0.4025	1.6257	-464.9512	-0.3843	0.4546	3148.0044	0.5985	0.3613	0.9711	0.0042	1e-04	0.6375	1.861	1.1954
4	1.3732	2270.3716	393.786	-1.6318	6236.8844	168644736.4	10107600.62	6494.638	5.9693	3.986	0.3287	0.8395	0.4092	1.2542	-217.036	-0.2023	0.4301	1623.6595	0.6067	0.7351	1.2396	0.0052	1e-04	0.6488	1.5606	0.9301
5	6.0688	2094.27	484.1893	-5.5338	6996.7899	188956887.5	6280681.980	5538.4876	7.344	4.6742	0.3414	0.96	0.3679	1.4527	-255.5274	-0.311	0.3953	1107.6975	0.5742	0.499	1.5206	0.0069	1e-04	0.8059	1.4439	0.8732
6	0.2059	2127.1181	519.8685	-4.7269	7853.8067	186998377.0	5623268.148	4568.3383	9.2651	5.6668	0.3295	0.9593	0.3664	1.0212	-12.4633	-0.0207	0.3695	761.3897	0.5477	0.2765	1.8057	0.0092	1e-04	0.8189	1.306	0.6242
7	0.7804	2262.6019	513.8555	-1.6942	8927.4784	159098436.9	2807379.333	3719.8043	12.2122	6.9594	0.3458	0.8305	0.398	1.0937	-49.6138	-0.0854	0.3499	531.4006	0.5396	0.3078	1.932	0.0052	1e-04	0.7403	1.1811	0.6081
Number_clusters	5	4	4	4	4	4	3	3	7	4	4	7	4	2	2	2	3	3	4	1	2	2	0	3	0	7
Value_Index	6.0688	2270.3716	642.5264	-1.6318	1676.911	49279925.34	29011342.53	2581.6228	2.9471	-0.5567	0.3287	0.8305	0.4092	0.6765	741.1774	0.4778	0.4546	4339.5011	0.6067		0.5655	0.0135	0	0.6375	0	0.6081

pbmc_4000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.0485	3532.149	19932.5615	6.2083	12057.3027	6083879531.	1933179110	1200835.2387	22.7413	1.9263	0.3176	0.6625	0.5841	0.1416	16968.3362	6.0623	0.4122	100417.6193	0.6169	0.0188	0.3027	0.4006	0	0.4835	5.7522	0.5032
3	15.7831	20959.0597	2538.8495	135.0128	20530.1053	1485356272.	205090313.6	32249.596	34.8664	11.9964	0.3077	0.3427	0.789	1.1896	-218.3942	-0.1592	0.553	10749.8653	0.9335	2.1246	0.3027	0.4766	0	0.1534	2.5374	0.0929
4	1.4566	24118.6943	256.0855	136.8985	23185.255	1316586608.	84200926.25	19357.3253	42.9745	19.9861	0.2251	0.4828	0.6687	0.7261	382.047	0.3768	0.4874	4839.3313	0.7884	4.5901	0.4326	0.0115	0	0.3531	1.9267	0.1471
5	3.4004	19363.4767	573.4775	114.2655	23584.4511	1852787051.	68980404.30	18138.4845	44.8528	21.3291	0.2197	0.6682	0.5519	5.7007	-986.2026	-0.8232	0.4367	3627.6969	0.7592	4.1391	0.4698	0.0058	0	0.5586	1.8371	0.139
6	1.3417	17932.4263	502.3355	105.6859	24581.1329	2054599071.	41327528.73	15765.4798	50.9919	24.5395	0.2023	0.7766	0.4874	1.7056	-389.2992	-0.4128	0.3999	2627.58	0.6654	4.5142	0.6205	0.0058	0	0.6361	1.6752	0.1546
7	0.3046	16993.7403	156.0393	99.9834	25626.5901	2126215220.	41048102.54	13928.564	59.4376	27.7759	0.1916	0.7222	0.5179	1.2055	-124.8074	-0.1701	0.3712	1989.7949	0.5904	-15.4805	0.7955	0.0038	0	0.7425	1.5114	0.1631
Number_clusters	3	4	3	4	3	3	3	3	3	4	7	3	3	3	3	3	3	3	3	1	3	3	0	3	0	3
Value_Index	15.7831	24118.6943	17393.7119	136.8985	8472.8027	4429753593.	1912670078	155693.3719	12.125	-6.6468	0.1916	0.3427	0.789	1.1896	-218.3942	-0.1592	0.553	89667.754	0.9335		0.3027	0.4766	0	0.1534	0	0.0929

pbmc_6000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.2613	6800.1187	24985.7357	-14.0345	8252.9363	3063427606(	3534974801.	247606.8346	3.8633	2.3946	0.2455	0.8443	0.5932	0.0449	41131.0093	21.2763	0.3476	123803.4173	0.6563	0.0516	0.2626	0.0016	0	0.6848	5.926	0.5362
3	6.7929	33308.774	7026.0634	115.4784	26134.1014	1763630070.	55328855.51	40430.7016	48.4469	14.6651	0.2938	0.3398	0.7621	2.8348	-1250.4711	-0.6462	0.5423	13476.9005	0.7902	1.949	0.2239	0.6649	0	0.2263	2.3739	0.0494
4	14.6733	56540.312	891.9798	158.8042	32246.8604	895485038.1	45884659.80	16561.5176	64.3512	35.8012	0.1748	0.4093	0.7262	0.1114	5023.0812	7.9605	0.4893	4140.3794	0.6637	0.2926	0.3013	0.0059	0	0.4166	1.5605	0.0698
5	2.909	50378.3771	1084.3551	146.6163	33576.5162	1065362554.	27640631.43	13999.5005	79.8305	42.3531	0.3224	0.3491	0.7249	1.5333	-738.3754	-0.3475	0.4383	2799.9001	0.6641	7.3999	0.2983	0.011	0	0.3935	1.5011	0.0433
6	0.2761	49477.6993	1226.656	143.5393	36552.3912	833516161.0	25667433.19	11451.3178	95.0142	51.7776	0.29	0.5504	0.6026	0.3936	1414.0447	1.5385	0.403	1908.553	0.5362	2.1117	0.4735	0.0055	0	1.0792	1.3229	0.0972
7	1.3889	51806.3322	418.0528	145.9419	38432.864	771590269.7	21179573.29	9148.0517	111.4996	64.814	0.2698	0.5551	0.6054	9.9862	-1673.7437	-0.898	0.3741	1306.8645	0.5116	15.786	0.5055	0.0064	0	1.2263	1.1774	0.0993
Number_clusters	4	4	3	4	3	3	3	3	4	4	4	3	3	3	3	3	3	3	3	1	3	3	0	3	0	5
Value_Index	14.6733	56540.312	17959.6723	158.8042	17881.1651	2800250095	3479645946.	183306.9489	44.5836	-14.5841	0.1748	0.3398	0.7621	2.8348	-1250.4711	-0.6462	0.5423	110326.5167	0.7902		0.2239	0.6649	0	0.2263	0	0.0433

