Published online 28 September 2021

Nucleic Acids Research, 2021, Vol. 49, No. 22 ***e127***
https://doi.org/10.1093/nar/gkab823

MLG: multilayer graph clustering for multi-condition scRNA-seq data

Shan Lu[1], Daniel J. Conn[2], Shuyang Chen[1], Kirby D. Johnson[3], Emery H. Bresnick[3] and Sündüz Keleş[1,2,*]

[1]Department of Statistics, University of Wisconsin, Madison, WI 53706, USA, [2]Department of Biostatistics and Medical Informatics, University of Wisconsin School of Medicine and Public Health, Madison, WI 53792, USA and [3]Wisconsin Blood Cancer Research Institute, Department of Cell and Regenerative Biology, University of Wisconsin School of Medicine and Public Health, Madison, WI 53705, USA

Received December 10, 2020; Revised August 13, 2021; Editorial Decision September 06, 2021; Accepted September 21, 2021

ABSTRACT

Single-cell transcriptome sequencing (scRNA-seq) enabled investigations of cellular heterogeneity at exceedingly higher resolutions. Identification of novel cell types or transient developmental stages across multiple experimental conditions is one of its key applications. Linear and non-linear dimensionality reduction for data integration became a foundational tool in inference from scRNA-seq data. We present multilayer graph clustering (MLG) as an integrative approach for combining multiple dimensionality reduction of multi-condition scRNA-seq data. MLG generates a multilayer shared nearest neighbor cell graph with higher signal-to-noise ratio and outperforms current best practices in terms of clustering accuracy across large-scale benchmarking experiments. Application of MLG to a wide variety of datasets from multiple conditions highlights how MLG boosts signal-to-noise ratio for fine-grained sub-population identification. MLG is widely applicable to settings with single cell data integration via dimension reduction.

INTRODUCTION

High-throughput single-cell RNA sequencing (scRNA-seq) captures transcriptomes at the individual cell level. While scRNA-seq is powerful for a wide range of biological inference problems, perhaps, its most common application thus far is cell type/stage identification. Specifically, clustering analysis of scRNA-seq enables identification of cell types in tissues (1–3), or discrete stages in cell differentiation and development (4,5) by leveraging similarities in the transcriptomes of cells. Although a plethora of methods, some of which repurpose existing *k*-means and Louvain algorithms for clustering, exist for cell type identification from single conditions (6), clustering of scRNA-seq data from multiple biological conditions (e.g. different treatments, time points, tissues) to elucidate cell types and subpopulation of cells has not been a major focus.

Joint clustering of scRNA-seq datasets across multiple conditions entails, in addition to standard normalization, feature selection, and dimension reduction, a consideration of whether or not data from cells across multiple stimuli need to be 'integrated' before the downstream analysis of clustering. This is because cell type-specific response to experimental conditions may challenge a joint analysis by separating cells both by experimental condition and cell type. Most notably, Seurat (v3) (7) uses canonical component analysis (CCA) (8) to perform data integration for high dimensional gene expression across conditions. Liger (9) achieves dimension reduction and data integration simultaneously by using penalized nonnegative matrix factorization (NMF) to estimate factors shared across conditions and specific to conditions. ScVI (10) and scAlign (11) both use deep neural networks to infer a shared nonlinear low-dimensional embedding for gene expression across conditions. Harmony (12) iteratively performs soft *k*-means clustering and condition effect removal based on clustering assignments. A common theme of these approaches is that majority of them (Liger, scAlign, scVI and Harmony) directly yield low-dimensional integrated data for downstream visualization and clustering.

While existing methods for joint analysis of scRNA-seq data (e.g. Seurat, scAlign, and Liger among others) regularly adapt modularity maximization algoritms such as Louvain graph clustering (13) with shared nearest neighbor (SNN) graph or shared factor neighborhood (SFN) graph built with their low-dimensional embeddings of the data, they vastly differ in their dimension reduction techniques as we outlined above. This leads to documented notable differences among these methods (14). To leverage strengths of different dimensionality reduction approaches, we develop an integrative framework named multilayer

*To whom correspondence should be addressed. Tel: +1 608 265 4384; Email: keles@stat.wisc.edu

graph (MLG) clustering as a general approach that borrows strength among a variety of dimension reduction methods instead of focusing on a single one that best preserves the cell-type specific signal. MLG takes as input a set of low-dimensional embeddings of all the cells, integrates them into a shared nearest neighbor graph with analytically provable improved signal-to-noise ratio, and clusters them with the Louvain algorithm, which has recorded outstanding performances in benchmarks (15,16) and is also the default clustering algorithm in many scRNA-seq analysis packages such as Scanpy (17), Seurat (v2, v3) (7,18) and Liger (9). MLG framework leverages a key observation that the nearest neighbor graphs constructed from different dimensionality reductions of scRNA-seq data tend to have low dependence. A consequence of this observation, supported by analytical calculations, is that the resulting multilayer integrative scheme yields a combined cell graph with higher signal-to-noise ratio. We further corroborate this result with computational experiments using benchmark data and illustrate that MLG clustering outperforms current best practices for jointly clustering cells from multiple stimuli and preserves salient structures of scRNA-seq data from multiple conditions. We illustrate this property of MLG clustering with an application to scRNA-seq data from mouse hematopoietic stem and progenitor cells (HSPCs) under two conditions (with or without a *Gata2* enhancer deletion)(19), and from mouse HSPCs under four conditions ($Gif1^{+/+}$, $Gif1^{R412X/R412X}$, $Gif1^{R412X/-}$, $Gif1^{R412X/-}$ $Irf8^{+/-}$) (20). Finally, we showcase how MLG enables robust analysis of recent SNARE-seq (21) data which generates two data modalities, accessible chromatin and RNA, within the same cells.

MATERIALS AND METHODS

Overview of multilayer graph clustering (MLG)

The MLG clustering algorithm is a general framework that aggregates shared nearest neighborhood graphs constructed from different linear and non-linear low-dimensional embeddings of scRNA-seq data (Figure 1A). It takes as input G sets of low-dimensional embeddings of the same dataset generated by different dimensionality reduction methods, with and/or without data integration for datasets across multiple conditions. Consequently, it constructs and then aggregates SNNs from each of the G embeddings and leverages Louvain modularity (13) maximization algorithm for the final clustering. Figure 1A depicts the workflow of the MLG with four existing scRNA-seq low-dimensional embedding methods as inputs. Here, we considered dimension reduction with PCA and consensus non-negative matrix factorization (cNMF) (22) as representatives of low-dimensional embeddings without data integration across multiple conditions, and Seurat (7) and Liger (9) as representatives with data integration. Next, we describe the construction of the SNN graphs in detail.

Let $D \in \mathbb{R}^{n \times d_0}$, where n and d_0 represent the number of cells and the dimension of latent factors, respectively, denote a low-dimensional embedding matrix. As in Figure 1 A, the matrix D can be obtained from principal component analysis (PCA), consensus nonnegative matrix factorization (cNMF) (22) or low-dimensional embeddings from scRNA-seq analysis packages such as scVI, Liger applied to scRNA-seq count matrices. For each cell i, let $L_k(i)$ denote the set of k nearest neighbors based on Euclidean distance across the d_0 latent factor vectors. An edge is added to the undirected SNN graph (i.e. A_{ij} is set to 1 in the corresponding adjacency matrix A) between cells i and j if their number of common neighbors is larger than $k\lambda$, where λ is a filtering threshold. The parameters k and λ determine the sparsity level of the SNN graph. A large k and small λ lead to a denser graph. While there is no optimal criteria to pick these parameters, Von Luxburg (23) suggests $k = O(n)$ to guarantee a connected graph. In the analyses provided in this paper, we used $k = 20$ and $\lambda = 1/5$. We also showed with our simulation study that clustering accuracy is relatively robust to the choice of k.

Next, given G adjacency matrices $\{A^{(g)}\}$, $g = 1, \cdots, G$, each corresponding to an SNN from a low-dimensional embedding, we aggregate them into a union adjacency matrix B as follows:

$$B_{ij} = \begin{cases} 1 & \text{if } A_{ij}^{(g)} = 1 \text{ for any g,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

MLG then uses this resulting adjacency matrix B for clustering with the Louvain modularity (13) maximization algorithm.

Benchmark datasets

We leveraged three public scRNA-seq datasets, summarized as (log-normalized) counts and with complementary characteristics, to benchmark MLG (Supplementary Tables S1 and S2). The first dataset, that we refer to as *Kowalczyk_1*, profiled transcriptomes of HSPCs among mice of different ages (*young* at 2-3 months and *old* at 22 months) and harbored cells from three hematopoietic stages: long-term (LT)-HSC, short-term (ST)-HSC and multi-potent progenitor (MPP) (24). We used this dataset for explicitly illustrating how different dimension reduction algorithms operate on scRNA-seq data across multiple conditions. An extended version of this dataset, labelled as *Kowalczyk_2*, included additional LT-HSC and ST-HSC cells from independent mice that were profiled months apart from the original dataset. A third dataset from the HSPC system (25) included the same cell types (LT-HSC, ST-HCS and MPP) from young and old mice and with and without LPS+PAM stimuli. This dataset (referred to as *Mann*) is an example with two factors at two levels and represents a setting with four experimental conditions. The third dataset, labelled as *Cellbench* (26), included scRNA-seq data of 636 synthetic cells created from three human lung adenocarcinoma cell lines HCC827, H1975 and H2228. RNA was extracted in bulk for each cell line. Then each cell line's RNA was mixed in at seven different proportions and diluted to single cell equivalent amounts ranging from 3.75 to 30 pg.

Aggregation of multiple SNNs boosts signal-to-noise ratio

Next, we define a notion of signal-to-noise ratio for graphs and show that MLG aggregation of SNNs boosts the signal-to-noise ratio of the sparse graphs typically found in scRNA-seq analyses. We base our theoretical analysis on stochastic block models (SBMs) (27), which are generative

