nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-0748-5

TooManyCells identifies and visualizes relationships of single-cell clades

Gregory W. Schwartz[1,2], Yeqiao Zhou[1,2], Jelena Petrovic[1,2], Maria Fasolino[3,4], Lanwei Xu[1,2], Sydney M. Shaffer[1,2], Warren S. Pear[1,2], Golnaz Vahedi[3,4] and Robert B. Faryabi[1,2] ✉

Identifying and visualizing transcriptionally similar cells is instrumental for accurate exploration of the cellular diversity revealed by single-cell transcriptomics. However, widely used clustering and visualization algorithms produce a fixed number of cell clusters. A fixed clustering 'resolution' hampers our ability to identify and visualize echelons of cell states. We developed TooManyCells, a suite of graph-based algorithms for efficient and unbiased identification and visualization of cell clades. TooManyCells introduces a visualization model built on a concept intentionally orthogonal to dimensionality-reduction methods. TooManyCells is also equipped with an efficient matrix-free divisive hierarchical spectral clustering different from prevalent single-resolution clustering methods. TooManyCells enables multiresolution and multifaceted exploration of single-cell clades. An advantage of this paradigm is the immediate detection of rare and common populations that outperforms popular clustering and visualization algorithms, as demonstrated using existing single-cell transcriptomic data sets and new data modeling drug-resistance acquisition in leukemic T cells.

Transcription is an important contributor to phenotypic and functional cell states. Emergent technologies such as single-cell RNA sequencing (scRNA-seq) have markedly improved identification and characterization of cell-state heterogeneity. To this end, algorithms for unsupervised delineation and visualization of cells with similar expression patterns have improved the understanding of cell-lineage complexity, tumor heterogeneity and diversity of response to oncology drugs[1–5]. Nevertheless, it remains challenging to simultaneously stratify rare and common cell populations and explore their relationships.

Clustering algorithms have been proposed to partition scRNA-seq data to enable identification of groups of cells with related transcriptional programs[1,6–10]. In most scRNA-seq analyses, the identified cell clusters are visualized using dimensionality-reduction algorithms such as *t*-distributed stochastic neighbor embedding (*t*-SNE) or uniform manifold approximation and projection (UMAP)[11–13]. These workflows produce and visualize single-resolution cell clusterings by using methods that mostly lack quantitative presentation of relationships among the clusters.

Resolution of cell-state stratification unduly influences findings in scRNA-seq experiments. For instance, a resolution separating lymphocytes from monocytes may not readily subdivide various lymphocyte lineages. Given that varying cell states are inherently nested, we postulated that algorithms delineating hierarchies of groups and visualizing their relationships can be used to effectively interrogate echelons of cell states. To this end, we developed TooManyCells for scRNA-seq data visualization and exploration. TooManyCells implements a suite of graph-based algorithms and tools for efficient, global, and unbiased identification and visualization of cell clades. TooManyCells maintains and presents cluster relationships within and across varying clustering resolutions, and enables delineation of context-dependent rare and abundant cell populations.

We demonstrated the effectiveness of TooManyCells in reliably identifying and clearly visualizing abundant and rare subpopulations using several analyses. Three publicly available scRNA-seq data sets, synthetic data, and controlled subsetting and mixing of single-cell population data were used for comparative benchmarking. TooManyCells outperforms other popular methods in detecting and visualizing rare populations, down to the smallest tested benchmark of 0.5% prevalence in several controlled cell admixtures and simulated data. Additionally, TooManyCells assisted in a fine-grained B-cell lineage stratification within mouse splenocytes, and was able to identify rare plasmablasts[14] that were overlooked by popular Louvain-based clustering and projection-based visualization algorithms.

We further used TooManyCells to explore the effect of dosage on acquiring resistance to a gamma-secretase inhibitor (GSI), a targeted Notch-signaling antagonist. While other popular methods failed, TooManyCells revealed a rare resistant-like subpopulation of parental cells. TooManyCells and its individual components are available through https://github.com/faryabib/too-many-cells.

Results

TooManyCells for visualization of cell-clade relationships. Clear visualization is critical for scRNA-seq data exploration, and is dominated by projection-based algorithms such as *t*-SNE and UMAP. For large and complex cell admixtures, projection methods suffer from rendering many overlapping cells, which overwhelms the single-cell-resolution visualization. More importantly, these algorithms generally do not report quantitative inter-cluster relationships and lack interpretable visualizations across clustering resolutions. To address these limitations, we developed TooManyCells for fully customizable visualization of inter-cluster relationships in a tree data abstraction (Fig. 1).

Multiple algorithms use traditional dendrogram plots to infer cell clades from scRNA-seq profiles[15–18]. Yet, robust cell-clade inference remains challenging. Alternatively, outputs of flat clustering algorithms at different resolutions can be related in a tree structure[18];

[1]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Abramson Family Cancer Research Institute Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. [4]Penn Epigenetics Institute, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: faryabi@pennmedicine.upenn.edu

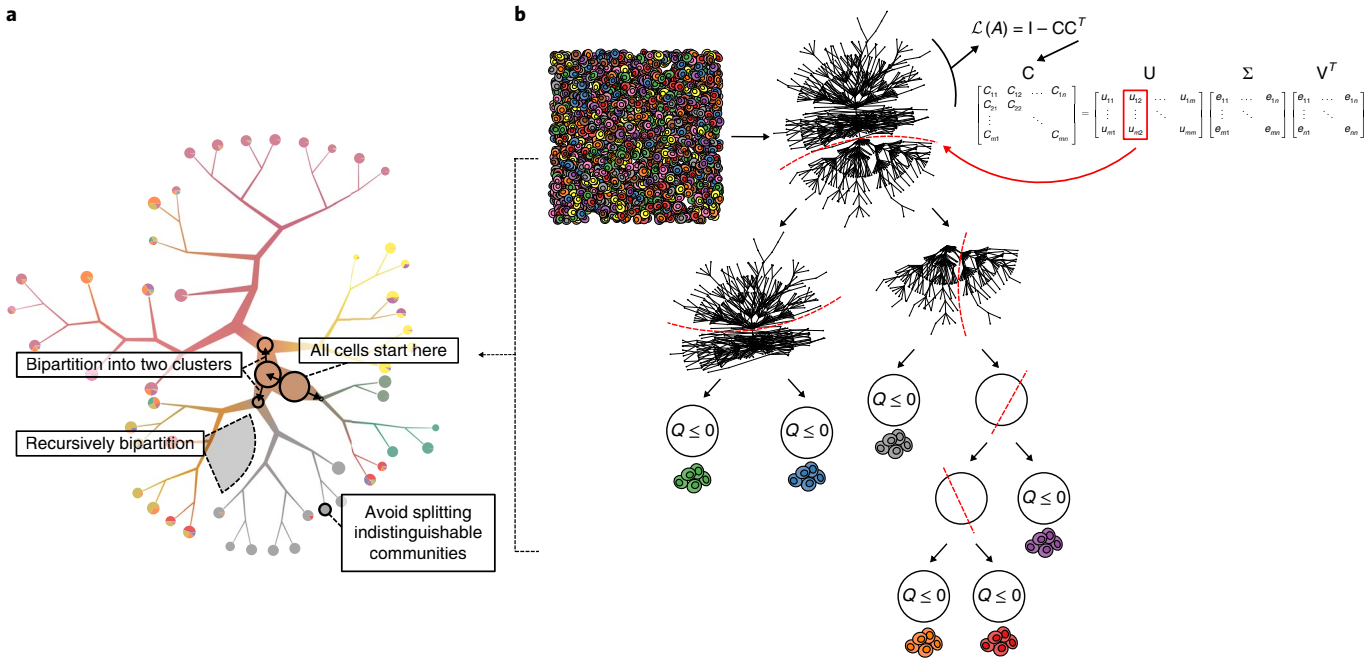

Fig. 1 | The TooManyCells visualization and clustering algorithms. a, TooManyCells visualizes intercluster relationships while providing many capabilities and options, including, but not limited to, weighted-average blending of colors, scaling branches, modularity overlays, smart tree pruning and several leaf-node visualizations. Cells from 11 mouse organs are color-coded on the basis of their organ of origin. **b**, TooManyCells matrix-free divisive hierarchical spectral clustering. TooManyCells is conceptually similar to recursive separation of cells on the basis of their color (representing state or type) similarities: it first separates green and blue from red, purple, orange and gray cells, followed by separation of green from blue, gray from red, purple, and orange, and so on. The network of cells (nodes) connected by their cosine similarities (edges) is recursively bipartitioned (red dashed lines) using truncated singular value decomposition (SVD) of the transformed matrix C that is directly calculated from the gene-expression matrix. Here, truncated SVD calculates only the first two left singular vectors corresponding to the two largest singular values instead of full matrix factorization. This 'matrix-free' process eliminates the need for the explicit calculation of cell–cell similarity (A) and the normalized Laplacian ($\mathcal{L}(A)$) matrices followed by full eigenvalue decomposition (calculation of all the matrices on the right-hand side of the equation instead of only the red-marked column) at each bipartitioning. Recursive bipartitioning is terminated when a candidate split results in non-positive Newman–Girvan modularity (Q). **I**, identity matrix; m, cell number; n, transcript number, T, matrix transpose; **C**, defined in the Methods; **U**, **Σ** and **V**, **C**'s column space bases, singular values and row space bases, respectively.

however, this method relies on arbitrary numbers of resolutions and tuning parameters. We reasoned that divisive hierarchical spectral clustering can overcome these limitations by using all information embedded in the cell–cell similarity graph. To enable efficient generation of the hierarchy, TooManyCells implements a transformation of the gene-expression matrix that eliminates the explicit calculation of cell–cell similarity and Laplacian matrices followed by full matrix factorization, which were otherwise required for finding the most informative bipartition of cells at each branching point (Fig. 1b). This novel 'matrix-free' approach substantially improves the memory and time requirements of divisive clustering and recursively identifies candidate bipartitions to create a hierarchy of cell clades. By using Newman–Girvan modularity[19] as a stopping criteria instead of an optimization parameter, TooManyCells bypasses limitations associated with heuristic global optimization-based clustering, such as Louvain-based algorithms[20,21], avoids creating arbitrarily small clusters, and allows simultaneous detection of large and small clusters (Methods and Supplementary Note 1).

For clear and interpretable displays of cell clades, TooManyCells is designed with many features that facilitate data exploration and assist with finding relevant populations, including branch scaling, weighted-average color blending and statistically driven tree pruning (Fig. 2 and Supplementary Note 1). To enhance data visualization versatility and complement existing single-resolution methods, TooManyCells can display any tree data structure and outputs of other clustering algorithms (Supplementary Figs. 1–4). To this end, TooManyCells produces visually informative hierarchies of nested cell clusters. Inner nodes are clusters at a given resolution, and leaf nodes are finer-grained clusters, for which additional bipartitioning would be as informative as random bipartitioning. To enable an end-to-end built-in scRNA-seq analysis solution, we also equipped TooManyCells with a suite of tools and functionalities including, but not limited to, data normalizations, data filtrations, similarity measure calculation, subtree generation, differential expression, data import/export and novel algorithms for scRNA-seq diversity quantification and rarefaction analysis (Supplementary Fig. 5, Methods and Supplementary Note 2).

TooManyCells efficiently identifies pure cell clusters. To assess the performance of TooManyCells, we first used the Tabula Muris data sets[22] to examine the extent of cell homogeneity in cluster identification. As part of the Tabula Muris, 11 organs from 3-month-old mice were profiled by scRNA-seq, and their cell-type composition was determined using organ-specific optimized analyses[22]. TooManyCells clusters were compared with the clusters generated by widely used Cell Ranger[23], Monocle[8], Phenograph[6], Seurat[7], RaceID[24], CIDR[16] and BackSPIN[15] algorithms, the latter two being agglomerative and divisive hierarchical algorithms, respectively (Fig. 3a–d and Supplementary Note 3).