Supplementary File 2

day1_norm

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	1.95	9072.2252	5181.8321	26.632	8607.1186	263462904.1	37855405.79	18768.0281	6.7664	3.4781	0.3353	0.5346	0.6502	0.3288	4835.627	2.0403	0.597	9384.014	0.8318	0.8665	0.306	0.3302	0	0.6087	1.9996	0.2635
3	84.3743	13543.8072	1255.554	71.3422	14162.4251	130092092.6	20083970.13	7770.1069	15.5799	8.401	0.3296	0.5576	0.6505	1.3542	-353.8986	-0.2612	0.5417	2590.0356	0.8216	1.9897	0.476	0.0254	0	0.5687	1.3384	0.2061
4	0.1834	12541.7418	739.9585	64.7438	16388.7966	125940139.3	12232001.41	5785.4295	21.8027	11.2829	0.3224	0.7156	0.5839	1.1455	-122.096	-0.1268	0.4774	1446.3574	0.7246	1.6085	0.6596	0.0075	0	1.0144	1.1482	0.3346
5	0.4185	11490.3893	1172.6752	58.0858	17563.7471	142783860.2	7203796.562	4812.2497	25.4841	13.5647	0.3085	0.7161	0.5703	0.3575	1964.3939	1.7959	0.4304	962.4499	0.7005	1.3576	0.7136	0.0022	0	1.0424	1.0302	0.1678
6	1.69	12370.3189	879.2112	62.7657	19742.2537	113436387.9	3029789.705	3644.0474	35.4878	17.9132	0.3036	0.8801	0.465	2.3261	-531.3263	-0.5688	0.3967	607.3412	0.6227	1.1818	0.9109	0.0025	0	1.1501	0.8969	0.1532
7	15.0212	12929.9776	504.7753	65.3847	21153.9958	105018613.3	1823542.315	2937.7559	42.7084	22.2199	0.264	0.8992	0.4575	0.6774	381.5331	0.4757	0.3694	419.6794	0.5618	1.4379	1.0954	0.0038	0	1.3664	0.8074	0.1897
Number_clusters	3	3	3	3	3	3	3	3	6	3	7	2	3	3	3	2	2	3	2	1	2	2	0	3	0	6
Value_Index	84.3743	13543.8072	3926.2781	71.3422	5555.3065	129218858.1	17771435.65	9013.2437	10.0037	-2.041	0.264	0.5346	0.6505	1.3542	-353.8986	2.0403	0.597	6793.9784	0.8318		0.306	0.3302	0	0.5687	0	0.1532

day7_hypo

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.4386	1326.0921	2633.1835	-3.4047	2715.5466	176082207.3	144783819.2	16007.3125	2.1187	1.6818	0.373	1.1499	0.3963	0.5185	1387.2903	0.9278	0.4506	8003.6563	0.485	-0.4387	0.547	0.0063	0	0.9539	2.5584	1.0058
3	2.343	2875.8047	1479.7757	18.8738	5737.2947	83921268.51	5874288.422	6800.5624	7.2969	3.9586	0.3716	0.6182	0.5575	0.4728	1155.2217	1.1141	0.5002	2266.8541	0.7785	0.8729	0.4967	0.0059	1e-04	0.4999	1.6858	0.4076
4	4.7287	3867.8489	662.5686	33.5746	7655.9125	55691113.43	2343529.376	3861.32	13.0009	6.972	0.3222	0.6964	0.5542	0.6421	449.1882	0.5565	0.4619	965.33	0.712	0.9314	0.8239	0.0085	1e-04	0.6609	1.2508	0.2752
5	0.5907	4053.6521	715.9756	35.0692	8822.8455	47787231.85	1212815.593	2879.4271	17.9208	9.3494	0.2694	0.7178	0.5373	0.9514	21.4485	0.0509	0.4227	575.8854	0.6474	0.1623	1.0513	0.0072	1e-04	0.8442	1.0607	0.3449
6	16.5369	4579.3563	342.9136	40.5379	9888.0272	39817753.73	1094708.925	2103.7994	23.6447	12.7964	0.3249	0.6481	0.556	0.9703	13.2284	0.0305	0.3918	350.6332	0.6519	0.8956	1.0251	0.0095	1e-04	0.855	0.9296	0.2196
7	0.2158	4545.1297	368.8832	39.5985	10527.6914	39020058.87	789465.0685	1787.9287	28.3727	15.0571	0.3139	0.6277	0.5461	1.3955	-138.8811	-0.2827	0.365	255.4184	0.627	0.789	1.1054	0.0032	1e-04	0.9035	0.862	0.2491
Number_clusters	6	6	3	6	3	3	3	3	6	6	5	3	3	4	4	2	3	3	3	1	3	6	0	3	0	6
Value_Index	16.5369	4579.3563	1153.4079	40.5379	3021.7481	63930783.77	138909530.8	6267.5077	5.7238	-1.1862	0.2694	0.6182	0.5575	0.6421	449.1882	0.9278	0.5002	5736.8021	0.7785		0.4967	0.0095	0	0.4999	0	0.2196

