scTriangulate: Decision-level integration of multimodal single-cell data

Guangyuan Li[1,2,3]*, Baobao Song[4], H.L. Grimes[2,4], V. B. Surya Prasath[1,2,3,5], Nathan Salomonis[1,2,3,5]*

Affiliations
[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[2]Department of Pediatrics, University of Cincinnati School of Medicine, Cincinnati, Ohio, USA
[3]Department of Biomedical Informatics, College of Medicine, University of Cincinnati, OH, 45267 USA
[4]Division of Immunobiology, University of Cincinnati, OH 45221 USA
[5]Department of Electrical Engineering and Computer Science, University of Cincinnati, OH 45221 USA

*To whom correspondence should be addressed:
li2g2@mail.uc.edu to G. L.
nathan.salomonis@cchmc.org to N. S.

Code and Data Availability
scTriangulate application is available as a Python3 package (https://pypi.org/project/sctriangulate). The source code is available at (https://github.com/frankligy/scTriangulate). The scripts and data for reproducing the results are available at (https://github.com/frankligy/scTriangulate/tree/main/reproduce)

The CITE-Seq processed results, scripts, relevant outputs and metadata has been deposited in Synapse (https://www.synapse.org/#!Synapse:syn26320566)

ABSTRACT

Hundreds of bioinformatics approaches now exist to define cellular heterogeneity from single-cell genomics data. Reconciling conflicts between diverse methods, algorithm settings, annotations or modalities have the potential to clarify which populations are real and establish reusable reference atlases. Here, we present a customizable computational strategy called scTrianguate, which leverages cooperative game theory to intelligently mix-and-match clustering solutions from different resolutions, algorithms, reference atlases, or multi-modal measurements. This algorithm relies on a series of robust statistical metrics for cluster stability that work across molecular modalities to identify high-confidence integrated annotations. When applied to annotations from diverse competing cell atlas projects, this approach is able to resolve conflicts and determine the validity of controversial cell population predictions. Tested with scRNA-Seq, CITE-Seq (RNA + surface ADT), multiome (RNA + ATAC), and TEA-Seq (RNA + surface ADT + ATAC), this approach identifies highly stable and reproducible, known and novel cell populations, while excluding clusters defined by technical artifacts (i.e., doublets). Importantly, we find that distinct cell populations are frequently attributed with features from different modalities (RNA, ATAC, ADT) in the same assay, highlighting the importance of multimodal analysis in cluster determination. As it is flexible, this approach can be updated with new user-defined statistical metrics to alter the decision engine and customized to new measures of stability for different measures of cellular activity.

INTRODUCTION

Single-cell genomics has led to the identification of discrete and transitional cell states that have redefined our understanding of cellular hierarchies in both health and disease [1,2,3]. Such discoveries entail the clear delineation of cell populations, their alignment to known cell types, new strategies for purification, and an improved understanding of underlying gene regulatory networks. As the catalog of known cell populations reaches closer to saturation, there is a growing emphasis on the creation of comprehensive and reliable cell atlases to reproducibly document and compare cells across studies and disease states [4,5,6]. Ideally, such atlases are informed by multiple molecular modalities, simultaneously measured in the same cells or nucleus (e.g., CITE-Seq, Multiome), to further resolve and functionally describe heterogeneity [7,8,9]. There are currently >250 computational methods for defining cell clusters from single-cell genomics data [10], with reasonable arguments that no single method is appropriate for all research questions (e.g., developmental transitions, clonal variation). The majority of these methods allow for extensive parameterization, resulting in different clusters with different boundaries, for different resolutions. While reference-based methods for supervised assignment of cell labels enable more reproducible cell assignments, the differing composition of biological specimens makes it unlikely that any given single reference will be sufficiently comprehensive to capture all present cell heterogeneity [11,12]. Similarly, ensemble-based approaches [13,14] rely on the assumption that unique populations are not valid, which is frequently not true. Hence, as single-cell multimodal genomics becomes more commonplace, we require adaptive approaches to define cell populations based on a diversity of unsupervised and supervised methods that can leverage multiple molecular measurement types. A potentially promising field of mathematics to

address such decision-making challenges is game theory, which has been leveraged extensively in multiple domains, including economics, military strategy, transportation, and genetics [15,16].

Results

Statistical framework

To address the aforementioned challenges, we introduce a new strategy for the integration of single-cell populations, called scTriangulate (**Figure 1**). This approach considers diverse user-supplied cell-cluster labels from any source, including unsupervised and supervised algorithms, along with molecular readouts from one or many simultaneous modalities (e.g., CITE-Seq, Multiome). Different from other multimodal methods that integrate at the data-level, through either a low-dimensional latent space [17,18] or through geometric graph [7], scTriangulate integrates results at a decision-level to reconcile conflicting cluster label assignments (**Supplementary Figure 1**).

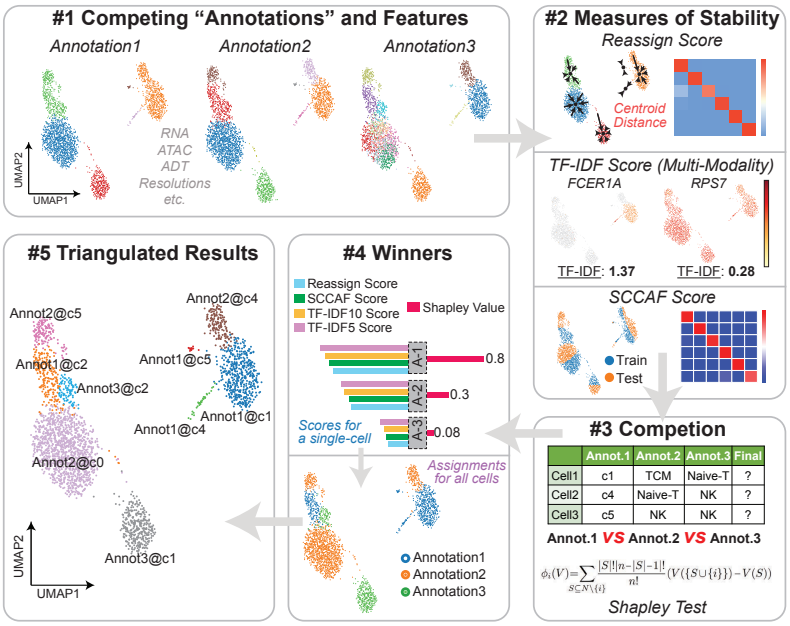

Figure 1. Overview of scTriangulate. An analysis begins with competing annotation sources within a single dataset (Step 1). Different statistical metrics stability are computed for each cluster and cell (Step 2). The reassign score considers the extent to which cells within a cluster can be reclassified to its own centroid based on nearest-neighbor. The TF-IDF score corresponds to the statistical strength of the nth ranked feature (gene, ADT, peak) in a cluster. The SCCAF score implements a prior reliability metric (Single Cell Clustering Assessment Framework) that leverages multivariate multi-class logistic regression. scTriangulate assigns every single cell to the best cluster label based on the Shapley value (Step 3). A given example of Shapley is shown for three competing labels from the different annotation sources (e.g., A-1) (Step 4). Final triangulated cluster-labels are reported after pruning the results for inconsistent assignments and a final reclassification step (Step 5). See Methods for details.

