LCA robustly reveals subtle diversity in large-scale single-cell RNA-seq data

Changde Cheng[1], John Easton[1], Celeste Rosencrance[1], Yan Li[2], Bensheng Ju[1], Justin Williams[1], Heather L Mulder[1], Wenan Chen[1] & Xiang Chen*[1]

Author affiliation:

[1]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

[2]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, USA

*Correspondence should be addressed to Xiang Chen (xiang.chen@stjude.org).

Abstract

Single-cell RNA sequencing has emerged as a powerful tool for characterizing the cell-to-cell variation and dynamics. We present Latent Cellular Analysis (LCA), a machine learning–based analytical pipeline that features a dual-space model search with inference of latent cellular states, control of technical variations, cosine similarity measurement, and spectral clustering. LCA has proved to be robust, accurate, scalable, and powerful in revealing subtle diversity in cell populations.

Keywords

Clustering, Single-cell RNA-seq

Backgrounds

Single-cell RNA sequencing (scRNA-seq) enables the characterization of cell-to-cell variation in transcript abundance, leading to a deep understanding of the diversity of cell types and the dynamics of cell states in populations of thousands to tens of thousands of single cells [1-3]. Although scRNA-seq offers enormous opportunities and has inspired a tremendous explosion of data-analysis methods for identifying heterogeneous subpopulations, significant challenges can arise because of the inherently high noise associated with data sparsity and the ever-increasing number of cells sequenced. The current state-of-the-art algorithms have significant limitations. The biological meaning of the cell-to-cell similarity learned by most machine learning–based tools (such as Seurat [4], Monocle2 [5], SIMLR [6], and SC3 [7]) is difficult to interpret. Several methods require the user to provide an estimation of the number of clusters in the data, and this may not be readily available. Furthermore, most methods have a high computational cost that will be prohibitive for datasets representing large numbers of cells. Lastly, although certain technical biases (e.g., cell-specific library complexity) have been recognized as major confounding factors in scRNA-seq analyses [8], other technical variations (e.g., batch effects and systematic technical variations that are irrelevant to the biological hypothesis being evaluated) have received minimal attention, even though they present major challenges to the analyses [9]. Most methods employ a gene-selection step before clustering analysis, based on the assumption that a small subset of highly variable genes is most informative for revealing cellular diversity. Although this assumption may be valid in many scenarios, it fails when informative genes are not

most variable. This can happen when the difference among subpopulations is subtle, or there is a strong batch effect, while most variable genes differ by batch.

To address these issues, we developed Latent Cellular Analysis (LCA), an accurate, robust, and scalable computational pipeline that will facilitate a deep understanding of the transcriptomic states and dynamics of single cells in large-scale scRNA-seq datasets. LCA makes a robust inference of the number of populations directly from the data (a user can specify this with a priori information), rigorously models the contributions from potentially confounding factors, generates a biologically interpretable characterization of the cellular states, and better reveals the underlying population structures. Furthermore, LCA addresses the scalability problem by learning a model from a subset of the sample, after which a theoretical scheme is used to assign the remaining cells to identified populations.

Results

Overview of the LCA method

LCA employs a heuristic dual-space search approach (**Fig. 1**) to provide an optimal model for subpopulation structure identification (including the removal of confounding factors, the inference of the number of clusters and informative states, and a mapping function from the expression vector to cluster membership). In the primary latent cellular (LC) space, LCA bypasses gene selection and performs LC state inference from the global gene expression matrix, removes cellular states representing known confounding factors (e.g., cell-specific library complexity and batch information), and measures cell-

70 to-cell similarity by using the cosine of the angle between the low-dimensional cellular-
71 state vectors. Spectral clustering is employed to derive a set of candidate clustering
72 models with a range of cluster numbers. Meanwhile, in the dual principle component
73 (PC) space, cells are projected into a low-dimensional space based on principal
74 component analysis (PCA) of the expression vector. The list of significant components
75 is determined by the Tracy-Widom test [10], and cell-to-cell similarity is measured by
76 correlation similarity in the PC space. LCA then ranks candidate clustering models
77 derived in the LC space by the silhouette index [11] measured in the PC space. Finally,
78 LCA retains informative cellular states from the selected clustering model(s) and uses
79 these states to update the final clustering solution(s). Although the "optimal model" is
80 not necessarily the solution with the best numerical score, it should be among the top-
81 ranked models. Therefore, LCA provides users with multiple top-ranked solutions for
82 biology-based evaluations.
83
84 **LCA accurately identifies cellular states in simulated datasets**
85 We benchmarked LCA with four commonly used state-of-the-art scRNA-seq clustering
86 algorithms (SC3 [7], Seurat [4], Monocle2 [5], and SIMLR [6]) in 3000 simulated
87 datasets. LCA outperformed competing methods in subpopulation number inference,
88 cluster membership inference, and running time.
89
90 We simulated three equally spaced cell subpopulations of various sizes by shuffling a
91 predetermined number of genes across cells from high-quality, Unique Molecular
92 Identifier (UMI)-based scRNA-seq data. We simulated seven scenarios with increasing

signal strength, varying the number of simulated differentially expressed genes (DEGs), the fold change among groups, and the average expression levels of the simulated DEGs. We also investigated the effect of the sample size (from 250 to 4000 cells) on the performance of the algorithms.

A typical clustering analysis partitions data points into meaningful subgroups without a priori knowledge of the number of subsets or their compositions. An essential component of clustering analysis is inferring the number of clusters from the data [12], which presents an analytical challenge distinct from that posed by clustering with a given number of clusters. Therefore, we first analyzed the accuracy and robustness of the subpopulation number inference. Monocle2 and SIMLR were excluded from this analysis because both algorithms require an input of the number of clusters from end-users. **Fig. 2A** shows that although LCA started with a slightly off-target estimation of the cluster number in datasets with very weak signal strength, it stably converged to the correct number with increasing signal strength from the number of cells, the number of DEGs, the fold difference, or the expression level. In contrast, using SC3 or Seurat resulted in an upward bias in the estimation of the number of clusters, and this bias increased with both the sample size and the signal strength. For example, in simulated model #6, with the strongest signal strength, the actual cluster number remained the same (three) for all sample sizes, but SC3 estimated medians of four, eight, and sixteen clusters in datasets with 250, 1000, and 4000 cells, respectively. Seurat estimated medians of three, four, and nine subpopulations in the same respective datasets. Combining the results of all simulations, we find that LCA arrived at the correct number

of clusters in 70.2% of simulations, which is a significantly higher percentage than was achieved with SC3 (5.6%, $P < 2.2 \times 10^{-16}$ [proportion test]) and Seurat (22.7%, $P < 2.2 \times 10^{-16}$ [proportion test]).

We used the variation of information (VI [13], 0 for perfect matching) (**Fig. 2B**) and normalized mutual information (NMI [14], 1 for perfect matching) (**Fig. 2C**) to measure the clustering accuracy. For methods with a feature for inferring the number of clusters from the data (LCA, SC3, and Seurat), we compared the ground-truth labels with inferred cluster membership from the optimal clustering models. For methods without this inference feature (Monocle2 and SIMLR), the optimal models were derived with the correct number of clusters (three). The comparison showed that LCA outperformed other methods in most simulations (**Fig. 2**). When all 3000 simulation datasets were combined, LCA achieved the highest overall accuracy (average NMI = 0.756; VI = 0.555). Specifically, when LCA correctly inferred the number of clusters (in 70.2% of all simulations), it produced near-perfect accuracy (average NMI = 0.941, VI = 0.122). The accuracies of the competing algorithms were consistently lower. Notably, SC3 achieved a relatively good performance (NMI = 0.807, VI = 0.389) when it happened to find the correct number of clusters (mostly in small datasets). However, its performance deteriorated for other datasets, primarily due to an overestimation of cluster numbers. The overall accuracy of SC3 was significantly lower than that of LCA (NMI = 0.583, $P < 2.2 \times 10^{-16}$; VI = 1.203, $P < 2.2 \times 10^{-16}$ [Wilcoxon signed-rank test]). More strikingly, even if provided with the correct cluster number, Monocle2 attained high accuracy only for the dataset with the strongest signal strength (simulated model #6), and its