sc68_vehi

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	2.9926	9595.1413	802.0216	19.1466	7339.6167	251789173.5	58796220.98	15974.5751	5.4619	3.4609	0.3695	0.6324	0.6067	0.5711	1400.8773	0.7507	0.4768	7987.2875	0.8027	5.7953	0.3781	0.0197	0	0.6072	1.8671	0.3764
3	2.9182	6183.852	1673.961	1.8766	8766.5329	392972716.1	48746449.89	13249.2198	6.1404	4.1728	0.3522	0.9424	0.477	0.5869	1430.0498	0.7034	0.4521	4416.4066	0.6747	0.7213	0.6359	0.0115	0	1.0184	1.6381	0.4078
4	0.4248	6449.2998	1969.8278	7.1988	12446.8306	271966838.7	19194748.38	9268.8119	9.0888	5.9648	0.2978	1.0402	0.4144	1.247	-299.1326	-0.1979	0.448	2317.203	0.6478	0.6411	0.8649	0.0063	0	0.9193	1.3864	0.3421
5	1.9775	7772.3964	341.2797	20.8163	14994.8093	221142403.2	11884458.64	6156.7445	14.4148	8.9799	0.2824	0.895	0.452	1.5	-412.6604	-0.3328	0.4119	1231.3489	0.6069	2.8087	1.0863	0.0047	0	0.9247	1.1364	0.3458
6	1.324	6829.0935	348.9227	12.9063	15623.5607	271041784.5	8969981.011	5660.867	15.7098	9.7665	0.2756	1.029	0.4114	1.2462	-306.8563	-0.1974	0.3788	943.4778	0.5688	2.2142	1.2622	0.0032	0	1.3002	1.0754	0.345
7	0.4351	6257.2625	1353.0272	7.6949	16378.9258	303973538.8	8779978.231	5195.4474	17.4454	10.6414	0.2703	0.958	0.4012	0.6734	585.9425	0.4846	0.3534	742.2068	0.5447	0.4816	1.3931	0.0037	0	1.2717	1.0177	0.3493
Number_clusters	2	2	5	5	4	5	4	5	5	5	7	2	2	4	4	2	2	3	2	2	2	2	0	2	0	4
Value_Index	2.9926	9595.1413	1628.5481	20.8163	3680.2978	100723816.7	29551701.50	2616.1899	5.3261	-2.2284	0.2703	0.6324	0.6067	1.247	-299.1326	0.7507	0.4768	3570.8809	0.8027	5.7953	0.3781	0.0197	0	0.6072	0	0.3421

sc68_cisp

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.8757	2964.2762	1844.8118	7.6838	3955.1525	160103562.5	16332059.96	14136.7364	3.3738	2.2984	0.3496	0.8312	0.5089	0.4891	1519.6784	1.0437	0.5139	7068.3682	0.6878	0.6465	0.4446	0.0066	0	0.6918	2.2366	0.5609
3	42.0242	3600.6297	745.8691	18.3012	6229.1545	133161703.2	13095705.59	7818.7115	6.8437	4.1557	0.3446	0.788	0.5009	0.649	523.998	0.5401	0.496	2606.2372	0.7057	0.9119	0.7725	0.0096	1e-04	0.6385	1.6847	0.5973
4	0.0337	3432.1127	210.872	15.7386	7483.5112	136725567.7	10220477.38	5892.6918	9.4089	5.514	0.3082	0.8353	0.4772	3.2219	-769.617	-0.6883	0.4489	1473.1729	0.6653	3.2423	0.9887	0.0072	1e-04	0.721	1.4514	0.4712
5	3.0487	2863.5135	347.3762	6.4012	7881.2252	179505993.6	7666269.000	5394.029	10.4834	6.0237	0.2975	0.8939	0.4522	7.7702	-918.3527	-0.8685	0.4054	1078.8058	0.6207	1.224	1.1767	0.0061	1e-04	1.134	1.3589	0.4439
6	0.3117	2708.121	767.9391	3.7037	8542.0193	193574783.2	4776225.668	4680.8623	12.4627	6.9415	0.3147	0.8372	0.4334	3.382	-414.1393	-0.7017	0.3754	780.1437	0.5916	0.4545	1.3418	0.0045	1e-04	1.0576	1.269	0.4964
7	36.9176	3143.8248	335.7631	11.2391	9950.8451	142225141.6	4557182.639	3501.1153	17.8001	9.2805	0.3095	0.9294	0.4006	1.9377	-216.7965	-0.4821	0.3557	500.1593	0.5527	0.5324	1.642	0.0093	1e-04	0.9127	1.1265	0.4613
Number_clusters	3	3	3	3	3	4	3	3	7	4	5	3	2	3	3	2	2	3	3	1	2	3	0	3	0	5
Value_Index	42.0242	3600.6297	1098.9426	18.3012	2274.002	39216561.55	3236354.374	4392.0052	5.3373	-0.8485	0.2975	0.788	0.5089	0.649	523.998	1.0437	0.5139	4462.131	0.7057		0.4446	0.0096	0	0.6385	0	0.4439

pbmc_4000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.13	4365.9018	9118.2269	20.6108	9711.1544	8113543008.	1010123019	152446.5924	11.7905	2.145	0.3139	0.6648	0.5946	0.2532	8138.872	2.9478	0.3675	76223.2962	0.6894	0.0972	0.2975	0.3956	0	0.3579	5.0962	0.443
3	4.5007	11959.1322	3248.7998	86.1798	17302.8379	2495548394.	215100715.4	44951.5627	25.3064	7.2745	0.298	0.41	0.7202	3.7064	-1214.3113	-0.7289	0.5316	14983.8542	0.9124	1.6989	0.3433	0.0129	0	0.1645	2.9764	0.1601
4	13.063	15846.358	1099.0374	103.4502	20592.4891	1873069891.	119487528.5	24268.5477	33.9097	13.4742	0.2522	0.5696	0.6297	0.3608	1861.8417	1.7698	0.4793	6067.1369	0.7495	1.6258	0.5819	0.002	0	0.3173	2.1727	0.1231
5	0.201	15582.8336	1586.2417	98.2824	22001.167	2023075100.	76806967.77	18836.4013	37.6284	17.3599	0.2907	0.6835	0.574	1.9517	-628.0707	-0.4867	0.4332	3767.2803	0.7004	1.1978	0.6667	0.0026	0	0.4014	1.8888	0.1522
6	18.1771	17968.9585	679.3488	106.3012	24859.8626	1377033262.	40012168.51	13299.3837	57.8484	24.5875	0.2669	0.7041	0.5453	1.1702	-162.4235	-0.1452	0.3992	2216.5639	0.624	1.6397	0.7983	0.0037	0	0.4014	1.606	0.1234
7	0.1406	17753.3849	877.8809	103.5326	25986.4271	1395055271.	32808704.87	11286.4117	66.0161	28.9728	0.2442	0.7495	0.5066	1.606	-344.4894	-0.3764	0.3709	1612.3445	0.5773	0.4416	0.9093	0.004	0	0.5289	1.4995	0.1563
Number_clusters	6	6	3	6	3	3	3	3	6	6	7	3	3	3	3	3	3	3	3	1	2	2	0	3	0	4
Value_Index	18.1771	17968.9585	5869.4271	106.3012	7591.6835	4995516109.	9886129477.	86812.0147	20.2199	-2.8423	0.2442	0.41	0.7202	3.7064	-1214.3113	-0.7289	0.5316	61239.442	0.9124		0.2975	0.3956	0	0.1645	0	0.1231