As its core decision engine, scTriangulate resolves conflicting clusters as a “game”, where each cluster label from single annotation is a player. Contrary to the intuition of letting each annotation compete with each other (competitive game theory), we design this method as a cooperative scenario in which all annotations are jointly considered to maximize overall goodness. Overall goodness is based on cluster stability measurement, with the emphasis on biological plausibility (i.e., marker genes), combined with technical artifact detection (i.e., cell doublets, cell-cycle effects) to down weight certain clusters. The Shapley Value is used to gauge the relative contribution of each annotation to guarantee fair credit allocation [19]. The relative contributions are used to assign single cells to the optimal cluster label. In particular, the univariate accuracy of cell population reclassification upon random sampling (hereafter, reassign score) [20], the exclusivity of marker gene expression via term frequency-inverse document frequency (hereafter TF-IDF score) [21], and a combined multivariate single-cell clustering assessment framework (hereafter SCCAF score) [22]. For multi-modality datasets, measurements from each modality are jointly considered (e.g., RNA, ATAC, ADT), with the contribution of each modality to each final cluster reported. Such metrics can be easily replaced or augmented with new metrics (e.g., secondary or tertiary TF-IDF scoring metrics). Following the assignment of the “best” single-cell clusters, weakly evidenced assignments can be optionally pruned and cells reclassified against the surviving clusters. For ease-of-use and optional integration with diverse analysis methods (e.g., batch-effects correction), scTriangulate is integrated with the SCANPY python toolkit, allowing for the streamlined processing and clustering of mRNA, antibody derived tag (ADT), and scATAC-Seq data along with interactive visualization of Shapley statistics, cluster assignments and assigned markers through a portable publication-ready web-browser.

scTriangulate can effectively merge clusters and annotations from diverse sources

To initially evaluate scTriangulate’s ability to integrate distinct clustering solutions and verify scTriangulate defined clusters are biologically meaningful, we leveraged a well-adopted reference cell atlas for peripheral blood mononuclear cells (PBMC) from the software Azimuth. In this silver standard reference, populations were defined from over 200 well-documented cell surface markers from CITE-Seq and matched transcriptomes in over 200,000 cells [7]. Unsupervised clustering of PBMC scRNA dataset (11,769 cells from 10X genomics) with low-resolution clustering results in extremely broad clusters that combine major immune lineages while high-resolution frequently sub-divide immunophenotypic populations (**Figure 2a,b** and **Supplementary Figure 2a-c**). From these different resolutions, scTriangulate selects a combination of clusters from each, which results in excellent agreement with the Azimuth references gauged by homogeneity (0.84), completeness (0.63), and V-measure (0.72) (**Figure 2c,d** and **Supplementary Figure 2d-f**). Importantly, scTriangulate finds stable subtypes of T-cells, typically not well-identified without cell surface ADT profiles (e.g. gamma T cells, MAIT, CD4+, and CD8+ T effector memory cells) (**Figure 2c**). scTriangulate revealed further heterogeneity in CD14 Monocytes, corresponding to well-established classical (*CLEC4D*, *CLEC5A*) and intermediate monocytes (*FCGR3A*, *CLEC10A*, *HLA-DRA*)[23], not annotated by Azimuth (**Supplementary Figure 2g**).

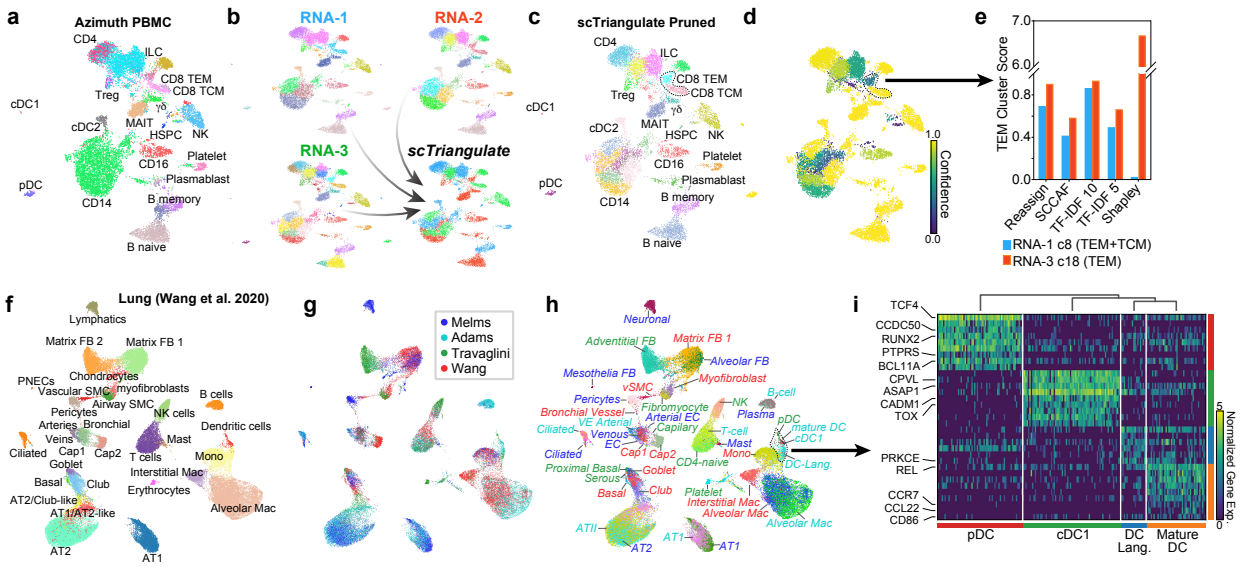

Figure 2. scTriangulate reconciles conflicts between diverse methods, algorithm settings, annotations. a-d) Scanpy Leiden clustered UMAP of an open-access PBMC scRNA-Seq dataset. a) Projected labels from a PBMC CITE-Seq reference (Azimuth, Level 2). b) Integration of leiden cluster resolutions of 1 (RNA-1),2 (RNA-2) and 3 (RNA-3) into a combined set of annotations (lower right) colored by the winning annotation source. c) scTriangulate defined clusters after pruning and reclassification. d) The confidence of each scTriangulate defined cluster (Methods). e) Comparison of each stability metric by Shapley between a broader T cell memory cluster (light blue) and a more granular subset (TEM, red). f-h) UMAP of lung snRNA-Seq from 9 donors (Wang et al. 2020). g-h) scTriangulate winning clusters after integrating annotations from three other lung atlas studies projected onto Wang (Melms, Adams, Travaglini) (see **Supplementary Figure 5**). The cell (g) or label (h) color corresponds to the source dataset annotations in panel h. i) Heatmap displaying the top discriminating Dendritic Cell markers for Wang cluster subpopulations determined from scTriangulate (literature implicated gene symbols displayed).

To confirm its ability to integrate annotations from reference atlases and software applications, we applied scTriangulate on two prior curated collections of human bone marrow and lung single-cell RNA-Seq data. Bone marrow contains dozens of cell populations intermixed with discrete developmental transitions among rare-progenitors. We previously compared a series of clustering algorithms (ICGS2, Seurat3, Monocle3) and carefully curated reference cell populations from 8 bone marrow healthy donors [20]. Triangulating these cluster predictions and annotations, we find that nearly all curated populations can be identified in the integration, in addition to more granular subtypes that correspond to well-established dendritic, monocytic and erythroid functional subsets[20] (**Supplementary Figure 3**). These include more discrete, prior established dendritic cell subsets, naive T-cells and follicular B-cell populations (**Supplementary Figure 4**). Similar to bone marrow, in the lung, diverse clustering algorithms,

single-cell technologies, cluster resolutions and extensive manual curation have been used to identify and characterize dozens of distinct cell types. Projecting 142 labels from four lung atlas-level studies onto a single dataset (**Methods**) demonstrates significant diversity and conflicts in cell-type heterogeneity, cluster boundaries and labels (**Supplementary Figure 5a**). In the scTriangulate integration, diverse reference atlas annotations were found to preferentially associate with distinct cell lineages, suggesting annotation bias from each (**Figure 2f-h**). Notably, relative to the original source dataset publication[24], scTriangulate was able to subdivide dendritic cells and alveolar macrophages into multiple populations delineated by well-defined literature markers [25,26] (**Figure 2i, Supplementary Figure 5b-f**).

