LINEAGE: Label-free identification of endogenous informative single-cell mitochondrial RNA mutation for lineage analysis

Li Lin[a,1], Yufeng Zhang[b,1], Weizhou Qian[a,1], Yao Liu[c], Yingkun Zhang[a], Fanghe Lin[a], Cenxi Liu[b], Guangxing Lu[b], Di Sun[d], Xiaoxu Guo[a], YanLing Song[a], Jia Song[d,2], Chaoyong Yang[a,d,2], and Jin Li[b,e,2]

[a]State Key Laboratory for Physical Chemistry of Solid Surfaces, Key Laboratory for Chemical Biology of Fujian Province, Key Laboratory of Analytical Chemistry, and Department of Chemical Biology, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, People's Republic of China; [b]State Key Laboratory of Genetic Engineering and School of Life Sciences, Fudan University, Shanghai 200433, China; [c]Department of Endocrinology and Metabolism, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai 200072, China; [d]Institute of Molecular Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China; and [e]Institute of Cell Biology and Biophysics, Leibniz University Hannover, Hanover 30167, Germany

Edited by David Weitz, Department of Physics, Division of Engineering and Applied Science, Harvard University, Cambridge, MA; received November 3, 2021; accepted December 20, 2021

Single-cell RNA-sequencing (scRNA-seq) has become a powerful tool for biomedical research by providing a variety of valuable information with the advancement of computational tools. Lineage analysis based on scRNA-seq provides key insights into the fate of individual cells in various systems. However, such analysis is limited by several technical challenges. On top of the considerable computational expertise and resources, these analyses also require specific types of matching data such as exogenous barcode information or bulk assay for transposase-accessible chromatin with high throughput sequencing (ATAC-seq) data. To overcome these technical challenges, we developed a user-friendly computational algorithm called "LINEAGE" (label-free identification of endogenous informative single-cell mitochondrial RNA mutation for lineage analysis). Aiming to screen out endogenous markers of lineage located on mitochondrial reads from label-free scRNA-seq data to conduct lineage inference, LINEAGE integrates a marker selection strategy by feature subspace separation and de novo "low cross-entropy subspaces" identification. In this process, the mutation type and subspace–subspace "cross-entropy" of features were both taken into consideration. LINEAGE outperformed three other methods, which were designed for similar tasks as testified with two standard datasets in terms of biological accuracy and computational efficiency. Applied on a label-free scRNA-seq dataset of BRAF-mutated cancer cells, LINEAGE also revealed genes that contribute to BRAF inhibitor resistance. LINEAGE removes most of the technical hurdles of lineage analysis, which will remarkably accelerate the discovery of the important genes or cell-lineage clusters from scRNA-seq data.

single-cell RNA-seq | lineage analysis | BRAF inhibitor resistance

Lineage analysis is an important assay for developmental biology, cancer biology, etc. Classical lineage analysis in developmental biology study is hypothesis driven and relies on the "pulse-chase" model with an inducible Cre-LoxP system (1, 2). This system labels the progenitors permanently so the source of a mature cell type can be identified via lineage tracing. Lineage analysis also can be used to identify the genes correlated to the clonal evolution of cancer cells upon treatment, which can be done on either primary cancer samples or cancer cell lines based on the somatic mutations. Traditionally, lineage analysis is time consuming, technically challenging, and in demand of much preexisting knowledge. A tool is desired to simplify the lineage analysis on complex tissues.

Single-cell RNA-sequencing (scRNA-seq) has become a powerful tool for biomedical research (3–6). Initially employed as an assay to identify clusters of cells with distinct transcriptomic features, scRNA-seq data now can provide additional information with the advancement of computational tools (7–10). Because scRNA-seq can profile many cell types simultaneously, potentially, it is a useful tool for lineage analysis. However, this application normally relies on exogenous barcodes created by transforming barcode libraries (11) or Cas9-based genome-editing tools (12). These cellular barcodes therefore are used as reference for clonal clustering. Regardless of the complexity of performing such experiments, it is impossible to directly analyze the lineage information of clinical samples with this strategy. We therefore aim to simplify the lineage analysis by developing a user-friendly computational algorithm based on endogenous markers of scRNA-seq data.

The whole-genome RNA single nucleotide polymorphisms (SNP) has been used as endogenous markers for lineage study (13), though the requirement of high sequencing coverage limits its application. In comparison, the size of mitochondrial genome is relatively small. Thus, mitochondrial RNA variant is a great

Significance

Lineage analysis is an important assay for developmental biology, cancer biology, etc. Traditional tools in this field are time consuming, technically challenging, and in demand of preexisting knowledge. By integrating exogenous barcodes into cells, single-cell RNA-sequencing (scRNA-seq) can be used to conduct such tasks, but these assays required significant expertise in both wet- and dry-laboratory experiments. We developed a user-friendly algorithm to conduct cell-lineage inference solely based on endogenous markers of label-free scRNA-seq. This algorithm is able to identify lineage-informative mutations from a bunch of interfering mitochondrial RNA variants with high accuracy and efficiency. With this algorithm, we removed most of the technical hurdles of lineage analysis on scRNA-seq and will dramatically accelerate its application in biological research.

Author contributions: L.L., Yufeng Zhang, W.Q., J.S., C.Y., and J.L. designed research; L.L., Yufeng Zhang, W.Q., Y.L., Yingkun Zhang., F.L., C.L., G.L., D.S., X.G., Y.S., J.S., C.Y., and J.L. performed research; L.L., Yufeng Zhang, W.Q., J.S., C.Y., and J.L. analyzed data; and L.L., Yufeng Zhang, W.Q., Y.L., Yingkun Zhang, F.L., C.L., G.L., D.S., X.G., Y.S., J.S., C.Y., and J.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]L.L., Yufeng Zhang, and W.Q. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: li_jin_lifescience@fudan.edu.cn, songjiajia2010@shsmu.edu.cn, or cyyang@xmu.edu.cn.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2119767119/-/DCSupplemental.