pbmc_6000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	1.3368	14161.267	20543.1989	50.2729	20220.4117	5798563155.	515237896.6	181227.5221	36.0469	3.9043	0.2477	0.5674	0.7439	1.0334	-115.583	-0.0323	0.5923	90613.761	0.8955	0.2757	0.2133	0.7245	0	0.351	4.941	0.3116
3	0.4936	47174.1522	-3794.4615	183.207	31272.3765	1353731602.	69112779.46	34763.6997	71.3478	20.3535	0.2918	0.2055	0.8583	149.6827	-2932.2783	-0.9917	0.5632	11587.8999	0.9657	-2.2205	0.1802	1.2583	0	0.1183	2.1951	0.0344
4	0.9089	5704.7424	68789.9711	-3.4869	23212.9388	1255890953	712371420.1	156841.275	62.6015	4.5113	0.2713	0.7928	0.4902	0.0291	64267.6207	33.3511	0.4285	39210.3188	0.502	-0.0363	0.8944	0.0016	0	1.1712	4.0329	0.5562
5	115.4654	81845.3432	1595.9715	214.3002	39702.5784	667849459.5	25374671.63	10377.4527	132.4281	68.1827	0.2586	0.5915	0.6232	0.8946	145.3919	0.1177	0.4437	2075.4905	0.6339	1.8078	0.3821	0.0051	0	0.7959	1.287	0.1495
6	1.1073	87221.9056	1415.3692	214.2172	42707.0173	519459975.6	23269943.28	7817.2129	191.0017	90.5134	0.2374	0.5505	0.6474	0.4784	1418.3993	1.0894	0.4058	1302.8688	0.6115	2.3574	0.3918	0.005	0	0.9301	1.1149	0.1187
7	3.0586	94017.2787	896.0082	216.4618	45124.5316	430736155.7	10631165.23	6057.4558	254.5114	116.8086	0.2047	0.6827	0.543	0.7414	384.3206	0.3484	0.3761	865.3508	0.5605	3.3787	0.4302	0.005	0	0.831	0.9783	0.0974
Number_clusters	5	7	4	7	5	3	5	3	5	5	7	3	3	2	2	2	2	3	3	1	3	3	0	3	0	3
Value_Index	115.4654	94017.2787	72584.4326	216.4618	16489.6396	1565000949	686996748.4	268541.3977	69.8266	-41.3407	0.2047	0.2055	0.8583	1.0334	-115.583	-0.0323	0.5923	79025.8611	0.9657		0.1802	1.2583	0	0.1183	0	0.0344

Supplementary File 3

	Silhouette k_2	Silhouette k_3	Silhouette k_4	Silhouette k_5	Silhouette k_6	Silhouette k_7
day1_norm	0.461411259	0.531113415	0.370549786	0.394185179	0.347939816	0.412120419
day7_hypo	0.682073491	0.532468182	0.635421138	0.597010772	0.729945212	0.534758528
sc68_vehi						
sc68_cisp						
pbmc_4000	0.904972291	0.803128323	0.713574035	0.643987297	0.558917001	0.658264645
pbmc_6000	0.687110344	0.568379471	0.719959268	0.69145881	0.746599617	0.662569735

Supplementary File 4

day1_norm

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	10.0725	7548.3142	2136.0562	2.2422	5283.0158	55724103.30	22421305.98	7814.777	3.5894	3.0618	0.3477	0.5694	0.6126	0.6445	1609.2286	0.5515	0.465	3907.3885	0.7741	2.8377	0.2082	0.004	1e-04	0.9345	1.3353	0.715
3	4.2569	7042.3403	1284.6308	2.8417	8477.8763	52412504.52	3879848.920	4935.246	7.7179	4.8483	0.3308	0.9368	0.4252	0.6982	799.2335	0.432	0.4463	1645.082	0.5913	1.2139	0.6681	0.0053	1e-04	1.4521	1.072	0.8809
4	0.3698	6769.1239	756.6096	3.5596	10890.9931	48217779.03	4389081.063	3653.0534	9.5803	6.55	0.2951	0.8725	0.422	0.9208	113.6905	0.086	0.4373	913.2634	0.5261	0.9384	1.0048	0.0046	1e-04	1.5792	0.9143	0.7999
5	0.2453	6314.061	1451.9969	0.8522	12533.2544	48119001.42	2418328.415	3027.1069	11.6935	7.9044	0.279	0.9144	0.412	2.0259	-570.204	-0.5055	0.4061	605.4214	0.4875	0.3635	1.2486	0.003	1e-04	1.773	0.8219	0.5386
6	48.8357	7344.6692	595.2103	11.3009	14739.6355	37938923.58	1526756.476	2166.9597	17.4472	11.0419	0.3135	0.8045	0.4301	2.0802	-639.7475	-0.5183	0.3793	361.1599	0.4759	0.9108	1.3411	0.0066	1e-04	1.6924	0.7042	0.4189
7	0.0492	7213.9631	827.6646	10.8047	15525.6533	41666434.65	935146.0402	1863.6358	21.6827	12.8391	0.294	0.8525	0.411	1.1291	-107.1571	-0.1142	0.3516	266.2337	0.4472	0.4975	1.5324	0.0086	1e-04	1.9874	0.6472	0.3968
Number_clusters	6	2	6	6	3	6	3	3	6	6	5	2	2	2	2	2	2	3	2	3	2	7	0	2	0	7
Value_Index	48.8357	7548.3142	856.7866	11.3009	3194.8606	13907588.91	18541457.06	1597.3384	5.7536	-1.3404	0.279	0.5694	0.6126	0.6445	1609.2286	0.5515	0.465	2262.3065	0.7741	1.2139	0.2082	0.0086	0	0.9345	0	0.3968

day7_hypo

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.641	1548.6404	572.3534	-5.4792	2297.3801	21024393.62	7596352.065	4805.7792	2.2311	1.7962	0.3192	1.0895	0.411	0.9157	134.8146	0.092	0.3162	2402.8896	0.5568	0.8354	0.492	0.007	2e-04	1.4583	1.4127	1.43
3	1.1923	1287.693	984.9167	-20.9178	3131.2698	30824780.08	3742864.009	3713.1222	2.5838	2.3248	0.2911	1.1129	0.3602	1.5835	-373.2904	-0.3677	0.4227	1237.7074	0.5585	0.6004	0.8293	0.0026	2e-04	1.2772	1.2637	1.486
4	5.0522	1620.8683	424.8306	-10.9142	4734.3201	24054913.39	1335898.446	2464.4981	5.2186	3.5026	0.3079	0.9755	0.3621	0.9764	17.9995	0.0241	0.4116	616.1245	0.5362	0.3843	1.3607	0.0052	2e-04	1.3278	1.0351	1.1589
5	0.53	1586.8399	504.1772	-11.6517	5523.1095	25065583.25	760246.1858	2022.3236	6.6192	4.2685	0.3065	0.9689	0.3546	2.8679	-367.992	-0.6485	0.3859	404.4647	0.5286	0.2582	1.5798	0.0094	2e-04	1.363	0.9392	0.9271
6	4.5623	1699.0092	299.1066	-8.5553	6443.9497	22492580.44	592905.123	1605.5061	8.5599	5.3766	0.3006	0.8722	0.3931	2.168	-300.0862	-0.5365	0.3667	267.5844	0.5267	0.5669	1.7417	0.0074	3e-04	1.2667	0.8311	0.6863
7	0.4152	1683.0045	336.1904	-8.913	6944.1739	23678473.68	491206.4247	1391.1335	10.3048	6.2052	0.2923	0.8939	0.3694	1.7944	-175.752	-0.4404	0.3434	198.7334	0.5024	0.2704	2.0007	0.0083	3e-04	1.5852	0.7735	0.6469
Number_clusters	4	6	4	2	4	4	3	4	4	4	3	6	2	2	2	2	3	3	3	1	2	5	0	6	0	7
Value_Index	5.0522	1699.0092	560.086	-5.4792	1603.0503	7780536.550	3853488.055	806.4496	2.6348	-0.412	0.2911	0.8722	0.411	0.9157	134.8146	0.092	0.4227	1165.1822	0.5585		0.492	0.0094	0	1.2667	0	0.6469