For each algorithm, default or suggested filters and parameters were considered (Methods). The first comparative analysis was performed on the basis of an increased level of cell-mixture complexity, in which the first 3, 6, 9 and finally all 11 data sets from thymus, spleen, bone marrow, limb muscle, tongue, heart, lung, mammary gland, bladder, kidney and liver were considered (Fig. 3a and

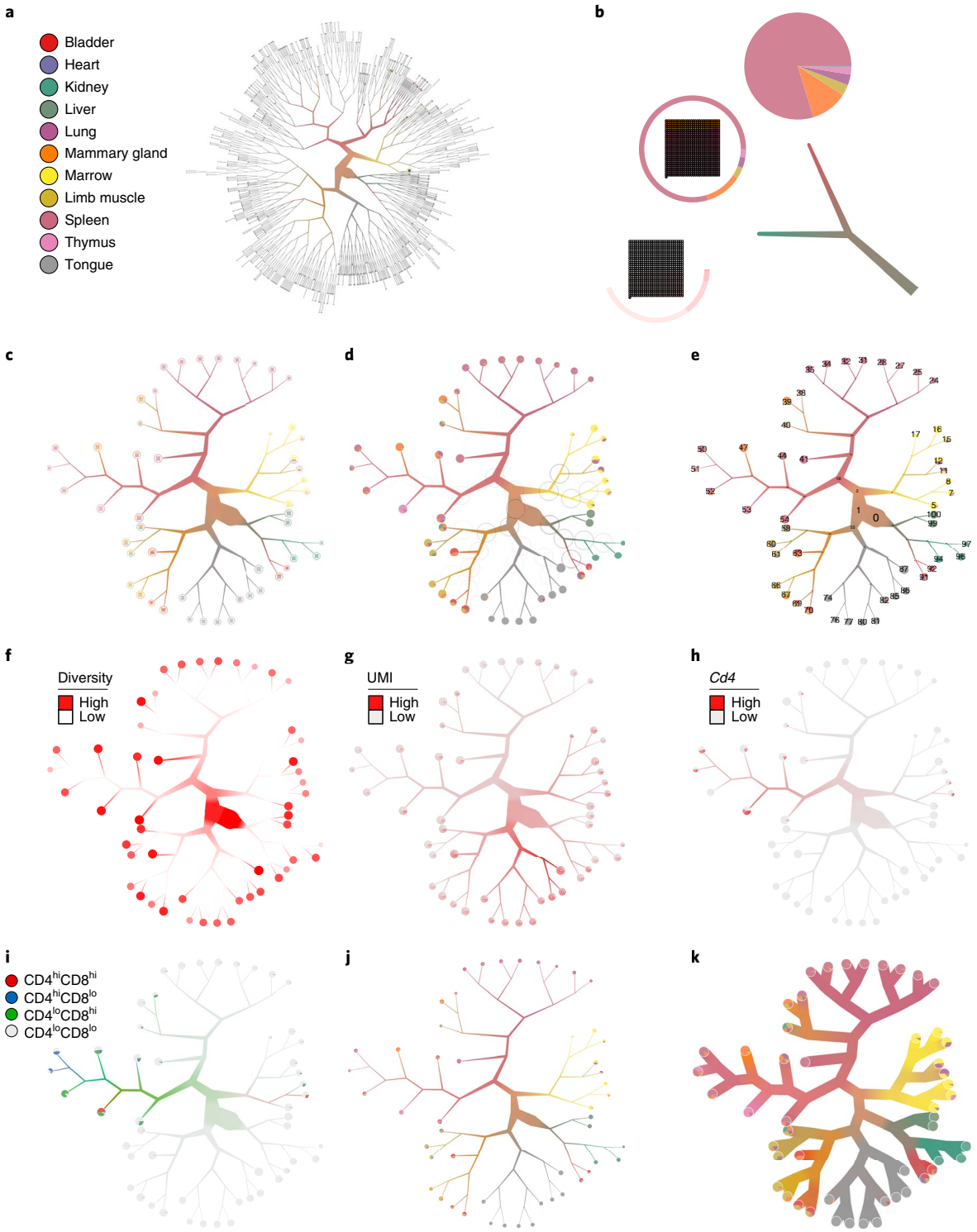

Fig. 2 | Example of TooManyCells visualization capabilities using 11 mouse organs. a, The complete tree with default settings. **b**, Different leaf rendering options (clockwise from bottom: gene expression, 'pie ring', pie chart), and an example of scaling and average-weighted color blending for branches. **c**, Tree from **a** pruned with median(node size) + 3 × MAD(node size), where MAD is median absolute deviation, which is used in **d**–**k**. **d**, Tree with modularity of bipartitioning at each internal node displayed as black circles; higher modularity is represented by darker circumference intensity. **e**, Tree with numbered nodes. **f**, Color-coded tree with a continuous variable (for example, cell diversity of organs; increasing color intensity represents increasing diversity). For clarity, inner and leaf nodes use different intensity scales. **g**, Color-coded tree with a discrete variable presenting unique molecular identifier (UMI) counts. **h**, Color-coded tree with expression level of a specific gene (*Cd4* expression level). **i**, Color-coded tree with expression level of multiple genes (*Cd4* and *Cd8* expression levels). **j**, Tree with nondefault scaling width. **k**, Tree with disabled branch scaling.

Supplementary Fig. 6). Further comparisons were carried out using three additional data sets of cell lines or fluorescence-activated cell sorting (FACS)-purified cells: CD14$^+$ monocytes, CD19$^+$ B cells and CD4$^+$ T cells[23] (Fig. 3b), seven cancer lines[17] (Fig. 3c) and B lymphocytes/natural killer, megakaryocyte-erythroid and granulocyte-monocyte progenitors[25] (Fig. 3d).

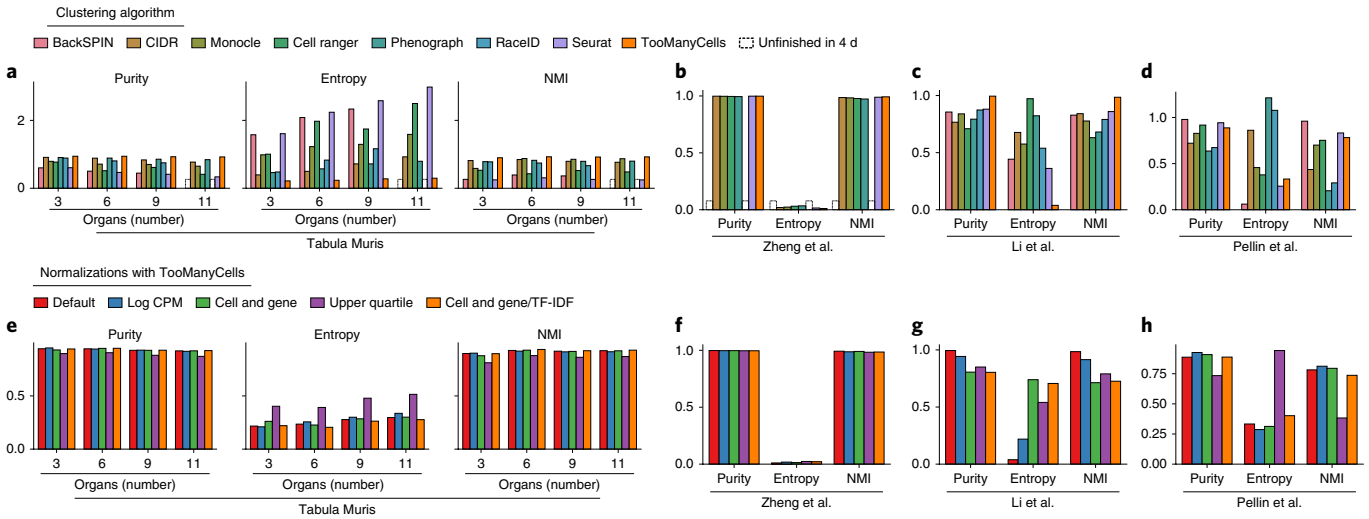

Fig. 3 | Comparative analysis of clustering performance and scalability. a, The population analyzed from the Tabula Muris[22] data set ($n = 41{,}689$ cells) is represented by the number of organs on the *x* axis. The first 3, 6, 9 or 11 organs from the following ordering were considered: thymus, spleen, bone marrow, limb muscle, tongue, heart, lung, mammary gland, bladder, kidney, liver. Purity (left), entropy (middle) and normalized mutual information (NMI, right) performance measures are calculated. Higher purity and NMI represent higher accuracy, while lower entropy indicates better performance. Default or suggested filterings and parameters were used for all algorithms (Methods). A dashed border indicates the algorithm did not finish within 4 d. **b**–**d**, The same measures as in **a** are used to compare clustering accuracy based on data from Zheng et al.[23] ($n = 23{,}910$ cells) (**b**), Li et al.[17] ($n = 561$ cells) (**c**) and Pellin et al.[25] ($n = 2{,}815$ cells) (**d**). **e**–**h**, Data sets equivalent to those in **a**–**d** subjected to TooManyCells normalization procedures: the TooManyCells default normalization (default) term frequency-inverse document frequency (TF-IDF), log counts per million (log CPM), total and gene median count normalization (cell and gene), upper quartile normalization (upper quartile) and total and gene median count normalization followed by TF-IDF normalization (Cell and Gene/TF-IDF).

Rare-cell-clustering RaceID and hierarchical CIDR and BackSPIN methods failed to finish analyses of the high complexity data sets of ~30,000 to 40,000 cells within 4 d (Fig. 3a,b). Across all complexities and evaluation metrics in the Tabula Muris data sets, TooManyCells was the most successful in separating cell-type labels (Fig. 3a). All the scalable algorithms that clustered the immune cells generally performed well. However, TooManyCells again marginally outperformed all others (Fig. 3b). Similarly, TooManyCells performed the best in separating seven distinct cancer-cell lines (Fig. 3c). However, the performance of TooManyCells was close with that of Seurat and Cell Ranger in separating lineage-negative hematopoietic progenitor cells (Fig. 3d). We note that these cells are highly heterogeneous, and their population structures, defined by a few cell-surface markers, remain enigmatic[25,26]. Comparison of different normalization procedures showed that the performance of TooManyCells was only marginally influenced by normalization choice (Fig. 3e–h and Supplementary Note 4).

While not scalable to large data sets (Fig. 3a), BackSPIN, another divisive-clustering algorithm, exhibited the best performance in separating highly diverse hematopoietic progenitor cells (Fig. 3d). Importantly, all the scalable algorithms only report single-resolution cluster outputs at a time, while the multilayer output of TooManyCells identifies context-dependent clades from the entire presented cluster hierarchy. The TooManyCells-rendered cluster tree further guides the choice of clustering granularity by contextualizing cluster features such as relative size, modularity (Fig. 2d) and distance from the root. This unique TooManyCells feature sets it apart from existing visualization algorithms that lack interpretable rendering of relationships across varying clustering resolutions. Furthermore, the run time of TooManyCells' multiresolution clustering was comparable to run times of single-resolution clustering algorithms for small data sets (Supplementary Fig. 7 and Supplementary Note 5), and markedly outperformed them for large data sets (Fig. 3a,b). Together, these data show that in contrast to rare-cell detection (RaceID) and hierarchical clustering (BackSPIN, CIDR), TooManyCells provides accurate and scalable clustering.