Intelligent integration of clusters from multiple modalities

Multimodal single-cell data can improve the delineation of rare and intermediate cell-populations with similar transcriptional identities, but unique epigenomes or cell-surface repertoires [27,28]. As an alternative to existing single-cell integration approaches, which rely heavily on user supplied resolutions, we applied scTriangulate to human blood immune populations assays with three distinct multimodal technologies: CITE-Seq (ADT+RNA), multiome (ATAC+RNA) and TEA-Seq [29] (ADT+ATAC+RNA). For each modality, we again examined the diversity of cell population relative to Azimuth predictions as well as known and novel markers for newly detected cell-populations, the confidence of population predictions and the preferential contribution of each modality to distinct cell lineages. For each modality we again produced clusters from three independent resolutions (leiden) for scTriangulate, using a single or multiple TF-IDF scores to bias for more discrete populations (determined based on the reported software cluster confidence). To ensure CITE-Seq predictions were not biased based on the Azimuth reference, we produced our own deep CITE-Seq dataset of human blood total nuclear cells (TNC) in two healthy donors. The donors were treated with granulocyte colony-stimulating factor (G-CSF) (G-CSF) to mobilize hematopoietic stem cells and progenitors (HSCP). scTriangulate includes specialized methods for processing ADTs, in particular, a modified Centered Log Ratio (CLR) normalization approach [30,31]. As expected, scTriangulate finds nearly all Azimuth defined cell populations, including rare MAIT, double-negative T-cells, CD56+ NK cells and other diverse T-cell subsets, independently in both replicates (**Figure 3a-d** and **Supplementary Figure 6**,**9**). Similar to scTringulate of PBMC scRNA-Seq, this analysis again delineated multiple subtypes of monocytes, with common gene and ADT profiles (**Supplementary Figure 7**,**8**,**10**). We delineate unique CD56 high and low MAIT-cell populations (**Figure 3c,d** and **Supplementary Figure 9e-g**) where CD56 high MAIT cells have increased ability to innate cytokine stimuli [32]. Overall, scTriangulate had improved measures of Homogeneity, Completeness and V-Measure than of its input ADT or RNA clustering results (**Figure 3e**). While the majority of the populations were also observed using an independent multimodal integration method (WNN in Seurat 4), no single WNN solution matched the Azimuth or scTriangulate clusters, with a high degree of variability among each tested resolution (**Supplementary Figure 9a**).

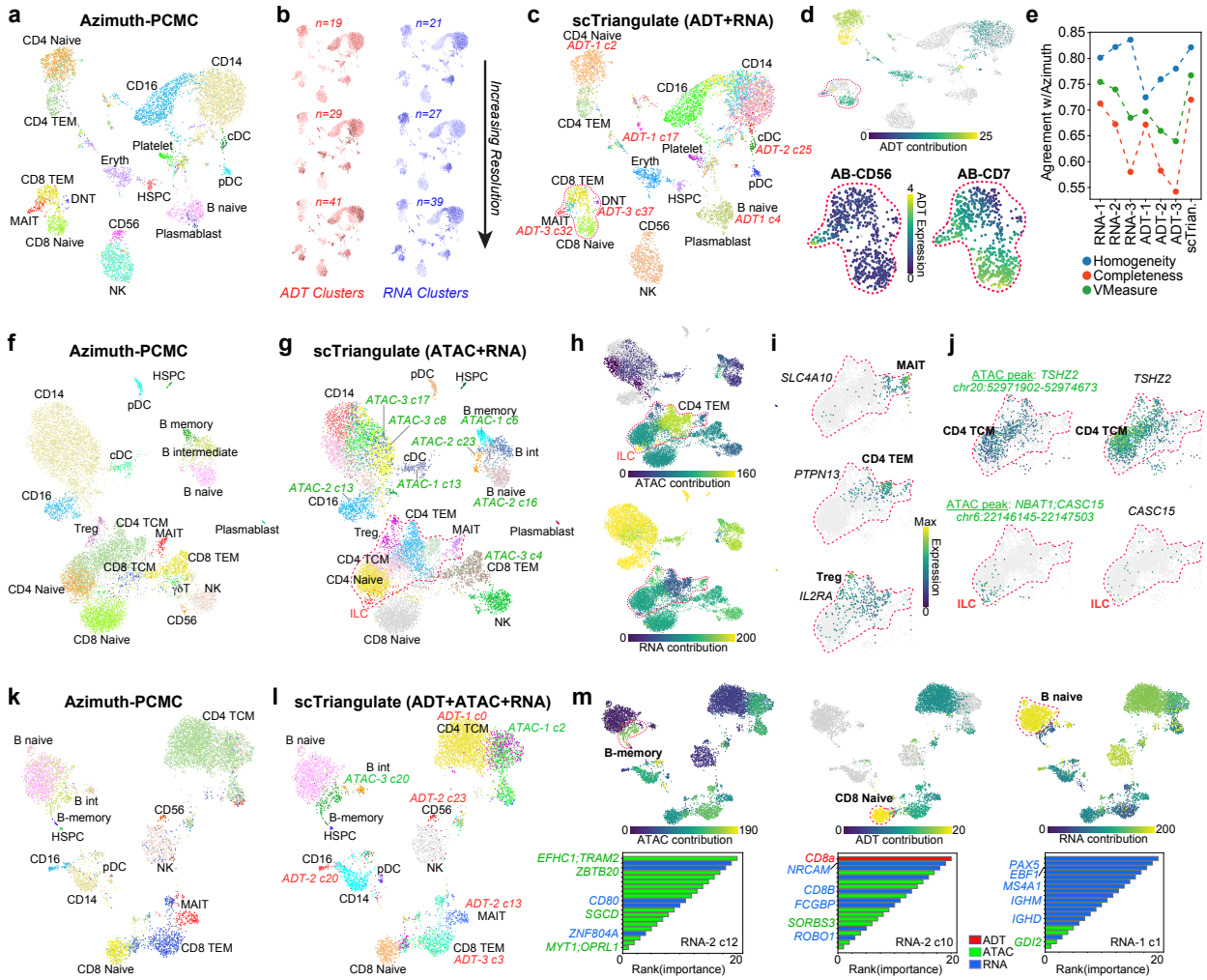

Figure 3. Multimodal single-cell integration enables new regulatory insights and novel cell populations in human blood. Scanpy Leiden UMAP for multimodal analysis of human blood by CITE-Seq (a-e), multiome GEX+ATAC (f-j), and TEA-Seq GEX+ADT+ATAC (k-m). UMAP coordinates are associated with ADT clusters (CITE-Seq, TEA-Seq) or RNA (Multiome). a,f,k) Projected labels from a PBMC CITE-Seq reference (Azimuth, Level 2). c,g,l) scTriangulate winning clusters after integration of multiple clustering resolutions (n=3) for each assayed modality relative to Azimuth annotations. Clusters specifically derived from ADT or ATAC cluster resolutions are labeled in the plot. d,h,m) The selective contribution of each indicated modality is overlaid on the dataset UMAP, based on the frequency of associated features among the top-20 markers of each final cluster (Methods). d) In the lower panel, the specificity of specific ADTs (CD56, CD7), demarcates the cluster boundaries of scTriangulate MAIT cell and CD8 T Effector Memory cell subpopulations respectively. e) Agreement with Azimuth for the assigned cell-to-cluster associations with the projected Azimuth labels, measured by Homogeneity, Completeness, and V-Measure for each individual or integrated set of clustering solutions. i-j) Visualization of the top multiome RNA (i) or ATAC (j) markers for indicated rare or novel T-cell subsets. J) Correspondence with the expression of the RNA of the same genomic loci. m) In the lower panels, the top-ranked features for example clusters are indicated by their modality (color), with example features called out.