sc68_vehi

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	3.7824	2737.8959	1728.6196	-7.041	4620.3362	41760651.66	7481316.608	6992.5819	2.0005	1.7018	0.2971	1.1933	0.3671	0.8474	421.7928	0.1799	0.3942	3496.291	0.493	0.4106	0.6068	0.005	1e-04	1.7035	1.2174	1.6875
3	3.2989	2839.141	1360.4645	-22.0342	6880.3838	52658666.05	5656423.485	4845.4538	2.8833	2.456	0.2884	1.0107	0.361	1.4732	-563.0736	-0.3209	0.4411	1615.1513	0.5391	0.4218	1.0116	0.0042	1e-04	1.3036	1.0251	1.6874
4	1.9059	3005.743	945.1954	-18.2439	9260.695	50872773.56	3892881.062	3592.3196	4.7851	3.3127	0.274	0.9897	0.3485	1.2982	-359.7241	-0.2295	0.4134	898.0799	0.5391	0.4197	1.3905	0.0043	2e-04	1.3929	0.8865	1.2946
5	0.3039	3036.3161	874.8381	-17.4561	10907.6641	52124746.04	1998970.054	2891.3892	6.3103	4.1158	0.2951	0.9798	0.34	1.2481	-238.1726	-0.1985	0.3866	578.2778	0.5259	0.3276	1.6881	0.0036	2e-04	1.5258	0.7935	0.9494
6	9.438	3148.3712	639.713	-15.1277	12513.2251	49745353.29	1508494.611	2361.4116	8.0189	5.0395	0.2816	0.8863	0.3609	1.4408	-309.9081	-0.3054	0.3649	393.5686	0.5153	0.424	1.928	0.005	2e-04	1.5214	0.7183	0.766
7	0.206	3160.1351	603.5983	-14.8393	13678.0189	50238562.07	997374.5453	2028.4336	9.6638	5.8667	0.2796	0.8834	0.3541	1.3252	-224.5349	-0.2448	0.3435	289.7762	0.4974	0.224	2.1689	0.0042	2e-04	1.5832	0.6619	0.6344
Number_clusters	6	7	4	2	4	4	5	3	4	6	4	7	2	2	2	2	3	3	3	1	2	2	0	3	0	7
Value_Index	9.438	3160.1351	415.2692	-7.041	2380.3112	3037864.967	1893911.007	893.9939	1.9018	-0.0965	0.274	0.8834	0.3671	0.8474	421.7928	0.1799	0.4411	1881.1397	0.5391		0.6068	0.005	0	1.3036	0	0.6344

sc68_cisp

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	9.6287	1361.8943	1071.2253	-9.0031	2435.1676	13248415.45	5715412.312	3929.051	1.8415	1.5965	0.2536	1.2942	0.3518	0.9666	53.0526	0.0346	0.3261	1964.5255	0.4365	0.353	0.6228	0.0017	2e-04	2.0179	1.1506	1.6445
3	0.6241	1535.3998	663.8067	-19.9574	3877.0719	15859644.93	2041315.215	2674.246	2.6789	2.3457	0.2575	1.0206	0.3519	1.0712	-73.0391	-0.0664	0.436	891.4153	0.4907	0.4117	1.0779	0.0031	2e-04	1.4928	0.9571	1.4537
4	0.9732	1541.9458	479.9353	-19.6034	5058.9408	16808970.97	1102090.710	2071.6327	4.0508	3.028	0.2337	1.0451	0.3306	1.0682	-50.7248	-0.0638	0.4091	517.9082	0.4906	0.3423	1.4492	0.0053	3e-04	1.5839	0.8438	1.3866
5	2.1178	1519.1025	415.6812	-20.2243	5885.1201	18295101.92	767456.7236	1711.5193	5.3207	3.6651	0.2385	1.0372	0.3252	1.3293	-224.1725	-0.2472	0.3797	342.3039	0.4862	0.3672	1.7087	0.0025	3e-04	1.6072	0.7684	1.0023
6	0.9948	1519.3173	419.8103	-20.2075	6750.761	18037271.06	630114.9985	1447.5985	6.7859	4.3333	0.2253	0.9346	0.3403	1.9079	-323.5947	-0.4748	0.3584	241.2664	0.4738	0.1598	1.9822	0.0069	4e-04	1.6404	0.6997	0.8837
7	0.1496	1568.603	616.3137	-18.7462	7454.0874	18046289.66	341205.5969	1222.419	8.2368	5.1315	0.2401	0.9548	0.3513	1.6166	-234.5744	-0.3803	0.3389	174.6313	0.4768	0.2045	2.0685	0.0076	4e-04	1.9864	0.6558	0.7052
Number_clusters	2	7	3	2	3	4	3	3	6	3	6	6	3	2	2	2	3	3	3	1	2	7	0	3	0	7
Value_Index	9.6287	1568.603	407.4186	-9.0031	1441.9043	536804.9163	3674097.097	652.1917	1.4652	-0.0668	0.2253	0.9346	0.3519	0.9666	53.0526	0.0346	0.436	1073.1102	0.4907		0.6228	0.0076	0	1.4928	0	0.7052