Published January 27, 2022.

resource as endogenous makers. We have shown that it is possible to perform lineage analysis with mitochondrial RNA variants based on scRNA-seq data on human islet (14). Unfortunately, the challenge for lineage analysis with mitochondrial RNA variants is to identify the informative variants (mutations) for following inference, without knowing the features of clones in the cells. Ludwig et al. (15) tried to solve this problem by analyzing bulk assay for transposase-accessible chromatin with high throughput sequencing (ATAC-seq) data, which can provide reliable and informative mitochondria genome variants as reference, from the same sample at first. Clustering is then performed on scRNA-seq data based on these variants identified from the bulk ATAC-seq data. The requirement of parallel bulk ATAC-seq obviously limits the application of this strategy. A tool solely based on scRNA-seq is more desirable.

The bottleneck of performing de novo and label-free lineage analysis solely based on scRNA-seq is to identify the informative variants for clones, so called clonal features, without preexisting knowledge. The clonal features are rather sparse and more sensitive to sequencing error and coverages in comparison to expression features. Due to their unique properties, the clonal features cannot be identified with the similar strategies as expression features of scRNA-seq data.

Thus, we developed an algorithm to call the clonal features efficiently based on "low cross-entropy subspace" separation and identification. Instead of using whole-genome SNPs as endogenous markers (13), we employed the mitochondrial RNA SNP for the analysis to avoid the limitation of sequencing coverage. The initial feature calling still showed an unignorable level of noise, due to sequencing errors and low coverage. We therefore improved the feature-selection process by integrating mutation type and subspace–subspace cross-entropy into consideration. Cross-entropy is a loss function that can be used to quantify the difference between two probability distributions (16). Based on this concept from information theory, we defined a 'cross-entropy' measure to quantify the difference of the embedded cluster structures among subspaces. The "low cross-entropy subspaces" discovered in this strategy could well capture lineage information from scRNA-seq. Meanwhile, we also integrated an optimized consensus-clustering process with a refinement step to further fully capture and refine the lineage structure from the "low cross-entropy subspaces."

This computational algorithm, called label-free identification of endogenous informative single-ccll mitochondrial RNA mutation for lineage analysis (LINEAGE), is one for de novo label-free lineage analysis of scRNA-seq based on endogenous lineage markers selection. We applied LINEAGE on a label-free BRAF inhibitor resistance study to identify and validate the genes associated to the resistance in melanoma. We expect the application of LINEAGE to dramatically accelerate lineage analysis–related studies.

Results

Working Principle of the Lineage Analysis in LINEAGE. Full-length scRNA-seq data generated by Smart-seq2 (17) protocol was used for lineage analysis. The mitochondrial RNA variants were called and the allele frequency ($AF_{x,b}$) was calculated to produce the variants frequency matrix (Fig. 1*A*; our preprocessing codes can be downloaded at https://github.com/songjiajia2018/ppl).

Due to its nature, variant frequency has many more noises than gene expression. These noises may come from many aspects of data including sequencing errors and low coverage. Therefore, variant-related analysis (18–22) normally needs bulk sequencing data with high coverage. However, scRNA-seq data are highly sparse, which makes it difficult to do feature selection. Here, we performed feature selection on the variant frequency from cells containing distinct clones with Seruat version 3.0/version 4.0 (8, 23) and Entropy subspace separation-based clustering for noise reduction (ENCORE). Unfortunately, neither of these computational algorithms managed to identify features with clonal information (*SI Appendix*, Fig. S1). These results suggested that traditional feature-selection strategies, which were designed for screening expression features, need to be revised for the analysis of variants frequency.

To address this point, LINEAGE (https://github.com/songjiajia2018/LINEAGE) developed a feature-selection strategy to efficiently pick out lineage-informative variants (defined as clonal features) from scRNA-seq datasets. Firstly, the variant-frequency matrix is separated into 12 submatrices according to the mutation types, and highly variable sites are discovered in each submatrix. A merged-frequency matrix with 12*20 highly variable variants is generated. This process guarantees that the initial selected highly variable features contain variants with different mutation types. So, the downstream analysis may avoid being misled by mutational type–specific systemic noises such as sequencing errors. Then, we applied the same hypothesis as ENCORE that features with similar dynamic pattern tend to capture similar cell cluster structures. Thus, subspace separation was performed on the merged matrix to generate 20 subspaces based on the dynamic patterns of variants frequency. In this way, variants with similar frequency dynamic patterns would be clustered into the same subspace, and cluster signals resulted from different events (noise, lineage, or other events) tend to be clearly separated. Then to find out informative feature subspaces for lineage inference, LINEAGE used a method to define the "cross-entropy" among subspaces and pick out "low cross-entropy subspaces" as informative subspaces. In detail, the consensus status among subspaces is indicated by "cross-entropy," which is defined based on the adjusted rand index (ARI, detailed in *SI Appendix*, Supplementary Note 1). This is based on the hypothesis that cluster structures with more consensus information among subspaces are more likely generated by informative events as lineage structures. By default, six subspaces with lowest cross-entropy among subspaces are selected for downstream analysis (*Materials and Methods* and Fig. 1*B*).

Then LINEAGE performed consensus clustering to get the initial clusters with clonal information based on these "low cross-entropy subspaces." Candidate endogenous markers were subsequently identified for each candidate cluster. To improve the accuracy, LINEAGE applied a refinement process by refining the distance between cells based on these marker variants (Fig. 1*C*). The final result of clonal identification was presented as t-distributed stochastic neighbor embedding (t-SNE)/Uniform Manifold Approximation and Projection (UMAP) plot as well as heatmap. The details of variant-frequency matrix generation, subspace separation, low cross-entropy subspace selection, and consensus clustering are described in *Materials and Methods*.