Similar to the CITE-Seq, scTriangulate of multiome and TEA-Seq PBMCs were found to have strong agreement with Azimuth predictions (**Figure 3f-j**, **Supplementary Figures 11-14**). For the analysis of snATAC-Seq (ATAC), scTriangulate adopts a modified version of epiScanpy [33] to collect peak-level information for ATAC cell clusters. Beyond these known populations, multiome GEX+ATAC uniquely identifies four CD4 T-cell subsets, not clearly found in Azimuth predictions (**Figure 3f,g**). These include ILC and CD4 TCMs, that were attributed to ATAC-Seq population-specific peaks, including a canonical Naive/CM marker TSHZ2 [34] and others which were only reflected at the level of gene accessibility (e.g., CASC15 [35]) (**Figure 3h-j**). We further denote CD4 TEM based on the expression of *ITGB1* [36], along with a PTPN13 high CD4 population. Given the regulatory role of PTPN13 in Th1 and Th2 differentiation [37], it may suggest a unique population poised toward a specific helper cell type in response to inflammatory stimuli. Intriguingly, scTriangulate of TEA-Seq PBMCs finds that distinct modalities (gene accessibility, transcriptomic or cell-surface features) often distinguish highly related cell-populations (**Figure 3k-m**). These include B memory(ATAC) naive B-cells (RNA) and CD8 T-cell subsets (ADT versus ATAC) (**Figure 3m**). Both multiome and TEA-Seq reproducibly identify similar Monocyte subsets identified in all TNC and PBMC analyses, with similar gene, ADT or ATAC features delineated in each (**Supplementary Figures 2**, **7**, **8**, **10**, **12**, **14**). These include subdivision of MAIT and NKT cell with biologically informative cell surface markers (TCR-Va7.2 and TCR-Va24-Ja18, respectively) and subdivision of CD4+ T-cells by the unique accessibility of a critical early T cell progenitor regulatory factor (ZBTB17 [38], **Supplementary Figure 14b**).

Runtime and scalability

While demonstrated here with up-to nine clustering solutions, scTriangulate is able to scale to dozens of annotation-sets or types (**Supplementary Figure 15**). The software automatically uses multiple cores, according to the number of annotation-sets integrated. A typical scRNA-Seq dataset with four provided annotation-sets can run in ~10 minutes, while datasets with ~100k features (e.g., GEX+ATAC), analyses will typically complete in <1 hour, supporting both person and high-performance compute environments.

DISCUSSION

As the number of single-cell modalities increase, it will be necessary to not only integrate cellular heterogeneity from each molecular readout but combine such predictions with reference cell annotations from existing studies. Such inputs can include multimodal predictions from independent workflows and varying resolutions. Importantly, scTriangulate is not limited to only three modalities and can be updated to include support for additional data types as well as stability metrics. These can potentially include additional scoring metrics to optimize the selection of meta-stable intermediates. We hope this approach will enable improved insights into single-cell biology informed by diverse new and existing single-cell methods for heterogeneity detection.

ACKNOWLEDGMENTS

We would like to thank Kang Jin, Balaji Iyer, Daniel Schnell and Harinder Singh for their valuable feedback on the approach and manuscript. We also thank Kelly Rangel in the CCHMC Gene Expression Core for generating CITE-Seq libraries and Andre Olsson for his assistance. This work was partially supported by the National Heart Lung and Blood Institute (R21HL150678 to H.L.G and N.S., U24HL148865), National Institute of Diabetes and Digestive Health (RC2DK122376 to H.L.G. and N.S.), the Chan Zuckerberg Pediatric Network for the Human Cell Atlas (N.S.) and the Cincinnati Children's Pediatric Cell Atlas Center (N.S. and H.L.G).

AUTHOR CONTRIBUTIONS

G.L. implemented the method and performed evaluations. B.S. generated the CITE-Seq data. N.S., L.G., H.L.G. and V.B.S.P designed the study and wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

METHODS

scTriangulate Algorithm

scTriangulate is organized into 5 principle steps (**Figure 1a**). In the first step, counts or normalized data matrices are supplied to the software for quality control, filtering and scaling (optional). This step may also include running Leiden clustering sequentially through scTriangulate when no prior annotations exist. In the second step, cluster stability is computed for all input annotation-sets. In the third step, a Shapley Value is computed to assign each cell to a supplied or scTrangulate computed set of labels. In the fourth step, the final annotation assignments are pruned according to the fraction of final versus original cell-cluster assignment (user defined) and re-classified using a nearest-neighbor strategy. Finally, these results are reported to the end-user using a series of interactive or static (pdf) visualizations to explore locally or over the web. Additional details and tutorials for scTriangulate are provided online with the software.

Step 1: Annotation assignment and quality control

Input data matrices and options are supplied to the software to enable automated processing of a dataset without annotations (i.e., clusters) or with externally produced annotations. This is performed using the preprocessing module wrapper function scanpy recipe. scTriangulte begins with two data matrices $\mathbf{X}$ and $\mathbf{A}$, where $\mathbf{X} \in \mathbf{R}^{\mathbf{I}\times\mathbf{F}}$ denotes the expression matrix and $\mathbf{X}_{\mathbf{if}}$ is the scaled expression value of feature **f** in cell **i,** with the total number of features as **F** and total number of cells as **I**. The features denote a gene, ADT, ATAC-Seq peak or custom modality. By default, scTriangulate assumes that the input expression data is log scaled values (counts per 10,000 reads per cells + 1). Let $\mathbf{A} \in \mathbf{R}^{\mathbf{I}\times\mathbf{A}}$, where $\mathbf{A}_{\mathbf{ia}}$ denotes the cluster label of cell **i** in annotation **a**.

Step2: Inference of cluster stability

To decide which cluster assignments are optimal, the software computes a series of metrics to gauge the reliability of each cluster for each annotation. scTriangulate uses four embedded scores: the **reassign** score, the **TFIDF10** score, the **TFIDF5** score and the **SCCAF** score. These default scores can be augmented by users with custom metrics to suit to either promote or disqualify clusters with certain properties (doublet, cell cycle, metabolic pathway, etc). Here we describe the four default scores used in the program:

Reassign Score: The reassign score quantifies the reliability of a cell-to-cluster assignment when compared to its own centroid (nearest neighbor distance). This function defines the centroid of each cluster based on the top ranked 30 markers, where each marker is defined using an integrative strategy considering both log-fold and two-sided t-test p-value in that population versus all other cells. Specifically, let $\mathbf{r^f_{pval}(c)}$ be the rank of feature **f** in cluster **c** based on the t-test p-value. The most significant feature will receive a rank=0. Then let $\mathbf{r^f_{lfc}(c)}$ be the rank of feature **f** in cluster **c** based on the log-fold change. Likewise, the features with largest positive log fold change will receive rank=0. The combined rank will then be:

$$r_{combined} = (r_{pval} + r_{lfc})/2 \qquad (1)$$

We define the function **NC(·)** to find the nearest centroid of each cell **i**, and the Reassign score for each cluster **c** as:

$$S^a{}_{reassign}(c) = \frac{\Sigma_{i \in c}(NC(i) = c)}{|c|} \qquad (2)$$

Where **|c|** denotes the size of cluster **c**, the reassign score effectively measures the fraction of cells in each cluster that can be reclassified to its own centroid. A higher value indicates greater stability.