pbmc_4000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	2.6413	15997.6639	10664.2582	44.3209	12914.9879	117589240.7	188916235.3	16526.7781	28.4475	5.1956	0.3219	0.3257	0.7639	0.2368	9012.3466	3.2221	0.3456	8263.3891	0.9071	1.3077	0.1657	0.6655	0	0.5058	1.6811	0.1243
3	17.8706	35692.8564	1746.979	126.7779	19905.8571	42336887.57	5553982.752	4352.7997	37.8584	19.7266	0.3034	0.2979	0.765	8.1577	-1481.0789	-0.8759	0.5456	1450.9332	0.8795	4.4432	0.2247	0.0552	0	0.4432	0.9446	0.0318
4	0.6	35273.2875	309.2018	123.2335	23227.7491	31509148.66	1012469.192	2984.8777	66.6056	28.767	0.2855	0.773	0.5675	0.6111	646.0395	0.6359	0.4803	746.2194	0.6518	27.9718	0.4507	0.0066	0	1.4799	0.7868	0.1342
5	0.8769	28671.1389	1594.6362	106.6544	23991.223	40303601.91	848178.9179	2760.8766	80.1317	31.1009	0.2831	1.0828	0.4281	0.573	1011.3827	0.7446	0.43	552.1753	0.5936	1.5804	0.5626	0.0042	0	2.1763	0.7411	0.1478
6	2.9749	32847.2225	924.6192	115.2519	26065.6103	33694644.96	1094788.574	1946.2808	92.4197	44.1179	0.2608	0.9445	0.4434	0.9648	38.7005	0.0364	0.3979	324.3801	0.5225	1.5741	0.682	0.0048	0	2.3063	0.6405	0.2391
7	13.2613	34162.425	522.7044	117.1209	27754.0836	29460555.27	680736.3194	1566.1132	120.4415	54.8274	0.2451	0.9281	0.435	1.2549	-162.9294	-0.2028	0.37	223.7305	0.4795	1.8001	0.7692	0.0051	0	2.4375	0.5804	0.2859
Number_clusters	3	3	3	3	3	3	3	3	4	4	7	3	3	3	3	3	3	3	2		2	2	0	3	0	3
Value_Index	17.8706	35692.8564	8917.2792	126.7779	6990.8693	64424614.22	183362252.5	10806.0564	28.7472	-6.7064	0.2451	0.2979	0.765	8.1577	-1481.0789	-0.8759	0.5456	6812.4558	0.9071		0.1657	0.6655	0	0.4432	0	0.0318

pbmc_6000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	23.5828	13067.263	2708.1268	17.1322	8556.503	140256158.4	29067660.51	12126.6719	4.6609	3.6799	0.3253	0.4856	0.6504	0.616	2194.0216	0.6231	0.3923	6063.3559	0.8248	2.4849	0.242	0.0111	0	0.7144	1.3993	0.324
3	0.5048	11514.1071	2138.1482	18.1761	12879.7707	130076394.7	16479544.22	7796.501	12.0669	5.7237	0.2923	0.9881	0.4594	0.736	775.041	0.3584	0.3643	2598.8337	0.6425	0.9411	0.6663	0.0045	0	1.3268	1.1375	0.4334
4	2.5258	11753.0587	1415.0693	23.2919	16389.3647	112619189.2	6467405.107	5419.5265	12.7073	8.2341	0.2804	0.8133	0.491	2.255	-1251.1026	-0.5559	0.4298	1354.8816	0.5951	0.8355	0.8738	0.0016	0	1.2958	0.9656	0.541
5	0.4722	11725.3559	600.6253	24.8445	18725.7451	108998754.3	4760677.493	4200.1083	17.8271	10.6248	0.2942	0.8014	0.4902	1.2379	-256.9752	-0.192	0.3957	840.0217	0.5621	1.8307	1.0143	0.0037	0	1.489	0.8529	0.3968
6	1.5931	10654.3969	592.625	18.827	19929.5929	122632035.6	2974522.300	3739.2271	19.9718	11.9343	0.2805	0.8731	0.4594	1.4747	-344.4086	-0.3214	0.3682	623.2045	0.5237	0.8771	1.1917	0.005	0	2.0053	0.8016	0.3421
7	1.4309	10055.3621	549.6484	15.3367	21196.455	128738128.8	2247874.893	3333.7172	21.9816	13.386	0.2803	0.8799	0.4548	1.3872	-357.0031	-0.2787	0.3472	476.2453	0.5106	1.2919	1.258	0.005	0	2.109	0.7597	0.3963
Number_clusters	2	2	5	5	3	5	3	3	3	5	7	2	2	3	3	2	4	3	2	2	2	2	0	2	0	2
Value_Index	23.5828	13067.263	814.444	24.8445	4323.2677	17253716.21	12588116.29	1953.1964	7.406	-1.0811	0.2803	0.4856	0.6504	0.736	775.041	0.6231	0.4298	3464.5023	0.8248	2.4849	0.242	0.0111	0	0.7144	0	0.324

Supplementary File 5

day1_norm

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	4.6746	10649.6541	3239.4566	16.9684	6896.7023	376003775.0	141071653.3	20880.8354	5.1683	3.9089	0.3376	0.5152	0.6527	0.4328	3060.8727	1.3097	0.5899	10440.4177	0.8075	1.3568	0.3023	0.2649	0	0.5	2.1595	0.287
3	31.7074	11653.0747	1180.8289	31.8478	11140.008	265630376.4	25438762.39	11078.2146	14.7409	7.3678	0.3539	0.6985	0.5719	1.6734	-612.8853	-0.4019	0.5072	3692.7382	0.7327	1.8224	0.5378	0.0098	0	0.6726	1.6029	0.3201
4	3.5203	10665.8387	904.3448	28.9352	13178.9818	270650824.9	13131074.55	8375.8931	16.7493	9.7449	0.3377	0.8141	0.5342	0.5385	1133.6542	0.8562	0.4628	2093.9733	0.6478	1.5342	0.7552	0.0047	0	1.0591	1.3934	0.447
5	0.4332	10199.7735	708.6697	27.5317	15361.0852	233084677.6	10276403.35	6715.9933	20.7045	12.1534	0.3378	0.9928	0.4306	1.2169	-221.1821	-0.178	0.4255	1343.1987	0.5732	1.1202	1.0083	0.0054	0	1.0879	1.2385	0.3775
6	0.1391	9879.6647	455.0524	26.3855	16233.6888	264495408.4	7560057.978	5626.0504	24.5705	14.5079	0.3139	0.9689	0.3992	1.244	-199.4691	-0.1958	0.39	937.6751	0.5371	1.4772	1.1565	0.0079	0	1.1356	1.1333	0.3045
7	0.361	9330.8095	988.0232	23.3579	16932.8079	297455773.5	6036576.504	5003.4543	28.3549	16.3131	0.2997	0.9324	0.3946	0.6662	459.9931	0.5004	0.362	714.7792	0.5073	0.5753	1.3011	0.0079	0	1.2892	1.0674	0.2902
Number_clusters	3	3	3	3	3	3	3	3	3	3	7	2	2	3	3	2	2	3	2	6	2	2	0	2	0	2
Value_Index	31.7074	11653.0747	2058.6277	31.8478	4243.3057	115393847.0	115632890.9	7100.2994	9.5727	-1.0818	0.2997	0.5152	0.6527	1.6734	-612.8853	1.3097	0.5899	6747.6795	0.8075	1.4772	0.3023	0.2649	0	0.5	0	0.287