LINEAGE Identified Clones from ScRNA-Seq Data in a De Novo and Label-Free Fashion. We firstly tested LINEAGE on a simulated dataset, which consists of two human melanoma cell lines, A375 and 451Lu (24) (dataset description is detailed in *SI Appendix*, Table S1). LINEAGE can separate the cells from different cell lines accurately (*SI Appendix*, Fig. S2 *A* and *B*). To further test the performance of LINEAGE on datasets with cells from close lineages, we tested LINEAGE on a scRNA-seq dataset (named as TF1 clones) containing 70 cells with exogenous barcoding from three clones and a more-complicated scRNA-seq dataset (named as TF1 barcoding) containing 158 cells with exogenous barcoding from 11 clones (15) (dataset description is detailed in *SI Appendix*, Table S1). In both cases, data from six subspaces were selected. We found that the lineage information of different clones was captured in different subspaces (Fig. 2). In comparison, the unselected subspaces

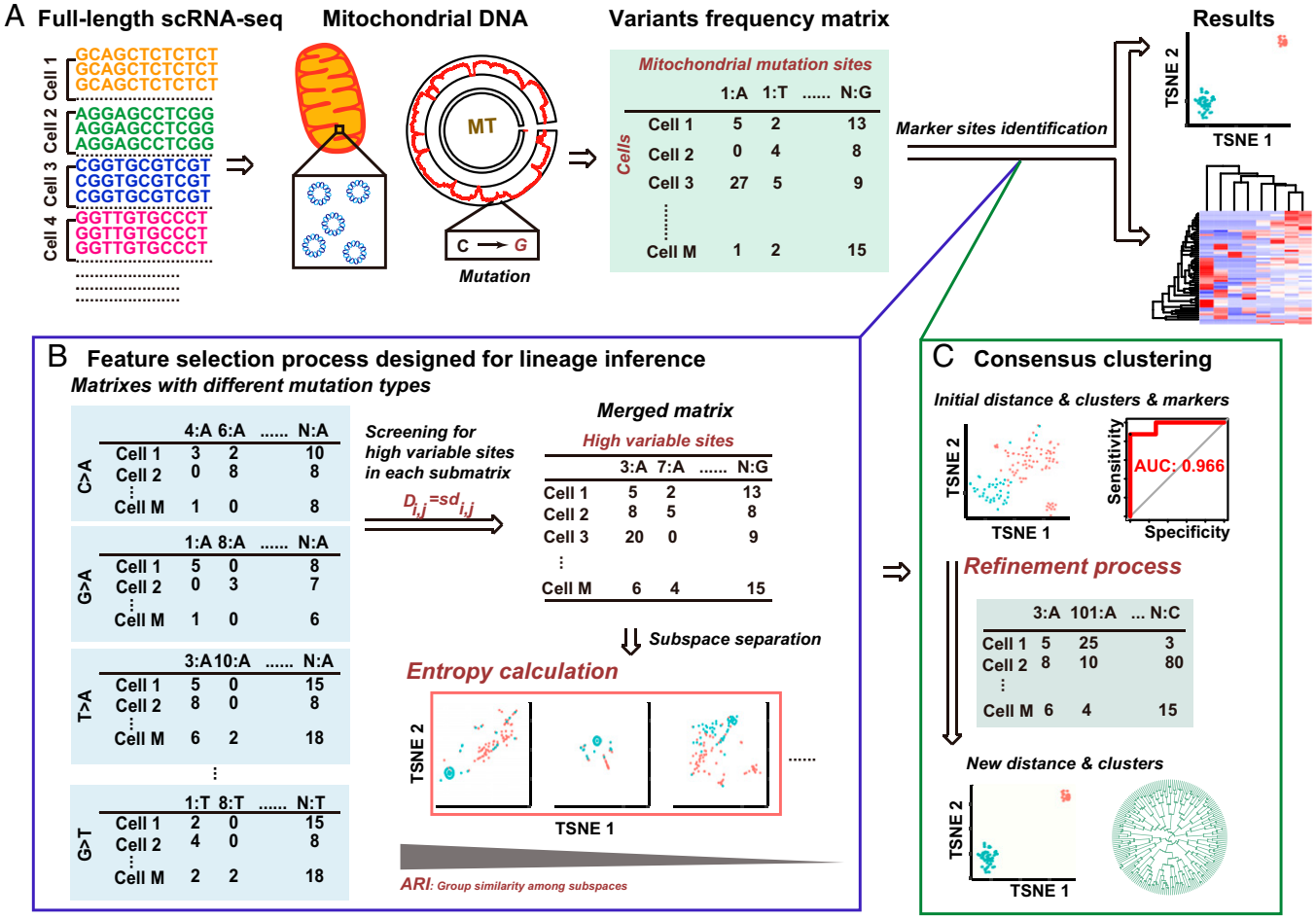

Fig. 1. A schematic representation of LINEAGE. (*A*) The whole analysis process of LINEAGE. Using full-length scRNA-seq dataset as input, mitochondrial RNA variants are called and the variant-frequency matrix is generated for lineage inference. (*B*) Feature selection. LINEAGE firstly screens highly variable variants across cells with different mutation types and then separates the merged highly variable variant-frequency matrix into subspaces according to their dynamic frequency patterns across cells. Subspace–subspace cross-entropy calculation is then conducted based on ARI calculation among clusters from different subspaces to find out the "low cross-entropy subspaces," which show higher consensus among subspaces than other subspaces. (*C*) Consensus clustering. LINEAGE learns a strong, informative similarity matrix by using similarity and cell group information from selected low–cross-entropy subspaces. LINEAGE then applies the learned similarity for initial cell-clustering and group marker identification. The group markers are then used as lineage-related mutations to refine the inference.

(high–cross-entropy subspaces) showed little lineage structure (*SI Appendix*, Fig. S3).

The capability of accurate lineage inference was compared among LINEAGE and three other methods, including Ludwig et al. (15) (mitochondrial SNPs from both bulk ATAC-seq + scRNA-seq), trajectory inference based on SNP information (TBSP) (13) (whole-genome RNA SNPs + expression), and Seurat version 3 (23) (gene expression). In the dataset containing 70 cells/three clones, LINEAGE outperformed the other methods as correctly sorting all the cells into corresponding clones (Fig. 3*A* and *SI Appendix*, Fig. S4*A*). For the dataset containing 158 cells/11 clones, LINEAGE identified clones with comparable accuracy as Ludwig et al. (15) and much-higher accuracy than the other two methods (Fig. 3*B* and *SI Appendix*, Fig. S4*B*). Especially, there is a large overlap of the markers selected by Ludwig et al. (15) and LINEAGE (*SI Appendix*, Fig. S5 and Table S2), although LINEAGE does not require high-depth bulk ATAC-seq data from same samples.