TFIDF(n) Score: The TFIDF score measures the number of exclusively expressed features in a cluster. This score is based on the observation that genes or other features that are uniquely expressed in one cluster typically demarcate the boundary of a potentially novel cell population. Unlike the reassign score which utilizes the expression values, the TFIDF score focuses on expression frequency (zero and nonzero) as a binary value. To calculate the TFIDF score for a cluster, we first define the Term Frequency (**TF**) and Inverse Document Frequency (**IDF**) of each feature **f** in each cluster **c**. A pseudo value ϵ = 1e−5 is added to avoid numerical issue:

$$TF_f{}^a(c) = \frac{\Sigma_{i \in c} Indicator(X_{if} > 0)}{|c|} + \epsilon \qquad (3)$$

$$IDF_f{}^a(c) = -log(\frac{\Sigma_{i \in I} Indicator(X_{if} > 0)}{|I|} + \epsilon) \qquad (4)$$

Where **I** denote cells in the whole dataset but **c** only contains cells in the selected cluster. Then we define the exclusivity, which is equal to the TFIDF score of a feature in the cluster **c** as:

$$Exclusivity_f{}^a(c) = TFIDF_f{}^a(c) = TF_f{}^a(c) \times IDF_f{}^a(c) \qquad (5)$$

Next this function ranks the features in each cluster based on their exclusivity in descending order. We can define the **TFIDF(n)** score of a cluster as the exclusivity value of the nth feature:

$$S^a{}_{TFIDF(n)}(c) = Exclusivity_f{}^a(c)^n \qquad (6)$$

We consider **TFIDF5** and **TFIDF10** scores by default. As an example, $S_{TFIDF(5)}(c1)$means the TFIDF score of the top 5th exclusively expressed feature in the cluster c1. If a cluster expresses few unique marker genes (e.g., n=3) but is considered biologically valid, the user can add or replace the existing TFIDF score with alternative rank, such as TFIDF1 to prioritize such features in the downstream Shapley analysis.

<u>*SCCAF Score*</u>: We use the previously described Single Cell Clustering Assessment Framework (SCCAF)[22] to serve as an added metric to gauge the biological reliability of the cluster. Different from the reassign score where we leverage the univariate markers from a two-sided t-test, SCCAF considers the marker definition challenge as a multivariate multi-class logistic regression problem, which inherently takes into account the interactions and dependencies between different features. In the originally proposed model, The expression matrix and the Annotation will serve as the data **X** and the label **Y** respectively. After logistic regression model inference, a confusion matrix $\Psi \in R^{C \times C}$ will be produced. This matrix stores the number of times the model correctly and incorrectly predicts each instance's cluster label. Using the confusion matrix, we define the SCCAF score of a cluster **c** as:

$$S^a{}_{SCCAF}(c) = \frac{\Psi_{cc}}{\Sigma_{c \in C} \Psi_{cc}} \qquad (7)$$

Step3: Shapley Value determination and cell-cluster assignment

Following determination of stability metrics in **Step 2** for each single cell, a resulting matrix $D \in R^{A \times M}$ will be produced, where D_{am} denotes the calculated metrics of annotation **a** under metrics **m**. For instance, if one selected cell has three conflicting annotations, namely label-1, label-2 and label-3. By default, matrix **D** will be of the shape 3 x 4 as scTriangulate compute 4 scores.

We consider each annotation here as one "player" in a cooperative game. In order to assess the relative importance of a player, the Shapley Value **Φ** has been proposed to guarantee fair credit allocation, which considers the amount of added contribution from each player (one label) to the collective performance of a team/coalition (multiple labels). With a proper Value function **V(T)** to represent the team's performance, the Shapley value of a label a can be computed as:

$$\Phi_a = \sum_{T \subseteq A \setminus \{a\}} \frac{|T|! \times |A - |T| - 1|!}{|A|!} \left(V(T \cup \{i\}) - V(T)\right) \qquad (8)$$

In brief, shapley value iterates every possible formation of the teams and measures the marginal contribution of each player (label) to the team, normalized by the size of team. scTriangulate defines the added performance (surplus) as below:

Algorithm 1 scTriangulate Computing ($V(T \cup \{i\}) - V(T)$)

1: Input: annotation a and metric matrix D
2: V=0
3: **for** $m \leftarrow \{\text{Score1}, \text{Score2}, \text{Score3}, ...\}$ **do**
4: **if** $\text{Rank}(D_{am} + 0.01) = |T \cup \{i\}|$ **then**
5: $V \leftarrow V + |T \cup \{i\}|$

scTriangulate uses the $\mathbf{\Phi_{a(i)}}$ to denote the relative importance of each label for each single cell **i**, and assigns this single cell **i** to the label with highest **Φ**:

$$arg\ max_{a \in A}\ \Phi_a(i) \qquad (9)$$

Step 4: Results pruning and cell reclassification
We denote the new cell label based on the highest Shapley value as $\mathbf{A^{raw}}$. We further filtered out unstable clusters based on two criteria:

(1) winning fraction: $wf_c = \frac{|A_c{}^{raw}|}{|A_c{}^{ori}|}$.Stable clusters have a higher winning fraction than unstable clusters, which may win out purely by chance. By default $\mathbf{wf_c < 0.25}$ will be labelled as unstable clusters. The winning fraction of raw clusters indicates the confidence that the program possesses for this prediction.

(2) By default, $|A_c{}^{raw}|$<10 will be labelled as unstable clusters.

All cells within the unstable clusters will be reassigned to its nearest neighbor cluster label.

<u>*Contributions of modality*</u>: The contribution of each modality {RNA, ADT, ATAC} in each cluster **c** is defined as a weighted sum of its markers (gene for RNA, antibody for ADT, peak for ATAC) in the top-20 marker list for the cluster **c**, the weight is the rank of the marker in the list such that the top1 marker will receive importance as 20. Specifically,

$$Contributionn_{RNA}(c) = \sum_{i \in top20} Indicator(i = gene) \times Rank \qquad (10)$$

$$Contributionn_{ADT}(c) = \sum_{i \in top20} Indicator(i = antibody) \times Rank \qquad (11)$$

$$Contributionn_{ATAC}(c) = \sum_{i \in top20} Indicator(i = peak) \times Rank \quad (12)$$

where **i** refers to each marker in the top-20 list, which can be either a gene, antibody or ATAC peak. Rank is the rank importance defined above.

Step 5: Results visualization

scTriangulate results can be exported en masse to an html archive to interactively explore marker expression, cell population boundaries, doublet cell prediction, quality control metrics, stability metrics, Shapley values, modality contributions and cell-population prediction confidence. These plots can be viewed as scatter plots (e.g., UMAP), heatmap or bar charts. In addition, these plots can be individually exported from the software. Such visualization is an essential step to guide optimization of scTriangulate results to use alternative stability strategies (e.g., TFIDF5) or alter the cutoffs for pruning and reclassification in the stored scTriangulate object.