139 performance was suboptimal compared to that of LCA (NMI = 0.190, $P < 2.2 \times 10^{-16}$;
140 VI = 1.719, $P < 2.2 \times 10^{-16}$ [Wilcoxon signed-rank test]). Similarly, the overall accuracies
141 of Seurat (NMI = 0.415, $P < 2.2 \times 10^{-16}$; VI = 1.545, $P < 2.2 \times 10^{-16}$ [Wilcoxon signed-rank
142 test]) and SIMLR (NMI = 0.231, $P < 2.2 \times 10^{-16}$; VI = 1.353, $P < 2.2 \times 10^{-16}$ [Wilcoxon
143 signed-rank test]) were significantly lower than that of LCA.
144

145 **Scalability to large data sets**

146 In addition to the superior accuracy of LCA with respect to both cluster number and
147 membership assignment, its implementation renders exceptional scalability, which is an
148 attractive feature for scRNA-seq analysis of an ever-increasing number of cells. LCA
149 can derive a cell subpopulation structure by using a relatively small representative set
150 (training cells) sampled from the full data. LCA provides mathematical formulae with
151 which to project the remaining cells (testing cells) directly to the inferred low-
152 dimensional LC space, after which individual cells are assigned to the subpopulation
153 with the best similarity. Consequently, LCA runs at a low level of computational
154 complexity. The median overall running times for a 4000-cell dataset with 10% or 25%
155 training cells were 40.6 s and 202.9 s, respectively. We compared the accuracy of
156 clustering between the training cells (5%–25% of all cells) and the remaining testing
157 cells in two simulated models (**Fig. 2C, D**, models #2 and #6). As expected, the
158 accuracy of the clustering model improved with the increase of training cells in the
159 simulated dataset #2 (weak signal strength) and remained at a near-perfect level in
160 simulated dataset #6 (strong signal strength). Strikingly, LCA achieved comparable

accuracies in the testing cells in both models and outperformed SC3 in terms of accuracy and running time.

Evaluation of LCA on published data sets

By using publicly available large-scale scRNA-seq datasets, we demonstrated that LCA identified orthogonal LC states of biological importance and produced parsimonious and accurate models that were highly consistent with biological knowledge.

We evaluated the performance of LCA with large-scale datasets published by Zheng et al. [15] and Tirosh et al. [16]. Using the GemCode platform (10x Genomics, Pleasanton, CA), Zheng et al. generated reference scRNA-seq transcriptomes for subpopulations of peripheral blood mononuclear cells (PBMCs) purified via well-established cell surface markers [15]. We generated a large-scale purified T-cell dataset (representing 55,000 cells) by combining the CD4+ T-helper (CD4+ helper), CD4+/CD25+ regulatory T (T_{reg}), CD4+/CD45RO+ memory T (T_{mem}), CD4+/CD45RA+/CD25− naïve T (CD4+ ab T), CD8+ cytotoxic T (cytotoxic T), and CD8+/CD45RA+ naïve cytotoxic T (CD8+ ab T) cell subpopulations. We inferred the cellular states and subpopulation structure by using 10% and 25% of the full dataset, and we assigned the remaining cells to one of the inferred subpopulations. We evaluated the top three models for both runs, which were the 3-population, 4-population, and 5-population models. Despite a 9.6-fold difference in running time, both the 3-population and 5-population models from the two runs achieved high consistency (NMI = 0.87 and 0.84, respectively). Although the 4-population models differed between runs, they represented two different subpopulation-merging orders

184 from the 5-population model to the 3-population model. We selected the 5-population
185 model learned from 25% cells (with 19 LC states) for further biological inference.
186 Whereas the purified T_{mem}, CD4+ ab T, and CD8+ ab T cells contained cells mostly from
187 a single subpopulation (Clusters 1, 3, and 4, respectively), different levels of
188 heterogeneity were detected in the remaining purified populations, especially the CD4+
189 helper and cytotoxic T cells. Given the single surface-marker settings in the purification
190 of CD4+ helper and cytotoxic T cells, it is not surprising that we found substantial
191 heterogeneity in the two populations. Nevertheless, LCA inferred a parsimonious
192 subpopulation structure in this large dataset (compared to the 1108 subpopulations
193 predicted by SC3). We selected representative genes encoding surface markers,
194 transcription factors, and secreted effector molecules for 19 usual T-cell subsets [17]
195 and derived a PCA projection of the six purified populations and five inferred clusters
196 based on the average population/cluster expression level of individual marker genes.
197 The first PC largely described the differentiation status, and the second PC represented
198 the difference between the CD4+ subsets and CD8+ subsets. As expected, Clusters 3
199 and 4 were found next to the CD8+ ab T and CD4+ ab T cells. Cluster 1 was found near
200 the T_{mem} population. Cluster 2 consisted mostly of T_{reg} cells and a smaller fraction of
201 CD4+ helper cells that was located adjacent to the T_{reg} population. Both the CD4+
202 helper and cytotoxic T-cell populations were split approximately equally between naïve
203 and differentiated cells and were spotted between their corresponding naïve clusters
204 (Cluster 3 for CD8+ cells and Cluster 4 for CD4+ cells) and differentiated clusters
205 (Clusters 1 and 2 for CD4+ cells and Cluster 5 for CD8+ cells). Analysis of selected
206 genes validated the separation of naïve and differentiated cells in the CD4+ helper and

207 cytotoxic T cells (**Fig. 3**). An evaluation of the first LC state inferred from the full
208 expression data revealed a striking approximation to the expected distribution of naïve
209 and differentiated cells in both purified populations and inferred clusters. Similarly, the
210 second LC state recaptured the CD4+ and CD8+ difference in the dataset. These
211 results showed that LCA identified LC states of biological importance and consequently
212 produced a parsimonious and accurate model that was highly consistent with biological
213 knowledge.
214
215 We evaluated the performance of LCA in a second dataset published by Tirosh et al.
216 [16], who employed a stepwise approach to analyze separately the malignant and
217 stromal cells of 19 melanoma tumors captured on the C1 Fluidigm platform. We applied
218 those authors' cell-selection criteria [16] to generate a dataset with 1169 malignant cells
219 from eight tumors and 2588 nonmalignant (stromal) cells. LCA inferred 18 clusters with
220 53 LC states (**Fig. 4**). Malignant cells dominated eight clusters, which were further
221 separated by the patient origin of the tumors (**Fig. 4A**). Among the stromal cells, distinct
222 clusters were identified for B cells, macrophages, cancer-associated fibroblasts, and
223 endothelial cells (**Fig. 4B**). Moreover, LCA divided tumor-infiltrating T cells and natural
224 killer cells into six clusters, which was concordant with the supervised analysis of T cells
225 based on surface markers by Tirosh et al. [16]. Using the pre-defined marker genes for
226 T-cell subsets and *MKI67* for cell-cycle activities, we revealed characteristics of these T-
227 cell populations (**Fig. 4C**). Cluster 1 was enriched for naïve CD4+ cells, whereas Cluster
228 2 harbored mostly T_{reg} and T_{FH} cells. Although Clusters 3 and 5 both had enriched
229 signatures for cytotoxic T and exhaustive T cells, Cluster 3 had greater signal strength

230 in both signatures, consistent with the reported high correlation between exhaustion
231 markers and cytotoxic markers [16]. Cells in active cell cycles were grouped in Cluster
232 4, which showed unique enrichment of *MKI67* expression. Lastly, natural killer cells and
233 T cells with weak cytotoxic activity and no exhaustion signatures were found in Cluster
234 6. As a comparison, we applied SC3 to the same data, which resulted in an estimation
235 of 43 clusters. Although SC3 clustering of malignant cells was consistent with the
236 separation by patient, accuracy in the stromal component was lower than with LCA
237 (**Supplementary Table 1**). These results demonstrate the power of LCA to reveal
238 subtle diversities in different subpopulations of tumor-infiltrating T cells in the presence
239 of strong transcriptomic variations among malignant cells (from different patients) and
240 various stromal cells from different lineages (cancer-associated fibroblasts,
241 macrophages, B cells, and endothelial cells).
242
243 **LCA work robustly on a data set with strong batch effects**
244 LCA groups cells based on orthogonal LC states aligned with major differences among
245 cells, which enables control of technical variations (e.g., batch effects) without an
246 explicit need of gene filtering. We evaluated the efficiency of batch effect removal in a
247 cancer cell dataset and further experimentally validated the subpopulations revealed by
248 LCA. Rhabdomyosarcoma is the most common soft-tissue tumor in children and has
249 two major histologic subtypes with different genomic landscapes: *PAX3/PAX7-FOXO1*
250 fusion–positive alveolar rhabdomyosarcoma (FP-ARMS) and fusion-negative embryonal
251 rhabdomyosarcoma (ERMS) [18]. LCA identified two subpopulations in a scRNA-seq
252 dataset for Rh41 cells (a commonly used human *PAX3-FOXO1* FP-ARMS cell line).