day7_hypo

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.2973	1373.5223	7.6053	-2.2094	2619.6097	270815538	7.1366019868	60380.3831	2.7356	1.7062	0.3984	0.5972	0.5171	0.4093	516.6912	1.4393	0.3793	30190.1915	0.5449	-4.7161	0.2269	0.328	0	0.8291	4.7549	0.4081
3	0.4056	692.8927	12090.3271	-40.7294	2635.6163	6043460531.	1361998382.	60145.2036	2.7538	1.7129	0.4059	0.752	0.4164	41.7607	-1011.192	-0.9709	0.3124	20048.4012	0.4876	0.3014	0.2625	9e-04	0	1.8973	4.6964	0.8845
4	20.5598	7361.1228	56.442	69.0312	10434.0659	195728631.7	40496081.94	8331.1636	31.8689	12.3656	0.2799	0.4892	0.664	9.3024	-517.6508	-0.89	0.4795	2082.7909	0.7377	-5.0242	0.5831	0.0059	0	0.4445	1.8153	0.2309
5	1.7421	5692.3958	651.5533	52.736	10683.625	269034617.6	40262928.07	8095.9841	35.243	12.7248	0.2849	0.6642	0.5906	0.9942	4.2032	0.0058	0.4294	1619.1968	0.7009	1.2747	0.6617	0.0019	0	1.2495	1.7567	0.2672
6	0.7277	6208.8965	19.3062	56.0305	11546.1467	248758566.9	20594993.61	6062.1083	42.4207	16.9941	0.2345	0.7013	0.559	0.6687	101.0888	0.4929	0.396	1010.3514	0.6248	-2.7936	0.8196	0.0019	0	1.2759	1.474	0.2424
7	3.8562	5226.0684	357.1645	46.5091	11622.1601	325623846.9	20575584.40	6002.4051	43.6986	17.1631	0.2368	0.7315	0.5494	1.0353	-22.7217	-0.034	0.3667	857.4864	0.6133	1.7346	0.8587	0.0036	0	1.6974	1.4511	0.2272
Number_clusters	4	4	3	4	4	4	4	4	4	4	6	4	4	3	3	2	4	4	4	1	2	2	0	4	0	7
Value_Index	20.5598	7361.1228	12082.7218	69.0312	7798.4496	5921037885.	1321502300.	51578.8604	29.1151	-10.2935	0.2345	0.4892	0.664	41.7607	-1011.192	1.4393	0.4795	17965.6103	0.7377		0.2269	0.328	0	0.4445	0	0.2272

sc68_vehi

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	2.1522	8813.1248	4371.0805	23.546	8488.1716	523327937.5	339064624.9	25537.8591	5.5141	3.2604	0.3223	0.6593	0.6196	1.232	-531.2731	-0.1882	0.5788	12768.9296	0.8191	0.6076	0.353	0.0097	0	0.6195	2.3112	0.3874
3	13.2608	11529.2421	1466.8438	57.4776	13594.1035	318066611.3	27075493.72	12040.0416	12.4414	6.9155	0.3431	0.548	0.6564	0.6047	1385.3343	0.6535	0.5287	4013.3472	0.8455	2.1664	0.4302	0.0142	0	0.4773	1.613	0.2029
4	1.0895	11064.6247	834.2919	54.1981	16116.1189	296225247.4	21826289.77	8748.0798	16.7147	9.5178	0.3329	0.8581	0.4954	1.2493	-308.9254	-0.1994	0.4721	2187.02	0.6883	0.9213	0.7576	0.0064	0	0.8718	1.3738	0.3179
5	1.2761	10280.9859	886.8905	48.753	17788.703	301463650.6	20148815.99	7205.488	21.9506	11.5554	0.3343	0.8788	0.485	0.8335	284.1803	0.1995	0.4264	1441.0976	0.6711	1.4535	0.8048	0.008	0	0.8802	1.2381	0.233
6	0.9703	10271.8313	543.7912	48.5692	19133.3822	307535770.6	9378006.359	5869.3757	27.2485	14.1859	0.3399	0.8368	0.4809	0.6606	555.364	0.5133	0.3919	978.2293	0.6089	1.8705	0.9961	0.0083	0	0.9138	1.1199	0.2836
7	0.4542	9843.028	340.0972	45.5542	20230.3876	315980793.6	5781731.466	5150.3252	32.6124	16.1665	0.3166	0.9862	0.3909	1.2002	-125.7806	-0.1665	0.3645	735.7607	0.5495	1.0255	1.2362	0.0048	0	1.2216	1.0477	0.3227
Number_clusters	3	3	3	3	3	3	3	3	3	3	7	3	3	2	2	2	2	3	3	1	2	3	0	3	0	3
Value_Index	13.2608	11529.2421	2904.2366	57.4776	5105.9318	183419962.2	311989131.2	10205.8558	6.9274	-1.0528	0.3166	0.548	0.6564	1.232	-531.2731	-0.1882	0.5788	8755.5824	0.8455		0.353	0.0142	0	0.4773	0	0.2029