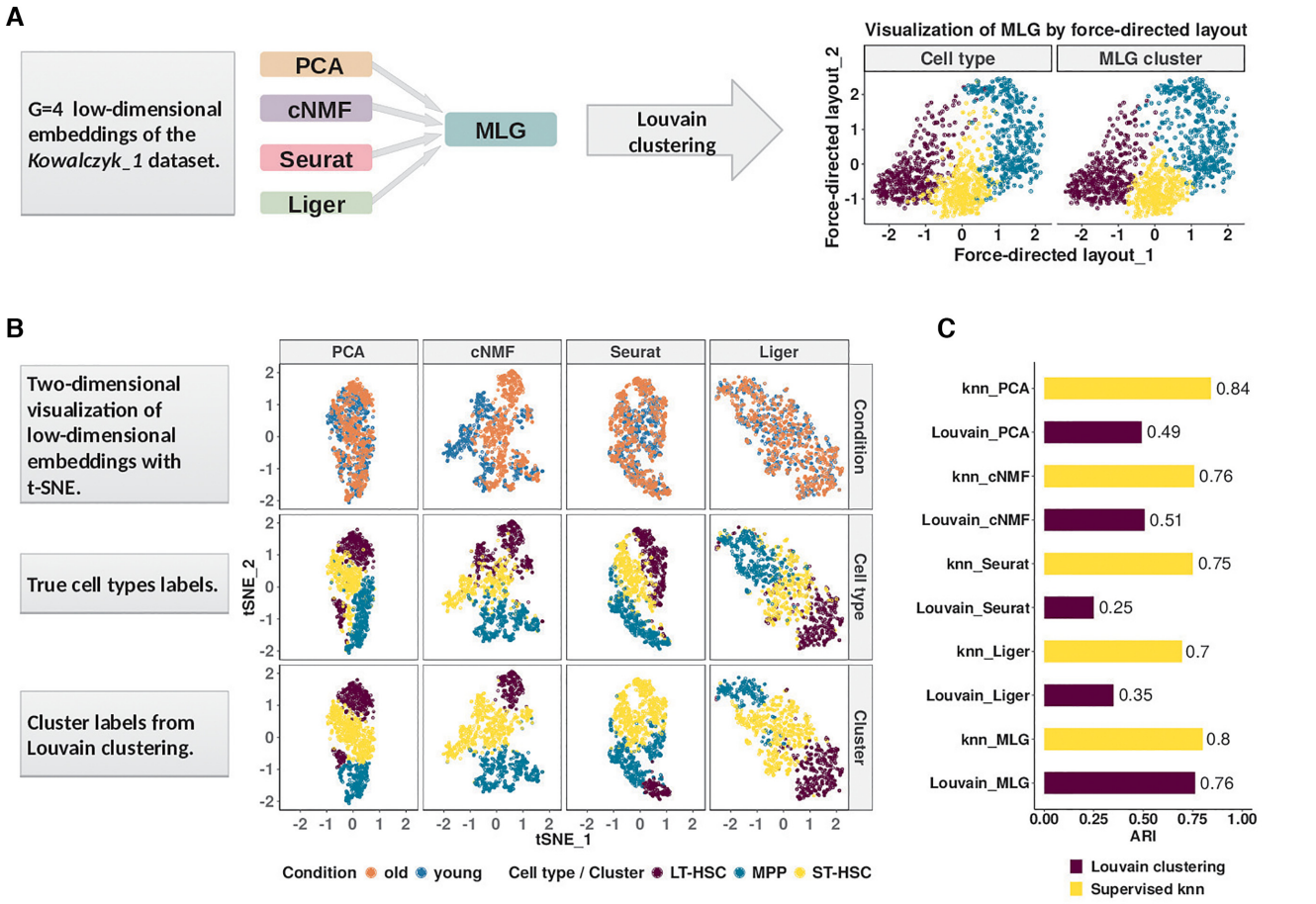

Figure 1. Multilayer graph (MLG) clustering workflow with an illustration on the *Kowalczyk_1* dataset. (**A**) MLG takes as input G different low-dimensional embeddings to construct SNN graphs, aggregates the resulting graphs, and applies Louvain algorithm for clustering the aggregated graph. The two scatter plots on the right visualize the MLG clustering results with a force-directed layout where the cells are labeled with their true labels (left) and MLG clustering labels (right). (**B**) Visualization of the PCA, cNMF, Seurat integration and Liger factors with t-SNE coordinates. Cells are labeled according to underlying experimental conditions (top row), cell types (middle row), and Louvain clustering assignments from SNN graphs constructed with each dimension reduction method (bottom row). (**C**) Adjusted Rand Index (ARI) between true cell type labels and labels from clustering/knn classification.

random graph models that serve as canonical models for investigating graph clustering and community detection. In the view of SBMs, cells are represented as vertices in a binary graph with edges representing similarity between cells. The cells are assumed to reside in distinct communities determined by cell-type, and this community structure determines the probability of an edge between cells. We denote the total number of cells (vertices) by n and the number of communities (clusters) by K. Let σ be a cluster assignment function where $\sigma(i)$ corresponds to the cluster label for cell i, and $\theta_{\sigma(i)\sigma(j)}$ denotes the connectivity probability (i.e. edge probability) between cells i and j. Furthermore, let θ_{in} and θ_{out} denote the minimum in-cluster and maximum out-of-cluster connectivity probability. We define the signal-to-noise ratio of the SBM as

$$\tilde{I} := \frac{(\theta_{\text{in}} - \theta_{\text{out}})^2}{\theta_{\text{in}}}. \tag{2}$$

Intuitively, if the difference between the minimum in-cluster and maximum out-of-cluster connectivity probabilities is large, it will be easier to distinguish communities from one another. Furthermore, this intuition is supported by theory, as Zhang *et al.* (28) prove that the performance of an optimal SBM estimator depends heavily on $\tilde{I}$. We provide a rigorous explication of this theoretical result in the Supplementary Materials Section 1.1.

We utilize this specific notion of signal-to-noise ratio to investigate the impact of aggregation. Given adjacency matrices A_1 and A_2 of two independent SBM graphs on the same set of cells and their union adjacency matrix B, we use superscripts to indicate parameters of each graph, e.g., $\theta_{k\ell}^{A_1}$ represents the connectivity probability of cells in cluster k and cells in cluster ℓ in graph A_1. In the realm of scRNA-seq data A_1 and A_2 are usually quite sparse, i.e., with small $\theta_{\text{in}}^{A_1}$ and $\theta_{\text{in}}^{A_2}$. If we further assume that $\theta_{\text{in}}^{A_1} = \theta_{\text{in}}^{A_2}$ and $\theta_{\text{out}}^{A_1} = \theta_{\text{out}}^{A_2}$, we arrive at the primary result of this section

$$\tilde{I}^B \geq \tilde{I}^{A_1} \times \frac{(2 - \theta_{\text{in}}^{A_1} - \theta_{\text{out}}^{A_1})^2}{2 - \theta_{\text{in}}^{A_1}}, \tag{3}$$

which implies that the signal-to-noise ratio of the aggregated SBM is nearly twice that of either of the base graphs,

A_1 and A_2. We provide a detailed derivation of this result in the Supplementary Materials Section 1.1.

This result on improved signal-to-noise ratio due to aggregation is foundational for establishing operational characteristics of MLG. It assumes that the SBMs are both sparse and independent of one another. As we show in the Results section, empirical observation also supports this assumption as SNN graphs are sparse, and we see only small proportions of overlapping edges between SNN graphs of different low-dimensional projections. In the Supplementary Materials Section 1.2, we further show that dimension reduction leads to perturbation in the local neighborhood structure of the full-data (e.g. data prior to dimension reduction) SNN graph. This, in turn, indicates that the dependence between SNNs derived from different low-dimensional projections is weak, and thus the overall assumptions of our result on the benefits of aggregation are approximately met. This result also explains the empirical behavior concerning low proportions of overlapping edges between SNN graphs that we observe across our benchmarking datasets in the Results section.

Signal-to-noise ratio estimation

For benchmark datasets with true cell type labels, an empirical version of signal-to-noise ratio for a given cell graph can be calculated as:

$$\mathrm{SNR} = \frac{(\hat{\theta}_{\mathrm{in}} - \hat{\theta}_{\mathrm{out}})^2}{\hat{\theta}_{\mathrm{in}}},$$

where $\hat{\theta}_{\mathrm{in}}$ stands for the estimated minimum 'within cell type' connectivity probability and $\hat{\theta}_{\mathrm{out}}$ stands for the estimated maximum 'out-of-cell type' connectivity probability. Let $A \in \{0, 1\}^{n \times n}$ denote the adjacency matrix of a cell graph with n cells from K cell types and let C_k denote the set of cells in cell type k. Then, $\hat{\theta}_{\mathrm{in}}$ and $\hat{\theta}_{\mathrm{out}}$ are given by

$$\hat{\theta}_{\mathrm{in}} = \min_{k \in \{1,\ldots,K\}} \frac{2 \times \sum_{i,j \in C_k, i<j} A_{ij}}{\left(|C_k| \times (|C_k| - 1)\right)},$$

$$\hat{\theta}_{\mathrm{out}} = \max_{k \neq \ell, k, \ell \in \{1,\ldots,K\}} \frac{2 \times \sum_{i \in C_k, j \in C_\ell, i<j} A_{ij}}{\left(|C_k| \times (|C_\ell|)\right)}.$$

Details on the execution of different scRNA-seq analysis methods

PCA. PCA is implemented through the `Seurat (version 3.1.4)` package. Specifically, we used `Seurat` function `NormalizeData` to scale counts with a size factor of 10,000 and then performed log-transformation while adding a count of 1 to avoid taking log of 0. The functions `FindVariableFeatures` and `ScaleData` were used to find highly variable features and adjust for technical and biological effects, respectively. The condition labels and total gene counts were regressed out with `ScaleData` for all simulated, benchmark, and application datasets. The percentage of mitochondrial gene counts were also regressed out for the dataset *Johnson20*. Each gene is scaled by its standard error of counts across cells before performing PCA with the `RunPCA` function.

cNMF. cNMF is applied through the code provided by (22) on Github https://github.com/dylkot/cNMF. Count matrices were provided as input for cNMF. Normalization and feature selection were carried out within the cNMF pipeline. The consensus analysis used 50 NMF runs.

Seurat-integration. Following the `Seurat` tutorial, function `SplitObject` was used to split each dataset by conditions, then `FindIntegrationAnchors` and `IntegrateData` were applied to integrate gene expression across conditions. The integrated gene expression matrix was scaled with function `ScaleData`. *RunPCA* was used to reduce the dimensionality of the scaled and integrated gene expression matrices.

Liger. Gene expression count matrices were normalized with the R package `rliger(1.0.0)` function `normalize`. Highly variable genes were chosen with `selectGenes`. Functions `scaleNotCenter`, `optimizeALS`, `quantile_norm` and `louvainCluster` were utilized to scale gene expression, integrate data and cluster cells. We used the low-dimensional embedding in data slot `@H` by scaling it to have row sum of 1 in SNN graph construction for MLG.

scAlign. Our application followed the tutorial of `scAlign` on Github https://github.com/quon-titative-biology/scAlign. To reduce computation time, we used PCA factors as input for encoder neural networks.

Harmony. We applied `Harmony` through the *Seurat* workflow with the `RunHarmony` function. `nclust` parameter was set to the true numbers of clusters for both the simulated and benchmark datasets.

scVI. `ScVI` is implemented through the `scVI` python package, available at https://scvi.readthedocs.io/en/stable/. We followed its 'basic tutorial' for model training, and used sample latent factors as input for k-means or Louvain clustering.

Parameters common to all methods. We kept 3000 genes for all the benchmark scRNA-seq datasets and 2500 genes for the simulated datasets. We kept 15 latent factors from all dimension reduction or data integration methods for the analysis of the benchmark scRNA-seq datasets. We applied k-means with the `R` function `kmeans` and set the numbers of clusters to the true number of clusters for all the simulated and benchmark datasets. Louvain clustering was applied through `Seurat::FindClusters`. The resolution was chosen through grid search until the target number of clusters was met. The target number of clusters were set to the true numbers of clusters for the benchmark datasets. The numbers of clusters for *Johnson20* and *Muench20* was chosen by the eigengap heuristic (23).

Differential expression analysis

Differential expression analysis was carried out with function `Seurat::FindMarkers` using the 'MAST' algorithm. In the case of multiple conditions per dataset, cluster

markers were identified with a differential expression analysis of cells from the same condition across different clusters and the *P*-value was set to be the smallest among all conditions.

The precision-recall analysis was performed based on gold standard cell type marker genes and cell type-specific DE genes across conditions, defined as having Bonferroni corrected *P*-values <0.01 in the analysis with the true cell labels. The precision-recall values were reported at cutoffs of 0.2, 0.1, 0.05, 0.01 and 0.001 for Bonferroni adjusted *P*-values in the cell type marker gene and condition DE gene identification analysis.

Average mixing metric

We utilized the mixing metric proposed in (7), implemented as function `MixingMetric` in `Seurat`, to quantify how well cells among different groups are mixed. The metric operates by constructing an ordered nearest neighbor list S_i for each cell i with the default number of neighbors set as 300. It then computes the 5th nearest neighbor of cell i in each group j, and evaluates its rank in the ordered list S_i. The mixing metric for cell i is the median of these ranks over all groups. The average mixing metric is the mean of the mixing metrics for all cells in the dataset. The range of average mixing metric is from 5 to 300. A small average mixing metric indicates well mixed cells among groups.

