# scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods

Luyi Tian[1,3*], Xueyi Dong[1,4], Saskia Freytag[2,3], Kim-Anh Lê Cao[5,6], Shian Su[1], Daniela Amann-Zalcenstein[1,3], Tom S. Weber[1,3], Azadeh Seidi[7], Shalin H. Naik[1,3], Matthew E. Ritchie[1,3,5*]

**1** Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 3052, Australia.
**2** Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 3052, Australia.
**3** Department of Medical Biology, The University of Melbourne, Parkville, 3010, Australia.
**4** College of Life Science, Zhejiang University, 866 Yuhangtang Road, Hangzhou, Zhejiang Province, 310058, P.R. China.
**5** School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010, Australia.
**6** Melbourne Integrative Genomics, The University of Melbourne, Parkville, 3010, Australia.
**7** Australian Genome Research Facility, 1G Royal Parade, Parkville, 3052, Australia.

* Corresponding authors: tian.l@wehi.edu.au; mritchie@wehi.edu.au

## Abstract

Single cell RNA sequencing (scRNA-seq) technology has undergone rapid development in recent years, bringing with it new challenges in data processing and analysis. This has led to an explosion of tailored analysis methods for scRNA-seq to address various biological questions. However, the current lack of gold-standard benchmarking datasets makes it difficult for researchers to evaluate the performance of the many methods. Here, we designed and carried out a realistic benchmark experiment that included mixtures of single cells or 'pseudo-cells' created by sampling admixtures of cells or RNA from 3 distinct cancer cell lines. Altogether we generated 10 datasets using a combination of droplet and plate-based scRNA-seq protocols, with varying data quality, population heterogeneity and noise levels. Using these benchmark datasets, we compared different protocols, evaluated the spike-in standard and multiple data analysis methods for tasks ranging from normalization and imputation, to clustering, trajectory analysis and data integration. Evaluation of methods across multiple datasets revealed some that performed well in general and others that suited specific situations. Our dataset and analysis provide a comprehensive comparison framework for benchmarking most popular scRNA-seq analysis tasks.

The rapid development of transcriptomic technology for single cell analysis has created a need for systematic benchmarking in order to understand the strengths and weaknesses of different platforms and computational methods. To date, there have been several comparison studies of different protocols and computational methods for single cell RNA sequencing (scRNA-seq). Svensson *et al.* [47] compared the quality of many publicly available datasets using spike-in controls, together with in-house data generated using mouse embryonic stem cells and human brain total RNA. Another recent study by Tung *et al.* [49] focused on the assessment of batch variation in scRNA-seq data and highlighted the importance of experimental designs that avoid confounding of biological and technical effects. In addition to protocol comparisons, several studies have assessed the performance of different scRNA-seq data analysis methods for tasks including normalization [4], feature selection [54], differential gene expression analysis [45], clustering [6, 7] and trajectory analysis [42]. These studies compare methods using either experimental data where cell type labels are available or simulated datasets. Such ground truth is however imperfect, and simulations rely on assumptions that may not reflect the true nature of scRNA-seq data.

Considering the heterogeneity between scRNA-seq datasets in terms of the number of clusters (cell types/states) and the presence of various technical artifacts, we set out to design a realistic gold-standard scRNA-seq control experiment that combines ground truth with varying levels of biological complexity. Two strategies are commonly employed to create such gold-standard gene expression datasets. The first uses small collections of exogenous spike-in controls (such as ERCCs [19]) that vary in expression in a predictable way, which have been widely adopted in scRNA-seq studies [47]. The second involves either the dilution of RNA from a reference sample or mixing of RNA or cells from two or more samples to induce systematic genome-wide changes. An early example of an scRNA-seq control dataset was presented in Brennecke *et al.* [3] and involved a dilution series to explore sensitivity of the Smart-seq protocol. Grün *et al.* [8] generated a benchmark dataset using single mouse embryonic stem cells (mESC) together with bulk RNA extracted from the same population, diluted to single cell equivalent amounts to quantify biological and technical variability. A limitation of these experiments is their lack of biological heterogeneity which makes them less useful for comparing analysis methods. Mixture designs, in which RNA or cells are mixed in different proportions to generate biological heterogeneity with in-built truth have been successfully used to benchmark microarray [5], RNA-seq [44] and scRNA-seq data [49].

To combine the strengths of these approaches, we designed a series of experiments using mixtures of either cells or mRNA from 3 cancer cell lines and included a dilution series to simulate variations in the mRNA content of different cells as well as ERCC spike-in controls wherever possible. Data were generated across four single-cell platforms (CEL-seq2, 10X Chromium, Drop-seq and SORT-seq). Our scRNA-seq mixology design simulates varying levels of biological noise, and contains known population structure to allow benchmarking of different analysis tools.

In this article we specifically highlight data processing, quality control, normalization and imputation, clustering, trajectory analysis and data integration to showcase the broad range of tasks that our unique collection of datasets allows us to benchmark. Our analyses across multiple datasets allows evaluation of the generalizability of different methods to help inform best practice in data analysis.