TooManyCells accurately delineates both rare and common subpopulations of controlled admixtures. Simultaneous detection of rare and common cell populations is a major challenge in scRNA-seq analysis. While many clustering algorithms claim to identify rare populations, few have explicitly benchmarked this ability. To rigorously assess each algorithm's affinity to delineate rare populations, we simulated different levels of rare and common populations based on cells from different mouse organs. An accurate clustering is expected to not only detect the rare populations from the common, but also distinguish the rare populations from each other. To this end, two equal-size rare populations were mixed with a common cell population. TooManyCells recapitulated known relationships between cell types within mouse organs (Supplementary Figs. 8–18) and showed that T cells were dissimilar from both macrophages and dendritic cells, as expected (Supplementary Fig. 19). On the basis of these data, ten different cell admixtures with different ratios of common T-cell and rare macrophage and dendritic-cell populations were generated (Methods).

Visual inspection of *t*-SNE projections showed discrepancies between the actual cell types and their cluster labels (Fig. 4a,b and Supplementary Fig. 20). Regardless of the clustering algorithm, *t*-SNE plots were limited in clearly distinguishing the two rare populations in an admixture. Visual inspections of *t*-SNE plots identified numerous small islands (Fig. 4a,b, left columns, and Supplementary Fig. 20). However, it was impossible to visually localize the true rare populations in the absence of cell-type labels. This issue is inherent to *t*-SNE, in which distance and density are converted to local density. UMAP projections had similarly poor performance (Supplementary Fig. 21). By contrast, TooManyCells is specifically designed to plot cluster relationships, and thus readily presented the rare populations (Fig. 4c and Supplementary Figs. 22 and 23). In the 10% rare populations admixture,

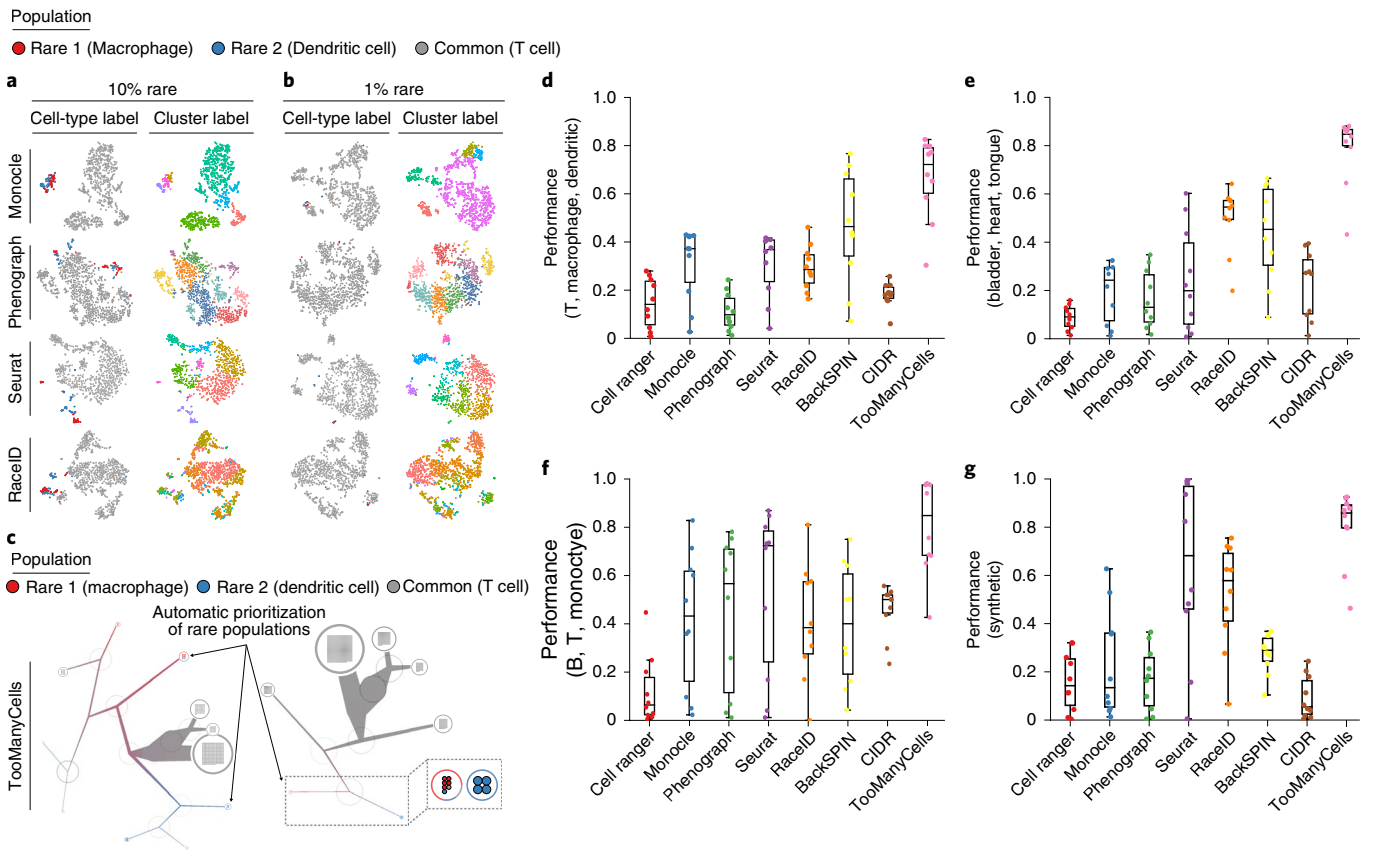

Fig. 4 | Detection of cells from two rare populations mixed with a common population was benchmarked for widely used clustering algorithms. **a**,**b**, Columns from left to right: cells labeled by actual cell types and assigned clusters of a clustering algorithm. Rows from top to bottom: Monocle, Phenograph, Seurat and RaceID *t*-SNE projections. Each projection used the corresponding package's implementation of *t*-SNE with the same seed (Methods). Analysis for 900 common (T cells) and 100 rare (50 macrophages and 50 dendritic cells) cells (**a**) and 990 common and 10 rare cells (5 macrophages and 5 dendritic cells) (**b**) are presented. **c**, TooManyCells priority given to rare cells (pruning median(modularity) + 5 × MAD(modularity)). Left, 900 common and 100 rare cells. Right, 990 common and 10 rare cells. Magnified rare-population-containing subtree showed in insert. Black to white colored circles represent high to low modularity. **d**–**g**, Box-and-whisker plots (center line, median; box limits, upper (75th) and lower (25th) percentiles; whiskers, 1.5× interquartile range; points, outliers) quantifying accuracy of rare population detection in admixtures from various data sets ($m = 10$ admixtures, $n = 1{,}000$ cells): **d**, T cells (common), macrophages (rare1) and dendritic cells (rare2); **e**, mouse bladder (common), heart (rare1), and tongue cells (rare2); **f**, human PBMC FACS-purified CD19$^+$ B (common), CD14$^+$ monocytes (rare1) and CD4$^+$ T (rare2) cells; **g**, three subpopulations of synthetic data. Each point represents the average performance of 10 experiments from an admixture (10 admixtures overall, from 90% common to 99% common). Performance indicates true rare pairs (that is, rare1 with rare1 in the same cluster)/total rare pairs (true rare pairs, and rare1 with rare2). TooManyCells was evaluated with default normalization. To accommodate the Splatter model in **g**, TooManyCells was run with PCA and relaxed modularity cutoff to account for the transformation.

TooManyCells separated the rare and common populations, and it then split the two rare groups, keeping the common cells in large clusters (Fig. 4c, left panel). Rare populations would have been easily identifiable even in the absence of cell-type labels as the branch thickness and modularity values (shown by black circles) pointed out the rare subpopulations (Fig. 4c, left panel). In 1% rare-population mixing experiment, TooManyCells again delineated the rare populations and readily presented them with the help of a drastically smaller subtree (Fig. 4c, right panel). Similar observations were made for eight other mixing experiments with different admixture ratios (Supplementary Figs. 22 and 23).

We next quantitatively compared the performance of TooManyCells in the detection of rare populations (Supplementary Figs. 20–23) with other commonly used clustering algorithms (Methods). These analyses showed that regardless of the purity benchmark (Fig. 3a), TooManyCells frequently outperformed other algorithms (Fig. 4d).

Given that the organ of origin would provide unbiased cell labeling, we further quantified how TooManyCells and other algorithms simultaneously segregated common and rare subpopulations in controlled admixtures consisting of cells from distinct mouse organs. In both the common bladder cells with rare cells from heart and tongue (Fig. 4e) and common tongue cells with more dissimilar (Supplementary Fig. 19) rare bone marrow and mammary gland cells (Supplementary Fig. 24), TooManyCells more accurately separated common and rare cells from different mouse organs.

Furthermore, controlled admixtures of FACS-purified CD14$^+$ monocytes, CD19$^+$ B cells and CD4$^+$ T cells from healthy human peripheral blood mononuclear cells (PBMCs)[23] confirmed that TooManyCells produces the best segregation of common B cells, and rare monocytes and T cells (Fig. 4f). More importantly, while *t*-SNE and UMAP embeddings lacked clear guidance toward the location of rare cells (Supplementary Figs. 25 and 26), structural features of the TooManyCells tree highlighted the rare subpopulations (Supplementary Figs. 27 and 28).

Lastly, we sought to characterize performance using synthetic data. Not only did TooManyCells accurately identify the number of populations in a controlled synthetic admixture (Supplementary Fig. 29), the algorithm also outperformed all other tested methods (Fig. 4f, Supplementary Fig. 30 and Supplementary Note 6).

Together, these data suggest that TooManyCells robustly outperformed the other algorithms in stratifying both common and rare subpopulations, and further revealed that the performance of BackSPIN, RaceID and Seurat markedly varied across benchmarking experiments.

TooManyCells identifies rare plasmablasts in mouse spleen. To further demonstrate the ability of TooManyCells to simultaneously stratify rare and common cell populations de novo, we analyzed the immune-cell composition of the C57BL/6 mouse spleen. With a restricted modularity-pruning threshold (Supplementary Fig. 31), TooManyCells readily separated B cells, T cells, macrophages and dendritic cells (Fig. 5a). As expected, B and T cells composed the majority of profiled splenocytes, and were mostly separated at the first bifurcation. The macrophages were less abundant and were separated from the T cells and further subgrouped. High modularity throughout the macrophage subtree suggested heterogeneity of splenic resident macrophages, confirming outcomes of flow-cytometry analysis[27,28]. Similarly, heterogenous and relatively rare dendritic cells were also partitioned in high-modularity locations (Fig. 5a), as expected[29].

Given the diversity of lymphocytes, we repeated the TooManyCells analysis with a less-restricted modularity-pruning threshold (Fig. 5b and Supplementary Fig. 31). Traversing further along the TooManyCells clustering hierarchy, T and B cells separated into more refined clusters (Fig. 5b). TooManyCells successfully separated $CD4^+$ and $CD8^+$ T cells (Fig. 5b and Supplementary Fig. 32), and stratified more common marginal-zone, germinal-center and follicular B lymphocytes (Fig. 5c). Labeling of the splenic TooManyCells tree by B-cell-subtype signatures[30] identified two branches enriched for rare splenic[14] IgJ-expressing plasma and plasmablast B cells (Fig. 5b–d and Methods). Together, these analyses showed the ability of TooManyCells to stratify both rare and common cell types in mouse spleen, and showcased TooManyCells-enabled multilayer exploration of single-cell clades de novo.