253 The *CD44* gene, which encodes a commonly used cell surface marker with great
254 prognostic and therapeutic potential [19, 20], appeared at the top of the differentially
255 expressed genes (DEGs) of the two subpopulations (**Supplementary Figure 1A**). Flow
256 cytometry confirmed a bimodal expression pattern of CD44 in Rh41 cells
257 (**Supplementary Fig. 1B**). We first used bulk RNA-seq to profile unsorted populations
258 and $CD44^{high}$ and $CD44^{low}$ subpopulations sorted by fluorescence-activated cell sorting.
259 In addition to the differences among the sorted $CD44^{high}$, $CD44^{low}$, and unsorted
260 populations, the analysis revealed strong batch effects. Specifically, samples in Batch 1
261 and those in Batches 2 and 3 were separated on the first PC, whereas the biologically
262 different populations (the unsorted population and sorted subpopulations) were
263 separated on the second PC (**Supplementary Fig. 1C**). We collected scRNA-seq data
264 for the unsorted populations in Batch 1 and the sorted $CD44^{high}$ and $CD44^{low}$
265 subpopulations in Batch 2 (10x Genomics, Pleasanton, CA). The libraries were
266 sequenced at different depths for the three samples. The median numbers of UMIs
267 captured per cell were 8278, 6678, and 12,850 for the $CD44^{high}$ subpopulation, $CD44^{low}$
268 subpopulation, and unsorted population, respectively (Kruskal-Wallis rank sum test,
269 $P < 2.2 \times 10^{-308}$). With both batch effects and biological difference among cell
270 populations, LCA inferred five clusters, with cells in Batch 1 being represented by
271 Clusters 1 and 2 and cells in Batch 2 being grouped in Clusters 3 to 5 (**Supplementary**
272 **Fig. 1D**). Similarly, SC3 inferred a 12-cluster structure. Consistent with the strong batch
273 effects observed in bulk RNA-seq analysis, when required to infer a two-cluster model,
274 both LCA and SC3 achieved a perfect separation based on the batch information
275 (NMI = 1 and 0.99, respectively). Next, we evaluated the association between the batch

276 covariate and individual LC states. When controlled at FDR ≤ 0.05, the first LC state was
277 the only one significantly associated with the batch information ($P < 2.2 \times 10^{-308}$)
278 (**Supplementary Fig. 1E**). After excluding this state, LCA retrieved an optimal structure
279 with two clusters. Of 7005 cells profiled in the sorted CD44^{low} subpopulation, 6733
280 (96.1%) were clustered into Cluster 2, suggesting that the subpopulation was relatively
281 pure. However, 1796 (26.6%) of 6757 of cells profiled in the sorted CD44^{high}
282 subpopulation were grouped in Cluster 2 with the CD44^{low} cells (**Fig. 5A**). We evaluated
283 the expression signatures for DEGs identified in bulk RNA-seq ($\log_2$FC ≥ 1, FDR ≤ 0.05,
284 FPKM ≥ 1 in at least one subpopulation). Of 1709 DEGs (**Supplementary Table 2**), 356
285 were captured in at least 10% of the cells in one sequenced population or inferred
286 cluster. As expected, Clusters 1 and 2 had higher average expression levels of genes
287 overexpressed in the bulk-sorted CD44^{high} and CD44^{low} subpopulations, respectively
288 (**Fig. 5B**). Importantly, Cluster 2 cells from the sorted CD44^{high} subpopulation had
289 significantly lower expression of those DEGs overexpressed in the bulk CD44^{high}
290 subpopulation than did Cluster 1 cells from the sorted CD44^{high} subpopulation
291 ($P < 2.2 \times 10^{-308}$) or the sorted CD44^{low} subpopulation ($P = 7.5 \times 10^{-71}$). Analysis of DEGs
292 overexpressed in the bulk CD44^{low} subpopulation showed the same pattern (**Fig. 5C)**.
293 These results suggested that Cluster 2 cells in the sorted CD44^{high} subpopulation more
294 closely resembled CD44^{low} cells. Moreover, the differential expression pattern was
295 essentially captured by the first remaining LC state (Spearman correlation
296 coefficient = −0.90, $P < 2.2 \times 10^{-308}$), confirming the biological importance of the inferred
297 LC states.
298

299 Of the three established molecular markers (*TFAP2B*, *MYOG,* and *NOS1*) for FP-
300 ARMS[21], *TFAP2B* and *MYOG* were overexpressed in the sorted CD44^{low}
301 subpopulation (**Supplementary Table 2**). Moreover, known *PAX3-FOXO1* target genes
302 were significantly enriched in DEGs overexpressed in the CD44^{low} subpopulation
303 (**Supplementary Table 3**). These results suggest that the CD44^{high} subpopulation
304 represents a less-differentiated, stem-like cell subpopulation. Although the exact
305 mechanism by which the distinct subpopulations develop warrants further investigation,
306 we conclude that LCA can control technical variations and reveal reliable transcriptome-
307 based heterogeneity.
308

309 ## Discussion and conclusions

310 The rapid technological advance in scRNA-seq platforms has inspired a tremendous
311 explosion of data-analysis methods for identifying heterogeneous subpopulations. Most
312 methods employ a gene-selection step before clustering analysis, based on the
313 assumption that a small subset of highly variable genes is useful for revealing cellular
314 diversity. Although this assumption is valid in most scenarios and reduces the data
315 dimensionality, it potentially excludes genes that are informative for separating
316 subpopulations with subtle diversity. Also, in datasets with strong batch effects, it may
317 result in a small set of retained genes being dominated by batch effects. Moreover,
318 several methods require the user to provide an estimation of the number of clusters in
319 the data, and this may not be readily available.
320

Using a dual-space search strategy, LCA bypasses the gene selection, learns biologically informative cellular states directly from the raw gene expression matrix, reduces potential technical variations, and measures the cell-to-cell distance by using cosine similarity in the low-dimensional and informative cellular-state space in a data-driven and unsupervised fashion. Cosine similarity has been widely used in information retrieval and text mining to reveal the relation between text documents, a process that shares many similarities with scRNA-seq analysis [22]. Furthermore, LCA provides a mathematical solution for assigning new cells to inferred clusters in a model learned from a subset of cells, a capability that is urgently needed to handle the ever-increasing sample sizes in scRNA-seq. We have demonstrated through extensive simulation and the use of large-scale scRNA-seq datasets that LCA is an efficient, scalable, and robust clustering algorithm that outperforms other tools without the explicit need for gene selection or an estimation of the number of clusters in the data.

Methods

Latent cellular states

The input to LCA is a gene expression matrix in which each column is a cell, and each row is a gene/transcript. In UMI-based platforms, the expression level of a gene in a cell is divided by the total expression in that cell to generate a relative expression matrix (T). In read-count based platforms, T can be derived from size factor normalized expression measures. The relative expression matrix is then log-transformed after adding a zero-correction term:

$$\mathrm{X} = \log(\mathrm{T} + \epsilon),$$

where ϵ is an arbitrarily small number.

We obtain the LC states from a singular value decomposition (SVD) of $\mathbf{X}$.

$$\mathbf{X} = \mathbf{G}\,\mathbf{\Lambda}\,\mathbf{S}^T,$$

where $\mathbf{S}$ is a cell–by–LC states matrix. We note that under certain conditions, $\mathbf{S}$ is the same as the loading matrix from the principle component analysis (PCA) result of $\mathbf{X}$ [23, 24].