# Results

## scRNA-seq mixology provides ground truth for benchmarking

As summarised in Supplementary Table 1, the scRNA-seq benchmarking experiment spanned 2 plate-based and 2 droplet-based protocols and involved 3 different experimental designs with replicates, yielding 10 datasets in total. Our experiment used the 3 human lung adenocarcinoma cell lines H2228, H1975 and HCC827, included either mixtures of RNA or single cells from these cell lines. For the single cell designs, the 3 cell lines were mixed equally and processed by 10X Chromium, Drop-seq [31] and CEL-seq2 [13] (referred to as sc_10X, sc_Drop-seq and sc_CEL-seq2 respectively). For the *'pseudo cell'* designs, we used plate-based protocols to mix and dilute samples in 2 different ways. For the first, we created 9-cell mixtures from the 3 cell lines by sorting different combinations of cells and generating libraries using CEL-seq2. The material after pooling from 384 wells were sub-sampled to either 1/9 or

1/3 of the total mixture to simulate cells with varying mRNA content and using different PCR product clean up ratios (sample:beads) ranging from 0.7:1 to 0.9:1. These data are referred to as cellmix1 to cellmix4 (Supplementary Figure 1B; Supplementary Table 1). We also sorted wells with 90 cells to provide a pseudo bulk reference for each mixture (referred to as cellmix5). The second design created *'pseudo cells'* by mixing bulk mRNA obtained from each cell line, which were diluted to create single cell equivalents (ranging from 3.75, 7.5, 15 and 30 pg per well) to again create controlled variations in mRNA content. Data were generated for this RNA mixture design using CEL-seq2 and SORT-seq [33] (referred to as RNAmix_CEL-seq2 and RNAmix_Sort-seq, Supplementary Figure 1A; Supplementary Table 1).

The three designs incorporate ground truth in various ways. For the single cell mixture datasets, the ground truth is the cell line identity which can be determined for each cell based on known genetic variation. The single cell mixtures were also generated using three different technologies, allowing cross-platform comparisons and testing of data integration methods. The *'pseudo cell'* datasets contain more clusters than the single cell datasets, and these clusters are more similar to each other, giving a continuous structure that simulates what may be expected in cell differentiation studies, where cells are transitioning between states. For the cell and RNA mixtures, the composition of cells/RNA that make up each *'pseudo cell'* are known, which serves as our ground truth. Moreover, the RNA mixture dataset contains technical replication and a dilution series, which is ideal for benchmarking normalization and imputation methods that are intended to deal with such technical variability. The data characteristics and analysis tasks each experimental design is best suited to benchmark are summarized in Supplementary Table 2.

## Quality control metrics allow comparisons between platforms

By comparing a range of quality control metrics collected across datasets using *scPipe* [48], we observed that the data from all platforms were of consistently high quality in terms of their exon mapping rates and the total unique molecular identifier (UMI) counts per cell (Supplementary Figure 2A-C). After normalization, the Principal Component Analysis (PCA) plots from three representative datasets show that our 9-cell mixture and RNA mixture datasets successfully recapitulate the expected population structure induced by our design (Figure 11C). Comparison of the three single cell datasets shows that the 10X platform outperforms the others in almost all aspects, regardless of the variation in sequencing depth (Supplementary Figure 2D). The UMI counts for CEL-seq2 were less than that observed for 10X, but consistent among single cell, 9-cell mixture and RNA mixture datasets, suggesting the robustness of the protocol. The data quality for SORT-seq was slightly lower than for CEL-seq2, which is likely due to technical artifacts encountered during sample pooling. The difference in data quality between the 9-cell mixture datasets highlights the importance of choosing the right clean up ratio, as decreasing the ratio decreased the yield and complexity of the PCR library and reduced the UMI count per cell (Supplementary Figure 2D).

## Intron mapping rates vary between platforms and conditions

As the proportion of reads that mapped to introns has not been explored in detail in previous studies, we investigated the intron mapping rates across all datasets and platforms (Figure 1E). Interestingly, we found substantial differences in the percentage of reads mapping to intron regions in datasets generated from different protocols and experimental designs. The single cell datasets, although prepared from the same batch of cells, exhibit substantial variability in intron mapping proportions between the three protocols, with 10X showing a much lower proportion of intron reads. In contrast, Drop-seq had the highest proportion of reads mapping to introns. The clean up ratio after PCR alters the fragment length in the final sequencing library, with smaller fragment size resulting from increased clean up ratios. Interestingly, the clean up ratio also affected the intron mapping rate, with intron signal decreasing as the fragment size decreased, which is caused by the clean up ratio.