The performance of the four methods was also quantified by the Nearest Neighbor Error (NNE) (25), which represents the error neighbor relationships among cells captured by each method (Fig. 3*C*). The calculation process of NNE is detailed in *SI Appendix*, Supplementary Note 2. LINEAGE performed as good as Ludwig et al. (15) as their NNE scores were comparable. The running time and the required input data of these four methods are shown in *SI Appendix*, Table S3. Obviously, LINEAGE has the higher computational efficiency than other variant-based methods. In general, LINEAGE can perform lineage analysis well on label-free scRNA-seq solely, without the requirement of preexisting bulk ATAC-seq data and exogenous barcodes.

LINEAGE Reveals Transcriptomic Features of BRAF Inhibitor–Resistant Clones in Cancer Cells with BRAF V600E Mutation. We then applied LINEAGE on a scRNA-seq dataset of BRAF V600E mutated melanoma cells 451Lu, which contains parental cells and BRAF inhibitor–resistant cells (24). By performing lineage analysis, we aimed to analyze the clonal evolution with BRAF inhibitor treatment and identify the genes correlated to BRAF inhibitor resistance. Two clusters with distinct clonal features were discovered (Fig. 4 *A* and *B* and *SI Appendix*, Fig. S6), defined as either sensitive Cluster A (the majority are parental cells) or resistant Cluster B (the majority are BRAF inhibitor–resistant cells). The differential distribution of parental and resistant cells in Clusters A and B indicated that the clonal evolution process happened in the selection process. Gene-expression comparison was performed between Clusters

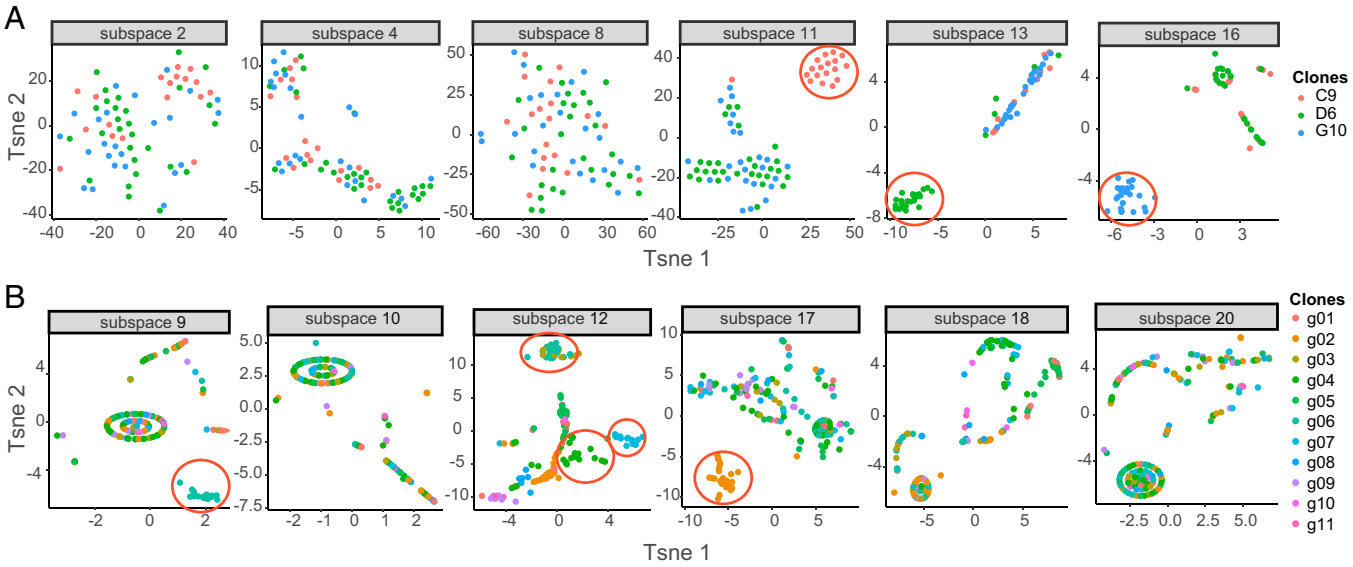

Fig. 2. Low–cross-entropy subspaces selected by LINEAGE. (*A*) Low–cross-entropy subspaces of a scRNA-seq dataset (TF1 clones) containing 70 cells with exogenous barcoding from three clones. The different clones are labeled in different colors. The distinctive clone groups in the subspaces are circled in red. (*B*) Low–cross-entropy subspaces of a scRNA-seq dataset (TF1 barcoding) containing 158 cells with exogenous barcoding from 11 clones. The different clones are labeled in different colors. The distinctive clone groups in the subspaces are circled in red.

A and B. In total, 64 significantly changed genes were found (*SI Appendix*, Figs. S7–S9; method is detailed in *SI Appendix*, Supplementary Note 3). The originally reported resistant gene *DCT* showed elevation in both sensitive and resistant clones, indicating that it may not be directly correlated to the clonal evolution (Fig. 4*C*).

Many of these genes were enriched in gene ontology (GO) term "GO_CC: MITOCHONDRION" with Gene Set Enrichment Analysis (Fig. 4*D*) (26). Since the connection between mitochondria and BRAF inhibitor resistance has been widely observed (27, 28), this result actually validated the function of LINEAGE. We then explored the detailed mechanism of the mitochondria–BRAF connection by focusing on the top differentially expressed gene *GSTP1* (Fig. 4*C*), which encodes an important redox regulator glutathione-S-transferase (GST). It is known that GST inhibitor can regulate the pigment generation in melanocytes (29). It has also been reported that the disruption of redox balance, which is an important function of GST, may conquer the resistance to BRAF inhibitor in melanoma cells (30). In consistence, combinational treatment of BRAF inhibitor Vemurafenib and GST inhibitor GSTO-IN-2 (31) synergistically decreased the cell viability of two melanoma cell lines (Fig. 4*E*). GST can be a target to induce synthetic lethality in BRAF V600E mutated cancer cells.

Discussion

Many lineage analysis–related studies, such as carcinogenesis studies, cancer resistance studies (32), or even developmental biology studies (33), have used scRNA-seq to understand the detailed mechanism. However, the requirement of exogenous barcode prevents many scientists from performing such kind of studies. Although several computational algorithms claimed that they can perform lineage analysis without exogenous barcodes, the requirement of preexisting knowledge such as parallel bulk WGS/ATAC-seq data on the same samples is certainly a technical challenge. In addition, the demand of extensive computational expertise and resource is also beyond many laboratories' capability.