Determination of significant LC states

We apply the Tracy-Widom test to associated eigenvalues to determine which LC states are significant [10, 25, 26]. The LC state associated with the eigenvalue λ is significant if it is significantly different ($P < 0.05$) from the Tracy-Widom distribution, with

$$\begin{aligned} \mu(n,m) &= \frac{(\sqrt{m-1}+\sqrt{n})^2}{m} \\ \sigma(n,m) &= \frac{\left(\sqrt{m-1}+\sqrt{n}\right)}{m}\left(\frac{1}{\sqrt{m-1}+\sqrt{n}}\right)^{\frac{1}{3}}, \end{aligned}$$

where n is the total number of genes and m is the total number of LC states. We then discard all the LC states that are not significant or strongly associated with known technical variations (e.g., cell-specific library complexity and batch information), leading to a lower-dimensional cell–by–LC states matrix $\mathbf{S}$.

Distance calculation

362 Distances between cells in $\mathbf{S}$ (the cell–by–LC states matrix) are calculated using cosine
363 distance:

364 $$K_{a,b} = 1 - \frac{\sum_{i=1}^{p} S_{a,i}\, S_{b,i}}{\sqrt{\sum_{i=1}^{p} S_{a,i}^2}\sqrt{\sum_{i=1}^{p} S_{b,i}^2}}$$

365 where $S_{a,i}$ and $S_{b,i}$ represent LC state i for cells a and b, and p is the total number of
366 retained LC states.

367

368 **Spectral clustering**

369 We perform spectral clustering [27] on the resulted distance matrix to derive a set of
370 candidate clustering models with a range of cluster numbers (i.e., 2 – 20 by default).

371

372 **Distance measure in the PC space**

373 The cell-gene relative expression matrix was decomposed by PCA and the significant
374 components were determined using the Tracy-Widom test. Distance between cells was
375 measured by the correlation distance of significant components.

376

377 **The optimal number of clusters and informative cellular states**

378 We rank the candidate clustering solutions (with different number of clusters) by the
379 silhouette score [11] measured in the PC space. With 2 or more clusters, the silhouette
380 measures the similarity of an individual to its cluster, as compared to other clusters. For
381 each cell, let d_b be the lowest dissimilarity to any other cluster and let d_w be the
382 average dissimilarity to other cells in its cluster. We calculate the silhouette as

$$Silhouette = \frac{d_b - d_w}{\max(d_b, d_w)}.$$

We assign a silhouette score of zero for the default solution of one cluster. An end-user may evaluate the top candidate solutions to determine the optimal number of solutions or specify it with a priori biological knowledge.

With the selected number of clusters, we retain LC states that show significant difference among candidate clusters, update the distance matrix and derive the final clustering solution.

Fast processing of large numbers of cells

We classify unknown cells $\mathbf{X}'$ efficiently by projecting them into space spanned by the informative LC states learned from a representative sample:

$$\mathbf{S}' = \mathbf{\Lambda}_l^{-1}\mathbf{G}_l^T\mathbf{X}',$$

where $\mathbf{\Lambda}_l$ is calculated from the original $\mathbf{\Lambda}$ by removing those rows and columns that are not associated with the informative LC states, and $\mathbf{G}_l$ is calculated from $\mathbf{G}$ by removing those columns that are not associated with the informative LC states. We can then calculate the cosine similarity between $\mathbf{S}'$ and an "average" cell from each cluster and find the cluster with the maximum similarity for each unknown cell.

Simulation of different cell types

403 We simulated different cell types from a relatively homogeneous population of cells. We
404 started from the scRNA-seq dataset for sorted CD44^{high} Rh41 cells. The data
405 represented 6757 cells with 20,709 genes detected, with 86% of the entries in the data
406 matrix being zeroes. We considered two factors that have substantial effects on the
407 performance of clustering analysis: the sample size and the degree of difference
408 between cell types.

409

410 We tested five different sample sizes (the total number of cells analyzed): 250, 500,
411 1000, 2000, and 4000 cells. We sampled the selected number of cells from the original
412 set of cells then randomly partitioned them into three sets of variable size, ranging
413 between 10% and 50% of the total number of cells.

414

415 To test the effect of the degree of difference between cell types, we varied the number
416 of DEGs and the simulated difference in gene expression. We chose three sets of
417 genes such that the average expression levels for each set were g_1, g_2, and g_3, with
418 $\frac{g_{i+1}}{g_i} = 2, i = 1,2$. We then rotated the labels of the three sets of genes to make three
419 different sets of cells with perturbed expression: g_1, g_2, g_3; g_2, g_3, g_1; and g_3, g_1, g_2. We
420 tested two different numbers of DEGs: 300 and 1200. The simulation setting is
421 summarized in **Supplemental Table 4**. For each set of parameters, we generated 100
422 random samples for testing.

423

424 **Normalized mutual information and variation of information**

425 We evaluate the performance of clustering against true labels of cells by using
426 normalized mutual information (NMI) [14] and variation of information (VI) [13]. We use
427 the R package *igraph* to calculate both metrics [28].

428

429 **Benchmarking**

430 We used SC3 (Version 1.4.2), Seurat (Version 2.0.0), Monocle2 (Version 2.4.0) and
431 SIMLR (Version 1.2.1) for comparisons with LCA. We used the default settings for all
432 the packages tested. For SIMLR, when the size of the sample was 1000 or larger, we
433 used SIMLR_Large_Scale() instead of SIMLR().

434

435 **Rh41 single-cell dataset**

436 The human alveolar rhabdomyosarcoma cell line Rh41 was grown in culture in a 5%
437 CO_2 incubator in 75-cm^2 vented flasks containing DMEM medium supplemented with
438 10% FBS and 2× glutamine until the cells reached 75% confluence at approximately
439 3.6×10^6 cells. The cells were detached from the flask with 7 mL of 1× citrate saline to
440 which 7 mL of DPBS was added. The cell suspension was then centrifuged at $300 \times g$
441 for 7 min, and the cell pellet was resuspended in 300 μL of blocking buffer (rat IgG/PBS)
442 and incubated on ice for 30 min. An aliquot of 50 μL of the cells in blocking buffer was
443 transferred to a separate tube for the isotype control. The cells were washed with 1 mL
444 of staining buffer (5% BSA/PBS) and centrifuged at $300 \times g$ for 5 min. The pellet, which
445 contained approximately 3×10^6 cells, was then incubated with rat IgG2B anti-CD44–
446 Alexa Fluor 488 antibody (R&D Systems) in staining buffer (15 μL antibody + 135 μL of

447 staining buffer) on ice for 30 min. For the isotype control, approximately 600,000 cells
448 were incubated with 5 μL of rat IgG2B–Alexa Fluor 488 (R&D Systems, Minneapolis,
449 MN) + 45 μL of staining buffer on ice for 30 min. After the incubation, both sets of cells
450 were collected by centrifugation, washed with 1 mL of staining buffer as described
451 above, and resuspended in staining buffer. Flow cytometric analysis was then used to
452 identify the fractions corresponding to the CD44 ^{high} and CD44 ^{low} populations.
453
454 For the single-cell experiment, Rh41 cells were grown in culture, harvested, and
455 washed in DPBS as described above. They were then resuspended in PBS/0.2% BSA
456 at a concentration of 1×10^6 cells/mL. The 10x Genomics Single Cell platform performs
457 3′ gene expression profiling by poly-A selection of mRNA within a single cell, which uses
458 a cell barcode and UMIs for each transcript. Single-cell suspensions were loaded onto
459 the Chromium Controller according to their respective cell counts to generate
460 approximately 6000 partitioned single-cell GEMs (Gel Bead-in-EMulsions). The library
461 was prepared using the Chromium Single Cell 3′ v2 Library and Gel Bead Kit (10x
462 Genomics) in accordance with the manufacturer's protocol. The cDNA content of each
463 sample after cDNA amplification for 12 cycles was quantified, and the quality was
464 checked by High-Sensitivity DNA chip analysis on an Agilent 2100 Bioanalyzer (Agilent
465 Technologies, Santa Clara, CA) at a dilution of 1:6. This quantification was used to
466 determine the final library amplification cycles in the protocol, which were calculated out
467 to 12 cycles. After library quantification and a quality check by DNA 1000 chip (Agilent
468 Technologies), samples were diluted to 3.5 nM for loading onto the HiSeq 4000
469 sequencer (Illumina) with a 2×75-bp paired-end kit, using the following read length: 26