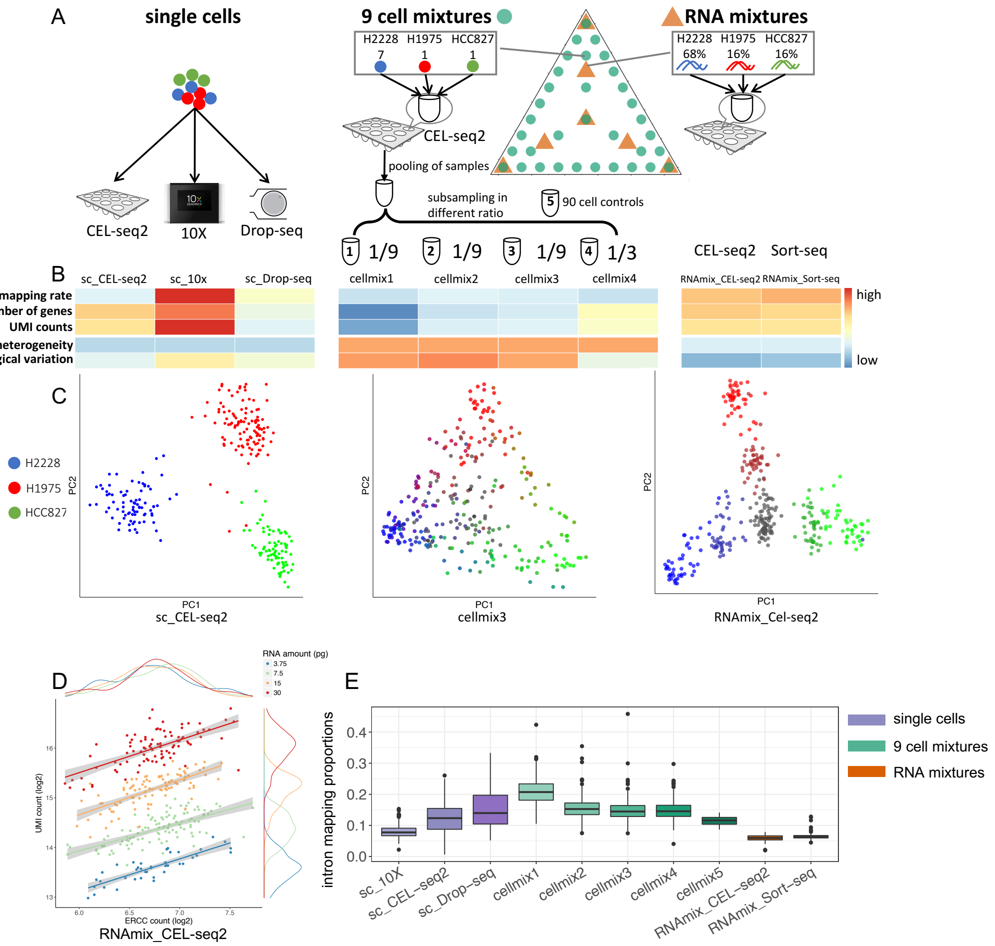


**Figure 1. Overview of our scRNA-seq mixology experimental design.** (**A**) The three different designs that used three distinct cell lines are shown from left to right. 1) Single cells from each line were combined in equal proportions and scRNA-seq was performed using the CEL-seq2 (sc_CEL-seq2), Drop-seq (sc_Drop-seq) and 10X Chromium protocols (sc_10x). 2) *'Pseudo cells'* were created by sorting different combinations of 9-cells from the three cell lines into 384-well plates and subsequently diluting them to obtain single cell equivalent amounts of RNA (cellmix1 to cellmix4). 90-cell mixtures were also include to create pseudo bulk references for each mixture (cellmix5). 3) *'Pseudo cells'* were created by mixing RNA obtained from bulk samples from the three cell lines in different proportions and diluting the samples to vary the mRNA amount from 3.75pg to 30pg (RNAmix_CEL-seq2, RNAmix_Sort-seq). (**B**) Several quality control metrics, including the number of genes detected and the number of UMI counts per cell are shown in a heatmap, together with other data characteristics such as population heterogeneity and the average biological coefficient of variation (BCV) obtained from an *edgeR* analysis. Additional quality control metrics are shown in Supplementary Figure 2. (**C**) PCA plots from representative datasets for each design (normalized using *scran*) highlight the structure present in each experiment. The single cell design is simplest, with 3 distinct groups related to cell line identity, while the 9-cell and RNA mixtures are more complex with more groups (34 or 7 respectively). This in-built truth can be used to benchmark different analysis methods. (**D**) Scatter plot of ERCC counts versus total UMI counts per cell, colour coded by the different RNA amounts. (**E**) The intron mapping rate across all datasets.

## ERCC spike-ins correlate with technical noise in the RNA mixture dataset

ERCC spike-in sequences have been widely used in scRNA-seq experiments to estimate technical noise, and the assumption that the biological variation of endogenous mRNAs does not affect ERCC spike-ins has been argued in previous studies. Considering that measuring the RNA abundance of single cells is difficult, we designed the RNA mixture experiment to have 4 different RNA amounts (3.75, 7.5, 15 and 30 pg) to create controlled variation in mRNA content. This dataset shows comparable quality control metrics to real single cell datasets, including the number of genes detected and the number of total UMI counts per cells (Supplementary Figure 2B-C). Under the assumption that the total UMI count for each cell is influenced by the mixture design, the RNA amount and technical noise which represents variations in capture efficiency for each sample, 3 different linear mixed models were proposed and compared (Supplementary Table 3). Model I used the mixture information as a covariate and included a random intercept for each RNA amount, so that samples with different RNA amounts have different intercept values. Model II expanded upon Model I by adding ERCC spike-in counts as a further covariate and assuming fixed effects, with the coefficients in the linear model not changing with different RNA amounts. Model III assumes random coefficients for ERCC spike-in counts, which allows for a distinct slope for the different RNA amounts. According to the AIC and $p$-value from model comparison using ANOVA, the addition of ERCC spike-in counts (Model II) greatly improved the model fit, while the random slope assumption of the ERCC spike-in counts (Model III) was unnecessary. Examination of the scatter plot of total UMI counts versus the ERCC spike-in counts show this clearly, with consistent slopes for different RNA amounts (Figure 1D). This relationship suggests that the ERCCs correlate with the variation of total count in the RNA mixture data given the same RNA amount. The ERCC counts have similar distributions among samples that have different RNA amounts, which invalidates the common assumption that ERCC spike-ins are less likely to be sampled when the amount of endogenous mRNA in a cell is high [47].