We created a "low cross-entropy subspace" separation and consensus clustering–based analysis as LINEAGE. LINEAGE uses informative mitochondrial RNA variants as endogenous markers, which has relatively small size, and its polymorphism is a frequent event across different tissues and ages (34). In comparison to the method from Ludwig et al. (15), LINEAGE simplifies the endogenous markers identification process and can be applied to scRNA-seq studies without preexisting bulk WGS/ATAC-seq data. In comparison to TBSP, LINEAGE requires shorter running time and has largely improved performance.

We tested LINEAGE on a classical clonal evolution study of BRAF inhibitor–resistant melanoma cells. By analyzing the label-free scRNA-seq dataset from this study, we discovered the sensitive and resistant clones. Differential expression analysis identified *GSTP1* as a BRAF V600E mutation resistance–related gene. GST inhibitor can sensitize melanoma cells to BRAF inhibitor treatment. Therefore, it may serve as the target to develop synthetic lethality therapies for BRAF V600E mutated cancer cells if this result can be validated by in vivo experiment.

In summary, LINEAGE removes most of the technical hurdles on performing lineage analysis by a "low cross-entropy subspace" separation and consensus clustering–based analysis. Due to the requirement of sequencing depth to call variants, it is still difficult to perform lineage analysis on scRNA-seq data from 3′/5′-end-directed scRNA-seq technologies. However, with LINEAGE, it is possible to perform lineage analysis on much existing Smart-seq2 data if desired. Biologists can spare their time and energy on answering biological questions instead of establishing complex labeling system or perform intensive computational analysis. The application of LINEAGE may remarkably accelerate the discovery of the important genes or cell clusters in the diverse context of biomedical research.

Materials and Methods

Data Preprocessing.

Read alignment. All scRNA-seq datasets used in this study were obtained from National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) with following accessions: Gene Expression Omnibus (GEO): GSE115218 (TF1_clones_scRNA, TF1_barcoding_scRNA) (15) and GEO: GSE108383 (scRNA-seq of A375 and 451Lu cell lines were used) (24). The reads were aligned to the GRCh38 human genome and its associated annotations (GRCh38.98) using Spliced Transcripts Alignment to a Reference (STAR) version 2.7.1a (35) with default parameters.

Mitochondrial genotype matrix generating. A bam file consisting of mitochondrial DNA records, which were extracted from the alignment result with Samtools (version 1.9) (36), was obtained. The total number of reads aligned

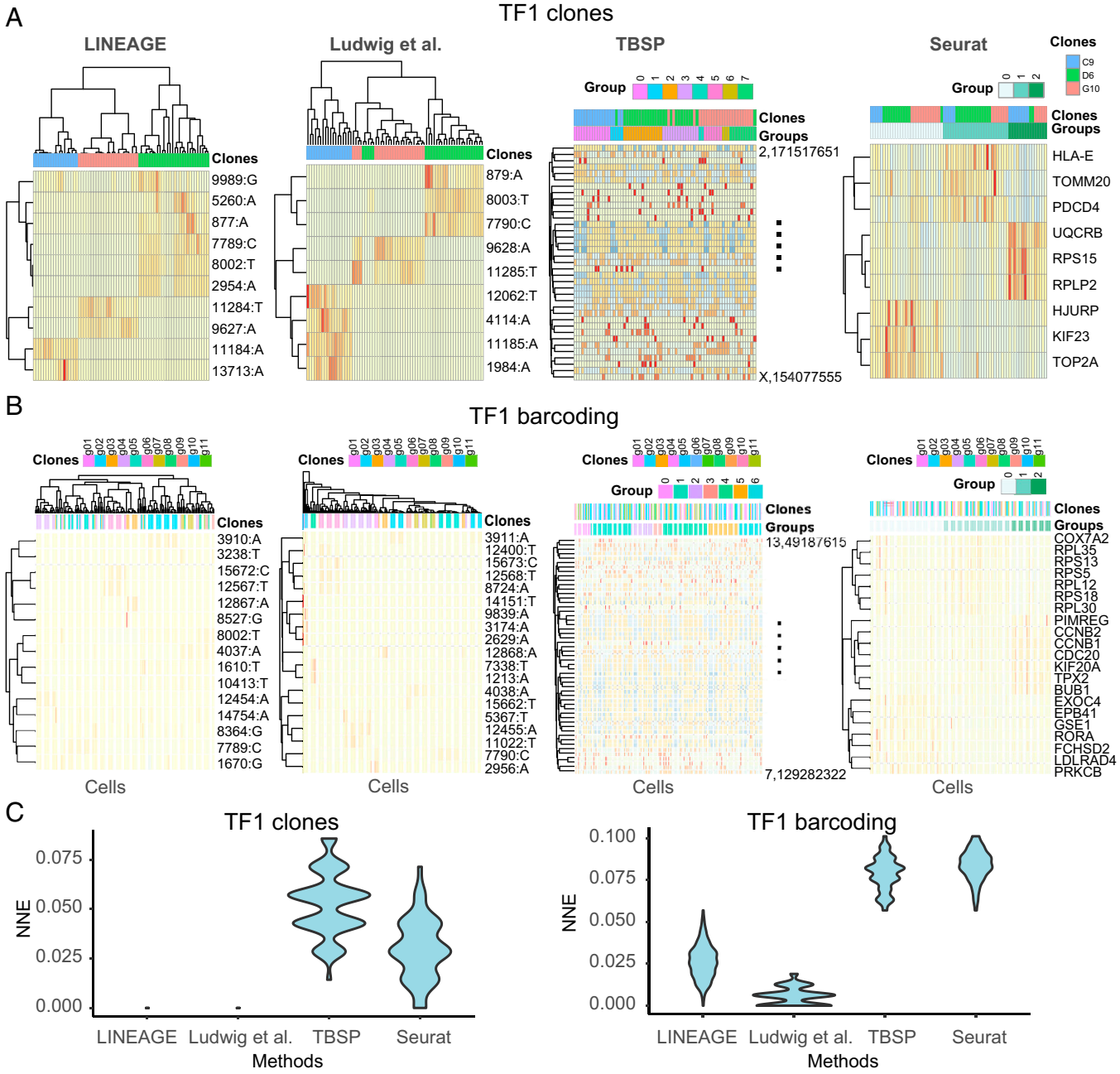

Fig. 3. Performance comparison among four methods. (*A*) Performance comparison on a standard dataset with three clones. Clone information is labeled by "Clones" annotation bars above the heatmap. The cluster groups inferred by TBSP and Seurat are also labeled by the "Group" annotation bars above the heatmap. (*B*) Performance comparison on a standard dataset with 11 clones. Clone information is labeled by "Clones" annotation bars above the heatmap. The cluster groups inferred by TBSP and Seurat are also labeled by the "Group" annotation bars above the heatmap. (*C*) Performance comparison based on NNE, which was inferred with the t-SNE distribution, resulted from all of the four methods.