To further assess the ability of popular methods to identify rare plasmablasts in mouse spleen, we used Seurat to generate *t*-SNE plots and cluster splenocytes. Overlaying cells in the *t*-SNE projection with their respective leaves from the TooManyCells tree (Supplementary Fig. 16) showed that, for the most part, cells nearby in the tree were nearby in the *t*-SNE projection (Fig. 5e). However, there were some discrepancies in which cells farther apart in the tree were proximal on the *t*-SNE plot (for example, mixing of green-labeled and pink-labeled cells on the top right of the *t*-SNE plot). Overlaying the B cell subtypes as defined by TooManyCells and validated by B-cell-subtype signatures (Fig. 5c,d) onto the *t*-SNE coordinates failed to visually separate plasmablasts from other B-cell subtypes (Fig. 5f). Furthermore, default Seurat clustering was unable to identify the distinct cluster of rare splenic plasmablasts (Fig. 5g). Together, these results further support the advantage of TooManyCells visualization and clustering over that of widely used algorithms in guiding simultaneous detection of rare and common splenocyte subpopulations.

Different GSI treatment regimens lead to distinct drug-resistant T-ALL populations. T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive malignancy in children and adults[31,32]. Identification of Notch family as the most frequently mutated genes in T-ALL led to clinical testing of gamma-secretase inhibitor (GSI), a targeted Notch-signaling antagonist[33]. However, GSI-resistance development may limit its clinical efficacy[34]. We generated new scRNA-seq data and used TooManyCells to investigate the effect of GSI on individual resistant DND-41 T-ALL cells that were selected under two distinct treatment regimens. Ascending-dose GSI-resistant cells (referred to as ascending resistant) were selected by gradually doubling the GSI dose from ~200% to 1,600% of the DND-41 half-maximum inhibitory concentration (IC_{50}), while sustained high-dose GSI-resistant cells (referred to as sustained resistant) were selected by a prolonged treatment with ~1,600% of the IC_{50} (Fig. 6a and Supplementary Fig. 33a). Transcriptomes of ~10,000 DND-41 cells from ascending resistant, sustained resistant, untreated parental and short-term (24 h) GSI-treated parental populations were profiled.

The TooManyCells tree of these four populations showed mixing of untreated and short-term treated parental cells and the heterogeneity of response to GSI in genetically homogeneous DND-41 parental cells (Fig. 6b), which was not due to technical biases (Supplementary Fig. 33b and independent bulk RNA-seq (data not shown)). While the sustained resistant population occupied a distinct part of the tree (Fig. 6b), the ascending resistant cells showed markedly diverse gene-expression profiles (Fig. 6b) and were significantly more heterogeneous (Supplementary Fig. 33c, $P = 0.0140$). Visualizing and quantifying relationships among the populations further showed that ascending resistant cells partially resembled both sustained resistant and parental cells (Fig. 6b and Supplementary Fig. 33d). TooManyCells revealed that ~40% of the ascending resistant cells were transcriptionally similar to the parental cells (Fig. 6b) and the remaining ascending resistant cells were more closely related to the sustained resistant population. Nevertheless, the expressions of several genes in this group of ascending resistant cells, including proto-oncogene *MYC* and anti-apoptotic gene activating transcription factor 5 (*ATF5*)[35–39], were significantly different from the sustained resistant population (Fig. 6c,d, Supplementary Fig. 33e,f and Supplementary Table 1). Together, these single-cell-resolution analyses identified a subpopulation of ascending resistant cells that, despite similarities with their sustained resistant counterparts, evolved differently to acquire GSI resistance and exhibited significantly lower expression of pro-survival genes—potentially enabling gradual adaptation to elevated GSI.

TooManyCells identifies a rare GSI-resistant-like subpopulation. To investigate the underpinning GSI-resistance mechanisms, we next focused on the sustained resistant cells (Fig. 6e), which were more distinct from the parental cells (Fig. 6b and Supplementary Fig. 33d). GSI treatment equally blunted expression of Notch and its known targets in drug-responsive and sustained resistant cells (Supplementary Fig. 33f–j and Supplementary Tables 2 and 3). By contrast, while short-term GSI treatment significantly reduced expression of *MYC* and its known targets in most of the parental cells (Supplementary Fig. 33f,i), it had no significant effect on their expression in the sustained resistant cells (Supplementary Fig. 33f,j). Together, these data imply that Notch-independent elevated *MYC* expression contributes to high GSI dosage tolerance.

To further test this hypothesis, we compared individual resistant and parental cells. Interestingly, this single-cell-resolution analysis revealed a rare (<1%) parental subpopulation that was transcriptionally similar to sustained resistant cells and localized at their encompassing subtree (Fig. 6e). This rare resistant-like subpopulation showed markedly elevated *MYC* levels compared with those of the other parental cells (Fig. 6f,g, 2.85 fold change, $P = 4.01 \times 10^{-8}$). Furthermore, gene set enrichment analysis (GSEA)[40] showed that known *MYC* targets[41] were the most differentially expressed pathways in the rare resistant-like cells compared with both other parental (Supplementary Fig. 33k and Supplementary Table 4) and sustained resistant cells (Supplementary Fig. 33l and Supplementary Table 5). Single-molecule RNA fluorescence in situ hybridization (FISH) analysis independently showed the prevalence and rarity of high *MYC* levels in sustained resistant and parental DND-41 cells, respectively (Fig. 6h and Supplementary Table 6).

Having verified the existence of high *MYC*-expressing resistant-like cells, we sought to find this rare parental subpopulation using other algorithms to compare against TooManyCells. These analyses showed that both *t*-SNE projection (Fig. 6i) and Seurat clustering (Fig. 6j) were unable to visually and algorithmically stratify this rare resistant-like subpopulation from the rest of the parental cells

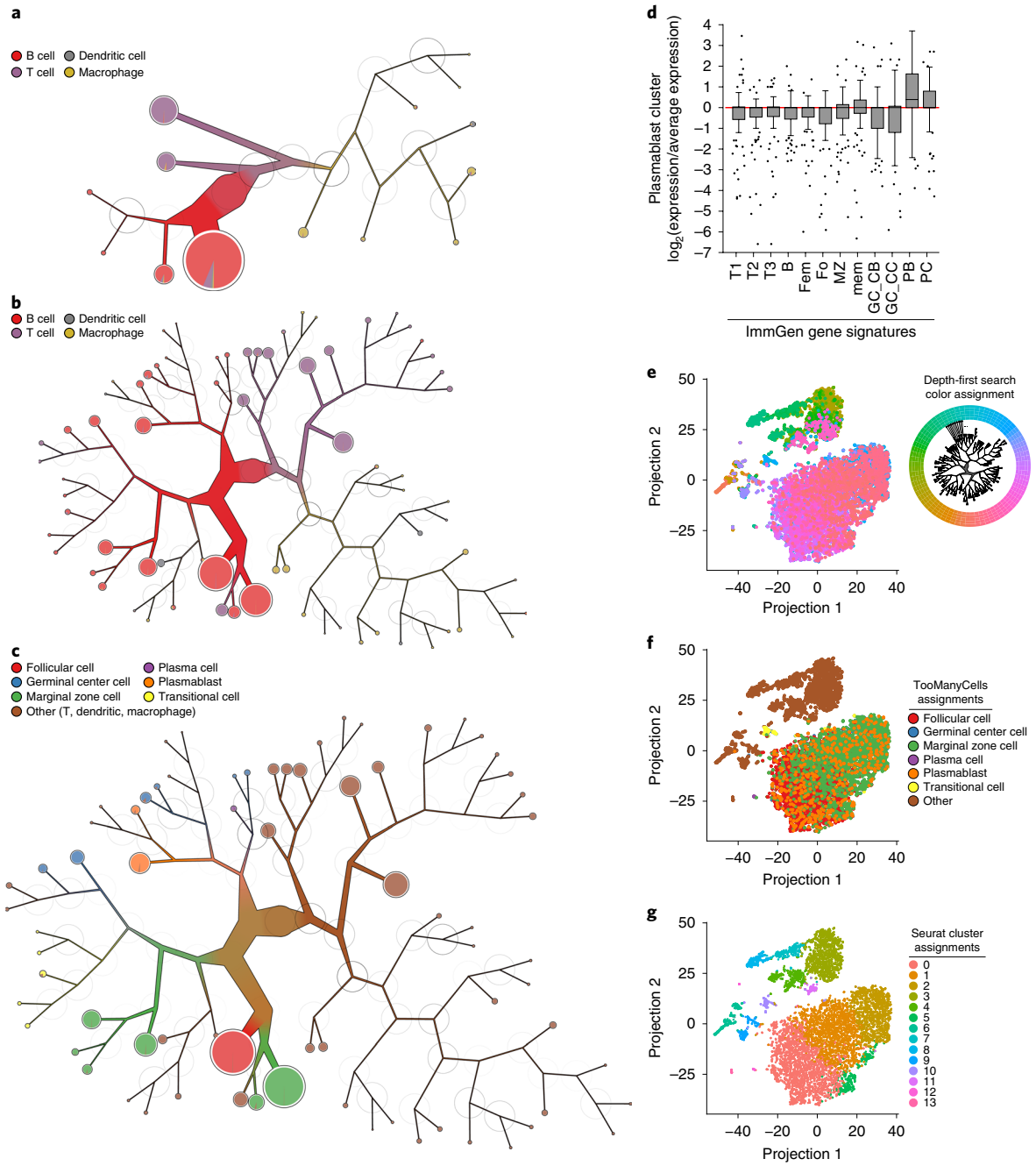

Fig. 5 | TooManyCells stratifies rare plasmablasts in mouse spleen. a,**b**, TooManyCells clustering tree of the mouse splenocytes labeled with major immune-cell lineages based on predefined lineage markers[22] with more (0.1 modularity) (**a**) and less (0.025 modularity) (**b**) restricted modularity-pruning thresholds, respectively. **c**, Tree from **b** colored with newly identified B cell subtypes (Methods). **d**, ImmGen MyGeneSet[30] gene expression for the top *n* = 100 differentially expressed genes of the plasmablast node from **c** compared with all other B-cell subtypes (box-and-whisker plots: center line, median; box limits, upper (75th) and lower (25th) percentiles; whiskers, 1.5× interquartile range; points, outliers). **e**, Cells from Supplementary Fig. 16 projected using Seurat's processing and *t*-SNE, colored by TooManyCells clustering tree leaves; each leaf is assigned a different color (top-right insert). Similar colors represent nearby locations within the tree; for example, pink and purple are closer in the tree than pink and green. **f**, Coordinates from *t*-SNE projection in **e**, colored by subset populations from **c**. Orange color-coded plasmablasts are indistinguishable from other B lymphocytes. **g**, Coordinates from the *t*-SNE projection in **e** colored by Seurat-generated cluster labels fails to separate plasmablasts. At each bipartitioining, the darkness of the circle circumference presents the modularity level. Definitions of *x* axis ticks from **d**: T1, splenic T1 (transitional); T2, splenic T2; T3, splenic T3; B: splenic B cells; Fem, female splenic B cells; Fo, splenic follicular; MZ, splenic marginal zone; mem, splenic memory; GC_CB, splenic germinal center centroblasts; GC_CC, splenic germinal center centrocytes; PB, splenic plasmablasts; PC, splenic plasma cells. *n* = 9,552 cells in all the panels.

(Supplementary Fig. 33m). Together, these analyses demonstrate the unique ability of TooManyCells to guide discovery of a rare DND-41 subpopulation that could potentially tolerate high GSI doses, and hint at underpinning resistance mechanisms.