Metric for evaluating how well clusters delineate established lineage markers

In order to evaluate how well different clustering results delineate known lineage-specific marker gene expression, we considered established lineage marker gene sets $\{g_i\}_{i=1}^{M}$ and cluster labels of the genes for clusters $\{1, 2, \ldots, K\}$. For each gene g_i, we denoted the cluster in which g_i is the most highly expressed as K_{g_i}. We then performed a two-sample differential expression test with null hypothesis that the expression of g_i in and out of cluster K_{g_i} are the same, and denoted the resulting gene-level p-value as p_{g_i}. Next, we combined the gene-level p-values using Fisher's combined probability test. The Chi-squared test statistic for this combined test is defined as

$$-2\sum_{i=1}^{M}\ln(p_{g_i}) \sim \chi^2_{2M}.$$

Clustering methods resulting in larger Chi-squared test statistic were considered as delineating the lineage markers better. Specifically, we used the R package `scran (version 1.20.1)` (29) to conduct the differential expression test instead of the R package `Seurat` to avoid truncation of the small *P*-values into 0 by `Seurat`.

Gene set enrichment analysis

Gene set enrichment analysis was carried out with the `R` package `topGO(version 2.36.0)` using the Fisher's exact test and `elim` (30) algorithm.

Simulations

Our two main simulation set-ups are based on the commonly used R package `Splatter` (31) for simulating scRNA-seq data and an adaptation of the simulation setting of (22) which takes advantage of `Splatter` and allows differential expression across conditions. In the latter set-up, the mean gene-expression profile for each cell is a weighted sum of *cell identity gene expression program* (GEP) and *activities* GEPs. Here, the cell identity GEP characterizes cell types whereas the activity GEP represents other sources of biological variations, like cell cycle effects or responses to specific stimuli. This set up also mimics the datasets with multiple cell types under stimuli (e.g., HSPCs from old and young mouse), that we have utilized in this paper. Specifically, each simulation replication involved the following data generation process.

1. **Simulate base mean expression.** For gene w ($w = 1, \ldots, W$), sample the magnitude of mean expression $\lambda'_w \sim \text{Gamma}(\alpha, \beta)$, the outlier indicator $p^o_w \sim \text{Ber}(\pi^o)$, and the outlier factor $\psi'_w \sim \text{Lognormal}(\mu^o, \sigma^o)$. The base mean expression for gene w is defined as $\lambda_w := p^o_w \psi'_w \text{Median}_w(\lambda'_w) + (1 - p^o_w)\lambda'_w$.
2. **Simulate gene expression program (GEP).** For gene w, sample differential expression (DE) indicator $p_w \sim \text{Ber}(\pi)$, DE factor $\psi_w \sim \text{Lognormal}(\mu, \sigma)$, and down regulation indicator $p^d_w \sim \text{Ber}(\pi^d)$. Let $\tilde{\psi}_w := \psi_w \mathbf{1}_{\{\psi_w > 1\}} + \frac{1}{\psi_w}\mathbf{1}_{\{\psi_w \le 1\}}$. The DE ratio for gene w is defined as $\delta_w := (1 - p_w) + p_w p^d_w \frac{1}{\tilde{\psi}_w} + p_w(1 - p^d_w)\tilde{\psi}_w$. The GEP is given by $\mathcal{G} := \left(\lambda_w \times \delta_w\right)_{w=1}^{W}$, where W is the total number of genes. Here, λ_w and δ_w denote the base gene mean and DE ratio of gene w, respectively. This procedure can be used to simulate identity and activity GEPs, with slightly different set of parameters (π, π^d, μ, σ).
3. **Simulate mean gene expression for cell type t and condition c ($t = 1, \ldots, T$, $c = 1 \cdots C$).** Following the procedure in step 2, simulate identity GEP $\mathcal{G}_1^{identity}, \ldots, \mathcal{G}_T^{identity}$ and activity GEP $\mathcal{G}_1^{activity}, \ldots, \mathcal{G}_A^{activity}$. Given activity GEP weights $\{u^{tc}_a\}_{a=1}^{A}$ for cell type t and condition c, the cell type and condition specific mean gene expression is defined as $\Lambda^{tc} := \sum_{a=1}^{A} u^{tc}_a \mathcal{G}_a^{activity} + (1 - \sum_{a=1}^{A} u^{tc}_a)\mathcal{G}_t^{identity}$. Activity GEP weights control the magnitude of condition effects.
4. **Correct for library size.** For cell k in cell type t and condition c, sample the library size for cell k $L_k \sim \text{Lognormal}(\mu^l, \sigma^l)$. The mean gene expression of cell k after correction for the library size is $\Lambda^L_k := \frac{\Lambda^{tc}}{(\Lambda^{tc})^T \mathbf{1}} \times L_k$.
5. **Correct for biological coefficient of variation (BCV).** Let ϕ be the BCV dispersion parameter, `df` be the degrees of freedom of the BCV inverse χ^2 distribution. The BCV for gene w of cell k is sampled through the formula $B_{k,w} = (\phi + 1/\sqrt{\Lambda^L_{k,w}})(\texttt{df}/\chi^2_{\texttt{df}})^{1/2}$. The BCV corrected mean for gene w of cell k is sampled by $\Lambda^*_{k,w} \sim \text{Gamma}(1/B^2_{k,w}, \Lambda^L_{k,w} B^2_{k,w})$.
6. **Generate counts.** Sample counts for gene w of cell k $Y_{k,w} \sim \text{Pois}(\Lambda^*_{k,w})$.

In order to make these simulations realistic, we set the parameters in a data-driven way. Specifically, the parameters of the simulation setting ($\alpha = 1.46, \beta = 1.48, \pi^o = 0.091, \mu^o = 2.82, \sigma^o = 0.84, \mu^l = 8.95, \sigma^l = 0.45, \phi = 0.11,$ `df` $= 36.57$) are estimated from the *Johnson20* dataset with the `Splatter` package. We considered four settings to explore the robustness of MLG to different magnitude of condition effects and its ability to identify rare cell types (condition specific parameters are summarized in the Supplementary Table S3). By varying the magnitudes of the stimulus effect with the weight of the activity GEPs, we evaluated operating characteristics in the the cases with large (settings 1, 3) and small (settings 2, 4) condition effects. Similarly, by varying the proportions of cells in simulated cell types, we assessed performances in cases with both balanced cell proportions (settings 1, 2) and rare cell types (settings 3, 4).

An additional set of simulations, settings 5-8, are generated by `Splatter`, with designs similar to settings 1-4 (Supplementary Table S4). However, the advantage of the set-up with settings 1-4 is its ability to generate more nuanced condition effects than the Splatter simulated 'batch effects'. Splatter enables generation of multiple conditions in the form of batch effects by multiplying gene expression with a 'differential expression ratio'. Such an effect tends to be easy to remove by regressing out the batch labels in the preprocessing step (Supplementary Figures S5 and S7).

RESULTS

The multilayer graph (MLG) clustering algorithm

We first present a detailed illustration of MLG on the *Kowalczyk_1* dataset (hemotapoitic stem and progenitor cells from young and old mice). The first and second rows in Figure 1B depict the two-dimensional visualizations of the low-dimensional embeddings of the *Kowalczyk_1* dataset with t-SNE (32) labeled by condition and cell types, respectively. The third row presents the Louvain clustering labels for each embedding. A direct comparison of the second and third rows illustrates the inaccuracies of each method for cell type identification. A similar visual illustration of the MLG results is provided in the right most panel of Figure 1A. To evaluate the performances of these low-dimensional embeddings independent of clustering, we followed a supervised approach. We leveraged a k-nearest neighbor (knn) classifier to classify the cells, and computed the commonly used metric adjusted Rand index (ARI, (33)) between predicted class labels and the true labels. The ARI values from the knn classification represent the best achievable performances with these low-dimensional embeddings. In addition to this, we also quantified the clustering performance of each method by ARI (Figure 1C).

The ARIs of the knn classifier with PCA and cNMF factors are 0.84 and 0.76, whereas the ARI of knn classifier with Seurat and Liger integrated data are 0.75 and 0.70, respectively. While PCA and cNMF keep cells from the same cell type close to each other in their respective low-dimensional spaces, Seurat and Liger perform similarly well in aligning data across conditions. However, all four methods exhibit cell type mix-up and, as a result, none of the methods have an ARI higher than 0.51 when these low-dimensional embeddings are clustered (Figure 1C). In the context of graph clustering, cells are partitioned so that cells within the same cluster are densely connected, while cells in different clusters are loosely connected. We observe for this dataset that, with dimension reduction without data integration, the clustering algorithm tends to separate cells under different conditions, e.g., old and young LT-HSCs with cNMF factors. However, after Seurat and Liger data integration, graph clustering tends to separate MPP cells into different clusters. To leverage strengths of different dimension reduction strategies, MLG clustering first constructs shared nearest neighbor graphs (SNNs) from each of the low-dimensional embeddings. Then, it aggregates the adjacency matrix of the resulting graphs with a union operation and employs modularity maximization (13) to cluster the resulting graph. By aggregating the SNNs obtained from each dimension reduction approach, MLG boosts the signal-to-noise ratio and improves the ARI from individual methods by 25% to 0.76.

In the following sections, we first provide explicit examples supporting the analytical underpinnings of MLG clustering as outlined in the Methods section and demonstrate how different low-dimensional embeddings can complement each other. We then evaluate MLG clustering on both simulated and benchmark datasets and compare it with state-of-the-art methods. Next, we discuss a weighting scheme to enable incorporation of additional low-dimensional embeddings into MLG and illustrate how this weighting scheme provides robustness. We showcase MLG in two separate scRNA-seq applications involving cells from two (19) and four different conditions (20), respectively. We further discuss how MLG can also be adapted beyond the analysis of scRNA-seq to SNARE-seq (21) which profiles transcriptome and chromatin accessibility from the same cells simultaneously.

Aggregating signal from shared nearest neighbor graphs of multiple low-dimensional embeddings boosts the 'signal-to-noise' ratio

A challenging aspect of generating low-dimensional embeddings of scRNA-seq data across multiple conditions is that, different dimensionality reduction methods might capture different aspects of the data. As a result, graph representations of the data constructed for downstream clustering might vary significantly between different lower-dimensional embeddings. In the HSCs of old and young mice (*Kowalczyk_1* from Figure 1), SNN graphs constructed from PCA and cNMF factors of the gene expression count matrix (with and without data integration across conditions) have small proportions of overlapping edges (Figure 2A). Specifically, SNN graphs from any two low-dimensional embeddings, except the pair PCA and Seurat-integration, which also performs PCA as the final dimension reduction, have at most 30% of their edges overlapping, indicating low dependence between these constructions. One possible reason for this, as we argue analytically in the Supplementary Materials Section 1.2, is that dimension reduction alters neighbors of cells. Furthermore, we also observe that all of the constructions tend to be sparse