to per allele on each site of mitochondrial genome were counted using a Python script (15). Here, nucleotides with minimum base quality or minimum read's alignment quality <30 were filtered out. The variant frequency ($AF_{x,b}$) was defined as the following:

$$AF_{x,b} = \frac{R_b}{\sum_{b\in\{A,G,C,T\}} R_b}.$$

R_b is the number of reads holding base b at position x; $\sum_{b\in\{A,G,C,T\}} R_b$ is the total coverage of position ×. Mitochondrial genotype matrix (M), where a column represented a single cell and a row represented variant frequency of a specific mitochondrial genotype, was thus generated.

Highly variable site identification. To screen out highly variable sites with different mutation types, M was split into 12 submatrices (Mi, i = 1, 2, 3 … 12) according to mutation types. The 20 highest-variable sites were then called by identifying the rows with highest SD across cells in each submatrix. Considering the sparsity of the matrix, which heavily affected the SD values, LINEAGE transformed the zeroes into ones when the median of the nonzero frequencies in the same row ≥0.6. Thus, a merged submatrix M_{240} of M was obtained by merging the resulted highly variable sites into a matrix.

Subspace Separation. Hierarchical clustering was carried out to reseparate the M_{240} into 20 submatrices according to the frequency dynamic patterns. In detail, the similarity between pairs of rows in M_{240} was commonly quantified by Pearson's correlation tests. The distance matrix (D_f) was then generated as:

$$D_f = J - S,$$

where J is an all-ones matrix and S is the similarity matrix. Hierarchical clusters were obtained based on this distance matrix, and the feature spaces with

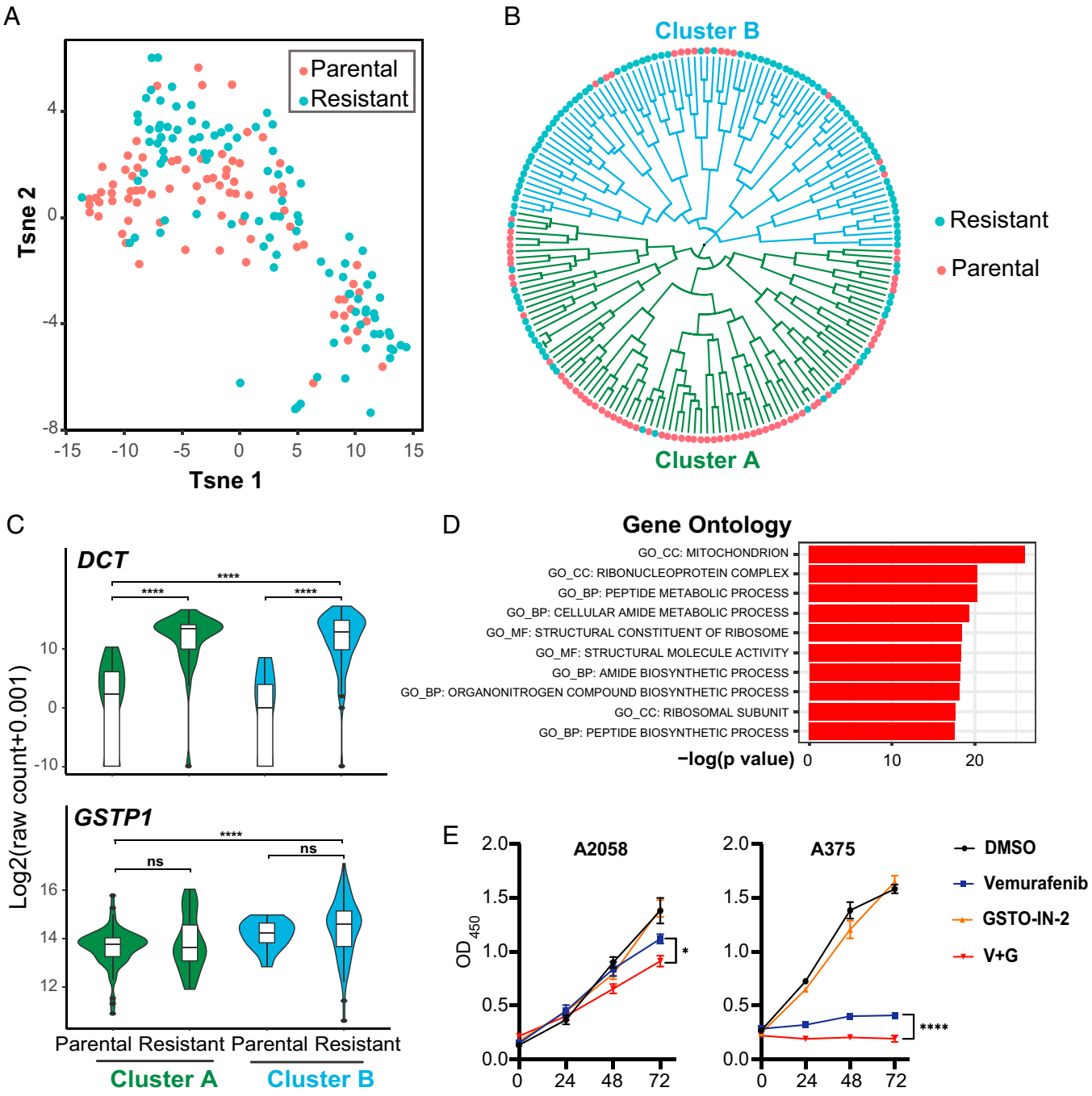

Fig. 4. LINEAGE identified important clonal evolution–related genes from a cancer dataset. (*A*) The lineage analysis result visualized by t-SNE plot from LINEAGE. Cells are labeled according to its BRAF inhibitor resistance status. (*B*) The lineage tree from LINEAGE. Cells are labeled according to its BRAF inhibitor resistance status as well as clonal status. (*C*) The expression levels of *DCT* and *GSTP1* across resistant and clonal status ($P = 0.39$ for "ns" in cluster A, $P = 0.22$ for "ns" in cluster B). (*D*) Gene ontology enrichment results of 64 differential expressed genes. (*E*) Cell viability determination under treatment of BRAF inhibitor Vemurafenib and GST inhibitor GSTO-IN-2 in two melanoma cell lines carrying BRAF V600E mutation (A2058: 4 μM V,1.25 μM G and A375: 8 μM V, 2.5 μM G. *$P < 0.05$, ****$P < 0.0001$).