Discussion

Popular single-cell clustering and visualization methods have been firmly set in variations of single-resolution clustering and projection-based visualization algorithms. While these methods are

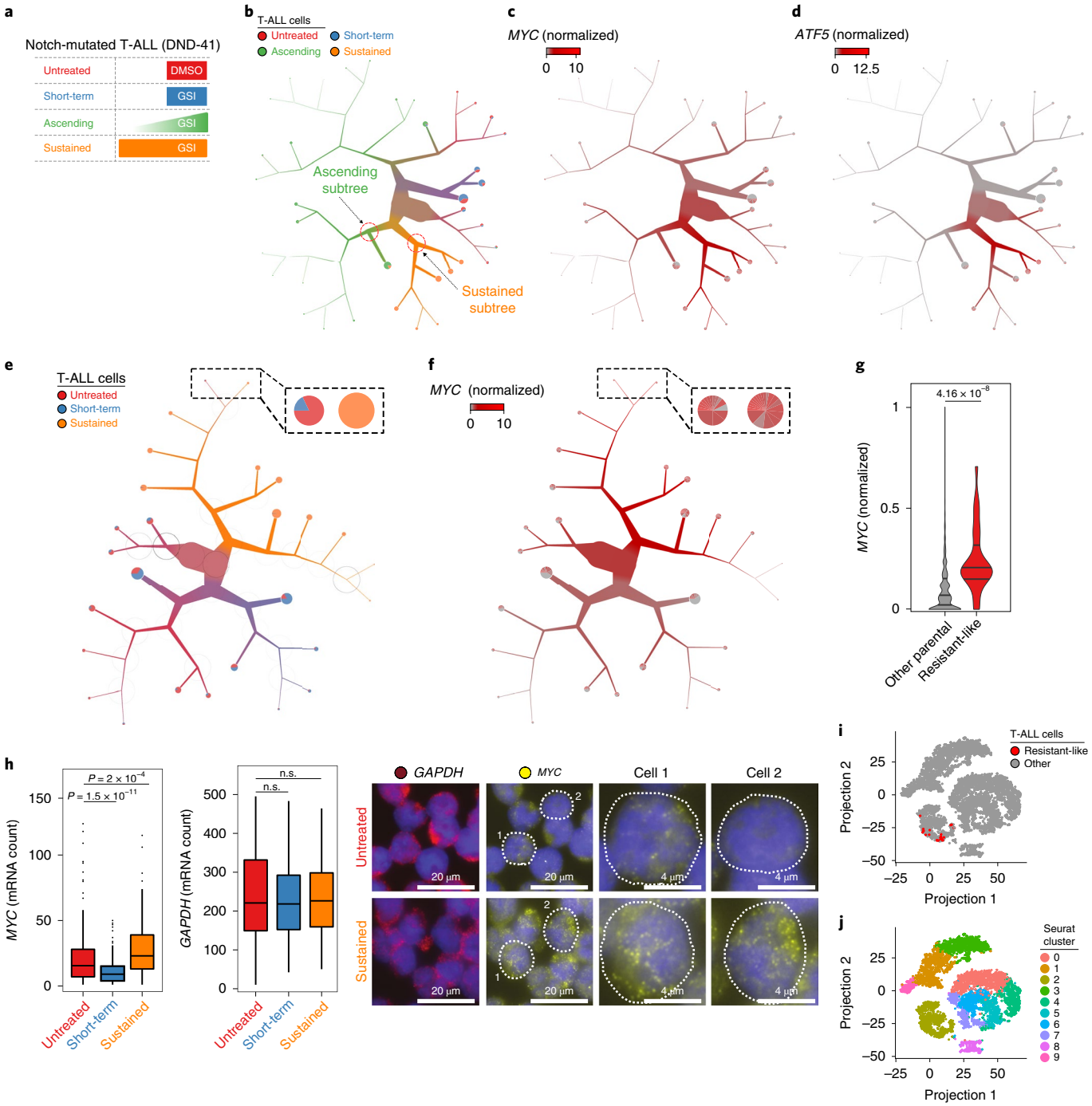

Fig. 6 | TooManyCells identifies GSI-resistant cell heterogeneity and detects resistant-like T-ALL cells. a, Treatment strategies for untreated ($n = 2,338$ cells), short-term ($n = 2,616$ cells), ascending ($n = 2,727$ cells) and sustained ($n = 2,417$ cells) DND-41 populations. **b**, TooManyCells tree showing distinct GSI-resistant populations ($n = 10,098$ cells). **c**,**d**, Upper quartile normalized (UQ) *MYC* (**c**) and *ATF5* (**d**) expression overlaid onto the tree in **b**. Gray to red, low to high expression. **e**, TooManyCells tree of parental and sustained populations ($n = 7,371$ cells). Magnified resistant-like subtree in insert. **f**, UQ *MYC* expression overlaid onto the tree in **e**. Magnified resistant-like subtree in insert. At each bipartitioining, the darkness of the circle circumference presents the modularity level. **g**, Violin plots (center line, median; upper and lower lines, 75th and 25th percentiles; lower and upper bounds, minimum and maximum) normalized *MYC* expression of resistant-like ($n = 28$ cells) and other parental ($n = 4,926$ cells) cells (two-tailed Mann–Whitney *U* test, $P = 4.16 \times 10^{-8}$). **h**, Box-and-whisker plots (center line, median; box limits, upper (75th) and lower (25th) percentiles; whiskers, 1.5× interquartile range; points, outliers) showing single-cell *MYC* (left) and *GAPDH* (center) RNA FISH signal distributions for untreated ($n = 250$ cells), short-term ($n = 261$ cells; two-tailed *t*-test, *MYC*: $P = 1.5 \times 10^{-11}$), and sustained ($n = 222$ cells; two-tailed *t*-test, *MYC*: $P = 2 \times 10^{-4}$) populations. Cell images (right) of RNA FISH signals for *GAPDH* (pseudo-color red) and *MYC* (pseudo-color yellow) in untreated (top) and sustained (bottom) cells. Top third and fourth columns showing two untreated cells with high *MYC* and low *MYC* expression, respectively. Bottom third and fourth columns showing two sustained cells with high *MYC* expression. Cell nuclei in purple. NS: $P > 0.005$. **i**, Cells from **e** projected using Seurat ($n = 4,954$ cells), colored by resistant-like population (red) from **e**. **j**, Coordinates from **i** colored by Seurat-generated clusters.

inherently useful for single-cell analysis, they may be unsuitable for certain applications as demonstrated in this study. Here, we developed TooManyCells, which provides complementary algorithms for clustering and visualization. TooManyCells uses a recursive technique to repeatedly identify subpopulations whose relationships are maintained in a tree. Compared with projection-based algorithms, the TooManyCells visualization model is different and, in conjunction with an array of visualization features, enables a flexible platform for cell-state stratification, exploration and rare-population detection. In addition to clustering and visualization, TooManyCells also provides other capabilities including, but not limited to, heterogeneity assessment, clumpiness measurement and diversity and rarefaction statistics. In addition to synthetic data, the superior performance of TooManyCells to simultaneously identify rare and common cell populations was demonstrated in three independent contexts. In controlled settings, TooManyCells not only separated the two rare cell populations from an admixture of common and rare cells, but successfully sequestered the two rare populations from each other. Applying TooManyCells to cell-lineage identification showed its ability to isolate rare plasmablasts from total mouse splenocytes, while a popular single-cell tool and visualization failed to do so. Lastly, TooManyCells was able to detect a resistant-like subclone in DND-41 cells with exceptionally high *MYC* levels that was separately verified by single-molecule RNA FISH and could potentially tolerate high doses of Notch inhibitor GSI, leading to the development of drug resistance in Notch-mutated T-ALL.

In addition to performance, scalability, and usability, we considered flexibility and versatility in the TooManyCells design. TooManyCells is a generic framework consisting of several algorithms that may be interchanged with other existing algorithms. The TooManyCells clustering and visualization modules, ClusterTree and BirchBeer, can be potentially used for analysis of other single-cell genomic or observation-feature data, respectively. Together, our studies suggest that further improvement of clustering and visualization techniques are warranted to fully explore outputs of various single-cell measurement technologies. TooManyCells is a step in that direction.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0748-5.

Received: 25 October 2019; Accepted: 15 January 2020;
Published online: 2 March 2020

References

1. Lafzi, A., Moutinho, C. & Picelli, S. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742 (2018).
2. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
3. Packer, J. & Trapnell, C. Single-cell multi-omics: an engine for new quantitative models of gene regulation. *Trends Genet.* **34**, 653–665 (2018).
4. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**, F1000 (2016).
5. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell rna-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
6. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
7. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
8. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
9. Azizi, E., Prabhakaran, S., Carr, A. & Pe'er, D. Bayesian inference for single-cell clustering and imputing. *Genomics Comput. Biol.* **3**, 46 (2017).
10. Ho, Y.-J. et al. Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome Res.* **28**, 1353–1363 (2018).
11. Van der Maaten, L. & Hinton, G. Visualizing data using T-SNE. *J. Mach. Learn. Res.* **9**(Nov), 2579–2605 (2008).
12. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
13. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
14. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody- secreting plasma cells. *Nat. Rev. Immunol.* **15**, 160–171 (2015).
15. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
16. Lin, P., Troup, M. & Ho, J. W. K. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59 (2017).
17. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
18. Zappia, L. & Oshlack, A. C. lustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* **7**, 7–9 (2018).
19. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).
20. Lancichinetti, A. & Fortunato, S. Limits of modularity maximization in community detection. *Phys. Rev. E* **84**, 066122 (2011).
21. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
22. The Tabula Muris Consortium. et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature* **562**, 367–372 (2018).
23. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
24. Herman, J. S. & Sagar and Grün, D. Fateid infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
25. Pellin, D. et al. Comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat Commun* **10**, 1–15 (2019).
26. Dahlin, J. S. et al. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in kit mutant mice. *Blood* **131**, e1–e11 (2018).
27. Borges da Silva, H. et al. Splenic macrophage subsets and their function during blood-borne infections. *Front. Immunol.* **6**, 480 (2015).
28. Den Haan, J. M. M. & Kraal, G. Innate immune functions of macrophage subpopulations in the spleen. *J. Innate Immun.* **4**, 437–445 (2012).
29. Hey, Y. Y. & O'Neill, H. C. Murine spleen contains a diversity of myeloid and dendritic cells distinct in antigen presenting function. *J. Cell. Mol. Med.* **16**, 2611–2619 (2012).
30. Jojic, V. et al. Identification of transcriptional regulators in the mouse immune system. *Nat. Immunol.* **14**, 633–643 (2013).
31. Winter, S. S. et al. Improved survival for children and young adults with t-lineage acute lymphoblastic leukemia: results from the children's oncology group AALL0434 methotrexate randomization. *J. Clin. Oncol.* **36**, 2926–2934 (2018).
32. Marks, D. I. et al. T-cell acute lymphoblastic leukemia in adults: clinical features, immunophenotype, cytogenetics, and outcome from the large randomized prospective trial (ukall XII/ECOG 2993). *Blood* **114**, 5136–5145 (2009).
33. Aster, J. C., Pear, W. S. & Blacklow, S. C. The varied roles of notch in cancer. *Annu. Rev. Pathol. Mech. Dis.* **12**, 245–275 (2017).
34. Knoechel, B. et al. An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia. *Nat. Genet.* **46**, 364–370 (2014).
35. Dluzen, D., Li, G., Tacelosky, D., Moreau, M. & Liu, D. X. BCL-2 is a downstream target of ATF5 that mediates the prosurvival function of ATF5 in a cell type-dependent manner. *J. Biol. Chem.* **286**, 7705–7713 (2011).
36. Yamazaki, T. et al. Regulation of the human chop gene promoter by the stress response transcription factor ATF5 via the AARE1 site in human hepatoma HepG2 cells. *Life Sci.* **87**, 294–301 (2010).
37. Liu, D. X., Qian, D., Wang, B., Yang, J.-M. & Lu, Z. P300-dependent ATF5 acetylation is essential for egr-1 gene activation and cell proliferation and survival. *Mol. Cell. Biol.* **31**, 3906–3916 (2011).
38. Angelastro, J. M. Targeting ATF5 in cancer. *Trends Cancer* **3**, 471–474 (2017).
39. Karpel-Massler, G. et al. A synthetic cell-penetrating dominant-negative ATF5 peptide exerts anticancer activity against a broad spectrum of treatment-resistant cancers. *Clin. Cancer. Res.* **22**, 4698–4711 (2016).
40. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
41. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Systems* **1**, 417–425 (2015).