## Comparison of normalization and imputation methods

Normalization is an important step in the analysis of scRNA-seq data, with the general goal of removing technical noise while retaining biological signal. Imputation on the other hand recovers missing data due to dropout events, which are excess zero counts caused by the limited capture efficiency of scRNA-seq protocols. We evaluated 12 normalization and imputation methods, with methods designed for bulk RNA-seq such as *TMM* [39], *CPM* [38] and *DESeq* [29], and methods designed for scRNA-seq, including *ZINB-WaVE* [37], *scone* [4], *kNN smooth* [52], *BASiCS* [50], *SCnorm* [1], *Linnorm* [55], *scran* [30], *SAVER* [16] and *DrImpute* [23]. Performance was evaluated using 2 metrics: the Pearson correlation coefficient of normalized gene expression among technical replicates for the RNA mixture data, and the silhouette width of clusters for all 3 designs. The silhouette width was calculated based on the PCA results obtained for each method. Example PCA plots for 3 methods (*scran* with *DrImpute* imputation, *BASiCS* and *kNN smooth*) applied to the RNA mixture data show dramatic differences between methods (Figure 2A). Methods such as *kNN smooth* failed to recover the designed population structure. In general, the gene expression correlation of technical replicates was lower for smaller amounts of RNA, due to the higher drop-out frequency (Figure 2B). Imputation methods such as *kNN smooth*, *SAVER* and *DrImpute* all systematically improved the correlations between technical replicates and reduced the differences in correlations between RNA amounts when compared to different normalizations alone. This clearly demonstrates that imputation can remove noise due to drop-out events as intended. A large diversity was seen in silhouette width among different methods in different datasets. We found *scran* and *Linnorm* to have good normalization performance among all datasets, and normalization when combined with imputation gave the best performance (Figure 2C). The *kNN smooth* method outperformed other methods in the single cell datasets but its performance degraded in the other datasets which had more complex population structure (Supplementary Figure 3). The reason for this may be because *kNN smooth* averages the UMI counts for each cell with its nearest neighbors iteratively, which guarantees a reduction in noise, but may introduce new biases when the kNN search is influenced by technical variation (i.e. if the neighbor of a cell is not biologically relevant but technically relevant, such as cells coming from the same batch or having the same library size). *ZINB-WaVE* on the other hand was able to preserve the biological structures of the mixtures, but had much lower correlations among technical replicates after normalization.