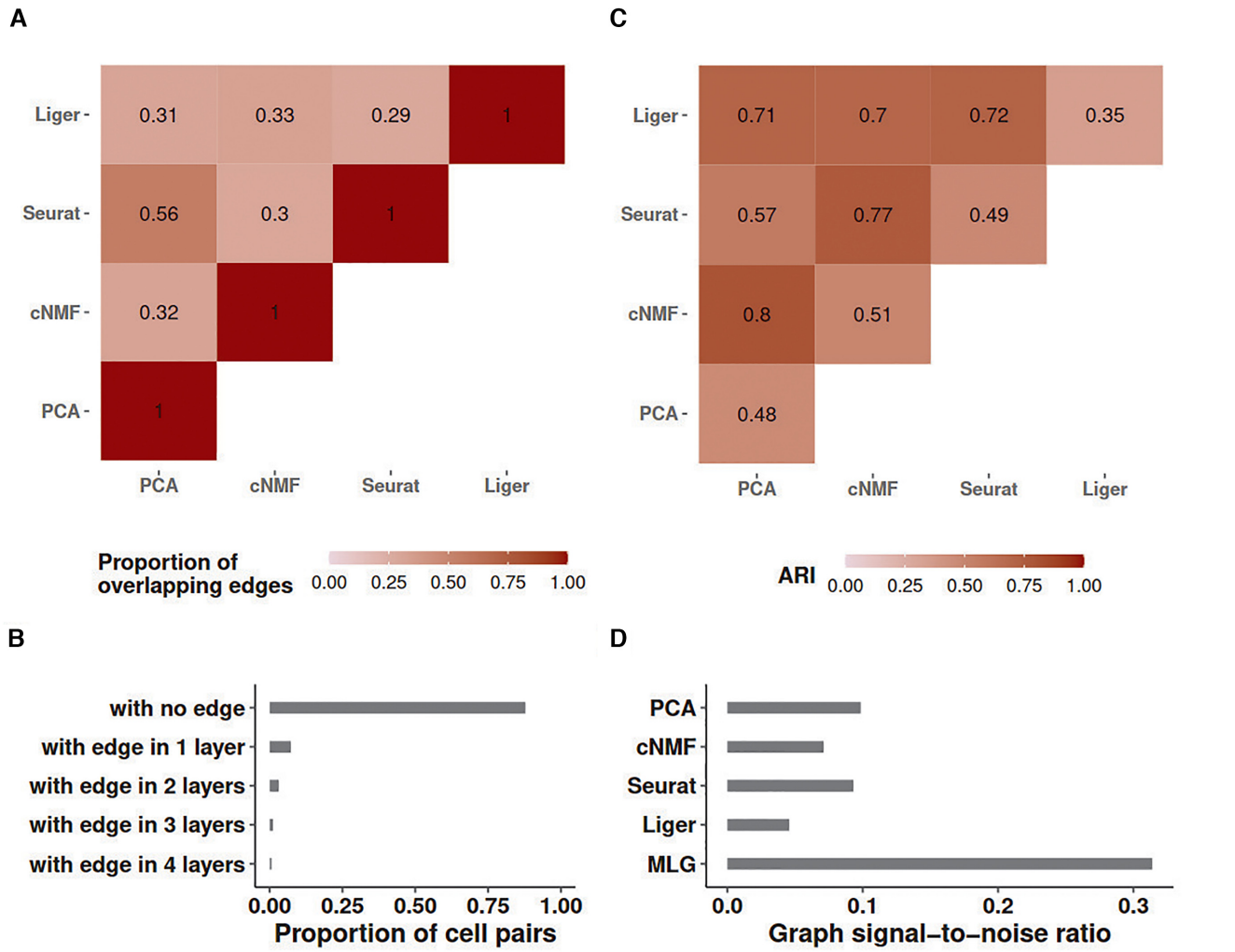

Figure 2. Graph characteristics of SNN graphs and the multilayer graph of the *Kowalczyk_1* dataset. (**A**) Heatmap of the proportion of overlapping edges between pairs of SNN graphs from different low-dimensional embeddings. (**B**) Proportion of cell pairs with edges across aggregation of individual SNN graphs from PCA, cNMF, Seurat and Liger low-dimensional embeddings. The number of layers represents the total number of individual SNN graphs that harbor edges between the cell pairs. (**C**) ARI of MLG clustering constructed by pairs of SNN graphs from the low-dimensional embeddings indicated as the rows and columns. Diagonal entries represent ARIs of SNN graph clustering from individual low-dimensional embeddings. (**D**) Estimated graph signal-to-noise ratios of individual SNN graphs and their multilayer graph.

as indicated by the proportion of cell pairs with no edges in the individual SNN graphs (bar labelled as 'with no edge' in Figure 2B).

Leveraging these empirical observations and recent advancements in stochastic block models, we show in the Supplementary Materials Section 1.1 that aggregating two sufficiently sparse graphs with independent edges by taking union of their edges leads to a graph with amplified 'signal-to-noise' ratio. Here, as we presented in the Methods section, we are using a notion of 'signal-to-noise' that refers to difference in connectivity of cells that are within the same cluster (i.e. cell type/stage) versus that are in different clusters. Figure 2C presents ARI values of MLG clustering based on pairs of SNN graphs (off-diagonal entries) and those of clusterings based on individual SNN graphs (diagonal entries). We observe that all the 2-layer MLG applications have improved clustering performance compared to SNNs from individual low-dimensional embeddings (e.g. Liger alone achieves an ARI of 0.35 whereas MLG with Liger and PCA low-dimensional embeddings as its two layers achieves 0.71). As expected, since the SNN graphs constructed with PCA and Seurat have more overlapping edges (56%), their 2-layer MLG results in the least improvement in ARI with a value of 0.57 compared to 0.48 and 0.49 with PCA and Seurat alone.

A majority of cell pairs (88%) do not have edges in any of the four layers of the SNN graph constructed from aggregation of PCA, cNMF, Seurat-integration, and Liger SNN graphs (Figure 2B). This is due to sparsity of individual SNN graphs. Furthermore, about 67% percent of the cell pairs with edges, have edges in only one layer, suggesting that different dimension reduction methods are capturing distinct features of the data. Aggregating the four layers with union operation increases the signal-to-noise ratio to three times of any single SNN graph layer (Figure 2D). A direct impact of this is increased clustering accuracy by MLG

compared to clustering of SNN graphs from individual low-dimensional embeddings. In fact, MLG clustering almost achieves the accuracy of supervised knn classifiers (0.76 versus 0.80, Figure 1C).

Increased 'signal-to-noise-ratio' by MLG aggregation translates into significant improvements in clustering accuracy and stability in computational experiments

To systematically evaluate the ability of MLG in improving clustering performance over clustering with individual graphs from specific low-dimensional embeddings, we conducted simulations from the settings described in Materials and Methods. We present here detailed results from setting 1 which is an adaptation of the general simulation setting from (22) where cell types exhibited condition specific effects. We generated multiple simulation replicates (100 for simulation settings 1 & 2 and 50 for settings 3-8), where each replicate included a total of 2500 cells from three cell types and across two conditions.

We considered four different dimension reduction procedures for MLG: PCA and cNMF, which do not perform comprehensive data integration, Seurat and Liger both of which perform data integration for cells from different conditions. We varied the apparent parameters of each method such as the numbers of PCA and NMF components and numbers of neighbors in the construction of SNN graphs. To compare with the MLG clustering, we applied both *k*-means and Louvain clustering on the SNN graphs constructed by the individual low-dimensional embeddings from these methods. Furthermore, we employed a supervised *k*-nearest neighbor classifier to establish the best achievable clustering performance for each graph in terms of ARI.

We first assessed the level of dependence between the SNN graphs constructed from these four low-dimensional embeddings across the simulation replicates and observed that the majority of the cell pairs were connected only in one of the SNN graphs. Specifically, only 3.6% of the pairs connected in at least one layer were common to all 4-layers (Figure 3A), a level comparable to 5.7% in the *Kowalczyk_1* benchmark dataset (5.7% is calculated by 'with edge in four layers' as a percentage of all pairs with edge in at least one layer in Figure 2B). Furthermore, the estimated signal-to-noise ratios of each graph supported the signal boost by MLG (Figure 3B). Figure 3D summarizes the ARI values of supervised knn, Louivan and *k*-means clustering with each individual low-dimensional embedding as a function of the numbers of PCA/cNMF components across the simulation replicates. We observe that MLG provides a median increase of 13%, 9%, 34% and 55% in ARI compared to Louvain clustering of individual PCA, cNMF, Seurat and Liger-based low-dimensional embeddings, respectively (first row of Figure 3D). Improvement in ARI by the MLG clustering with the Louivan algorithm is even higher (median levels of 85%, 9%, 80% and 45%) compared to *k*-means clustering of the low-dimensional embeddings by each of the four methods (first versus second rows of Figure 3C). Furthermore, MLG yields accuracy levels that are comparable to those of best knn accuracy in a supervised setting by these four methods (third row of Figure 3D). Since both the dimension reduction methods and SNN graph construction depend on key parameters such as the numbers of latent factors and numbers of neighbors, we varied these parameters in a wide range and observed robustness of MLG to the choice of these two parameters (Figure 3C, D).

Simulation setting 2 explores the case when the condition effect is relatively small (Supplementary Figure S2, with an average between conditions mixing metric of 13.4) and, therefore, data integration is not required. The MLG result with only two layers (PCA and cNMF) is presented in Supplementary Figure S1. There is an average of 15% and 11% increase in ARI over Louvain clustering using only PCA and cNMF factors (Supplementary Figure S1D). The clustering performance is also robust over a wide range of numbers of latent factors and numbers of neighbors (Supplementary Figure S1C, D). Simulation settings 5 & 6, with similar designs to settings 1 & 2, are generated using `Splatter` (Supplementary Figure S5) and include condition effects that can be captured and removed by PCA. In these settings, MLG and PCA have better performances than cNMF, Seurat and Liger (Supplementary Figure S6). In addition to simulations investigating large and small condition effects (settings 1, 2, 5 and 6), we also assessed the impact of rare cell types (settings 3 & 4, Supplementary Figure S3; settings 7 & 8, Supplementary Figure S7). We used both the Adjusted Rand Index (ARI) and probability of identifying rare cell types as performance metrics. For each simulated dataset, a method is deemed to successfully identify the rare cell type if the corresponding results included a cluster with more than 80% of its cells coming from the rare cell type. Then, the probability of identifying a rare cell type for a method is the average rate of rare cell type identification across all simulated datasets. As depicted in Supplementary Figures S4 and S8, MLG presents advantages over other methods in ARI and yields a rare cell types identification probability comparable to its best composite layer.

MLG clustering outperforms other methods in recovering known biological signal

We next compared MLG clustering to state-of-the-art scRNA-seq dimension reduction and data integration methods on the benchmark datasets with cells from multiple conditions and ground truth cell type labels, namely *Kowalczyk_1*, *Kowalczyk_2*, *Mann* and *Cellbench* (Figure 4). These datasets exhibit different levels of difficulty for clustering based on the average mixing metric (7), which ranges from 5 to 300 and quantifies how mixed different group of cells are (Figure 4 A). The separations between conditions are relatively small in datasets *Kowalczyk_1*, *Kowalczyk_2*, *Cellbench*, indicating that condition specific responses of cells do not dominate over their cell type-specific expression programs, and are markedly large in *Mann* with average mixing metrics of 10.3, 12.0, 13.7 and 27.0, respectively. In contrast, separations between cell types are low in *Mann* and high in *Cellbench*, with average mixing metrics of 41.8 and 278.7, respectively. This exposition of condition and cell type separation levels indicate that *Cellbench* is the easiest to cluster and *Mann* is the hardest. In these benchmarking experiments, we extended the set of methods we considered to include Harmony integration,