470 bp Read1 (10x cell barcode and UMI), 8 bp i7 Index (sample index), and 98 bp Read2
471 (insert). In total, 518 million, 237 million, and 154 million reads were obtained for
472 unsorted, $CD44^{low}$, and $CD44^{high}$ populations, respectively. The Cell Ranger 2.0.1
473 Single-Cell Software Suite (10x Genomics) was implemented to process the raw
474 sequencing data from the Illumina HiSeq run. This pipeline performed de-multiplexing,
475 alignment (GRCh38/STAR), and barcode processing to generate gene-cell matrices
476 used for downstream analysis.
477
478 After matrix generation, the ribosomal and mitochondria-related genes were filtered out.
479
480 **Rh41 bulk RNA-seq dataset**
481 RNA was isolated from the sorted subpopulations by using Trizol (Thermo Fisher
482 Scientific) in accordance with the manufacture's recommendations. RNA libraries were
483 prepared with the KAPA RNA HyperPrep Kit with RiboErase (Roche), using the
484 recommended conditions. Briefly, 200 ng of total RNA was used as input for
485 fragmentation, reverse transcription, and second-strand synthesis. After clean-up, end
486 repair, and A-tailing, NEXTflex adapters (Bioo Scientific, Austin, TX) were ligated to the
487 fragments, and this was followed by 12 cycles of PCR amplification on a C1000
488 Thermal Cycler (Bio-Rad). Paired-end sequencing was performed (151 bases per read)
489 on a HiSeq 4000 system (Illumina). Three replicates were generated. HTSeq [29] was
490 used to produce the count data, and edgeR [30] was used for the DE analysis with TMM
491 normalization. Each replicate was coded as a pair of $CD44^{high}$ and $CD44^{low}$ in the
492 analysis.

493

494 **Statistical analysis**

495 If not specifically stated, the statistical significance was evaluated by the Wilcoxon rank

496 sum test.

497

498 Abbreviations

499 **CD**: Cluster of differentiation **DEG**: Differentially expressed genes **ERMS**: fusion-

500 negative embryonal rhabdomyosarcoma **FP-ARMS**: fusion–positive alveolar

501 rhabdomyosarcoma **LCA**: Latent cellular analysis **NMI**: Normalized mutual information

502 **PBMC**: Peripheral blood mononuclear cell **PCA**: Principal component analysis **t-SNE**: t-

503 Distributed stochastic neighbor embedding **UMI**: Unique molecular identifier **VI**:

504 Variation of Information

505

506 Declarations

507 **Acknowledgements**

508 We thank Keith A. Laycock for editing the manuscript.

509

510 ***Funding***

511 This study was supported in part by the National Cancer Institute of the National

512 Institutes of Health under Award Number P30CA021765 and by ALSAC.

513

514 ***Availability of data and materials***

515 The Rh41 scRNA-seq dataset generated during the current study is available from the
516 corresponding author upon reasonable request. The bulk RNA-seq data for sorted
517 $CD44^{high}$ and $CD44^{low}$ subpopulations have been deposited in the European
518 Bioinformatics Institute (EMBL-EBI) under accession number EGAS00001002868. The
519 functions used for the data analysis are included in the single_cell_LCA package, which
520 can be installed from Bitbucket (https://bitbucket.org/changde_cheng/single_cell_lca).
521

522 ## References

523 1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J,
524 Tuch BB, Siddiqui A, et al: **mRNA-Seq whole-transcriptome analysis of a**
525 **single cell.** *Nat Methods* 2009, **6:**377-382.
526 2. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based**
527 **technologies will revolutionize whole-organism science.** *Nat Rev Genet*
528 2013, **14:**618-630.
529 3. Stegle O, Teichmann SA, Marioni JC: **Computational and analytical**
530 **challenges in single-cell transcriptomics.** *Nat Rev Genet* 2015, **16:**133-145.
531 4. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I,
532 Bialas AR, Kamitaki N, Martersteck EM, et al: **Highly parallel genome-wide**
533 **expression profiling of individual cells using nanoliter droplets.** *Cell* 2015,
534 **161:**1202-1214.
535 5. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ,
536 Livak KJ, Mikkelsen TS, Rinn JL: **The dynamics and regulators of cell fate**
537 **decisions are revealed by pseudotemporal ordering of single cells.** *Nat*
538 *Biotechnol* 2014, **32:**381–386.
539 6. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S: **Visualization and**
540 **analysis of single-cell RNA-seq data by kernel-based similarity learning.**
541 *Nat Methods* 2017, **14:**414-416.
542 7. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan
543 KN, Reik W, Barahona M, Green AR, Hemberg M: **SC3: consensus clustering**
544 **of single-cell RNA-seq data.** *Nat Methods* 2017, **14:**483-486.
545 8. Wagner A, Regev A, Yosef N: **Revealing the vectors of cellular identity with**
546 **single-cell genomics.** *Nat Biotechnol* 2016, **34:**1145-1160.
547 9. Hicks SC, Townes FW, Teng M, Irizarry RA: **Missing data and technical**
548 **variability in single-cell RNA-sequencing experiments.** *Biostatistics* 2017.
549 10. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS*
550 *Genet* 2006, **2:**e190.

11. Rousseeuw PJ: **Silhouettes - a graphical aid to the interpretation and validation of cluster-analysis.** *Journal of Computational and Applied Mathematics* 1987, **20:**53-65.
12. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic.** *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2001, **63:**411-423.
13. Meilă M: **Comparing clusterings—an information based distance.** *Journal of Multivariate Analysis* 2007, **98:**873-895.
14. Danon L, Diaz-Guilera A, Duch J, Arenas A: **Comparing community structure identification.** *Journal of Statistical Mechanics-Theory and Experiment* 2005.
15. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al: **Massively parallel digital transcriptional profiling of single cells.** *Nat Commun* 2017, **8:**14049.
16. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al: **Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.** *Science* 2016, **352:**189-196.
17. Dong C, Martinez GJ: **T cells: the usual subsets (poster).** 2010.
18. Chen X, Stewart E, Shelat AA, Qu C, Bahrami A, Hatley M, Wu G, Bradley C, McEvoy J, Pappo A, et al: **Targeting oxidative stress in embryonal rhabdomyosarcoma.** *Cancer Cell* 2013, **24:**710-724.
19. Li F, Tiede B, Massague J, Kang Y: **Beyond tumorigenesis: cancer stem cells in metastasis.** *Cell Res* 2007, **17:**3-14.
20. Yan Y, Zuo X, Wei D: **Concise Review: Emerging Role of CD44 in Cancer Stem Cells: A Promising Biomarker and Therapeutic Target.** *Stem Cells Transl Med* 2015, **4:**1033-1043.
21. Rudzinski ER, Anderson JR, Lyden ER, Bridge JA, Barr FG, Gastier-Foster JM, Bachmeyer K, Skapek SX, Hawkins DS, Teot LA, Parham DM: **Myogenin, AP2beta, NOS-1, and HMGA2 are surrogate markers of fusion status in rhabdomyosarcoma: a report from the soft tissue sarcoma committee of the children's oncology group.** *Am J Surg Pathol* 2014, **38:**654-659.
22. Dumais ST: **Latent semantic analysis.** *Annual Review of Information Science and Technology* 2004, **38:**189-230.
23. Pearson K: **On lines and planes of closest fit to systems of points in space.** *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1901, **2:**559-572.
24. Hotelling H: **Analysis of a complex of statistical variables into principal components.** *Journal of Educational Psychology* 1933, **24:**417-441.
25. Tracy CA, Widom H: **Level-spacing distributions and the Airy kernel.** *Physics Letters B* 1993, **305:**115-118.
26. Johnstone IM: **On the distribution of the largest eigenvalue in principal components analysis.** *Annals of Statistics* 2001, **29:**295-327.
27. Ng AY, Jordan MI, Weiss Y: **On spectral clustering: Analysis and an algorithm.** In *Advances in neural information processing systems*; *2002*. 2002: 849-856.