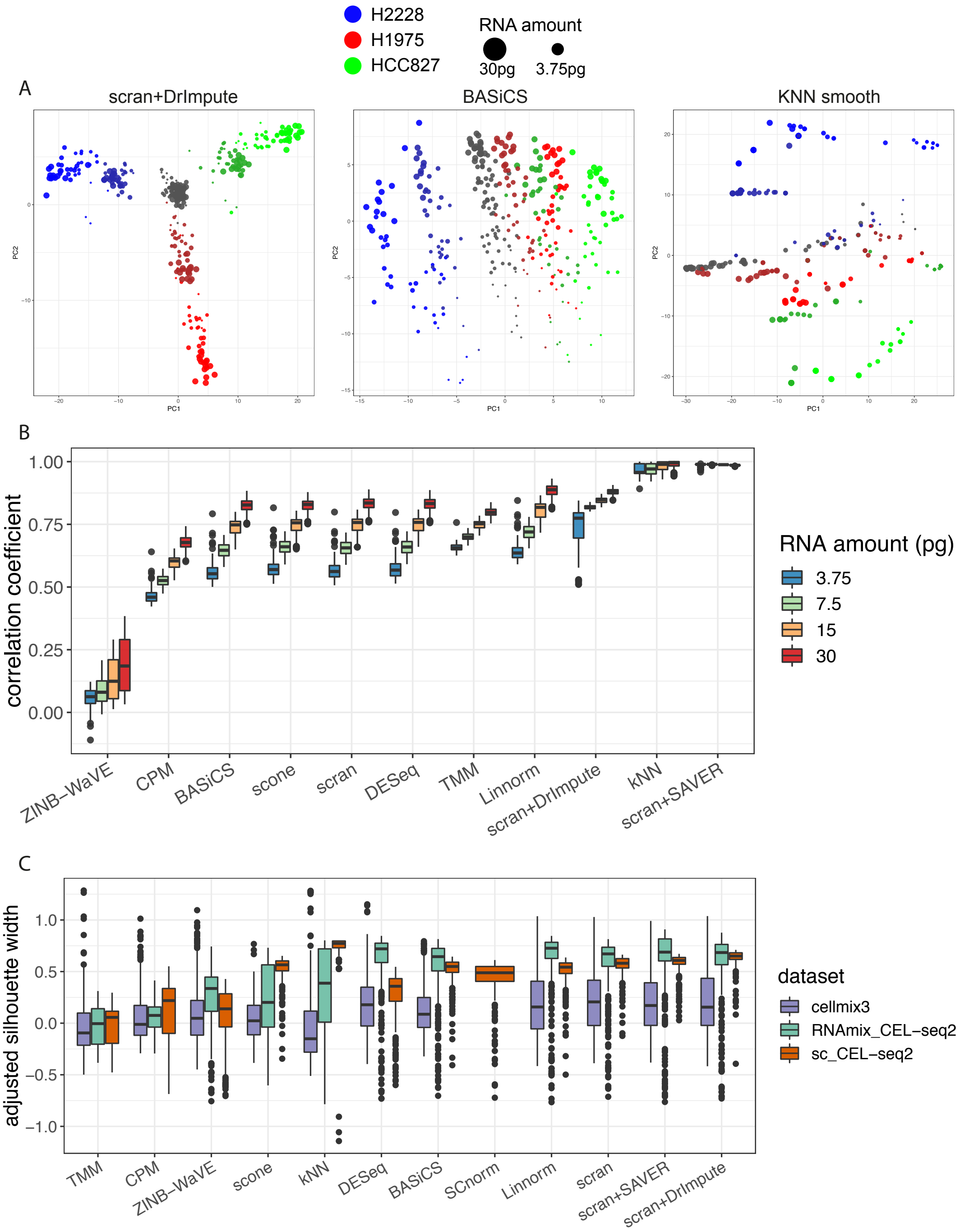


**Figure 2. Comparisons of normalization and imputation methods using multiple mixture datasets** (**A**) Example PCA plots after normalization and imputation by different methods using the RNAmix_CEL-seq2 dataset. (**B**) Pearson correlation coefficients of technical replicates in the RNAmix_CEL-seq2 dataset after normalization and imputation by different methods. (**C**) Silhouette widths calculated using the known cell/mixture groups after different normalization and imputation methods, using the datasets cellmix3, RNAmix_CEL-seq2 and sc_CEL-seq2. The distance is normalized against the baseline silhouette width obtained from the raw unnormalized read counts.

## Comparison of clustering methods

The benchmark datasets we designed vary in population heterogeneity, ranging from low in the single cell dataset to medium in the RNA mixture experiment, to high in the 9-cell mixtures. This allowed us to assess clustering performance in a variety of settings. Six methods, including *RaceID* [9], *RaceID2* [10], *RCA* [25], *Seurat* [31], *clusterExperiment* [35] and *SC3* [21], were evaluated using all benchmarking datasets. We measured the performance of the different clustering methods by calculating the entropy of cluster accuracy and cluster purity. The entropy of cluster accuracy is defined as the average of the true cluster labels within each cluster computed for each method. A low entropy of cluster accuracy indicates that the cells in a cluster identified by the method are homogeneous, which means that cells belonging to the same cluster are likely to be assigned to the same cell type. Some methods over-cluster which will produce low entropy of cluster accuracy; an extreme example would be if each cell is assigned its own individual cluster, in which case the entropy of cluster accuracy would be zero. Therefore, we use a second metric, entropy of cluster purity, to measure whether cells that have the same annotated group will have a similar cluster specification as calculated by the clustering method. In contrast to the entropy of cluster accuracy, the entropy of cluster purity lacks control of under-clustering and in an extreme case it can be zero when all cells are assigned to the same cluster. These two metrics were used together, to account for both under-clustering and over-clustering for each method (Figure 3A). We found good correlation between these two metrics and the Adjusted Rand Index (ARI) [17], which is a commonly used metric to evaluate the clustering performance by computing the similarity to the annotated clusters (Supplementary Figure 4). Unlike the ARI which only measures similarity, these metrics can capture both under-clustering and over-clustering and reveal more heterogeneity among different methods.

The results for 3 representative datasets are shown in Figure 3, with the remaining shown in Supplementary Figure 4. No method uniformly outperformed others across all situations under default settings. In general, *SC3* and *Seurat* achieved a good balance between under-clustering and over-clustering across all datasets, performing best when there was clear separation between cell types, as was the case in the single cell datasets (Figure 3B). The accuracy of all methods was lowest in the 9-cell mixture dataset (Figure 3D, these datasets also had the lowest ARI values as shown in Supplementary Figure 4) due to the continuous population structure which gives low separation between different clusters. *RaceID2* had high accuracy, but produced many more clusters than other methods, suggesting that it may be best suited to datasets where many small populations exist. On the other hand, *RaceID* significantly underestimated the number of clusters, returning fewer clusters than the optimal number. The *clusterExperiment* method failed to produce results for all RNA mixture and cell mixture datasets.