Methods

Clustering. TooManyCells implements a generalized adaptation of a matrix-free hierarchical spectral-clustering process originally proposed for text mining[42]. Spectral clustering using normalized cuts is a technique to partition data into groups, or clusters, in which the items in a cluster are more similar to each other than they are to items in other clusters[43]. This analysis is based on the pairwise similarity between items, leading to a computational complexity of $O(m^2)$ with m items[42]. Let $\mathbf{A}$ be a similarity matrix where $\mathbf{A}(i, j)$ represents the similarity between items i and j and $\mathbf{D} = \mathrm{diag}(\mathbf{A1})$ be the diagonal matrix where $\mathbf{1}$ is a column vector of 1s. Then

$$\mathcal{L}(\mathbf{A}) = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

defines the normalized Laplacian of $\mathbf{A}$. A partition into two clusters denoted by 0 and 1 labels can be defined as

$$C(i) = \begin{cases} 1, & \mathbf{V}(i) \geq 0 \\ 0, & \mathbf{V}(i) < 0 \end{cases}$$

where $\mathbf{V}$ is the eigenvector corresponding to the second smallest eigenvalue of $\mathcal{L}(\mathbf{A})$[43]. Alternatively, the eigenvector corresponding to the second largest eigenvalue of the shifted Laplacian,

$$\hat{\mathcal{L}}(\mathbf{A}) = \mathbf{I} + \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

can be used instead of the second smallest eigenvalue of the Laplacian matrix. While this process bipartitions the data into two clusters, its inefficiency in both time and space makes the algorithm impractical for recurrent clustering of a large number of single cells. To improve the speed of spectral clustering while retaining the original accuracy, TooManyCells implements a generalized version of an algorithm that was originally proposed for text mining[42] and can be used with sparse scRNA-seq matrices or any other observation/feature matrix. This implementation explicitly circumvents calculating $\mathbf{A}$ and the complete SVD of $\mathcal{L}(\mathbf{A})$.

To this end, let $\mathbf{B}_1$ be an $m \times n$ matrix with m rows of cells and n columns of read counts. TooManyCells takes as input a transpose of this matrix to conform to the current single-cell-matrix file-format standards in which the cells are columns. By default, TooManyCells offers the option to remove columns (genes) with no reads and rows (cells) with <250 read counts. Then, for all $1 \leq i \leq m$, $1 \leq j \leq n$,

$$\mathbf{B}_2 = \log(\mathrm{m}/\mathrm{d}_\mathrm{j})\mathbf{B}_1(\mathrm{i},\mathrm{j})$$

where $d_j = \sum_{k=1}^{m} \delta[\mathbf{B}_1(k, j)]$ and $\delta(x)$ is 1 if $x \geq 1$ and 0 if $x = 0$, for all $x \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ is a set of positive integers. This normalization transforms $\mathbf{B}_1$ into a term frequency-inverse document frequency (TF-IDF) matrix $\mathbf{B}_2$ (refs. [44,45]), where the importance of common genes is de-emphasized for clustering. Intuitively, a ubiquitously expressed gene is unlikely to be as important for cell clustering compared with a gene expressed only in a given subpopulation. Other data normalizations can be performed prior to this transformation, or replace the TF-IDF process entirely. For instance, one may normalize each cell based on its total read count followed by the normalization of each gene by that gene's median positive read count. In order to relate cells in a matrix-free manner, cosine similarity was used[46]. It has been shown[42] that the similarity matrix $\mathbf{A}$ can be derived from $\mathbf{B}_2$ with

$$\mathbf{A}(i, j) = \frac{\sum_{k=1}^{n} \mathbf{B}_2(i, k)\mathbf{B}_2(j, k)}{\sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(i, k)}\sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(j, k)}}$$

However, in order to lower the computational complexity, TooManyCells does not calculate this matrix. Instead, a new matrix $\mathbf{B}$ is defined as

$$\mathbf{B}(i, j) = e_i^{-1}\mathbf{B}_2(i, j)$$

where $e_i = \sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(i, k)}$ is the Euclidean norm of $\mathbf{B}_2$ row i.

To prepare the matrix as a form of a normalized Laplacian, let $\mathbf{D} = \mathrm{diag}(\mathbf{B}(\mathbf{B}^T\mathbf{1}))$ and $\mathbf{C} = \mathbf{D}^{-1/2}\mathrm{B}$, where T denotes matrix transposition. Then the eigenvector of $\mathcal{L}(\mathbf{A})$ corresponding to the second smallest eigenvalue is the second left singular vector corresponding to the second-largest singular value of $\mathbf{C}$, which can be found using truncated SVD[42]. It has been shown that the computation complexity of this process is $O(Jm)$, the number of non-zero entries of $\mathbf{C}$, where J is the average number of expressed genes within a cell. This bipartition can be recursively applied to each delineated cluster until a stopping criteria is reached, which results in a divisive hierarchical cluster structure.

In accordance with the original implementation[42], TooManyCells uses Newman–Girvan modularity (Q)[19] as a stopping criteria. Modularity is a measure from community detection which has also been used in single-cell clustering through optimization using the Louvain method[6,7,21]. Let $G = (V, E)$ be a weighted graph of m nodes (cells) with e edges. Then, as $\mathbf{A}$ represents the connectivity strength among nodes, Newman–Girvan modularity measures the strength of the partition of nodes. For a bipartition,

$$Q(C_1, C_2) = \sum_{k=1}^{2}\left(\frac{O_{kk}}{L} - \left(\frac{L_k}{L}\right)^2\right)$$

where $O_{kk} = \sum_{i \in C_k, j \in C_k} \mathbf{A}(i, j)$ is the total degree of nodes in cluster C_k, if $d_i = \sum_{j=1}^{m} \mathbf{A}(i, j)$ is the degree of node i then $L_k = \sum_{i \in C_k} d_i$ is the total degree of nodes in C_k, and $L = \sum_{i=1}^{m} d_i$ is the degree of all nodes in the network. Q measures the distance of edges within clusters to the random distribution of clusters, such that $Q > 0$ denotes non-random communities and $Q \leq 0$ demonstrates communities randomly found[19,42].

TooManyCells uses Q to assess a candidate bipartition of cells to determine whether to continue the recursion or stop as a leaf in the divisive hierarchical clustering. That is, at each bipartion, if $Q > 0$ then continue the recursion, otherwise stop. Thus, the end result of this top-down clustering is a tree structure of clusters, where each inner node is a cluster and the leaves are the most fine-grained clusters where any additional splitting would lead to random partitioning of cells. This process has $O(Jm\log m)$ computational complexity[42]. The code for the TooManyCells implementation of this algorithm is available at https://github.com/faryabib/too-many-cells.

Visualization. The TooManyCells clustering algorithm results in a tree structure, where each inner node is a coarse cluster and each leaf is the most refined cluster per modularity measure. The BirchBeer rendering method was developed for displaying single-cell-cluster hierarchies. To this end, BirchBeer utilizes graphviz for node coordinate placement and the Haskell diagrams library as rendering engine.

BirchBeer provides a multitude of graphical features to assist in the detection and interpretation of cell clusters. The tree leaves can be displayed in various ways. Single-cell-resolution exploration is facilitated by drawing color-coded individual cells at the tree leaves. Alternatively, a pie chart can be shown to visualize a summary of the cell composition of the clusters at the tree leaves. Both single-cell resolution and statistical summarization can be shown using a 'pie ring'. Each tree branch can be scaled to the relative number of cells within each subtree, allowing for quick inspection of cell-population sizes of various clustering levels and visualizing clusters of rare and common populations. Furthermore, colors can be applied to each branch such that the weighted average blend of the colors of each label in the subtree is used, allowing for immediate detection of subtrees with large differences or similarities. Cluster numbers can be displayed on each node, tracing the data back into a human-readable interpretation of differences between the clusters at various hierarchy levels. Furthermore, the modularity of each candidate split can be displayed at each node as a black circle with varying darkness to demonstrate the dissimilarity of cell populations encompassing that assay. Large trees may result in busy figures, much like large t-SNE plots, so options to prune the tree are available. Cutting the tree at certain levels, node sizes or modularity are some options, but additionally there is a statistically driven option called --smart-cutoff, which cuts the tree depending on the MAD. For instance, a stopping criteria of four MADs from the median node size to keep the structure of the tree but prune smaller branches. BirchBeer accepts JSON trees as a standard input. The code for BirchBeer is available at https://github.com/faryabib/birch-beer.

Differential expression. Given multiple cluster-identification numbers, TooManyCells can perform differential expression analysis to identify the difference between the gene expression of cells in these clusters. TooManyCells interfaces with edgeR for differential-expression analysis[47]. Cells were processed using the recommended edgeR settings for single-cell analysis: genes with at least one count per million (cpm) in at least two cells were kept, normalized with calcNormFactors, and analyzed with estimateDisp, glmFit and glmLRT, respectively. To visually facilitate this analysis, BirchBeer can label clusters with their identification numbers. All the presented differential expression analyses and statistics use this feature of TooManyCells.

Diversity analysis. While Shannon entropy is frequently used as a measure of 'diversity', the effective number of species is a more meaningful measure of diversity in biological settings. For example, a population with 16 equally abundant species should be twice as diverse as a population with 8 equally abundant species. Assuming each cell is an 'organism' belonging to a 'species' group defined by the clustering algorithm, then a diversity index can be applied to find the effective number of cell states in a population.

The diversity satisfying such a property can be defined as[48]

$${}^{q}D = \left(\sum_{i=1}^{R} p_i^q\right)^{(1/(1-q))} \quad (1)$$

where p_i is the frequency of species i, R is the total number of species in the population, and q is the 'order' of diversity. $q > 1$ gives additional weight towards common species, and more weight is given to rare species when $q < 1$. $q = 1$ gives equal weight to all the species regardless of their commonality, and is defined as

$${}^{1}D = \exp\left(-\sum_{i=1}^{R} p_i \ln p_i\right)$$

Several diversity measures can be derived from equation (1). For instance, ${}^{0}D$ defines richness, or the number of species, in the population. ${}^{1}D$ relates to

exp(Shannonentropy) and ${}^2D$ is the inverse of the Simpson index. Various diversity measures have been used previously in domains such as lymphocyte receptor repertoires and cell clones[49–51]. Here, we use the diversity in TooManyCells to quantitate the effective number of cell states within a population.

TooManyCells implements the concept of rarefaction curve from ecology[52] to estimate the number of detectable species in a given number of profiled single cells. Briefly, the estimated number of species in a population can be calculated from a given number of samples taken from a population through random subsampling. The estimated number of species in a subsample of size n representing X_n species can be calculated as

$$E[X_n] = R - \binom{N}{n}^{-1} \sum_{i=1}^{R} \binom{N - N_i}{n} \quad (2)$$

where N is the total number of cells, R is the total number of cell states in all samples and N_i is the number of cells belonging to state i. For the interval [0, R], equation (2) generates a rarefaction curve that shows the estimated number of species for a given number of profiled cells. The steepness of the rarefaction curve may represent the heterogeneity of a population. For a given number of subsamples, the estimated number of species across multiple populations can be compared based on their respective rarefaction curves. This property is useful for comparing populations with different sample sizes. A plateau in the curves indicates no substantial increase in the number of new cell states, implying a sufficient sampling to observe all the cell states in a sample. TooManyCells implements this procedure to rarefy populations.