sc68_cisp

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	0.0852	2384.0529	2343.2945	3.3726	3643.8351	123439879.8	31232943.35	12665.503	3.8564	2.0443	0.3363	0.7591	0.4985	2.7429	-951.8601	-0.6344	0.3374	6332.7515	0.6223	0.6971	0.3295	0.0132	1e-04	0.8423	2.1081	1.173
3	2.9595	3585.6106	1170.0128	20.5335	6279.4008	87642321.35	9888456.539	6250.2167	6.1754	4.1425	0.378	0.7765	0.5038	1.3876	-290.2134	-0.2788	0.4958	2083.4056	0.6665	0.6061	0.7638	0.0078	1e-04	0.7208	1.5376	0.821
4	7.0185	4004.2505	516.6236	26.271	8243.2247	65969112.08	3850199.326	4131.7907	10.2942	6.2664	0.3469	0.7897	0.483	1.4359	-311.1764	-0.303	0.4565	1032.9477	0.6425	1.1035	0.9861	0.0059	1e-04	0.8206	1.2356	0.4919
5	0.5426	3810.8643	364.333	23.202	9118.4226	70277729.65	3105569.987	3368.7929	13.1106	7.6857	0.3543	0.8941	0.4418	1.5226	-285.5531	-0.3425	0.4148	673.7586	0.5879	0.9501	1.2563	0.0079	1e-04	1.1281	1.1218	0.4772
6	0.5755	3607.1414	544.1983	20.0995	9850.0221	73473385.61	2136131.268	2904.6447	15.4303	8.9139	0.3307	0.8459	0.4307	0.726	221.4994	0.3766	0.3837	484.1075	0.5537	0.4827	1.456	0.0075	1e-04	1.185	1.0382	0.5678
7	1.4217	3812.7628	216.9446	22.8228	10785.802	66399723.44	1069696.888	2344.7468	20.2807	11.0424	0.3443	0.8403	0.4221	2.0105	-251.8023	-0.5008	0.3587	334.9638	0.5319	1.0244	1.5963	0.0078	1e-04	1.1872	0.9357	0.3597
Number_clusters	4	4	3	4	3	4	3	3	7	4	6	2	3	2	2	2	3	3	3	1	2	2	0	3	0	7
Value_Index	7.0185	4004.2505	1173.2817	26.271	2635.5657	25981826.83	21344486.81	4296.8603	4.8504	-0.7046	0.3307	0.7591	0.5038	2.7429	-951.8601	-0.6344	0.4958	4249.3459	0.6665		0.3295	0.0132	0	0.7208	0	0.3597

pbmc_4000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	4.0252	5746.8203	2291.0483	30.3345	12223.7172	1501426835.	1039939493.	79343.6251	18.8342	2.5072	0.3281	0.8082	0.6009	0.3052	5070.4072	2.2758	0.4911	39671.8125	0.7806	0.9223	0.3765	0.4618	0	0.6902	4.0378	0.4795
3	0.119	5743.9214	7876.7232	25.1885	15648.425	1376646862.	215932804.2	49563.3766	29.0172	4.0136	0.2998	0.6369	0.5865	0.0845	17154.6842	10.8231	0.4871	16521.1255	0.7633	0.1604	0.5691	0.0055	0	0.5147	3.1771	0.4165
4	5.3776	14363.5171	2471.0705	93.4995	20543.9759	678261360.6	28854543.93	16163.9204	37.6664	12.3069	0.2752	0.4847	0.6985	4.5659	-1404.9961	-0.7796	0.4771	4040.9801	0.8372	1.1605	0.5084	0.0083	0	0.3333	1.811	0.1743
5	3.5057	18370.6131	380.9699	108.8447	23227.683	524454106.4	23016088.26	9805.7959	47.1345	20.2867	0.2997	0.4941	0.6659	0.5868	705.6483	0.7035	0.435	1961.1592	0.7766	6.2185	0.5758	0.0073	0	0.446	1.4312	0.1199
6	0.2942	16237.9545	2390.8537	97.1846	24116.5865	598244068.0	15551428.62	8914.4238	54.4106	22.3152	0.2905	0.741	0.5555	0.708	641.6785	0.4121	0.3981	1485.7373	0.7082	0.913	0.7068	0.0061	0	1.1138	1.3411	0.1437
7	7.6091	22417.8233	782.3899	118.5735	27454.959	339418990.8	4550105.060	5476.7486	78.4088	36.3222	0.2943	0.7046	0.5527	0.8027	217.3302	0.2456	0.3724	782.3927	0.6467	1.7414	0.7722	0.0071	0	1.1038	1.0852	0.1427
Number_clusters	7	7	3	7	4	4	3	4	7	5	4	4	4	4	4	2	2	3	4	1	2	2	0	4	0	5
Value_Index	7.6091	22417.8233	5585.675	118.5735	4895.5509	544578247.7	824006689.1	27041.3317	23.9982	-5.9513	0.2752	0.4847	0.6985	4.5659	-1404.9961	2.2758	0.4911	23150.687	0.8372		0.3765	0.4618	0	0.3333	0	0.1199

pbmc_6000

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman	Rubin	Cindex	DB	Silhouette	Duda	Pseudot2	Beale	Ratkowsky	Ball	Ptbiserial	Frey	McClain	Dunn	Hubert	SDindex	Dindex	SDbw
2	2.4608	8159.3881	1557.7453	30.6363	13483.9572	4266045279.	8431427065.	103992.5178	9.2177	2.6734	0.281	0.773	0.6281	0.3653	5113.2109	1.7368	0.5373	51996.2589	0.812	2.1536	0.3284	0.409	0	0.5444	4.1419	0.407
3	0.6975	6160.6509	668.5144	10.3552	15507.7372	6339110159.	4690614266.	78813.7378	11.3595	3.5274	0.2331	0.7198	0.5663	10.0132	-3094.6524	-0.8997	0.4811	26271.2459	0.6931	12.6357	0.6183	0.0072	0	0.5552	3.2829	0.3018
4	0.1793	4892.1464	29074.4461	-6.2161	21656.3625	3195110768.	4391758846.	69309.2768	36.8244	4.0112	0.2302	0.6306	0.5751	0.0772	23094.5021	11.9537	0.4278	17327.3192	0.6107	-0.0213	0.8483	8e-04	0	0.6394	2.7824	0.2814
5	62.6655	32818.0066	1434.8288	147.717	32933.269	494662243.2	49980506.01	9950.777	64.5688	27.9386	0.2788	0.4429	0.7237	0.4037	2292.8896	1.4764	0.439	1990.1554	0.7469	2.4637	0.4993	0.0025	0	0.4744	1.2365	0.1131
6	1.5519	34264.8073	1115.6684	147.6309	35201.6427	447419689.3	28860045.81	7687.2943	84.177	36.165	0.2462	0.4898	0.686	0.4748	1446.8034	1.1053	0.4022	1281.2157	0.669	1.4857	0.6042	0.0053	0	0.8684	1.0683	0.1683
7	2.1699	35271.4624	836.3779	147.3635	37458.23	383443922.7	19750355.42	6254.9385	104.1293	44.4467	0.2086	0.6378	0.5933	1.3167	-337.6886	-0.2402	0.3736	893.5626	0.6093	1.654	0.6747	0.0053	0	0.958	0.9457	0.1304
Number_clusters	5	7	4	5	5	5	5	5	5	7	5	5	3	3	2	2	3	2	2	2	2	2	0	5	0	5
Value_Index	62.6655	35271.4624	28405.9318	147.717	11276.9065	2653205971.	4341778340.	57095.0172	27.7444	-15.7011	0.2086	0.4429	0.7237	10.0132	-3094.6524	1.7368	0.5373	25725.013	0.812	12.6357	0.3284	0.409	0	0.4744	0	0.1131