## Comparison of trajectory analysis methods

Five methods, including *Slingshot* [46], *Monocle2* [36], *SLICER* [53], *TSCAN* [18] and *DPT* [11] were evaluated using the RNA mixture and cell mixture datasets. These datasets were chosen as they both contain clear 'pseudo' trajectory paths from one pure cell line to another that are driven by controlled variations in RNA amount. For simplicity, we chose H2228 as the root state of the trajectory (Figure 4A). We evaluated the correlation between the pseudotime generated from each method and the rank order of the path from H2228 to the other cell lines (Figure 4B) to examine whether each method can position cells in the correct order. In addition, we calculated the coverage of the trajectory path (Figure 4C), which is the percentage of cells that have been assigned to the correct path, and assesses the sensitivity of the method. We randomly sampled highly variable genes as input, to assess the variation and robustness of each method. *Slingshot* and *Monocle2* showed robust results according to both metrics and generated meaningful representations of the trajectory, while *Slingshot* sometimes gave an extra trajectory path (Supplementary Figure 5). In contrast, *SLICER* places all cells in the correct path but was unable to order them correctly or recover the expected structure induced by the mixture designs.

Despite the similar performance of *Slingshot* and *Monocle2*, their results differ in terms of the way they position the cells. *Slingshot* does not perform dimensionality reduction itself and presents the result as is, whereas *Monocle2* uses DDR-tree for dimensionality reduction, and tends to place cells at the nodes of the tree rather than in transition between two nodes (Figure 4A). For example, the RNA

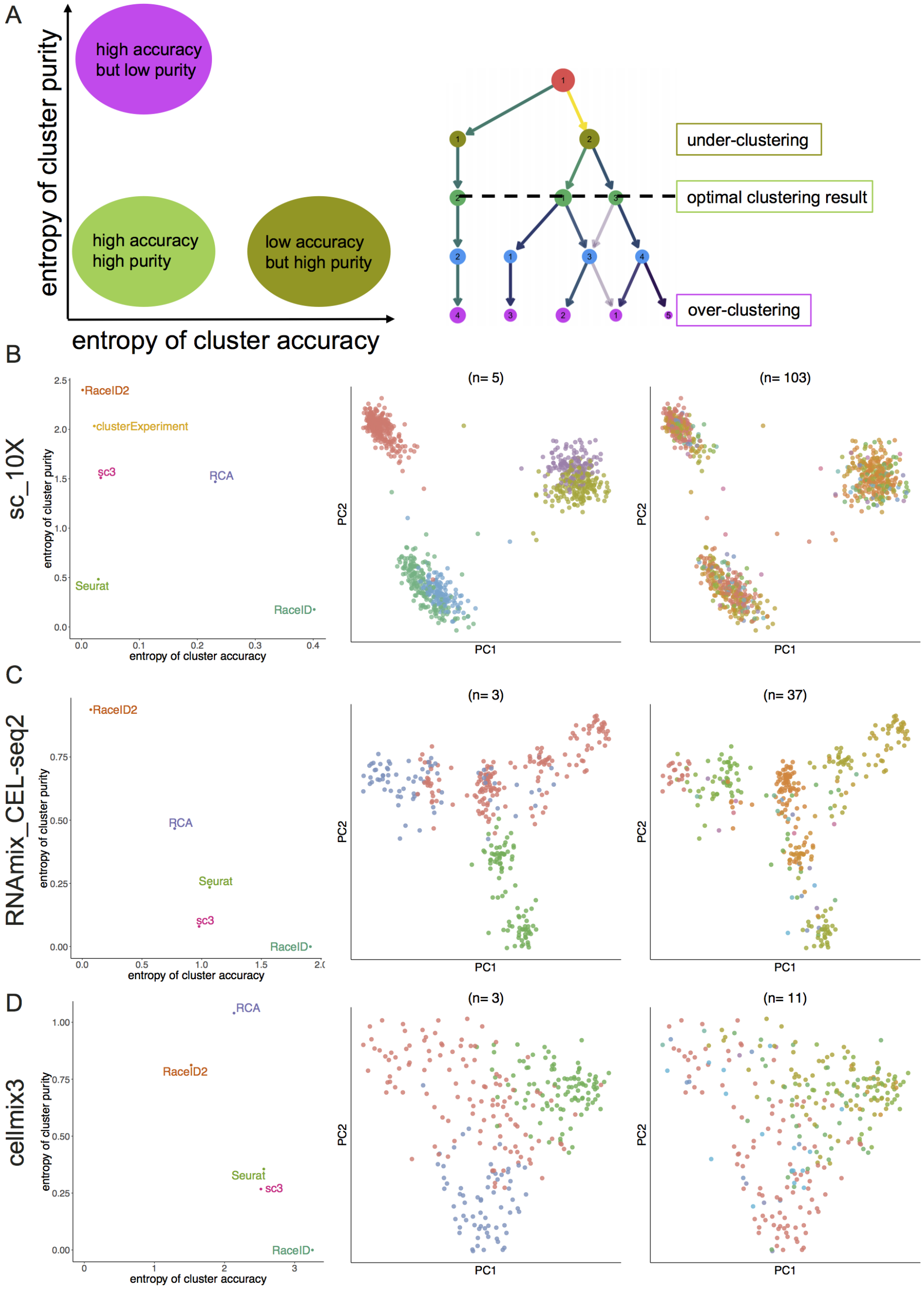


**Figure 3. Comparison of scRNA-seq clustering methods.** (**A**) An overview of the evaluation approach. High entropy of cluster accuracy measures the degree of over-clustering, while high entropy of cluster purity measures under-clustering. The clustering tree adopted from the package *Clustree* [56]. (**B,C,D**) Entropy of cluster purity versus entropy of cluster accuracy for representative single cell (sc_10X), RNA mixture (RNAmix_CEL-seq2) and cell mixture (cellmix3) experiments. The clustering results for two methods, *Seurat* and *RaceID2* were plotted for each design, with cells colour-coded by cluster assignment. The value of $n$ indicates the number of clusters found by each method.