596 28. Csardi G, Nepusz T: **The igraph software package for complex network**
597 **research.** *InterJournal* 2006, **Complex Systems:**1695.
598 29. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-**
599 **throughput sequencing data.** *Bioinformatics* 2015, **31:**166-169.
600 30. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for**
601 **differential expression analysis of digital gene expression data.**
602 *Bioinformatics* 2010, **26:**139-140.
603 31. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev
604 S, Jenkins SL, Jagodnik KM, Lachmann A, et al: **Enrichr: a comprehensive**
605 **gene set enrichment analysis web server 2016 update.** *Nucleic Acids Res*
606 2016, **44:**W90-97.
607
608

609 **Figure Legends**

610

611 **Figure 1**. Overview of the workflow of LCA. LCA infers LC states from full expression
612 matrices. Explicit gene filtering is not necessary. LCA converts the raw transcript count
613 data to gene fractions and then performs log transformations. The algorithm features a
614 dual-space model search, generating candidate clustering models based on the cosine
615 similarity matrix in the LC space. Candidate models are then ranked based on the
616 silhouettes measured in the PC space.

617

618 **Figure 2**. Benchmarking of LCA against four other methods with simulated datasets. (A)
619 The number of clusters inferred by LCA, SC3, and Seurat was compared for seven
620 simulated datasets: the signal strength increases from simulated datasets 1 to 6. Each
621 simulated dataset contained 100 randomly generated samples. (B) Comparison of the
622 clustering performance of LCA, SC3, Seurat, Monocle2, and SIMLR, as measured by
623 variation of information, across the seven simulated datasets. For LCA, SC3, and
624 Seurat, the number of clusters was inferred during the analysis. For Monocle2 and
625 SIMLR, we used the correct number of clusters. (C) Comparison of the clustering
626 performance as measured by normalized mutual information. (D) Clustering
627 performance scalability of LCA and SC3. Each panel represents the results obtained
628 with a different percentage of cells used for training (from 5% to 25%). The total number
629 of cells in the sample was 4000. The signal in this comparison was relatively weak. (E)
630 Comparison of the clustering performance scalability of LCA and SC3 when the signal
631 was strong. (F) Comparison of the running times of LCA and SC3.

632

633 **Figure 3.** Investigation of PBMC heterogeneity by using LCA. We applied LCA to data
634 from Zheng et al.[25] for 55,000 cells representing a combination of six sorted T-cell
635 populations. LCA identified five cell populations with biologically meaningful LC states.
636 (A) Cell distance matrix, sorted by source ID. (B). Cell distance matrix, sorted by LCA
637 clustering results. (C) The first LC state captured the difference between the naïve and
638 differentiated T cells, whereas the second LC state captured the difference between the
639 CD4+ and CD8+ cells. (D) LCA further revealed heterogeneity in the CD4+ helper T
640 cells and CD8+ cytotoxic T cells.

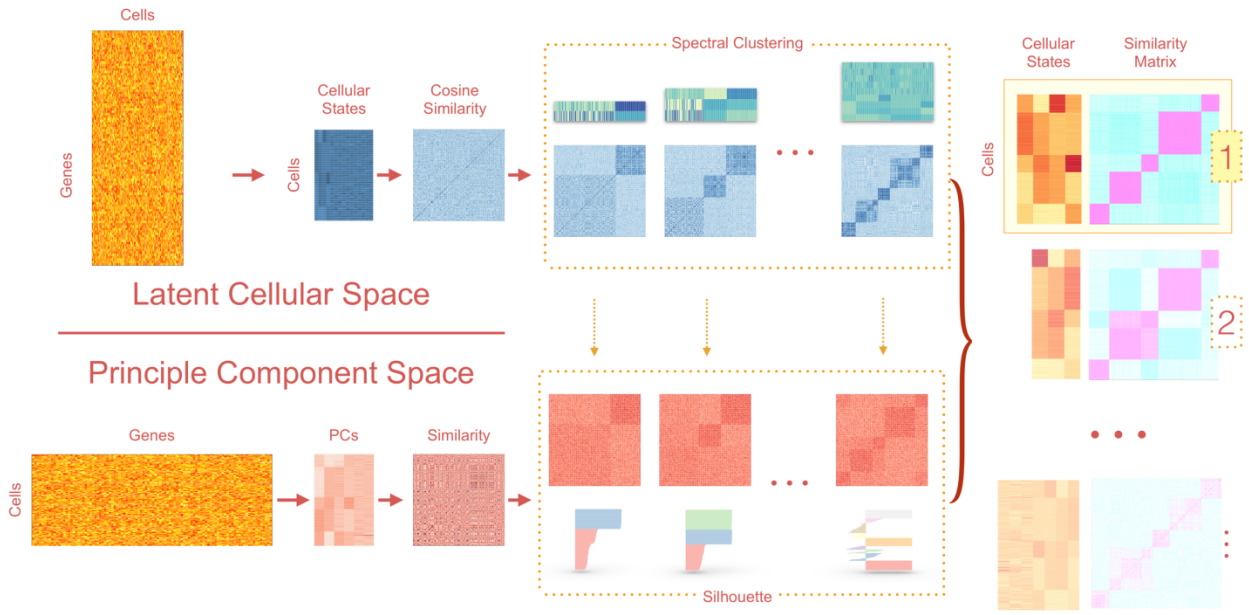
641

642 **Figure 4.** Reanalysis of melanoma cellular data [16] with LCA. We applied LCA to 1169
643 malignant cells from eight tumors and to 2588 stromal cells. LCA identified 18 clusters
644 with a clear biological interpretation. LCA further revealed subtle diversity among the
645 infiltrating T cells. (A) t-SNE plot of the clustering results for malignant cells, contrasting
646 the LCA results and original results. (B) t-SNE plot of the clustering results for stromal
647 cells, contrasting the LCA results and original results. (C) Enrichment of gene markers
648 in T-cell clusters.

649

650 **Figure 5**. LCA analysis of Rh41 cells, correcting for batch effects. (A) t-SNE plots of
651 clustering results, colored according to source ID or cluster ID. (B) Cellular expression
652 patterns of DEGs identified in bulk RNA-seq. (C) Violin plot of expression patterns of
653 DEGs from bulk RNA-seq in cells classified by source and cluster ID.