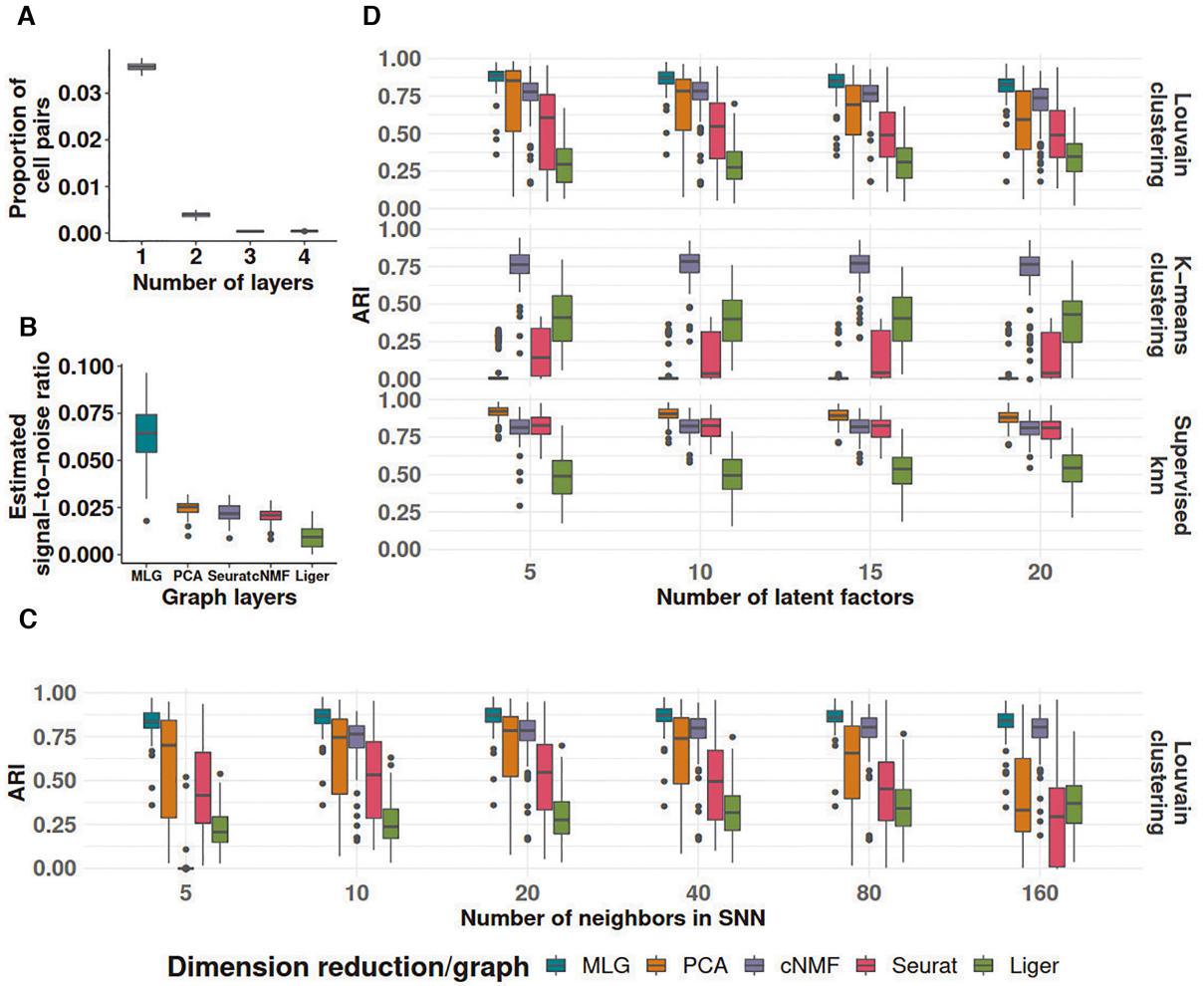

Figure 3. Simulation results for the 'large-condition-effect' setting. (**A**) Proportions of cell pairs with edges across different numbers of layers of MLG constructed from SNN graphs of PCA and cNMF, Seurat, and Liger (with 20 neighbors in SNN graphs and 10 latent factors in low-dimensional embeddings). The boxplots depict the proportions across all the simulation replicates. (**B**) Estimated signal-to-noise ratios of SNN graphs constructed from different low-dimensional embeddings and their multilayer graph across all the simulation replicates (with 20 neighbors in SNNs and 10 latent factors in low-dimensional embeddings). (**C**) Louvain clustering accuracy of SNN graphs as a function of numbers of neighbors in SNN graph construction (with 10 latent factors in low-dimensional embeddings). (**D**) Adjusted Rand index comparison of Louvain and *k*-means clustering of SNN graphs from different low-dimensional embeddings and their MLG as a function of number of latent factors in the low-dimensional projections (with 20 neighbors in SNN graphs). ARI values of supervised knn classifiers for individual SNN graphs are provided as reference.

scVI batch correction, and scAlign integration in addition to PCA and cNMF dimension reduction, Seurat integration, Liger integration. We performed both k-means clustering with the latent factors estimated from these dimension reduction/data integration methods and also Louvain clustering with graphs constructed from their low-dimensional embeddings (with package default weighted SNN for Seurat, package default SFN graph for Liger, unweighted SNN graphs for all other methods). As an overall measurement of the difficulty of the clustering task, we performed supervised knn classification using SNN/SFN graphs and observed largely similar supervised knn accuracy from different low-dimensional embeddings (Figure 4B) (with scAlign on *Cellbench* and *Mann* datasets as notable exceptions). The accuracies of k-means and Louvain clustering with graphs from individual low-dimensional embeddings are markedly lower than their corresponding knn classifiers for datasets *Kowalczyk_1*, *Kowalczyk_2*, and *Mann*. Harmony, Seurat and PCA perform well for *Cellbench* which is an easier dataset in terms of clustering since the separation between the cell types is large (Figure 4A). MLG clustering outperforms the alternatives across all datasets with a minimum of 13% and a maximum of 64% increase in ARI. Each cluster identified by MLG has a dominating cell type, as depicted in Supplementary Figure S12. Furthermore, as apparent from the Louvain clustering results of individual low-dimensional embeddings in Figure 4B, MLG does not require each individual SNN graph it aggregates over to perform well. For example, while Seurat, Liger, PCA and cNMF have individual ARIs of 0.35, 0.11, 0.10, 0.00, respectively, by combining SNN graphs resulting from these, and boosting signal-noise-ratio, MLG outperforms their individual performances with an ARI of 0.51 for the *Mann* dataset that appears to be the most challenging to cluster.

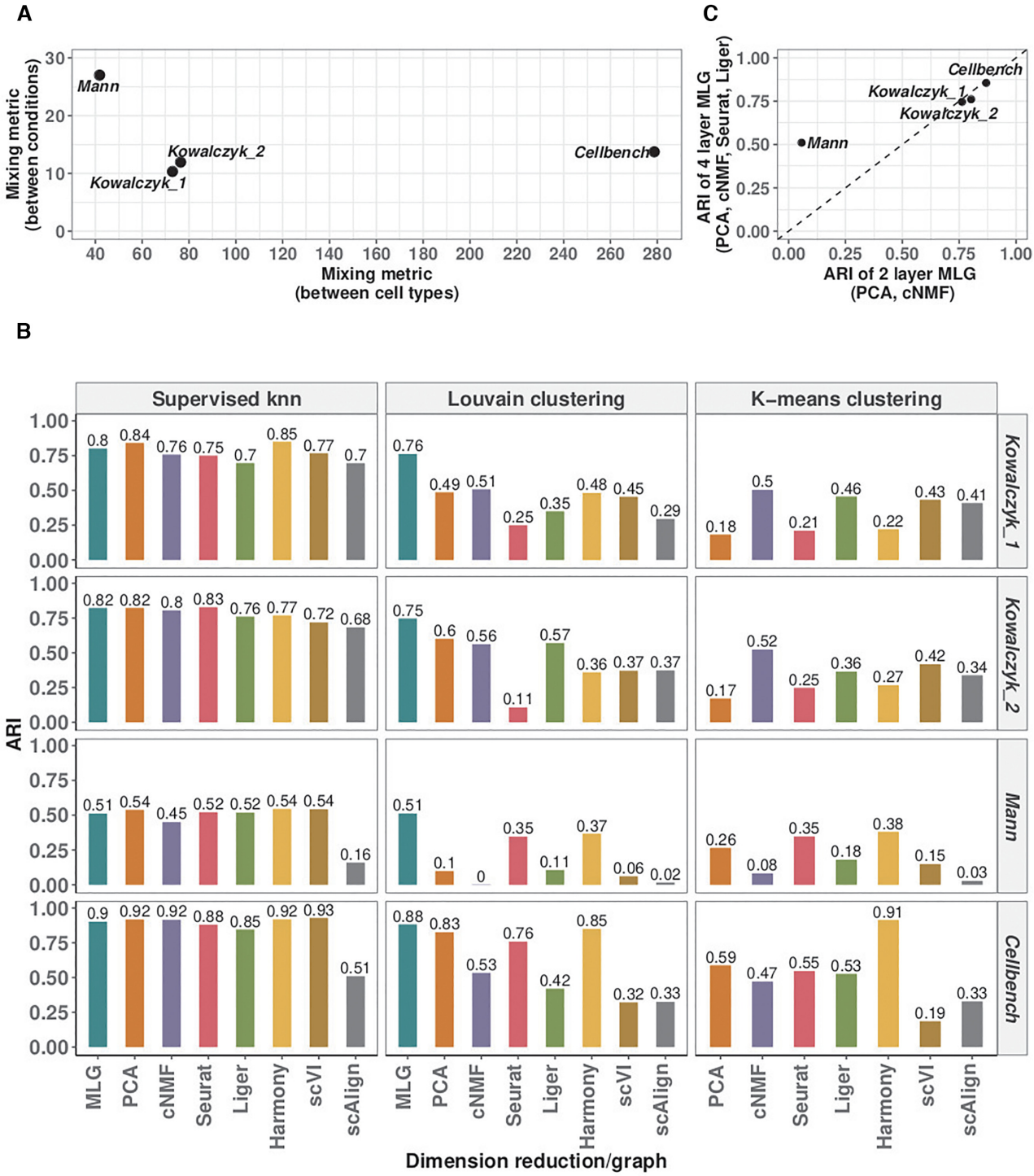

Figure 4. Evaluation with benchmark datasets. (**A**) Average mixing metric between cell types (x-axis) and conditions (y-axis) for the four benchmark datasets. The mixing metric ranges between 5 and 300, with larger values indicating large separation between the groups (e.g. cell types or conditions). (**B**) Louvain and k-means clustering accuracy of different methods across the benchmark datasets. Supervised knn classification for SNN graphs from each low-dimensional embedding is provided as a measure of the difficulty of the clustering task. (**C**) Comparison of the 4-layer MLG (PCA, cNMF, Seurat-integration, Liger) with the 2-layer MLG (PCA, cNMF) across the benchmark datasets.

We leveraged these benchmark datasets to further investigate the impact of different low-dimensional methods aggregated by MLG. Figure 4C displays the ARI values of 2-layer MLG which aggregates over only PCA and cNMF and 4-layer MLG which also aggregates over low-dimensional embeddings from data integration with Seurat and Liger. Empirically, when condition separation, as measured by the average mixing metric and also visualized by tSNE or UMAP visualizations of the data prior to integration, is low as in *Kowalczyk_1, Kowalczyk_2* and *Cellbench* datasets, the two MLG strategies perform similarly; however, for datasets with larger separation between conditions (*Mann*), MLG benefits significantly from aggregating over low-dimensional embeddings from data integration.