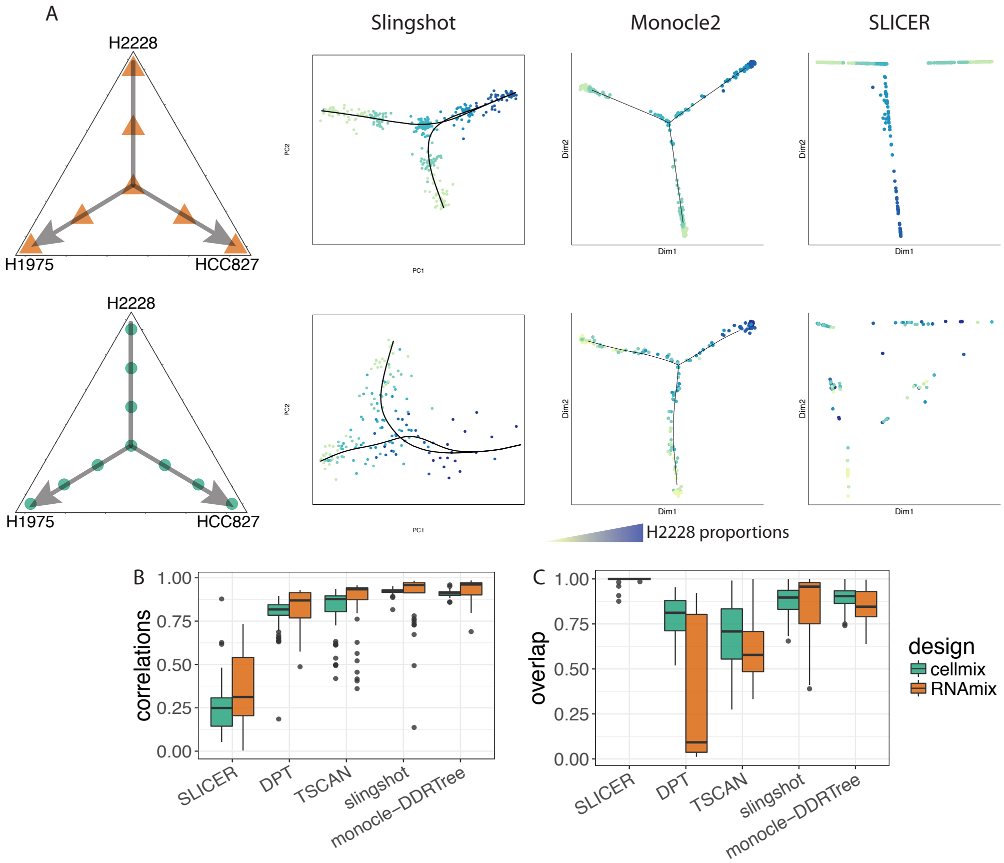


**Figure 4. Comparison of scRNA-seq trajectory analysis methods.** (**A**) The trajectory path chosen for the RNA mixture dataset (top) and cell mixture dataset (bottom) along with visualizations of the output from *Slingshot*, *Monocle-DDRTree* and *SLICER*. Cells are coloured by the proportion of H2228 RNA present, which was chosen as the root of the trajectory. (**B**) Boxplot showing the Pearson correlation coefficient between the calculated pseudotime and the ground-truth for each method, with genes randomly sub-sampled to assess robustness. (**C**) The percentage of cells that are correctly assigned to the trajectory, again with random sub-sampling of genes to assess robustness.

mixture dataset has 7 clusters by design which are equally distributed along the path between one pure cell line and another. *Monocle2* assigns most of the cells to the three terminal states, leaving only a few in between, which does not reflect the designed structure. Indeed, this feature might fit real situations in cell differentiation, where most cells are in defined cell states with only a small proportion in transition between different groups. However, such an assumption may not always hold and care is therefore needed when interpreting the results.

## Comparison of data integration methods

Whilst combining scRNA-seq data between studies is an attractive way to increase cell number and ensure reproducible results, there are many challenges such as the high drop-out rate, large technical noise introduced by library preparation and difference in sequencing depth. Several methods have been proposed to solve this problem, but havn't been compared by a well-designed benchmark dataset. Since the single cell and RNA mixture data were generated using multiple protocols, we used these datasets to compare these state-of-art methods including *MNN* [12], *Scanorama* [14], *scMerge* [28], *ZINB-WaVE* [37], *Seurat* [43] and *MINT* [40]. As expected, when naively combining the independent datasets (single cells