Preprocessing and Quality Control

scRNA-Seq PBMC

Pre-processed sparse-matrix counts (h5) were obtained from a human 10x Chromium (3' version 3 chemistry) PBMC sample, available from the 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3). Cell barcodes were filtered based on a min_gene > 300, min_count > 500, pct_counts_mt < 20%, to yield 11,022 single cells. The raw UMI count matrix was adjusted to logarithmic Counts Per Ten Thousand (CPTT), we then performed hypervariable genes selection (top 3000), neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=50), and Leiden clustering sequentially through the scTriangulate preprocessing module wrapper function scanpy_recipe. Separate leiden clusters (annotations) were obtained using a resolution of 1,2 and 3. All the preprocessing and Quality Control were run in scTriangulate version 0.9.0 and scanpy version 1.7.2. The AnnData after Leiden clustering was subjected to the scTriangulate program (win_fraction_cutoff=0.25, add_metrics='tfidf5') as input. Azimuth L2 annotations were derived from the filtered raw count data through the Azimuth PBMC online R-Shiny app (https://app.azimuth.hubmapconsortium.org/app/human-pbmc).

CITE-Seq Total Nucleated Cell

De-identified human granulocyte colony-stimulating factor (G-CSF) mobilized total nucleated cells (TNCs) were obtained through apheresis from Cell Processing Core (CPC) at Cincinnati Children's Hospital Medical Center (CCHMC). Two young adult donors (~30 years of age, male and female) were separately profiled. For each donor, 100,000 cells were stained with 31 human TotalSeq-A antibodies (BioLegend®, San Diego, CA) cocktail for 30 minutes on ice, and subsequently washed with flow buffer (1% FBS in PBS) and counted. ~15,000 cells were loaded on a 10x Chromium controller with 3' GEX version 3 chemistry. The GEX libraries were sequenced with a target depth of >50,000 reads/cell with PE150 on Illumina S4 flow cell. ADT

library was sequenced with a target depth of >10,000 reads/cell with PE50 on Illumina SP flow cell.

#	Antibody	BL Totalseq Cat#	Clone
1	CD2	309229	TS1/8
2	CD3	300475	UCHT1
3	CD4	300563	RPA-T4
4	CD7	343123	CD7-6B7
5	CD8	301067	RPA-T8
6	CD10	312231	HI10a
7	CD11b	301353	ICRF44
8	CD13	301729	WM15
9	CD14	301855	M5E2
10	CD15	125615	MC-480
11	CD16	302061	3G8
12	CD19	302259	HIB19
13	CD20	302359	2H7
14	CD24	311137	ML5
15	CD33	366629	P67.6
16	CD34	343537	581
17	CD38	303541	HIT2
18	CD45	304064	HI30
19	CD45RA	304157	HI100
20	CD49f	313633	GoH3
21	CD56	392421	QA17A16
22	CD64	305037	10.1
23	CD66b	392905	6/40c
24	CD71	334123	CY1G4
25	CD90	328135	5E10
26	CD117	313241	104D2
27	CD123	306037	6H6
28	CD135(FLT3)	313317	BV10A4H2
29	CD235a	349117	HI264
30	HLADR	307659	L243
31	CD110 MPL	XXXXXX	S16017A

Raw RNA-Seq FASTQ files were aligned with Cell Ranger 3.1.0 to the human reference genome (hg19) to produce a filtered counts matrix. Cell barcode filtering, normalization and Azimuth reference classification were performed as with the scRNA-Seq, to yield 6,406

(replicate 1) and 4,333 (validation) cells. For the ADT count, we conducted Center log-ratio (CLR) normalization[31] using scTriangulate.preprocessing.Normalization.CLR_normalization function. We then conducted hypervariable gene selection (top=3000), neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=50) and Leiden clustering (r=1,2,3) for RNA. In parallel, we performed neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=15) and Leiden clustering (r=1,2,3) for the ADT data. Finally, we concatenate both properly normalized RNA and ADT values to a combined AnnData through scTriangulate.preprocessing.concat_rna_and_other function. The combined AnnData serves as the input for the downstream scTriangulate analysis (win_fraction_cutoff=0.25, add_metrics=None). All the preprocessing and QC steps were conducted using scTriangulate version 0.9.0 and scanpy version 1.7.2. The Seurat WNN analysis was performed following the official vignette (https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis.html). Specifically, we chose 30 Principle Components (PC)s for RNA and 18 PCs for ADT, respectively. We adjusted the resolution arguments in FindClusters function in Seurat Version 4.0 to achieve different clustering results.

Multiome PBMC

Multiome PBMC sparse matrix count h5 data was downloaded from 10x Genomics website (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). We conducted quality control based on both RNA and ATAC peaks. We filtered out nuclei with min_genes < 300, min_counts < 500, pct_counts_mt > 20% for RNA data, together with an additional criteria for detecting >1000 peaks in ATAC data based on episcanpy[33] tutorial. Taken together, 10,991 nuclei were kept for further analysis. We normalized both RNA and ATAC count data to logarithmic CPTT, then performed hypervariable gene selection (top=3000), neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=15), and Leiden clustering (r=1,2,3) for the RNA expression. For the ATAC peaks, we performed hypervariable peak selection (top=100,000), neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=100) and Leiden clustering (r=1,2,3). The arguments for n_top_features and n_pc were performed as indicated in the episcanpy[33] tutorial. Finally, we concatenate both properly processed RNA and ATAC values to a combined AnnData through scTriangulate.preprocessing.concat_rna_and_other function. The combined AnnData serves as the input for the downstream scTriangulate analysis (win_fraction_cutoff=0.35, add_metrics='tfidf5'). The inclusion of the secondary TF-IDF score was based on a preliminary evaluation with both the default and additional tfidf5 parameters, which better reflected known cell population marker diversity. All the preprocessing and QC steps were conducted using scTriangulate version 0.9.0 and scanpy version 1.7.2.

TEA-Seq PBMC

TEA-Seq provides simultaneous measurement of RNA, ADT, and ATAC in the same single nuclei. The RNA- ATAC combined h5 files and ADT count CSV files were downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4949911). Customized scripts (available on Github) were used to find the common nuclei in these two files and Three different AnnData corresponding to three modalities were generated. We conducted the quality control for both RNA and ATAC, based on the same cutoffs mentioned above, nuclei with

min_gene < 300, min_counts < 500, pct_counts_mt > 20%, min_atac_peaks < 1000 will be removed. This step had 8,213 nuclei retained for further analysis. For the RNA, we normalized the RNA UMI count to logarithmic CPTT, we then performed hypervariable gene selection (top=3000), neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=15), and Leiden clustering (r=1,2,3). For the ATAC, we firstly normalized the ATAC peak count to logarithmic CPTT, we then performed hypervariable peak selection (top=60000), neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=100), and Leiden clustering (r=1,2,3). The arguments for n_top_features and n_pc base on both the episcanpy[33] tutorial and dedicated experiments. For the ADT, we conducted CLR normalization[31] using scTriangulate.preprocessing.Normalization.CLR_normalization function, followed by neighbor graph construction (n_neighbor=15), dimension reduction (n_pc=15) and Leiden clustering (r=1,2,3). Finally, we combine three processed AnnData objects using scTriangulate.preprocessing.concat_rna_and_other function. The combined AnnData serves as the input for the downstream scTriangulate analysis (win_fraction_cutoff=0.25, add_metrics=None). The Azimuth L2 annotation was obtained using post-QC RNA raw UMI count data through the web app (https://app.azimuth.hubmapconsortium.org/app/human-pbmc). All the preprocessing and QC steps were conducted using scTriangulate version 0.9.0 and scanpy version 1.7.2.