Cluster purity. To compare the accuracy of clustering algorithms, we used measures that quantify the extent of clustering output 'purity'. We considered cluster output 'purity' measures since they mitigate lack of information about markers accurately defining 'true' cell identity. Moreover, these measures are robust to cluster size variability. For instance, FACS-purified $CD4^+$ cells lack the resolution to accurately define 'ground truth' cell types, as these cells comprised of several functionally well-characterized subtypes (for example, various T_H1, T_H2, T_H17, T_{reg} and many more $CD4^+$ T cell types). To assess cluster 'purity', three measures were used: purity, entropy and NMI. All three measures are commonly used in scRNA-seq comparative analysis[53–55].

Purity is based on the frequency of the most abundant class (for example cell type) in a cluster. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ be the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ be the set of classes. Then purity is defined as

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

where N is the total number of cells, ω_k is the set of cells in cluster k and c_j is the set of cells in class j (ref. [45]). This measure ranges from 0, poor clustering, to 1, perfect clustering.

Entropy as a measure of cluster accuracy uses Shannon entropy to measure the expected amount of information from the clusters. The entropy of each cluster k is defined by

$$H(\omega_k) = \sum_j \frac{|\omega_{kj}|}{|\omega_k|} \log \frac{|\omega_{kj}|}{|\omega_k|}$$

where ω_{kj} is the set of cells from $\omega_k \cap c_j$. Then, according to ref. [56], the entropy for the entire clustering is

$$\text{entropy}(\Omega, \mathbb{C}) = \sum_k \frac{|\omega_k|}{N} H(\omega_k)$$

Here, lower entropy of a clustering indicates higher accuracy.

NMI measures the normalized dependency of the class labels on the cluster labels, or the amount of information about the class labels gained when the cluster labels are given. Mutual information is defined by

$$I(\Omega; \mathbb{C}) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|}$$

To compare mutual information across clusterings, $I(\Omega; \mathbb{C})$ is normalized to the interval [0, 1]. As $I(\Omega; \mathbb{C})$ is bounded by $\min[H(\Omega), H(\mathbb{C})]$, where

$$H(\Omega) = -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N}$$

is the entropy of Ω along with the analogous $H(\mathbb{C})$, total normalization NMI can be defined by

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{\min[H(\Omega), H(\mathbb{C})]}$$

where higher values indicate more accurate clustering based on $\mathbb{C}$ (ref. [57]).

For Tabula Muris, four data sets were generated on the basis of organ admixture complexity: either the first 3, first 6, first 9 or all 11 organs were considered from thymus, spleen, bone marrow, limb muscle, tongue, heart, lung, mammary gland, bladder, kidney and liver. Other data sets were not subsampled as the complexity was lower or controlled.

Each algorithm was run on each data set with default or suggested settings. Suggested settings: for Monocle, densityPeak method was used. For Seurat, Louvain clustering after K-nearest neighbor graph construction was used with ten dimensions from PCA (as in the PBMC3k vignette, which was followed as the recommended Seurat processes). More lenient filtering thresholds from the Tabula Muris organ annotation vignette were used for data sets with fewer cells. For BackSPIN, the number of levels was set to 4, as shown in the documentation.

Rare population benchmark. Rare population detection was determined by the ability of algorithms to separate two known rare populations from each other. Three cell types were considered for one common and two rare cell populations. As T cells were dissimilar from both macrophage and dendritic cells (Supplementary Fig. 19), T cells were chosen as the common population with macrophages and dendritic cells as the rare populations, all from mouse spleen. To benchmark clustering accuracy in separating rare cells, we also performed two additional experiments based on mixings cells from different mouse organs: (1) tongue (common), mammary (rare) and bone marrow (rare); and (2) bladder (common), heart (rare) and tongue (rare). Likewise, for the immune population data set: $CD4^+$ T (common), $CD14^+$ monocytes (rare), $CD19^+$ B (rare) cells were used[23]. There were 100 data sets of 1,000 cells generated by randomly subsampling from each cell type or organ. These 1,000 cells per data set ranged from 900 to 990 common cells and 100 to 10 rare cells (for example, half macrophages and half dendritic cells), with 10 runs each. For instance, the smallest common data set was comprised of 900 common cells (90%) and 100 rare cells (10%, 5% for each rare population). The largest common data set was comprised of 990 common cells (99%) and 10 rare cells (1%, 0.5% for each rare population). All algorithms were run on these data sets with default or suggested settings in the same fashion as in the cluster purity benchmark. These results were visualized using t-SNE for each package (Monocle, reduceDimension with t-SNE method; Phenograph, TSNE from scikit-learn, which is not included in Phenograph; BackSPIN, TSNE from scikit-learn which is not included in BackSPIN; Seurat, RunTSNE with dim.use of ten dimensions; CIDR, Rtsne froRtsne which is not included in CIDR; RaceID, comptsne; and Cell Ranger, output t-SNE projections). UMAP visualization was calculated with the UMAP-learn python package. TooMany output was visualized using BirchBeer trees and given rare population priority with --smart-cutoff 5 --min-distance-search 1.

To quantify these benchmarks, a contingency table of the fraction of pairwise labels was used. For all rare cell pairs, a true pair was called if the two cells were of the same cell type (for example a macrophage with another macrophage or a dendritic cell with another dendritic cell), while a false pair was called if the two cells were of different cell types (for example, a macrophage with a dendritic cell). Then, the measure for accuracy in this benchmark was the fraction of true pairs in all pairs.

For the simulated rare population benchmark, Splatter[58] with default settings was used to generate data sets of 1,000 cells in three groups, identical in composition to the previous subsampled rare population benchmark. Here, TooManyCells was run with --pca 50 (in concordance with Seurat) to account for the synthetic nature of the Splatter model, and --min-modularity –0.05 to accommodate the PCA transformation. BackSPIN, RaceID and Phenograph did not use dimensionality reduction by default, as with TooManyCells, so additional benchmarks were run with dimensionality reduction through the TooManyCells PCA matrix for BackSPIN and Phenograph (which do not have any function for reduction in their libraries), and CCorrect for RaceID.

Timing benchmark. There were 1,000 cells used to benchmark clustering algorithm times in order to accommodate RaceID, CIDR and BackSPIN, which did not finish on larger data sets from the purity benchmark after 4 d. Each algorithm was run 10 times to determine an average runtime.

Distribution-based pruning and stopping criteria. TooManyCells can prune the tree by including a stopping criteria in a variety of ways, including specific nodes, the minimum size of a node (that is, number of cells), and the proportion of cells in each child node. To simultaneously identify both rare and common cell populations, TooManyCells uses modularity to guide the tree pruning. TooManyCells quantifies the distribution of modularity for all non-leaf nodes and chooses a value of modularity on the basis of the specified number of median absolute deviations from the median (or a chosen value). The algorithm preserves all paths to all nodes of this value or greater, and cuts all levels below. This results in large nodes with low modularity in their descendants and small nodes with high modularity.

Clumpiness. The hierarchical structure generated from any hierarchical clustering, both divisive and agglomerative, holds cells in the leaf nodes. Each cell can be assigned a label, such as an organ of origin, cell type or expression level of high or low. In order to quantify the level of aggregation within the tree, a measure of

Table 1 | B-cell-subtype markers

Cell type	Genes
Plasma cell	*Igj*
Germinal center cell	Classified using ImmGen
Follicular cell	*Fcer2a, Klf2*
Marginal zone cell	*Tcf4, Crebl2*
Transitional cell	*Gfi1, Myb, Uhrf1*
Plasmablast	Classified using ImmGen

'clumpiness' is needed[59]. For instance, the degree of how 'clumped', or co-localized, are CD4 T cells and CD8 T cells within the tree. Here, one would expect those T cells to be grouped together more closely than CD4 T cells with B cells. A clumpiness measure enables the quantification of this similarity.

The clumpiness measure used here was specifically designed for hierarchical structures and was previously described in more detail[59]. Briefly, consider a rooted k-ary tree. The clumpiness of the set of leaves M when partitioned according to $L = \{L_1, L_2, \ldots, L_n\}$ is defined as

$$C(L) = \frac{1}{n}\left(\prod_{i=1}^{n}\frac{x}{y_i}\right)^{1/n} \quad (3)$$

This measure takes the geometric mean of x weighted by y_i. x represents the weighted number (weighted by distance to the descendant leaves) of 'viable' non-root inner nodes, and y_i is the frequency of leaves in L_i in all leaves not connected to the root node. Viable nodes are comprised of inner nodes that have at least one vertex of each label in their descendant leaves. The clumpiness of a label L_i with itself is simply considering an L' containing two sets — leaves in L_i and all other leaves. Then the clumpiness of L_i with itself is $1 - C(L')$ (ref. [59]).

Splenic cell markers. Branches of the TooManyCells tree were defined in two ways. First, differential expression analysis was carried out for each node, and the following lineage markers were used to designate enriched cell type in each leaf node. Second, populations listed in Table 1 were classified using ImmGen: the top 100 differential genes in those nodes were used as input to ImmGen MyGeneSet in order to find enrichment for markers from the designated cell type[30].

Lineage-specific transcription factors in addition to cell surface markers were used, since scRNA-seq cannot differentiate between cytoplasmic and surface expression of markers.

GSI-resistant T-ALL cell culture. DND-41 cells (DSMZ, cat. no. ACC525) were purchased from the Leibniz-Institute DSMZ-German Collection of Microorganisms and Cell Lines. Cells were cultured in RPMI 1,640 (Corning, cat. no. 10-040-CM) supplemented with 10% fetal bovine serum (Thermo Fisher Scientific, cat. no. SH30070.03), 2 mM L-glutamine (Corning, cat#. no. 25-005-CI), 100 U ml^{-1} and 100 µg ml^{-1} penicillin–streptomycin (Corning, cat. no. 30-002-CI), 100 mM nonessential amino acids (Gibco, cat. no. 11140-050), 1 mM sodium pyruvate (Gibco, cat. no. 11360-070) and 0.1 mM of 2-mercaptoethanol (Sigma, cat. no. M6250). All cells were grown at 37 °C and 5% CO_2, with medium refreshed every 3–4 d. Cells were regularly tested for mycoplasma contamination.

IC_{50} values for gamma-secretase inhibitor (GSI) compound E (Calbiochem, cat. no. 565790) were calculated from dose-response curves using CellTiter Glo Luminescent Cell Viability Assay (Promega, cat. no. G7571). Briefly, 1,000 treatment-naive DND-41 cells in 5 replicates per condition were plated in 96-well plates with vehicle or increasing concentrations of GSI (0.016, 0.031, 0.062, 0.125, 0.25, 0.5, 1, 2 µM). Luminescence was measured on day 7 with CellTiter Glo Luminescent Cell Viability Assay according to the manufacturer's instructions. DND-41 IC_{50} of GSI was determined to be 5 nM.

To generate ascending GSI-resistant cells, DND-41 treatment-naive cells were cultured in the presence of 10, 20, 40, 80 and 125 nM GSI, with concentration increasing every week for six weeks and maintained in 125 nM GSI. To generate sustained high-dose GSI-resistant cells, DND-41 treatment-naive cells were cultured in the presence of 125 nM GSI for at least 6 weeks. The establishment of GSI-resistance was determined with IC_{50} assay as described above. Both ascending and sustained high-dose GSI-resistant DND-41 cells can tolerate 10 µM GSI with less than 20% cell death. Short-term DMSO/GSI treatment was performed on treatment-naive DND-41 cells with 125 nM DMSO/GSI for 24 h.