features (variants) from each resulted cluster were defined as "feature subspaces." In this way, highly variable sites with similar frequency dynamic patterns would be grouped into the same feature subspace; thus, similar clonal cluster signals can be concentrated in same subspace.

Entropy evaluation. In each subspace, LINEAGE used t-SNE followed by k-means clustering to realize cell-clustering (we defined these cluster results as C_{subs}). The k, which indicated the number of clusters, in the k-means clustering process was set to 3 for datasets with cells >100 and otherwise was set to 2. To effectively identify subspaces that might contain clone lineage information, LINEAGE calculated ARI between pairs of subspaces as consensus indicator and got the consensus index matrix as A. LINEAGE defined entropy based on the cell distribution similarity among subspaces, so called "cross-entropy" here. The subspaces with lowest "cross-entropy" (I_i, $i = 1, 2, 3 \ldots 20$), which indicated highest consensus with other subspaces, were selected for subsequent consensus clustering. In this study, six low–cross-entropy subspaces were selected in all cases, and the number of subspaces can be adjusted by parameter. Here, I was defined as:

$$I_i = 1/\max(A_{ij}),\ i, j \in 20 \text{ subspaces},\ i \neq j.$$

Consensus Clustering. Then, LINEAGE generated a combined distance matrix as follows:

$$\tilde{D} = \sum D_i,\ i \in \text{selected subspaces},$$

where D_i is the distance matrix from a selected subspace.

Meanwhile, LINEAGE also characterized consensus information across subspaces by calculating a consensus-factor matrix based on the clustering results (C_{subs}) from the selected subspaces:

$$\tilde{W}_{ij} = \begin{cases} 0, & \text{if cell}_i \text{ and cell}_j \text{ are never located in same cluster} \\ 1, & \text{if cell}_i \text{ and cell}_j \text{ are located in same cluster in one subspace} \\ 2, & \text{if cell}_i \text{ and cell}_j \text{ are located in same cluster in two subspaces} \\ \cdots & \end{cases}.$$

By integrating distance information and consensus information, LINEAGE generated a more-integrative distance matrix D as follows:

$$D = \tilde{D} \cdot \frac{1}{3\tilde{W} + J},$$

where J is an all-ones matrix. Basing on distance matrix D, t-SNE followed by k-means clustering strategy was used to infer the initial consensus cell clone clusters. Here, an adaptive density peak detection algorithm implemented in the ADPclust package in R (37) was integrated to accurately infer the number of clusters.

Marker Variants Identification.

Group marker identification. To screen out the marker variants, LINEAGE transformed the consensus-clustering result into a binary cluster: for a cell group, if a cell is in this group, set its cluster label as 1; if not, set as 0. For each highly variable variant, a receiver operating characteristic (38) curve was built, and the area under the receiver operating characteristic (AUC) score was thus calculated with the frequency distribution as predictor and the binary cluster labels of each group as response. Pearson's correlation coefficient was also calculated between the binary cluster labels and the frequency distribution. Variants with P values <0.05 were considered as cell group markers. These markers were ranked by AUC scores, since higher AUC scores indicated more-reliable markers. Subsequently, 10 to 20 markers with the highest AUC scores were used to refine the consensus-clustering result.

Refinement based on marker variants. After the frequency submatrix consisting of frequencies of markers (M_m) was gotten, cells with zeroes on all markers were removed. LINEAGE integrated both t-SNE and UMAP (39) for dimension reduction and got refined consensus-clustering results based on the dimension-reduction results separately.

Iterative optimization. Considering the randomness from clustering and dimension-reduction processes, an iteration process was implemented in LINEAGE to guarantee a more-stable and -reliable cell-clustering/clone-tracing result. Based on the assumption that real clone clusters always show more-reliable markers and cell cluster information from more-effective subspaces with larger information capacity, a measurement S_{score} was defined for optimization as follows:

$$S_{score} = D_{score} + \frac{2.5 \cdot \sum_{i=1}^{10} AUC_i}{10},$$

where $\sum_{i=1}^{10} AUC_i$ is the sum of the 10 greatest AUC scores of inferred markers. D_{score} is a score calculated in the refinement process. Concretely, LINEAGE calculated ARIs between the refined clustering results (resulted from t-SNE or UMAP-Kmeans procedures) and the clustering results in the selected subspaces (C_{subs}). Subspaces with ARI > 0.1 were recorded as effective subspaces and the number of effective subspaces was labeled as n. To evaluate the information capacity of the consensus results, a D_{score} was defined to reflect the consensus among the refined results and the selected subspaces:

$$D_{score} = n + B_{max} + 2D_{ARI},\ \text{and}\ D_{ARI} = \begin{cases} 1 - 0.5, & \text{if } n = 1 \\ 1 - \overline{A_{subs}}, & \text{if } n > 1 \end{cases},$$

where B_{max} represents the maximum ARI between the refined consensus-clustering result and the clusters in effective subspaces; thus, n+B_{max} indicates the consensus information capacity from effective subspaces. Meanwhile, $\overline{A_{subs}}$ represents the average of a submatrix of A, which consists of ARI values among effective subspaces, and D_{ARI} represents the information overlap status among effective subspaces. Lower D_{ARI} means higher overlap. In this way, consensus result-containing cluster structures from various subspaces with low-overlap cluster information is more preferred.

Among the iteration with same or different parameters, the one with highest S_{score} was reserved as the best result.