MLG strategy is robust to imbalanced cell type representations across conditions

When considering scRNA-seq datasets across multiple conditions, a key challenge is the ability to identify distinct cell types in cases with varying levels of representation of cell types under different conditions. Among the benchmark datasets we considered, *Kowalczyk_1* and *Mann* have relatively balanced cell types in different conditions. Specifically, in the *Kowalczyk_1* dataset, the three different cell types (LT-HSC, MPP, ST-HSC) vary at proportions of (0.17, 0.18, 0.18) and (0.16, 0.16, 0.15) among the old and young mice, respectively. Similarly, in the *Mann* dataset, the three different cell types (LT-HSC, MPP, ST-HSC) vary at proportions of (0.11, 0.02, 0.09), (0.12, 0.05, 0.12), (0.04, 0.03, 0.07) and (0.10, 0.10, 0.14) among the old-no-stimuli, old-stimuli, young-no-stimuli and young-stimuli mice conditions, respectively. Taking advantage of these balanced datasets, we conducted a computational experiment to investigate performance under imbalanced cell type representations across the conditions. Specifically, we subsampled 'old MPP' cells in *Kowalczyk_1*, and 'old-stimuli-MPP' cells in *Mann*, at proportions 0, 0.10, 0.25, 0.50, 0.75 and 1 of the original size, where 1 corresponded to the original dataset. Figure 5A reports the mean, 5, and 95 percentiles of the ARI values across 20 sub-sampling replications for the 2-layer MLG (PCA and cNMF), 4-layer MLG (PCA, cNMF, Seurat, Liger) and Louvain and *k*-means clustering with PCA, cNMF, Seurat, Liger, Harmony latent factors. For *Kowalczyk_1*, 2-layer MLG outperforms other methods and is robust to the imbalance of the cell types between conditions. PCA and cNMF can accommodate the small separation between conditions of this dataset without any explicit data integration (Figure 4B). However, methods utilizing data integration are affected by the misalignment of cells, which in turn reduces their clustering accuracy (Seurat, Liger). The 4-layer MLG is also relatively robust to cell type imbalance despite its aggregation over Seurat and Liger, and outperforms clustering with all individual low-dimensional embeddings in accuracy and stability. Since the *Mann* dataset has larger separation between conditions, aggregation over only methods without data integration (2-layer MLG) results in similarly poor performance as PCA and cNMF. In contrast, the 4-layer MLG, by aggregating over data integration methods Seurat and Liger in addition to PCA and cNMF, has the highest ARI. Compared to the balanced case (labeled as 1 in Figure 5A), 4-layer MLG suffers from accuracy loss because of mis-aligned cells in integration; however, it still performs better than using just one layer of integrated data.

MLG clustering across multiple stimuli leads to more powerful downstream differential expression analysis

We evaluated the impact of improvement in clustering accuracy by MLG clustering on the downstream analysis of identifying marker genes of individual cell types and cell type-specific differentially expressed (DE) genes across conditions. We first generated 'gold standard' marker genes and lists of differentially expressed genes using the true labels of the cells in the benchmark datasets (Materials and Methods). Next, we identified cluster-specific marker genes and lists of DE genes across conditions using the cluster assignments obtained with MLG clustering and the alternative methods used for the benchmark datasets. We evaluated whether more accurate separation of the cell types by MLG leads to marker and DE gene identification that aligns better with the gold standard sets with precision-recall (PR) curves (Figure 5B for the *Kowalczyk_1* datasets, Supplementary Figures S10, S11 for the *Kowalczyk_2* and *Mann* datasets). Overall, MLG exhibits better PR values for both cell type marker and condition DE gene identification across multiple datasets. More specifically, MLG is the only method with moderate to high precision-recall in identifying cell type marker genes for ST-HSC, whereas all other methods perform poorly in identifying marker genes specific to this cell type (Figure 5B).

A weighting scheme enables robust incorporation of additional composite layers to MLG

We have considered four scRNA-seq dimension reduction methods to provide composite layers for MLG based on (i) how well they are already adapted by the community (PCA, Seurat); (ii) the type of dimension reduction they employ (NMF based cNMF and Liger in addition to PCA); and (iii) whether or not they have built in data integration (Seurat, Liger). The second reasoning is based on our analytical result (Supplementary Materials Sections 1.1, 1.2) that aggregating results from different dimension reduction methods that lead to sparse and independent shared nearest neighbor graphs improves the signal-to-noise ratio. We also evaluated construction of MLG with scVI, Harmony, and scAlign, and decided not to utilize these methods in constructing MLG. Harmony and scVI both show moderate dependencies with PCA in multiple datasets (their average proportion of overlapping edges with PCA-SNN graph are 63% and 45%, respectively, in the benchmark datasets), and are therefore excluded to ensure low dependence between the dimension reduction methods as required by the analytical calculations that support improved signal-to-noise ratio (Supplementary Materials Section 1.1). scAlign using PCA, or CCA as input are correlated with PCA and Seurat results. While scAlign allows using gene expression without initial dimension reduction as input, such an execution is computationally expensive (55 min of elapsed time, 25 h of CPU time for a dataset of 1000 cells).

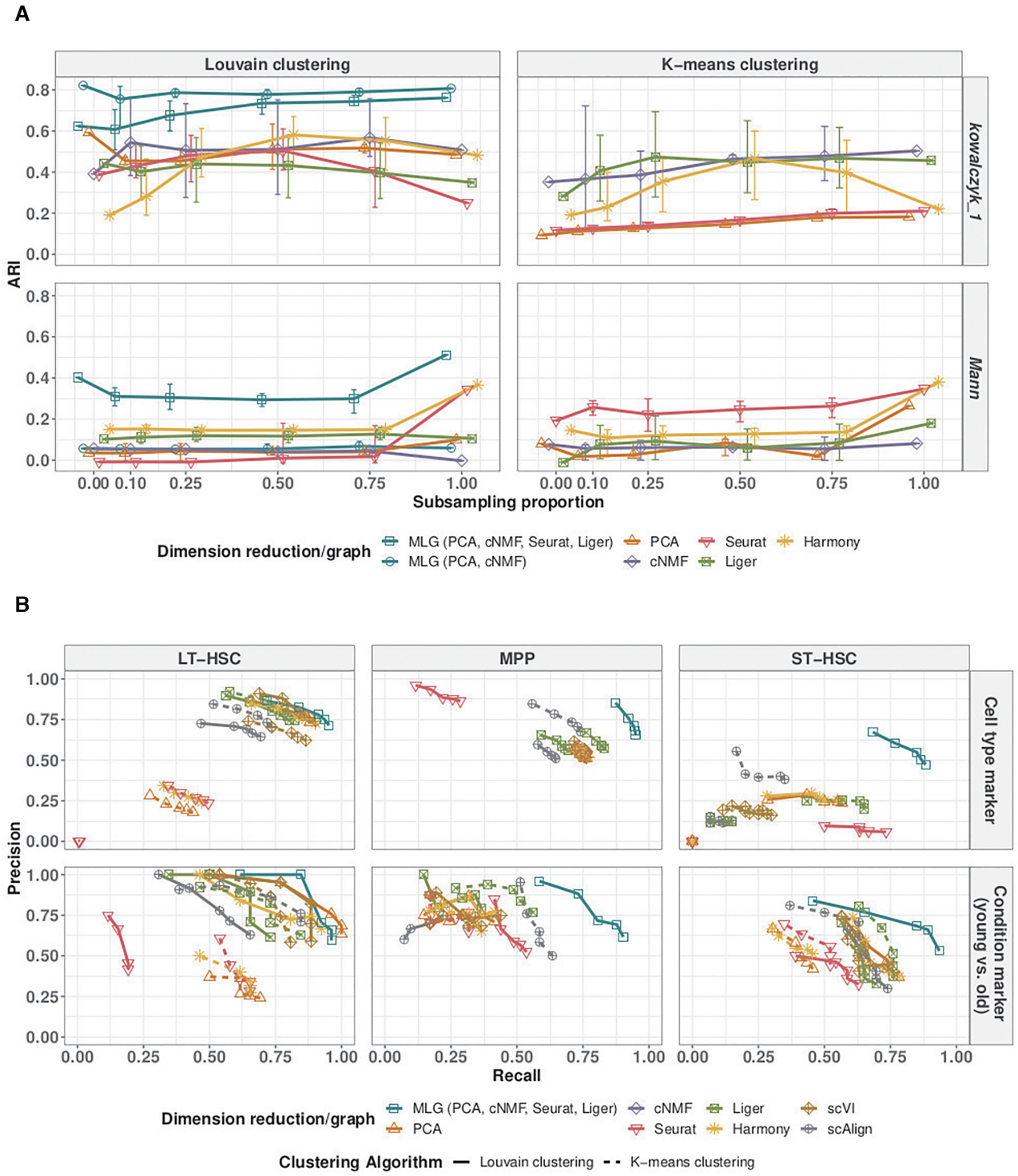

Figure 5. Robustness analysis of clustering over imbalanced samples and downstream impact of MLG clustering on marker gene identification. (**A**) Adjusted Rand index from Louvain and k-means clustering of low-dimensional embeddings with varying levels of imbalance in the proportion of cells from different cell types. For the *Kowalczyk_1* and *Mann* datasets, varying proportions of MPP cells from the 'old' and 'old, stimulated' conditions were subsampled for the respective analysis. Reported ARI values correspond to mean, 5 and 95 percentiles across 20 subsampling experiments. (**B**) Differential expression analysis for cell type/cluster marker gene identification and condition (young versus old) DE gene identification in the *Kowalczyk_1* dataset. The precision-recall (PR) values in the top panel evaluate inferred cluster (i.e. cell type) marker genes of each method against cell type marker genes defined with ground truth cell type labels as gold standard. The bottom panel evaluates the inferred condition (i.e. age) DE genes of each cell type against the gold standard. Gold standard cell type marker and age DE genes are defined as genes with Bonferroni corrected *P*-values less than 0.01 in the differential expression analysis with ground truth cell labels. The PR values are reported at cutoffs of 0.2, 0.1, 0.05, 0.01 and 0.001 for Bonferroni adjusted p-values in the cluster marker gene and age DE gene identification analysis.

In extending our benchmarking experiments to include additional layers, we observed that the improvement in accuracy of MLG starts to get 'saturated' after adding four layers with PCA, cNMF, Seurat, and Liger (Supplementary Figure S27). To enable potential integration of additional dimension reduction methods by users, we developed and evaluated a weighting scheme (Supplementary Materials Section 1.3). Weighted MLG evaluates each dimension reduction result using their knn graph's ability of incorporating cells from different conditions and predicting the gene expression profiles of neighboring cells. In simulation settings 3, 4, 7, and 8, it yields higher rare cell type identification probability (with an average improvement of 9% over MLG across all the rare cell type simulation settings) and better protection over irrelevant layers (Supplementary Figure S9). Collectively, this suggests that weighted MLG provides a robust way of incorporating additional layers to MLG beyond the four core methods (PCA, cNMF, Seurat and Liger) we have extensively utilized and evaluated.

MLG clustering confirms disrupted differentiation patterns in HSPCs lacking the murine *Gata2* -77 enhancer

We utilized MLG clustering to analyze scRNA-seq data of hematopoietic progenitor cells sorted from E14.5 fetal livers of $-77^{+/+}$ (wild type, WT) and $-77^{-/-}$ (mutant, MT) murine embryos (dataset *Johnson20* from (19)). The mutant condition corresponded to homozygous deletion of the murine *Gata2* −77 enhancer, located −77 kb upstream of the *Gata2* transcription start site, as described in (19). The samples from both WT and MT mice included a complex mixture of progenitors with diverse transcriptional profiles (34) from a pool of common myeloid progenitor (CMP) and granulocyte-monocyte progenitor (GMP) cells and resulted in 14,370 cells after pre-processing (19). Exploratory analysis with the data visualization tools t-SNE (32), UMAP and SPRING (35) revealed a small separation between the cells from the wild type and mutant conditions and an average mixing metric of 15.25. Furthermore, Johnson et al. (36) previously showed that $-77^{+/+}$ fetal liver CMPs have the potential to differentiate into erythroid and myeloid cells *ex vivo*. In contrast, the mutant $-77^{-/-}$ CMPs and GMPs generate predominantly macrophage progeny. This instigated us to proceed with clustering of the cells without data integration since the subsampling experiments with benchmark datasets highlighted that data integration may cause misalignment of cells in this setting with potentially imbalanced cell types.