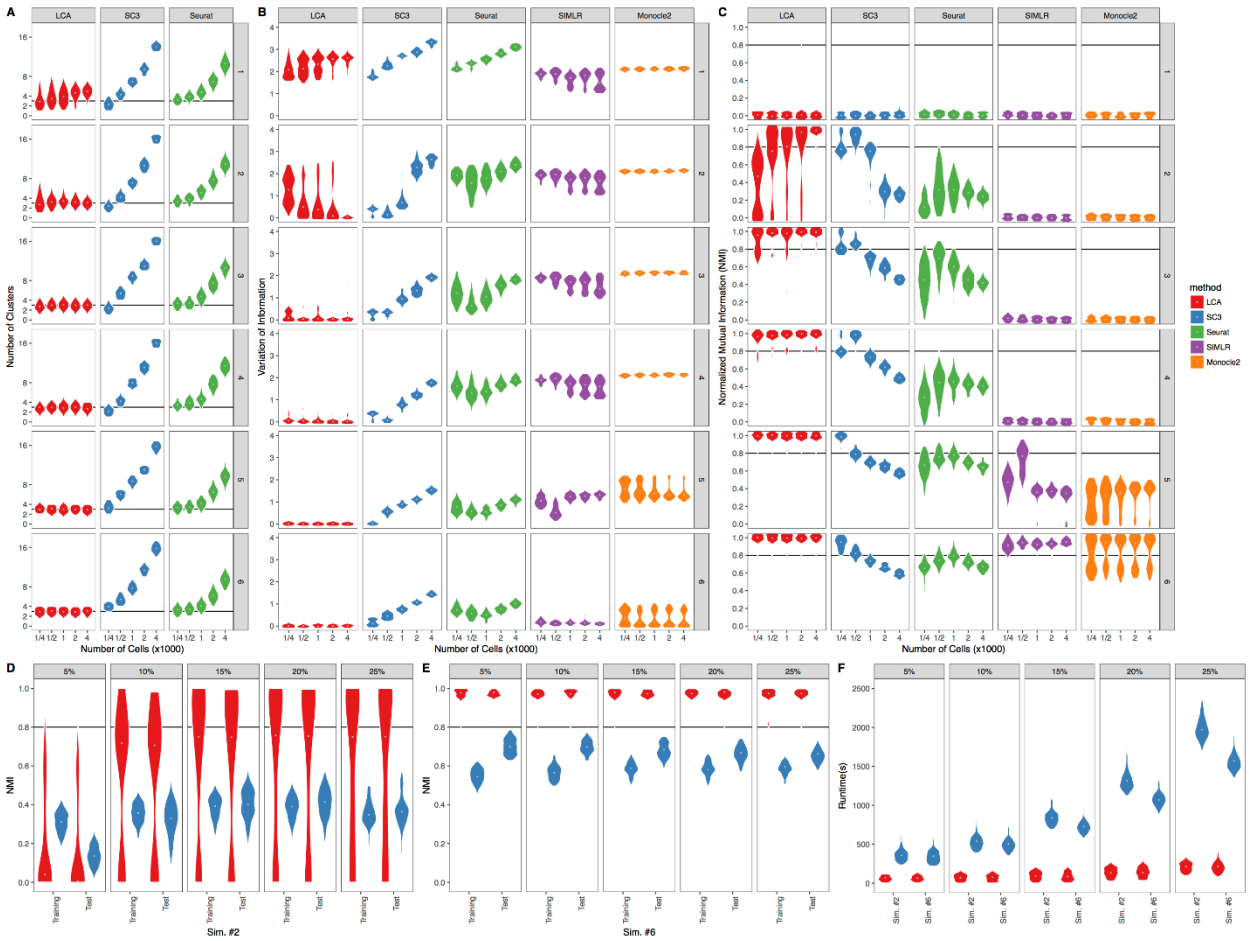
654 **Figure 1.** Overview of the workflow of LCA.



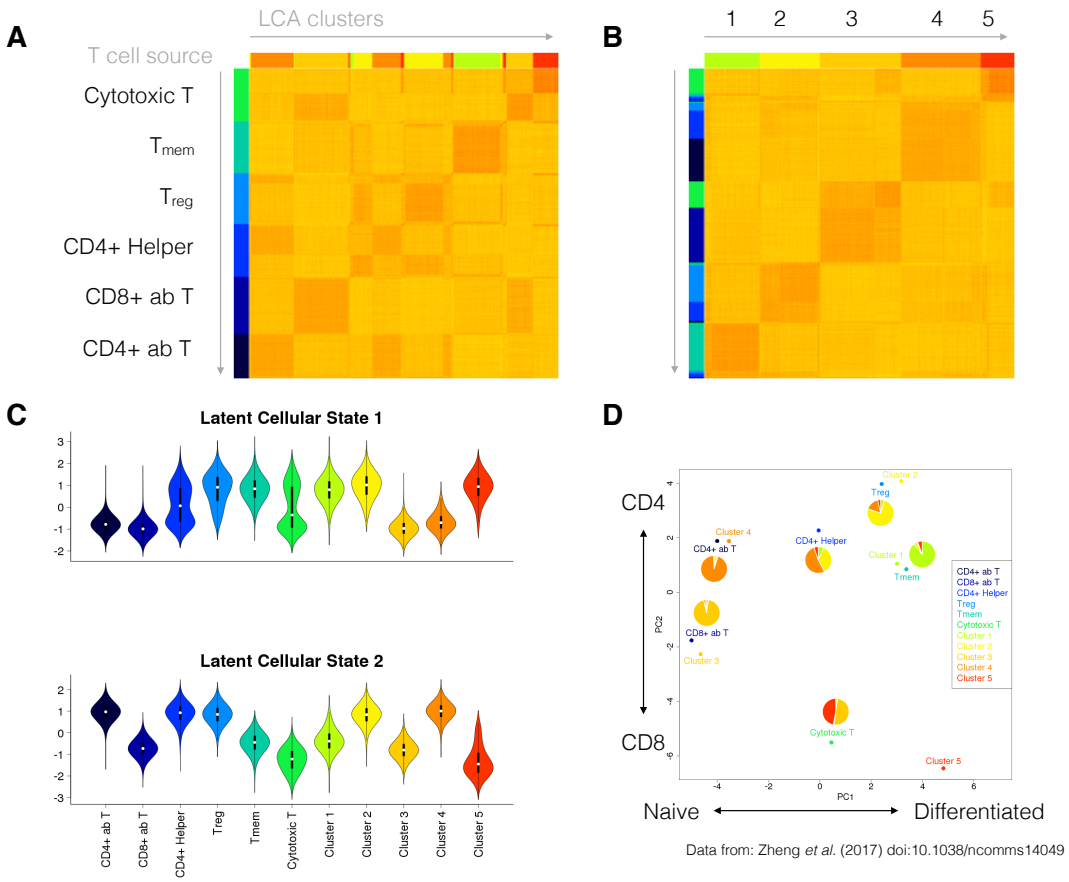
655

656

Figure 2. Benchmarking of LCA against four other methods with simulated datasets.



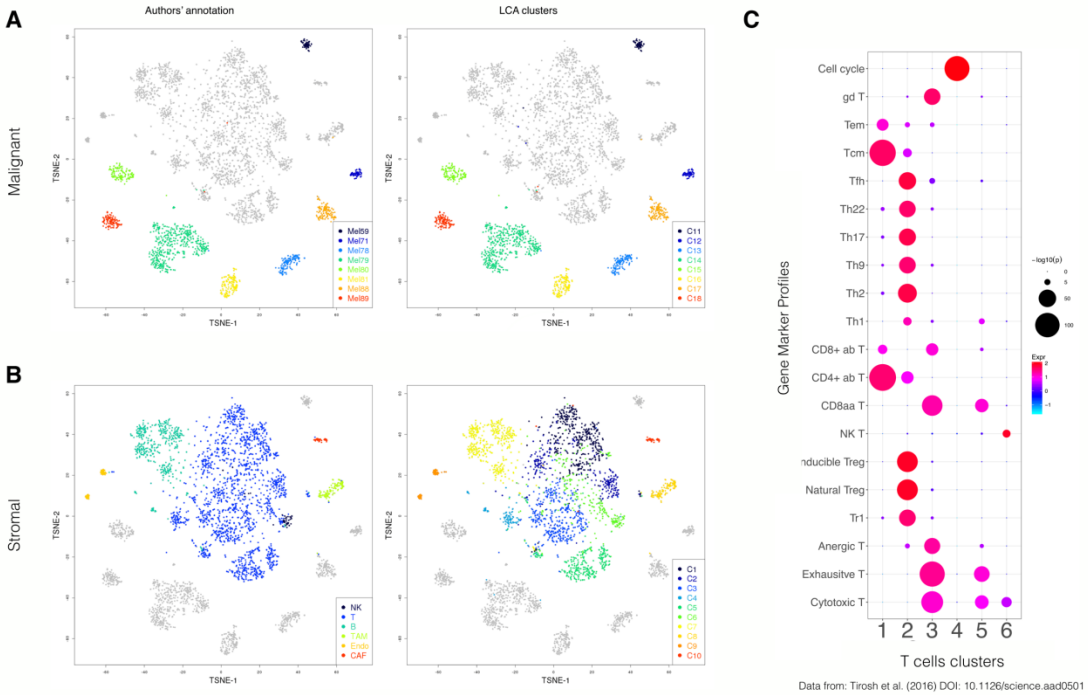
660 **Figure 3.** Investigation of PBMC heterogeneity by using LCA.