Lung Cell Atlas

scRNA-Seq and snRNA-Seq cell-population assignment and counts data were obtained from four independent studies [24,39,40,41]. Among these, a lung snRNA-Seq from 9 donors of varying age (newborn, ~3 years old and ~30 years old)[24] was selected as the base dataset, whereas the remaining were classified on to this query dataset using the software cellHarmony[11]. In brief, counts for each dataset were scaled to CPTT and log2 adjusted. For each dataset, reference centroids were initially computed from the author provided cell-type labels by identifying the top-60 marker genes (MarkerFinder algorithm) for non-diseased control samples (excluding cell-cycle gene enriched clusters). cellHarmony was run in two-steps: 1) centroid-based classification of each reference to Wang et al. 2020 with a Pearson correlation rho>0.2 and 2) self-reclassification of Wang et al. 2020 re-derived markers and centroids from the first classification (MarkerFinder and cellHarmony), to improve the accuracy of the classifications between scRNA-Seq and snRNA-Seq. The obtained annotations and UMAP coordinates[24] were added to the existing AnnData by scTriangulate.preprocessing.add_annotation and scTriangulate.preprocessing.add_umap functions respectively. The RNA raw UMI count was normalized to logarithmic CPTT. The processed AnnData was used as the input for running scTriangulate (win_fraction_cutoff=0.4, add_metrics='tfidf5'). The inclusion of the secondary TF-IDF score was based on a preliminary evaluation with both the default and additional tfidf5 parameters, which better reflected known cell population marker diversity.

Bone Marrow Cell Atlas

Alternative scRNA-Seq clustering algorithm results and curated annotations were obtained from our prior benchmarking study[20], using results deposited in the Synapse database (syn26320732). The obtained annotations and umap coordinates were added to the existing AnnData by scTriangulate.preprocessing.add_annotation and

scTriangulate.preprocessing.add_umap functions respectively. The RNA raw UMI count was normalized to logarithmic CPTT. The processed AnnData was used as the input for running scTriangulate (win_fraction_cutoff=0.3, add_metrics='tfidf5'). The inclusion of the secondary TF-IDF score was based on a preliminary evaluation with both the default and additional tfidf5 parameters, which better reflected known cell population marker diversity (see one versus two TF-IDFs in **Supplementary Figure 4d**,**e**).

Benchmark of Running time

We conducted three benchmarks to assess the contribution of increasing number of cells, features and annotation-sets (e.g., clustering solutions) on overall run time. The impact of varying number of cells was tested using Wang et al. 2020, which has 46,500 cells and 4 different annotations on randomly sampled 10,000, 20,000, 30,000 and 40,000 cells using the scTriangulate lazy_run wrapper function (assess_pruned=False, viewer_cluser=False, viewer_heterogeneity=False). The impact of varying number of features was tested in the multiome PBMC (GEX+ATAC) using 3 annotations on the 10,991 cells, downsampled to 40,000, 60,000, 80,000 and 100,000 features and the above test function. The impact of varying number of annotation-sets was tested using the TNC CITE-Seq (sample 1), using 3,6,9 and 12 annotation-sets of varying leiden resolutions. Runtime was evaluated using the default multi-core setting and compared with the actual cpu time if no parallelization was applied.

SUPPLEMENTARY FIGURES

Supplementary Figure 1. Defining stable versus unstable subclusters with scTriangulate. A theoretical example is outlined for the determination of stable subclusters via comparison of the Shapley Value statistic. a) One cluster (c5) in annotation-set A1 (left) is divided by an imaginary boundary from an overlapping annotation-set (A2) (right). b) Under the condition where either one of the subclusters (A2@c3, or A2@c6) is more stable than the broader cluster (A1@c5), based on the Shapley value derived from the reassign, TF-IDF and SCCAF scores, this boundary will be kept and serve as a valid division that demarcates two stable subpopulations. c) Alternatively, under the condition where none of the subclusters is more stable than the broader cluster, this boundary will be void and the program will reject this boundary.

Supplementary Figure 2. Integration of PBMC scRNA-Seq clusters from different software resolutions. Detailed clustering results and scTriangulate outputs from Figure 2a-e. a) Projected labels from a PBMC CITE-Seq reference (Azimuth, Level 2). B) Scanpy Leiden clustering results for three resolutions (RNA-1, RNA-2 and RNA-3). c) scTriangulate wrapper function (calling scrublet) to predict multiplet cells. d) scTriangulate final cluster source from panel b. e) scTriangulate final cluster labels. The prefix of each label indicates the annotation source and the suffix the original cluster label. f) Agreement with Azimuth for the assigned cell-to-cluster associations with the projected Azimuth labels, measured by Homogeneity, Completeness, and V-Measure for each individual or integrated set of clustering solutions. Azimuth-defined CD14 Monocytes are outlined highlighted in red. g) Canonical markers of

monocyte heterogeneity with Azimuth defined CD14 Monocytes; classical CD14 Monocyte = *S100A9, CLEC4D, CLEC5A*; intermediate CD14 Monocyte = *CLEC10A, FCGR3A, HLA-DRA*; and inflammatory monocyte = *MX1, MX2, IFI44*.

Supplementary Figure 3. Cooperative integration of diverse single-cell algorithms and curated annotations in Bone Marrow. a) UMAP of previously produced scRNA-Seq clustering results from three algorithms (ICGS2, Seurat3 and Monocle3) with the author source annotations (Hay et al. 2018) for an adult bone marrow atlas spanning ~100,000 cells, 8 donors and 64 captures (10x Chromium , 3' version 2). b) scTriangulate final cluster source from panel a. c) Cluster confidence for each final cluster (winning fraction, see Methods). d) scTriangualte final results when only one TF-IDF score (TFIDF10) was applied (conservative). e) scTriangualte final results when two TF-IDF scores (TFIDF5, TFIDF10, default setting) were applied, to obtain more granular clusters (marker gene-driven Shapley Value).

Supplementary Figure 4. scTriangulate improves the delineation of diverse hematopoietic lineages. Evidence for the delineation of new subclusters relative to the original Hay et al. bone marrow annotations. a) The original author-identified Dendritic Cell (DC) cluster (left) split into four subpopulations (middle) with support by canonical marker-gene expression (middle). Heatmap of the top marker genes for all scTriangulate DC subclusters (right). b) Same as panel a, for the original author-identified Follicular B cells. c) Same as panel a for the original author-identified Naive T cells.

Supplementary Figure 5. Leveraging Multiple Lung Cell Atlases to Create a Unified Collection of Stable Annotations. a) cellHarmony supervised mapping of scRNA-Seq reference labels (Adams et al. 2020, Melms et al. 2021, Travaglini et al. 2020) on to Wang et al. 2020 lung snRNA-Seq. c) scTriangulate integration finds winning cell populations of each source atlas. c) scTriangulate final clusters with labels colored according to the associated source study (panel b). d) Confidence for each final cluster. e-f) Heatmaps displaying the top discriminating Dendritic Cell (e) or Alveolar macrophage (f) markers for Wang cluster sub-populations determined from scTriangulate.

Supplementary Figure 6. Multimodal Integration of Antibody- and RNA-Defined Clusters. Detailed clustering results and scTriangulate CITE-Seq outputs from Figure 3a-e. a) Leiden clustering results from three RNA and three ADT resolutions (r=1,2,3). b) Doublet-score predictions. c) scTriangulate final clusters with labels colored according to the source Leiden cluster and modality. d) Final pruned scTriangulate clusters with source labels.

Supplementary Figure 7. Consistent Transcriptomic Monocyte Heterogeneity in CITE-Seq Clusters. UMAP visualization of the same monocyte RNA markers from Supplementary Figure 2g in CITE-Seq clusters.

Supplementary Figure 8. Cell-Surface Antigen-Associated Monocyte Heterogeneity in CITE-Seq Clusters. UMAP visualization of ADTs underlying the observed scTriangulate monocyte heterogeneity.