We constructed a multilayer graph from PCA (after regressing out total counts, mouse batch effects, and percentage count of mitochondrial transcripts) and cNMF factors. Figure 6A displays the SPRING visualization of the MLG clustering with four optimally chosen clusters based on the eigengap heuristic (23), and also highlights pseudotime trajectories of the wild type and mutant cells. We linked these clusters to established cell populations as (i) CMP, (ii) erythroid/megakaryocyte, (iii) bipotential GMP and monocyte, (iv) neutrophils, by leveraging established lineage defining markers of these HSPC populations (37), i.e., *Flt3* and *Hlf* for CMPs, *Hba-a2* and *Car1* for erythrocytes; *Ly86* and *Csf1r* for monocytes; *Gstm1* and *Fcnb* for neutrophils, and *Pf4* for megakaryocytes (Figure 6A). Overall, cells in different clusters exhibited expression patterns of documented lineage markers (37) consistent with their inferred cluster labels (Supplementary Figure S15). Furthermore, an unbiased marker gene analysis of each MLG cluster with MAST (38) (Supplementary Figure S14) yielded marker genes. A gene set enrichment analysis with top 50 marker genes of each cluster agreed with the MLG cluster labels based on known lineage markers (Supplementary Figure S23).

Next, we compared MLG clusters with Louvain clustering of the cell graphs constructed from PCA, cNMF factors, Seurat-integration, Liger, Harmony, scVI and scAlign (Figure 6B). cNMF, Seurat-integration, Liger, Harmony, scVI and scAlign tended to partition the cells based on their mitochondrial gene expression. Specifically, the clusters in cNMF, Seurat-integration, Liger, Harmony, scVI and scAlign are driven by the mitochondrial gene expression patterns apparent in the SPRING plot (Supplementary Figure S13). Concordant with these partitions, the top 10 marker genes for cNMF cluster 1, Seurat-integration cluster 4, Liger cluster 2, Harmony cluster 3, scVI cluster 4 and scAlign cluster 4 have 5, 7, 7, 2, 7 and 7 mitochondrial genes, respectively. While it is possible to adjust for mitocondrial gene expression of the cells for some settings such as PCA, the Seurat-integration framework does not enable adjustment for biological variables in its CCA factors. Methods including cNMF, Liger, scVI, scAlign do not encompass mechanisms to adjust for potential continuous confounders. While the clustering of the low-dimensional embedding from PCA is similar to the MLG clustering, it merges cells expressing *Gata2* and erythroid specific-gene *Car1* with cluster 1 that represents early progenitors (Supplementary Figures S15, S16) and results in a markedly smaller cluster of erythroid cells. In addition to these differences, established lineage markers do not delineate clusters from cNMF, Seurat-integration, Harmony, scVI and scAlign as expressing cell type specific marker genes as clearly as MLG clustering (Supplementary Figures S16, S17, S18, S19, S20, S21, S22). As a quantitative measure of how well lineage marker genes delineate clusters obtained with different methods, we provide the Chi-squared statistic of Fisher's combined probability test for these genes in Supplementary Table S7, which further highlights the better performance of MLG. Collectively, this analysis illustrates the power of MLG in identifying cell types/stages with low signal by its aggregation strategy.

MLG uncovers cell stages in mouse HSPC under four experimental conditions

Dataset *Muench20* (20) contains scRNA-seq of 813 mouse hematopoietic progenitor cells under 4 conditions: wild type ($Gif1^{+/+}$), heterozyous *Gfi1* R412X mutation ($Gif1^{R412X/-}$), homozygous *Gfi1* R412X mutation ($Gif1^{R412X/R412X}$), heterozygous *Gfi1* R412X mutation and one silenced *Irf8* allele ($Gif1^{R412X/-}Irf8^{+/-}$). Joint dimension reduction was conducted on all the cells, with PCA, cNMF, Seurat integration, Liger, Harmony, scVI and scAlign, followed by SNN graph construction with each individual dimension reduction results. Louvain graph clus-

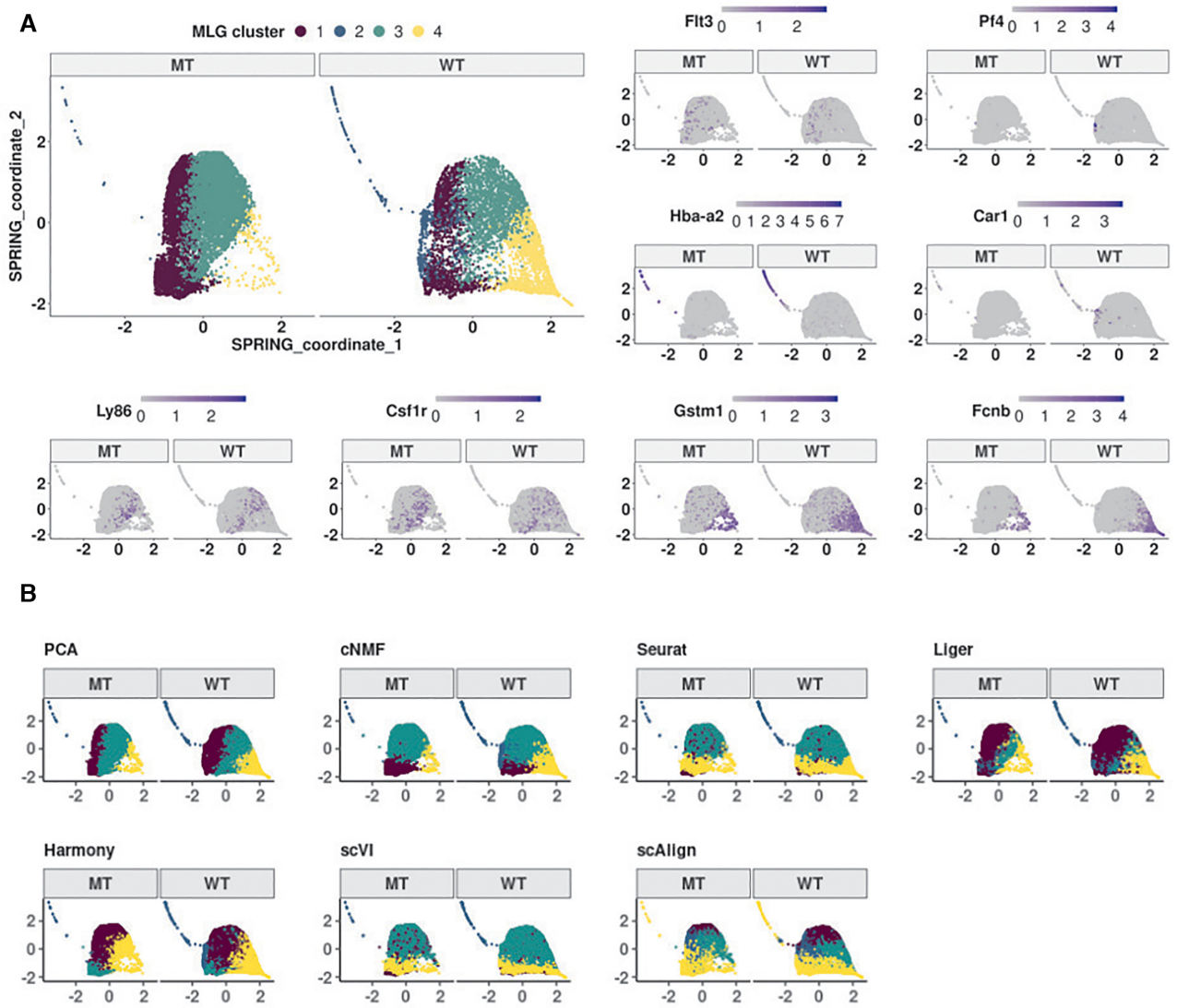

Figure 6. Clustering analysis of the *Johnson20* (19) dataset. (**A**) SPRING (35) visualization of mutant and wild type cells from *Johnson20*. Cells are labeled with MLG clustering (top left) and expression of linage marker genes. (**B**) SPRING visualization of mutant and wild type cells with cell labels from Louvain clustering of SNN graphs with PCA, cNMF factors, Seurat integrtion, Liger, Harmony, scVI and scAlign low-dimensional embeddings.

tering was applied to each SNN graph along with MLG with PCA, cNMF, Seurat, Liger SNN graphs as four layers. A total of seven clusters were identified with the eigengap heuristic (23) (Figure 7B). Through the expression patterns of HSPC-relevant transcription factors and granule protein-encoding genes (Figure 7A, gene set 1), surface marker genes (Figure 7A, gene set 2) and linage marker genes (Figure 7A, gene set 3), we labelled the clusters 1 to 7 as HSPC, Multi-Lin (multi-lineage progenitor), Mono-1, Mono-2 (Monocyte progenitor), Neu-1, Neu-2, Neu-3 (Neutrophil progenitor). The MLG clustering results also matched the author annotated cell labels from (20) with the exception of the extremely small clusters, e.g. NK(3 cells), DC(5 cells), missed by MLG. We compared the resulting clustering from each method using the author annotated cell labels as gold standard. Since there are 17 distinct cell labels in the author annotation set, we used a wide range of numbers of clusters in the comparison (Figure 7C). MLG yielded clustering results most closely in agreement with the gold standard cell labels. Moreover, it performed more robustly against the choice of numbers of clusters, highlighting the overall robustness of the MLG aggregation strategy.

DISCUSSION

We have introduced MLG clustering as a versatile method to identify cell types/stages in scRNA-seq data from multiple stimuli by aggregating cell graphs of multiple low-dimensional embeddings of the data. MLG construction capitalizes on the complementary information captured with SNN graphs from different dimension reduction methods such as PCA and cNMF in addition to data integration methods such as Seurat and Liger for scRNA-seq data from multiple stimuli. By aggregating sparse SNNs with low edge overlap, MLG amplifes the signal-to-noise ratio. Here the signal-to-noise ratio is quantified by comparing the within