Methods Performance Evaluation and Comparison. A simulated dataset was generated by mixing two human melanoma cell lines A375 and 451Lu. LINEAGE processes described as above (*Data Preprocessing*, *Subspace Separation*, *Consensus Clustering*, and *Marker Variants Identification*) were carried out on the simulated and the two standard benchmark datasets. Codes of Seurat version 3 and version 4 (8, 23), ENCORE (40), the method developed by Ludwig et al. (15), and TBSP (13) were downloaded and run according to their manuals on the two benchmark datasets.

Cell Culture. A2058 and A375 were cultured in Dulbecco's modified Eagle medium (Thermo Fisher Scientific) with 10% fetal bovine serum (Yeasen) and 5% penicillin/streptomycin (Gibco) at 37 °C with 5% CO_2. Cell viability was performed with Cell Counting Kit-8 (CCK-8; Yeasen) according to the manufacturer's instructions. Briefly, cells were seeded into 96-well plate at a density of 1,000 cells per well. Cells were treated with different chemical combinations (Vemurafenib, GSTO-IN-2; MCE) and examined at the time point of 0, 24, 48, and 72 hours. At each time point, CCK-8 (10%) was added to the wells, and after an incubation of 1 h at 37 °C, absorbance was measured at 450 nm with a Microplate Reader Infinite F50 (Tecan).

Data Availability. The scripts for mitochondrial genotype matrix preparation and examples are available at GitHub, https://github.com/songjiajia2018/ppl. The scripts for lineage analysis and example data are available at GitHub, https://github.com/songjiajia2018/LINEAGE.

ACKNOWLEDGMENTS. This work was supported by Chinese Ministry of Science and Technology (MOST) 2018YFA0801300 and 2020YFA0803601 and National Natural Science Foundation of China (NSFC) 32071138 to J. L., NSFC 22104080 to J. S., and NSFC 21735004 and 21927806 to C. Y. We thank Vijay G. Sankaran (Broad Institute of MIT and Harvard) for providing their open-source code for performing mitochondrial genotyping.

1. L. He *et al.*, Proliferation tracing reveals regional hepatocyte generation in liver homeostasis and repair. *Science* **371**, eabc4346 (2021).
2. P. Collombat *et al.*, The ectopic expression of Pax4 in the mouse pancreas converts progenitor cells into alpha and subsequently beta cells. *Cell* **138**, 449–462 (2009).
3. Z. Song, A. M. Xiaoli, F. Yang, Regulation and metabolic significance of *de novo* lipogenesis in adipose tissues. *Nutrients* **10**, 1383 (2018).
4. X. Fan *et al.*, Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res.* **28**, 730–745 (2018).
5. X. Han *et al.*, Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107.e17 (2018).
6. S. R. Srivatsan *et al.*, Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).
7. Y. Li, Q. Xu, D. Wu, G. Chen, Exploring additional valuable information from single-cell RNA-seq data. *Front. Cell Dev. Biol.* **8**, 593007 (2020).
8. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
9. W. Gong *et al.*, Benchmarked approaches for reconstruction of in vitro cell lineages and in silico models of C. elegans and M. musculus developmental trees. *Cell Syst.* **12**, 810–826.e4 (2021).
10. D. E. Wagner, A. M. Klein, Lineage tracing meets single-cell omics: Opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
11. C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, A. M. Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
12. J. J. Quinn *et al.*, Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, eabc1944 (2021).
13. J. Ding, C. Lin, Z. Bar-Joseph, Cell lineage inference from SNP and scRNA-seq data. *Nucleic Acids Res.* **47**, e56 (2019).
14. L. Lin *et al.*, Single-cell transcriptome lineage tracing of human pancreatic development identifies distinct developmental trajectories of alpha and beta cells. bioRxiv [Preprint] (2021). 10.1101/2021.01.14.426320 (Accessed 15 January 2021).
15. L. S. Ludwig *et al.*, Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
16. C. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
17. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
18. S. Behjati *et al.*, Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
19. D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, E. Shapiro, Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, e50 (2005).
20. G. D. Evrony *et al.*, Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
21. M. A. Lodato *et al.*, Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
22. S. Becattini *et al.*, T cell immunity. Functional heterogeneity of human memory $CD4^+$ T cell clones primed by pathogens or vaccines. *Science* **347**, 400–406 (2015).

23. T. Stuart *et al.*, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
24. Y. J. Ho *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).
25. B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
26. A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
27. J. Kaplon *et al.*, A key role for mitochondrial gatekeeper pyruvate dehydrogenase in oncogene-induced senescence. *Nature* **498**, 109–112 (2013).
28. M. R. Ruocco *et al.*, Metabolic flexibility in melanoma: A potential therapeutic target. *Semin. Cancer Biol.* **59**, 187–207 (2019).
29. Q. Cong *et al.*, HCV poly U/UC sequence-induced inflammation leads to metabolic disorders in vulvar lichen sclerosis. *Life Sci. Alliance* **4**, e202000906 (2021).
30. B. B. Paudel *et al.*, Disruption of redox balance enhances the effects of BRAF-inhibition in melanoma cells. bioRxiv [Preprint] (2019). 10.1101/818989 (Accessed 28 October 2019).
31. K. H. Chang, L. Lee, J. Chen, W. S. Li, Lithocholic acid analogues, new and potent alpha-2,3-sialyltransferase inhibitors. *Chem. Commun.* (6), 629–631 (2006).
32. C. E. Eyler *et al.*, Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome Biol.* **21**, 174 (2020).
33. C. Weng *et al.*, Single-cell lineage analysis reveals extensive multimodal transcriptional control during directed beta-cell differentiation. *Nat. Metab.* **2**, 1443–1458 (2020).
34. J. Naue *et al.*, Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA. *Mitochondrion* **20**, 82–94 (2015).
35. A. Dobin *et al.*, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
36. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
37. X.-F. Wang, Y. Xu, Fast clustering using adaptive density peak detection. *Stat. Methods Med. Res.* **26**, 2800–2811 (2017).
38. X. Robin *et al.*, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
39. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
40. J. Song *et al.*, Entropy subspace separation-based clustering for noise reduction (ENCORE) of scRNA-seq data. *Nucleic Acids Res.* **49**, e18 (2021).