Supplementary Figure 9. Computational and Experimental Validation of scTriangulate Cell Populations. a) UMAP visualization of clustering results from an independent integration analysis of the CITE-Seq dataset (Seurat-WNN) using a range of software clustering resolutions (r=0.25, 2, 3.75). b-g) UMAP projection of an independent replicate TNC CITE-Seq dataset (validation dataset). b) Projected labels from the PBMC CITE-Seq reference (Azimuth Level 2). c) Predicted doublet scores. d) scTriangulate integration results from Leiden clustering results from three RNA and three ADT resolutions (r=1,2,3). e) Final pruned scTriangulate clusters with source labels. f) Relative contribution of ADT features to scTriangulate final clusters. g) Independent evaluation of MAIT cell heterogeneity based on CD56 ADT expression in validation TNC CITE-Seq dataset.

Supplementary Figure 10. Cell-Surface Antigen-Associated Monocyte Heterogeneity in Validation CITE-Seq Clusters. UMAP visualization of ADTs underlying the observed scTriangulate monocyte heterogeneity in the CITE-Seq validation dataset.

Supplementary Figure 11. Multimodal Integration of ATAC- and RNA-Defined PBMC Clusters. Detailed clustering results and scTriangulate Multiome outputs from Figure 3f-j. a) Leiden clustering results from three RNA and three ATAC resolutions (r=1,2,3). b) Doublet-score predictions. c) scTriangulate final clusters with labels colored according to the source Leiden cluster and modality. d) Confidence for each final cluster.

Supplementary Figure 12. Consistent Transcriptomic Monocyte Heterogeneity in Multiome Clusters. UMAP visualization of canonical markers of CD14 Monocyte heterogeneity to that of CITE-Seq and scRNA-Seq by scTriangulate. Canonical markers of monocyte heterogeneity with Azimuth defined CD14 Monocytes; classical CD14 Monocyte = *S100A9, CLEC4D, CLEC5A*; intermediate CD14 Monocyte = *CLEC10A, FCGR3A, HLA-DRA*; and additional scTriangulate predicted top Monocyte population markers = *DYSF, JAK2, VCAN*.

Supplementary Figure 13. Multimodal Integration of ADT-, ATAC- and RNA-Defined PBMC Clusters. Detailed clustering results and scTriangulate TEA-Seq outputs from Figure 3k-m. a) Leiden clustering results from three RNA, three ADT and three ATAC resolutions (r=1,2,3). b) Doublet-score predictions. c) scTriangulate final clusters with labels colored according to the associated source Leiden cluster and modality. d) Confidence for each final cluster.

Supplementary Figure 14. Delineation of T-cell scTriangulate Subsets by Joint Cell Surface and ATAC-Seq Analysis. a) Dynamic expression profiles for 11 ADTs with significant heterogeneity of the Azimuth-defined CD4 T Memory cell compartment. b) Visualization of the top multiome ATAC marker for the CD95+ subset (left) and correspondence with expression of the RNA of the same genomic loci.

Supplementary Figure 15. Run Time and Scalability of scTriangulate. a) The impact of increasing numbers of features on runtime is shown for the multiome GEX+ATAC dataset (>100,000 features) in this study with randomly downsampling. b) The impact of increasing

numbers of cells on runtime is shown for the Wang et al. Lung dataset (~46,000 single-cells) with all four annotation sources, with random downsampling of cell barcodes. c) The impact of increasing numbers of annotation-sets on runtime is shown for the CITE-Seq (sample 1) dataset (~6,000 single-cells), for 3, 6, 9, 12 query annotation-sets (in parallel, by default, blue line), compared with CPU time (sequentially, orange line).

REFERENCES CITED

1. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
2. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).
3. Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702 (2016).
4. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
5. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
6. Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
7. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
8. Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00927-2.
9. Zeng, P., Wangwu, J. & Lin, Z. Coupled co-clustering-based unsupervised transfer learning for the integrative analysis of single-cell genomic data. *Brief. Bioinform.* **22**, (2021).
10. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, e1006245 (2018).
11. DePasquale, E. A. K. *et al.* cellHarmony: cell-level matching and holistic comparison of single-cell transcriptomes. *Nucleic Acids Res.* **47**, e138 (2019).
12. Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
13. Yang, Y. *et al.* SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics* **35**, (2019).
14. Huh, R., Yang, Y., Jiang, Y., Shen, Y. & Li, Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. *Nucleic Acids Res.* **48**, 86–95 (2019).
15. A review of transport market modeling using game-theoretic principles. *Eur. J. Oper. Res.* **291**, 808–829 (2021).
16. Zhu, X. *et al.* Integrating Evolutionary Game Theory into Mechanistic Genotype-Phenotype Mapping. *Trends Genet.* **32**, 256–268 (2016).
17. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
18. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 1–17 (2020).
19. Shapley, L. S. 17. A Value for n-Person Games. *Contributions to the Theory of Games (AM-28), Volume II* 307–318 (1953) doi:10.1515/9781400881970-018.
20. Venkatasubramanian, M., Chetal, K., Schnell, D. J., Atluri, G. & Salomonis, N. Resolving single-cell heterogeneity from hundreds of thousands of cells through sequential hybrid clustering and NMF. *Bioinformatics* **36**, (2020).
21. Young *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of

renal tumors. *Science* **361**, (2018).
22. Miao, Z. *et al.* Putative cell type discovery from single-cell gene expression data. *Nat. Methods* **17**, 621–628 (2020).
23. Sampath, P., Moideen, K., Ranganathan, U. D. & Bethunaickan, R. Monocyte Subsets: Phenotypes and Function in Tuberculosis Infection. *Front. Immunol.* **9**, (2018).
24. Wang, A. *et al.* Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. (2020) doi:10.7554/eLife.62522.
25. Sawai, C. M. *et al.* Transcription factor Runx2 controls the development and migration of plasmacytoid dendritic cells. *J. Exp. Med.* **210**, 2151–2159 (2013).
26. Guilliams, M. *et al.* Unsupervised High-Dimensional Analysis Aligns Dendritic Cells across Tissues and Species. *Immunity* **45**, 669 (2016).
27. Miao, Z. *et al.* Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nat. Commun.* **12**, 1–17 (2021).
28. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
29. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. (2021) doi:10.7554/eLife.63632.
30. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, (2017).
31. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
32. Dias, J., Leeansyah, E. & Sandberg, J. K. Multiple layers of heterogeneity and subset diversity in human MAIT cell responses to distinct microorganisms and to innate cytokines. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5434 (2017).
33. Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. EpiScanpy: integrated single-cell epigenomic analysis. doi:10.1101/648097.
34. Zhao, J. *et al.* Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discovery* **6**, (2020).
35. Ercolano, G. *et al.* Distinct and shared gene expression for human innate versus adaptive helper lymphoid cells. *J. Leukoc. Biol.* **108**, 723–737 (2020).
36. Nicolet, B. P., Guislain, A. & Wolkers, M. C. CD29 enriches for cytotoxic human CD4 T cells. doi:10.1101/2021.02.10.430576.
37. Regulation of Signal Transducer and Activator of Transcription Signaling by the Tyrosine Phosphatase PTP-BL. *Immunity* **26**, 163–176 (2007).
38. Cheng, Z.-Y., He, T.-T., Gao, X.-M., Zhao, Y. & Wang, J. ZBTB Transcription Factors: Key Regulators of the Development, Differentiation and Effector Function of T Cells. *Front. Immunol.* **0**, (2021).
39. Melms, J. C. *et al.* A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, (2021).
40. Adams, T. S. *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science advances* **6**, (2020).
41. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).