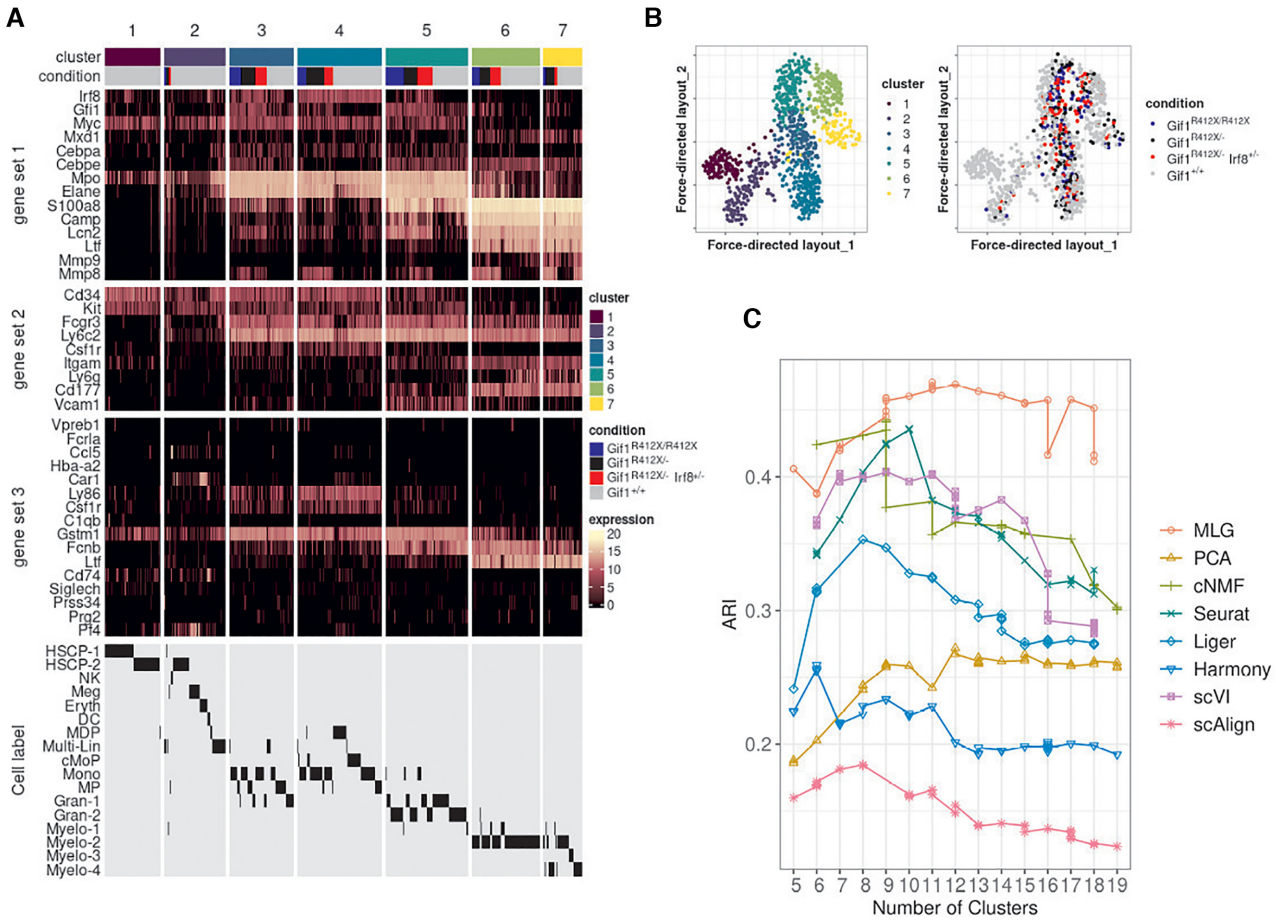

Figure 7. Clustering analysis of the *Muench_20*(20) dataset. (**A**) Heatmaps of the cell-level gene expression for HSPC-relevant transcription factors and granule protein-encoding genes (gene set 1), surface markers (gene set 2) and lineage defining markers (gene set 3). Cells are grouped by MLG clustering and cell conditions. MLG clustering is compared with author annotated cell labels from (20) in the last row of the heatmap. (**B**) Visualization of the dataset with the force directed layout of MLG. Cells are labeled with MLG clusters (left) and cell conditions (right). (**C**) ARIs are computed between author annotated cell labels and clustering results of MLG and individual SNN graphs constructed from low-dimensional embeddings by PCA, cNMF, Seurat-integration Liger, Harmony, scVI, and scAlign.

cluster connectivity of cells to the between cluster connectivity.

While the MLG clustering framework is flexible in the types of dimension reduction methods it can aggregate over, computational experiments with both simulated and benchmark datasets demonstrated that combining SNN graphs from low-dimensional embeddings of PCA and cNMF perform well when the stimuli is not the dominant effect and MLG benefits from aggregating over Seurat-integration and Liger when data integration is essential to accommodate the stimuli effect. Furthermore, both PCA and cNMF are robust, and easy to tune compared to other dimension reduction methods for scRNA-seq. Another guiding principle for building MLG is to avoid the simultaneous use of methods that are highly related to prevent amplification of false cell-to-cell connections. The mixing metric performed as a robust quantity for capturing the level of the stimuli effect to guide the aggregation task in terms of whether or not to include low-dimensional embeddings from data integration. Overall, we found that Louvain clustering of MLG exhibited superior performance over clustering with individual low-dimensional embeddings both in accuracy of the clustering and the downstream marker gene identification.

To enable integration of additional dimension reduction methods to MLG by users, we developed a weighting scheme. Weighted MLG evaluates each dimension reduction result using their knn graph's ability to incorporate cells from different conditions and predict the gene expression profile of neighboring cells. This weighting scheme performed robustly in a wide variety of simulation settings, making it feasible to include additional scRNA-seq dimension reduction methods in MLG than discussed here.

While our focus in this work has been predominantly scRNA-seq datasets from multiple conditions, we showcase how MLG can be readily applied to analyze multimodal SNARE-seq data from joint profiling of transcriptome and chromatin accessibility (21) and that it performs as well as the methods tailored for this data type in Supplementary Section 1.4 (Supplementary Figure S24, S25, Supplementary Tables S5 and S6). We expect MLG clustering to be generally applicable for SNARE-seq, Share-seq (39) and other single cell data modalities such as scHi-C

(40), joint profiling of transcriptome and DNA methylation (41). In these applications, each modality yields individual low-dimensional embeddings of the cells and aggregation of their graph products can increase the overall signal. Finally, the computational cost of MLG is naturally driven by the dimension reduction methods it aggregates over (Supplementary Materials Section 1.5, Supplementary Figure S26).

DATA AVAILABILITY

Datasets *Kowalczyk_1* and *Kowalczyk_2* are available at NCBI GEO with accession number GSE59114. The *Mann* dataset is available at NCBI GEO with accession number GSE100426. The *Cellbench* dataset is available from Github at https://github.com/LuyiTian/sc_mixology/blob/master/data/mRNAmix_qc.RData. The $-77^{+/+}$ and $-77^{-/-}$ datasets are available at NCBI GEO with accession number GSE134439. Dataset *Muench20* is available at https://www.synapse.org/#!Synapse:syn16806696. The SNARE-seq dataset, *Chen19*, is available at the Gene Expression Omnibus database under accession number GSE126074. Dataset *Census_of_immune_cells* is used to evaluate time and memory requirement of each program. It is available at https://data.humancellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79.

We implemented the MLG clustering in an open-source R package available on Github https://github.com/keleslab/mlg.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported by National Institutes of Health [HG011371 to S.K., HG003747 to S.K., R01 DK68635 to E.H.B.]; and Carbone Cancer Center [P30CA014520 to E.H.B.]. Funding for open access charge: National Institutes of Health [HG003747].
Conflict of interest statement. None declared.

REFERENCES

1. Zeisel,A., Muñoz-Manchado,A.B., Codeluppi,S., Lönnerberg,P., La Manno,G., Juréus,A., Marques,S., Munguba,H., He,L., Betsholtz,C. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.
2. Chen,R., Wu,X., Jiang,L. and Zhang,Y. (2017) Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep.*, **18**, 3227–3241.
3. Lambrechts,D., Wauters,E., Boeckx,B., Aibar,S., Nittner,D., Burton,O., Bassez,A., Decaluwé,H., Pircher,A., Van den Eynde,K. *et al.* (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.*, **24**, 1277–1289.
4. DeLaughter,D.M., Bick,A.G., Wakimoto,H., McKean,D., Gorham,J.M., Kathiriya,I.S., Hinson,J.T., Homsy,J., Gray,J., Pu,W. *et al.* (2016) Single-cell resolution of temporal gene expression during heart development. *Dev. Cell*, **39**, 480–490.
5. Mathys,H., Adaikkan,C., Gao,F., Young,J.Z., Manet,E., Hemberg,M., De Jager,P.L., Ransohoff,R.M., Regev,A. and Tsai,L.-H. (2017) Temporal tracking of microglia activation in neurodegeneration at single-cell resolution. *Cell Rep.*, **21**, 366–380.
6. Kiselev,V.Y., Andrews,T.S. and Hemberg,M. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
7. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck III,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
8. Hotelling,H. (1992) Relations between two sets of variates. In: *Breakthroughs in Statistics*. Springer, pp. 162–190.
9. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.
10. Lopez,R., Regier,J., Cole,M.B., Jordan,M.I. and Yosef,N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
11. Johansen,N. and Quon,G. (2019) scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.*, **20**, 166.
12. Korsunsky,I., Millard,N., Fan,J., Slowikowski,K., Zhang,F., Wei,K., Baglaenko,Y., Brenner,M., Loh,P.-R. and Raychaudhuri,S. (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.
13. Blondel,V.D., Guillaume,J.-L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.*, **2008**, P10008.
14. Tran,H. T.N., Ang,K.S., Chevrier,M., Zhang,X., Lee,N. Y.S., Goh,M. and Chen,J. (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.
15. Duò,A., Robinson,M.D. and Soneson,C. (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, **7**, 1141.
16. Freytag,S., Tian,L., Lönnstedt,I., Ng,M. and Bahlo,M. (2018) Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, **7**, 1297.
17. Wolf,F.A., Angerer,P. and Theis,F.J.(2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
18. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
19. Johnson,K.D., Conn,D.J., Shishkova,E., Katsumura,K.R., Liu,P., Shen,S., Ranheim,E.A., Kraus,S.G., Wang,W., Calvo,K.R. *et al.* (2020) Constructing and deconstructing GATA2-regulated cell fate programs to establish developmental trajectories. *J. Exp. Med.*, **217**, e20191526.
20. Muench,D.E., Olsson,A., Ferchen,K., Pham,G., Serafin,R.A., Chutipongtanate,S., Dwivedi,P., Song,B., Hay,S., Chetal,K. *et al.* (2020) Mouse models of neutropenia reveal progenitor-stage-specific defects. *Nature*, **582**, 109–114.
21. Chen,S., Lake,B.B. and Zhang,K. (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.*, **37**, 1452–1457.
22. Kotliar,D., Veres,A., Nagy,M.A., Tabrizi,S., Hodis,E., Melton,D.A. and Sabeti,P.C. (2019) Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, **8**, e43803.
23. Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.
24. Kowalczyk,M.S., Tirosh,I., Heckl,D., Rao,T.N., Dixit,A., Haas,B.J., Schneider,R.K., Wagers,A.J., Ebert,B.L. and Regev,A. (2015) Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.*, **25**, 1860–1872.
25. Mann,M., Mehta,A., de Boer,C.G., Kowalczyk,M.S., Lee,K., Haldeman,P., Rogel,N., Knecht,A.R., Farouq,D., Regev,A. *et al.* (2018) Heterogeneous responses of hematopoietic stem cells to inflammatory stimuli are altered with age. *Cell Rep.*, **25**, 2992–3005.
26. Tian,L., Dong,X., Freytag,S., Le Cao,K.-A., Su,S., JalalAbadi,A., Amann-Zalcenstein,D., Weber,T.S., Seidi,A., Jabbari,J.S. *et al.* (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.
27. Holland,P.W., Laskey,K.B. and Leinhardt,S. (1983) Stochastic blockmodels: first steps. *Soc. Networks*, **5**, 109–137.
28. Zhang,A.Y. and Zhou,H.H. *et al.* (2016) Minimax rates of community detection in stochastic block models. *Ann. Stat.*, **44**, 2252–2280.
29. Lun,A.T., McCarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Research*, **5**, 2122.

30. Alexa,A., Rahnenführer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
31. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
32. Maaten,L. V.D. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
33. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.
34. Olsson,A., Venkatasubramanian,M., Chaudhri,V.K., Aronow,B.J., Salomonis,N., Singh,H. and Grimes,H.L. (2016) Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, **537**, 698–702.
35. Weinreb,C., Wolock,S. and Klein,A.M. (2018) SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, **34**, 1246–1248.
36. Johnson,K.D., Kong,G., Gao,X., Chang,Y.-I., Hewitt,K.J., Sanalkumar,R., Prathibha,R., Ranheim,E.A., Dewey,C.N., Zhang,J. *et al.* (2015) Cis-regulatory mechanisms governing stem and progenitor cell transitions. *Sci. Adv.*, **1**, e1500503.
37. Giladi,A., Paul,F., Herzog,Y., Lubling,Y., Weiner,A., Yofe,I., Jaitin,D., Cabezas-Wallscheid,N., Dress,R., Ginhoux,F. *et al.* (2018) Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.*, **20**, 836–846.
38. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
39. Ma,S., Zhang,B., LaFave,L.M., Earl,A.S., Chiang,Z., Hu,Y., Ding,J., Brack,A., Kartha,V.K., Tay,T. *et al.* (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, **183**, 1103–1116.
40. Zhou,J., Ma,J., Chen,Y., Cheng,C., Bao,B., Peng,J., Sejnowski,T.J., Dixon,J.R. and Ecker,J.R. (2019) Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 14011–14018.
41. Clark,S.J., Argelaguet,R., Kapourani,C.-A., Stubbs,T.M., Lee,H.J., Alda-Catalinas,C., Krueger,F., Sanguinetti,G., Kelsey,G., Marioni,J.C. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 781.