GSI-resistant T-ALL single-cell RNA-sequencing. Prior to single-cell transcriptomic profiling, cells were washed with 1× PBS (Corning, cat. no. 21031CV) and stained with DAPI (Sigma-Aldrich, cat. no. D9542), and live cells were sorted on BD FACS Aria II using 100-µm nozzle. Cells were washed twice with RPMI, counted and single-cell RNA-seq was performed using 10x Genomics Single Cell 3′ Library and Gel Bead Kit v2 (10x Genomics, cat. no. 1000092) following the manufacturer's instruction. Briefly, cells were loaded onto independent channels of a Chromium Controller (10x Genomics) for targeted recovery of 3,000 cells per condition. Complementary DNA was synthesized and amplified with PCR for 13 cycles. Amplified cDNA was assessed for QC and quantified on Agilent TapeStation using High sensitivity D5000 chip and subsequently used for library construction. Libraries were quantified using KAPA Library Quantification Kits for Illumina platform (KAPA Biosystems, Roche, cat. no. KK4824) and pair-end sequenced on NextSeq 550 using 150 cycles High Output kit.

FASTQ file generation and alignment to GRCh38 were performed using Cell Ranger v2.1.1 with default arguments. In total, 10,109 cells passed the Cell Ranger QC and showed the typical 'knee' plots indicating high quality from untreated (2,340), short-term (2,618), ascending (2,734) and sustained high-dose (2,417). These cell were aggregated using Cell Ranger. The fraction of reads in cells was 94.1%. The total number of post-normalization reads was 786,185,264, with mean reads per cell at 66,768 and median genes per cell of 3,333. Multiplets were identified with Scrublet[60] and removed from the Cell Ranger filtered matrix, which was then used as input to TooManyCells or Seurat with default settings.

RNA FISH. Parental DND-41 cells treated with 125 nM DMSO or GSI and sustained GSI-resistant cells were harvested and resuspended in PBS at a concentration of 4.5×10^6 cells ml^{-1}. For each conditions, 80 µl of the cells were added to the same polysine microscope slide (Thermo Scientific, cat. no. P4981) using silicone isolators (Electron Microscopy Sciences, cat. no. 7033905) and adhered to the slide for 30 min at room temperature in a humidified chamber. Cells were then fixed in 4% formaldehyde (Fisher Scientific, cat. no. PI28908) in 1× PBS for 10 min, and then dipped in 1× PBS. Cells were permeabilized in 0.5% Triton (Sigma-Aldrich Roche, cat. no. 10789704001) in 1× PBS for 15 min and dehydrated with an ethanol row of 70%, 80% and 100% ethanol for 2 min each. Cells were washed in wash buffer containing 2× SSC, 10% formamide (Thermo Fisher, cat. no. 3442061L), in Nuclease-free water (Ambion, cat. no. AM9937) to remove remaining ethanol. 50 µl of hybridization mix (10% dextran sulfate, 10% formamide, 2× SSC) and 1 µl of RNA FISH probes against *MYC* (Alexa594) and *GAPDH* (Alexa 647) (gift from A. Raj) were added to a 24 × 50 mm coverslip, attached to the slide and sealed with no-wrinkle rubber cement (Elmer's). Hybridization was performed overnight in a 37 °C humidified chamber. Rubber cement was removed, and cells were washed for 30 min in wash buffer. Cells were then stained with 0.1 µg ml^{-1} DAPI in 2× SSC for 15 min in a coplin jar with shaking. Slide was allowed to completely dry before mounting on coverslip with Slowfade Gold Antifade Reagent (Invitrogen, cat. no. S36936) and sealing with transparent nail polish.

Imaging was carried out on a Nikon widefield fluorescent microscope (Nikon Ti-E with a ×60 Plan-Apo objective) and z stack size of 10 µM with a z step size of 330 nM (Nikon Elements software). DAPI signal was used for manual nuclei segmentation and the number of *MYC* or *GAPDH* mRNA in each cell were determined as described in ref. [61] (https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Home). 250, 261, and 222 DMSO-treated parental, GSI-treated parental and sustained resistant cells were analyzed, respectively. The number of *MYC* or *GAPDH* RNA FISH count were compared by t.test in R. Example images of DMSO-treated parental and sustained GSI-resistant cells were selected on the brightest z plane and adjusted in ImageJ such that the brightness of each channel is comparable across the two conditions.

Reporting Summary. Further information on research design is available in the Nature Life Sciences Reporting Summary linked to this article.

Data availability

The accession number for the new data sets reported in this paper is Gene Expression Omnibus: GSE138892. Microfluidics single-cell RNA-seq count data from 11 organs in 3 female and 4 male, C57BL/6 NIA, 3-month-old mice were obtained from https://figshare.com/articles/_/5715025, removing P8 libraries due to outlier cell counts[22]. FACS-purified CD14$^+$ monocytes, CD19$^+$ B and CD4$^+$ T cells were obtained from https://support.10xgenomics.com/single-cell-gene-expression/datasets (ref. [23]). Data for seven cancer-cell lines were obtained from GSE81861 (ref. [17]). FACS-purified B lymphocytes/natural killer, megakaryocyte-erythroid, and granulocyte-monocyte progenitors were obtained from GSE117498 (ref. [25]).

Code availability

TooManyCells is available at https://github.com/faryabib/too-many-cells or as a Docker image https://cloud.docker.com/repository/docker/gregoryschwartz/too-many-cells/. An R wrapper for TooManyCells is available at https://cran.r-project.org/web/packages/TooManyCellsR. BirchBeer is available at https://github.com/faryabib/birch-beer or as a Docker image https://cloud.docker.com/repository/docker/gregoryschwartz/birch-beer. Codes necessary to reproduce the presented analyses are available at https://github.com/faryabib/NatMethods_TooManyCells_analysis.

References

42. Shu, L., Chen, A., Xiong, M. & Meng, W. Efficient spectral neighborhood blocking for entity resolution. In *2011 IEEE 27th International Conference on Data Engineering* 1067–1078 (IEEE, 2011).
43. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 18 (2000).
44. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **28**, 11–21 (1972).
45. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge University Press, 2008).
46. Salton, G., Wong, A. & Yang, C. S. A vector space model for automatic indexing. *Commun. ACM* **18**, 613–620 (1975).
47. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
48. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427 (1973).
49. Schwartz, G. W. & Hershberg, U. Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics. *Phys. Biol.* **10**, 035005 (2013).
50. Schwartz, G. W. & Hershberg, U. Germline amino acid diversity in b cell receptors is a good predictor of somatic selection pressures. *Front. Immunol.* **4**, 357 (2013).
51. Meng, W. et al. An atlas of b-cell clonal distribution in the human body. *Nat. Biotechnol.* **35**, 879–884 (2017).
52. Heck, K. L., van Belle, G. & Simberloff, D. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* **56**, 1459 (1975).
53. Tian, L. et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* **16**, 479–487 (2019).
54. Ronen, J. & Akalin, A. netSmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res* **7**, 8 (2018).
55. Dai, H., Li, L., Zeng, T. & Chen, L. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.* **47**, e62 (2019).
56. Tan, P.-N., Steinbach, M., Karpatne, A. & Kumar, V. *Introduction to Data Mining* 2nd edn (Pearson, 2019).
57. Kvålseth, T. O. On normalized mutual information: measure derivations and properties. *Entropy* **19**, 631 (2017).
58. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
59. Schwartz, G. W., Shokoufandeh, A., Ontañón, S. & Hershberg, U. Using a novel clumpiness measure to unite data with metadata: finding common sequence patterns in immune receptor germline v genes. *Pattern Recognit. Lett.* **74**, 24–29 (2016).
60. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019).
61. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).

Acknowledgements

This work was supported by T32-CA-009140 (to G.W.S.), LLS-5456-17 (to J.P.), R01-CA-215518 (to W.S.P.), R01-HL-145754, the Penn Epigenetics pilot award and the Sloan Foundation Grant (to G.V.), Therapeutics Translational Medicine and Therapeutics program for Transdisciplinary Awards Program in Translational Medicine and Therapeutics, Concern Foundation's The Conquer Cancer Now Award, Susan G. Komen CCR185472448 and R01-CA-230800 (to R.B.F.).

Author contributions

Conceptualization: R.B.F., G.W.S.; Methodology: G.W.S., R.B.F.; Software: G.W.S.; Investigation: G.W.S., R.B.F., J.P., Y.Z.; Formal Analysis: G.W.S., R.B.F., J.P., M.F., S.M.S., L.X., Y.Z.; Resources and Reagents: R.B.F., G.V.; Writing, Review and Editing: G.W.S., R.B.F., W.S.P., J.P., Y.Z.; Writing, Original Draft: G.W.S., R.B.F.; Supervision: R.B.F.; Funding Acquisition: R.B.F.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0748-5.

Correspondence and requests for materials should be addressed to R.B.F.

Peer review information Nicole Rusk and Lin Tang were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Robert B. Faryabi

Last updated by author(s): 11/29/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Complete code and documentation for the software suite developed in this study (TooManyCells tool) is available on GitHub under the following weblink: https://github.com/faryabib/too-many-cells. Scripts corresponding to the analyses contained in this paper are further provided at: https://github.com/faryabib/NatMethods_TooManyCells_analysis. The following programs were used for data collection: Cell Ranger: 2.1.1 Nikon NIS Elements AR 5.20.0 64-bit
Data analysis	Complete code and documentation for the software suite developed in this study (TooManyCells tool) is available on GitHub under the following weblink: https://github.com/faryabib/too-many-cells. Scripts corresponding to the analyses contained in this paper are further provided at: https://github.com/faryabib/NatMethods_TooManyCells_analysis. The following programs were used for data analysis: Seurat: 2.3.4 Monocle: 2 Cell Ranger: 2.1.1 Phenograph: 1.5.2 edgeR: 3.18.1 RaceID: 0.1.3 CIDR: 0.1.5 BackSPIN: 0.2.1 Rtsne: 0.15 umap-learn: 0.3.9 TooManyCells: 0.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

GSI-resistant T-ALL data were deposited in GSE138892.
11 C57BL/6 NIA mouse organs data were obtained from https://figshare.com/articles/_/5715025.
FACS-purified CD14+ monocytes, CD19+ B, and CD4+ T cells data were obtained from https://support.10xgenomics.com/single-cell-gene-expression/datasets.
Seven cancer cell lines data were obtained from GSE81861.
FACS-purified B lymphocytes/natural killer, megakaryocyte-erythroid, and granulocyte–monocyte progenitors data were obtained from GSE117498.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen based on the sample size in the publicly available datasets. The sample size (number of cells) varies from 200 to 4000, which covers the range of the sample sizes of most single cell RNA-seq datasets.
Data exclusions	All datasets generated in this study were filtered using standard quality thresholds commonly used for scRNA-seq data. All filters used are per-selected and specified in the Methods. P8 libraries were removed from Tabula Muris data sets due unusually high bar-code counts compared to other samples.
Replication	We performed scRNA-seq on one replicate of each condition as the single cells themselves can already serve as technical replicate. We also performed bulk RNA-seq on the same conditions and saw strong correlation between the bulk and single-cell results for the same conditions. 1,000 treatment-naïve DND-41 cells in 5 replicates/condition were plated in 96-well plates with vehicle or increasing GSI concentrations to measure cellular viability. All attempts at replications were successful.
Randomization	Randomization was used if applicable. In general, randomization was achieved by using several publicly available data sets. For the analysis presented in Figure 3, different number of organs were selected. For the analysis reported in Figure4, 1000 controlled admixtures were generated for each spike-in ratio.
Blinding	Blinding was based on the publicly available datasets. For the GSI resistant data set, blinding is not applicable because treatment condition and responsiveness defines the cell phenotypes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	DND-41 cells (DSMZ, cat# ACC525) were purchased from the Leibniz-Institute DSMZ-German Collection of Microorganisms and Cell Lines
Authentication	Short tandem repeats (STRs) authentication was used three times during GSI-resistant cell selection.
Mycoplasma contamination	Cells were regularly tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line was used.