661

662

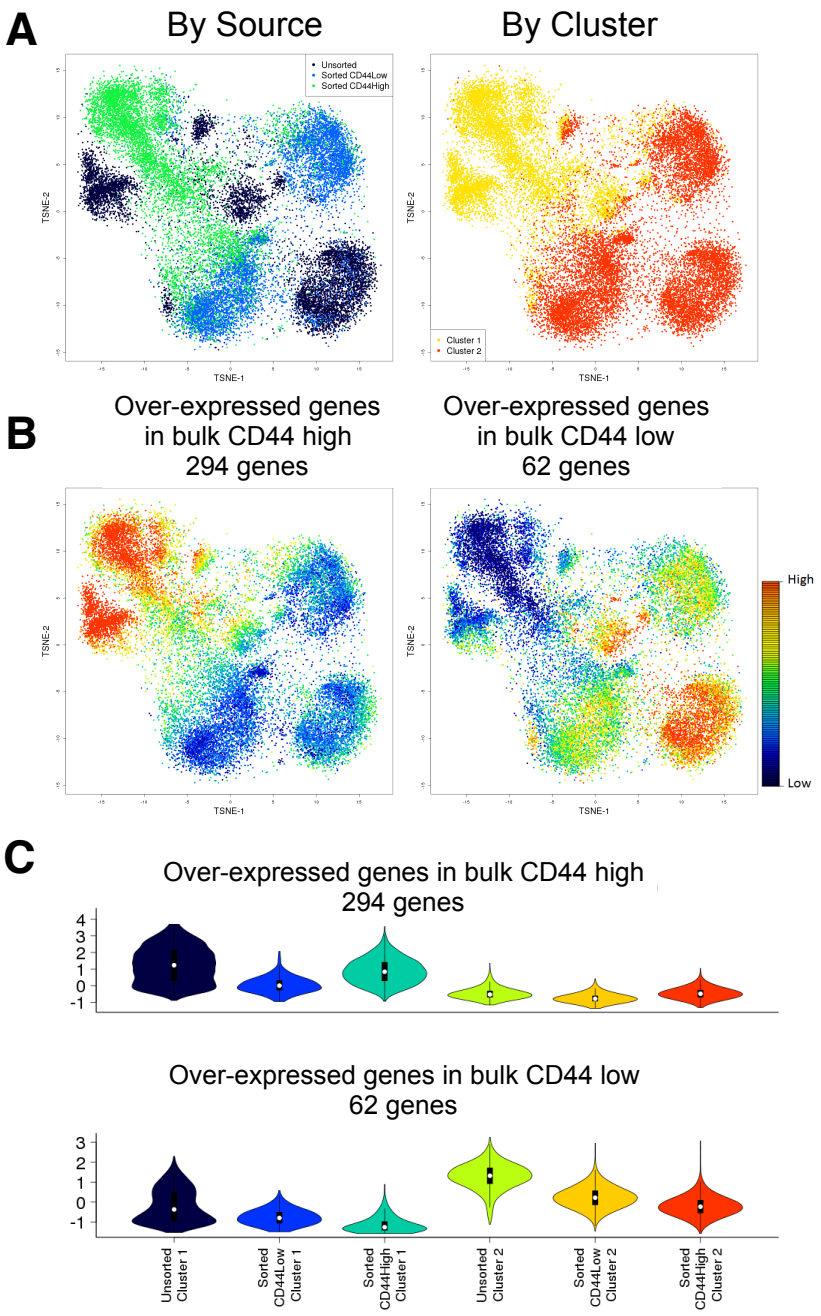
663 **Figure 4.** Reanalysis of melanoma cellular data with LCA.



664

665

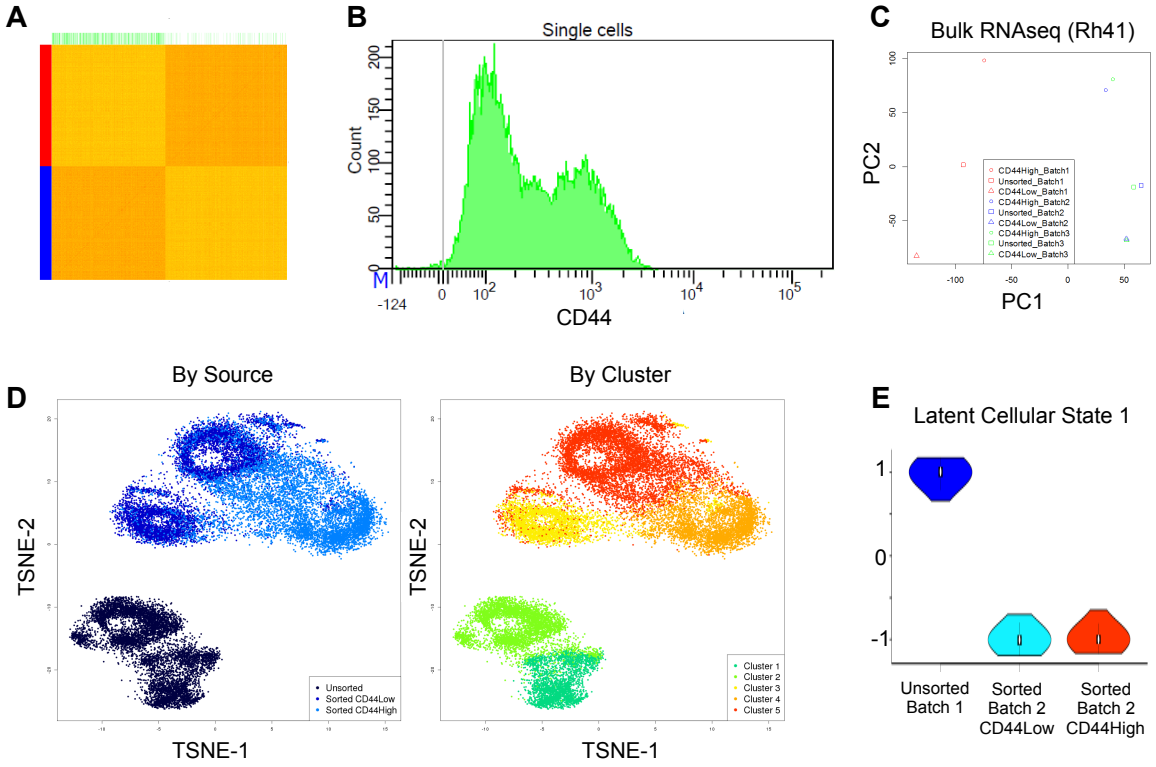
666 **Figure 5.** LCA analysis of Rh41 cells, correcting for batch effects.



667

668 **Supplemental Figure 1.** Single-cell analysis of Rh41 cells with strong batch effects. (A)
669 Distance matrix visualized as a heat map. (B) Flow cytometry confirmed a bimodal
670 expression pattern of CD44 in Rh41 cells. (C) The first two PCs for the bulk RNA-seq
671 data showed strong batch effects. (D) t-SNE plot of cells, including sorted and unsorted
672 samples, colored according to source/cluster ID. (E) The first LC state was significantly
673 associated with batch information.
674

675 **Supplemental Figure 1.** Single-cell analysis of Rh41 cells with strong batch effects.



676

677

678 **Supplementary Table 1**. The results of applying LCA and SC3 to the dataset of Tirosh

679 et al.[26], using normalized mutual information.

	LCA	SC3
All cells	0.77	0.69
Malignant cells	0.99	0.99
All stromal cells	0.53	0.42
All cells excluding T cells	0.97	0.92
Stromal excluding T cells	0.90	0.71

680
681

682

683

684 **Supplementary Table 2**. Expression profile of differential expressed genes (in a

685 standalone excel file).

686

687

688

689 **Supplementary Table 3**. ChEA analysis by Enrichr[31] of 361 overexpressed genes in

690 the $CD44^{low}$ subpopulation of Rh41 (in a standalone excel file).

691

692

693 **Supplementary Table 4**. Summary of simulated datasets.

Signal Scale	Number of Genes Perturbed	Mean Gene Fraction in Group 1	Mean Gene Fraction in Group 2	Mean Gene Fraction in Group 3
Sim. #1	300	1e-5	2e-5	4e-5
Sim. #2	300	2e-5	4e-5	8e-5
Sim. #3	300	4e-5	8e-5	1.6e-4
Sim. #4	1200	1e-5	2e-5	4e-5
Sim. #5	1200	2e-5	4e-5	8e-5
Sim. #6	1200	4e-5	8e-5	1.6e-4

694

695