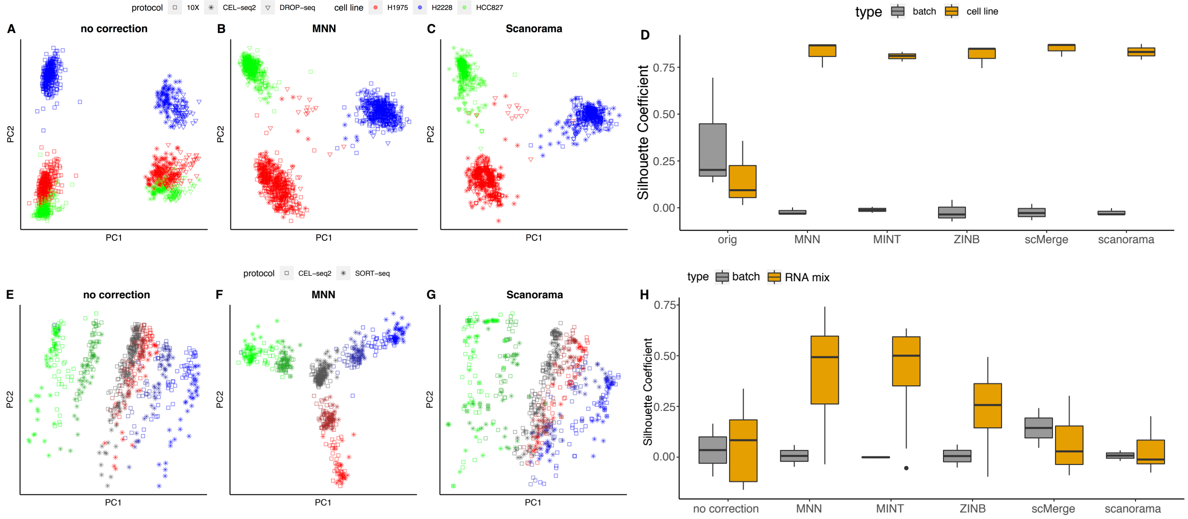


**Figure 5. Comparisons of data integration methods for batch effect correction for the three single cell experiments and the RNA mixture experiments.** (**A,E**) PCA sample plot when the protocols are naively combined, highlighting a technical effect (all common 13,575 genes are considered). (**B-C, F-G**) PCA on the most variable genes (861 for single cells and 1,136 for RNA mixtures) for *MNN* and *Scanorama* (see Supplementary Figure 6 for the other methods that were assessed). (**D, H**) Silhouette coefficients calculated on either the batch information or known sample group for the methods that output a batch-corrected data matrix; if effective at removing the technical effect, the coefficient should be low for the batch information and high for the cell groups as biological variation is retained.

Figure 5**A**, RNA mixtures Figure 5**E**) the largest source of variability was technical (related to single cell protocol) rather than biological (related to cell line/mixture group), resulting in a strong batch effect (Supplementary Figure 6**A,D**). For each design, we applied several integrative methods developed for combining independent batches of experiments. Recently proposed methods such as *MNN*, *Scanorama*, *scMerge* and *ZINB-WaVE* generate batch-corrected data which can then be analyzed using other downstream analysis tools. Diagonal Canonical Correlation Analysis combined with Dynamic Time Warping from *Seurat* and *MINT* [40] output a low-dimensional representation of the data. *MINT* includes an embedded gene selection procedure whilst projecting the data into a lower dimensional space (Supplementary Table 4). Dimension reduction of the results from different methods via PCA, t-SNE [51] for *Seurat* or resulting components (*ZINB-WaVe, MINT*) show variations in their ability to handle batch effects that depends on the complexity of the data. We assessed performance both graphically (Figure 5**B,C, F, G** and Supplementary Figure 6) and numerically by calculating the silhouette coefficient according to protocol and known cell line/mixture group information (Figure 5**D, H**). To reduce computational time, all methods with the exception of *scMerge* and *MINT* were run on the most highly variable genes (861 for single cells and 1,136 for RNA mixtures, 10,000 genes for *Scanorama*). For the single cell design, most methods perform comparably well and are effective at removing or adjusting for differences between protocols. For the RNA mixture design that includes a larger number of groups (7) and a smaller number of cells per group, most methods, with the exception of *MNN* and *MINT*, were unable to successfully remove technical variation from the data. In most studies where cell types are unknown, data integration is followed by clustering analysis and differential expression analysis to identify marker genes. Therefore, choosing the most appropriate integrative method to either correct or adjust for technical variation is crucial for downstream analysis.

# Discussion

We designed and generated a comprehensive scRNA-seq benchmarking dataset with varying levels of biological noise and in-built ground truth via population structure that ranges from simple to complex. These datasets incorporate various mixture designs processed using multiple single cell technologies to facilitate comparisons of both protocols and methods. Although our comparison shows the 10X Chromium platform to produce the highest quality data, both Drop-seq and CEL-seq2 are very flexible protocols, with various parameters that can be optimized and tuned. When fully optimized or under different conditions, these protocols may well be comparable to 10X. These datasets allowed us to study the intrinsic properties of scRNA-seq data and scRNA-seq protocols, such as the systematic differences in intron reads between protocols, which has been underexplored in previous studies. As new methods such as RNA velocity [24] and pipelines such as zUMIs [34] start to incorporate intron reads into their analysis, it is important for researchers to be aware of protocol-specific biases that may influence intron abundance and potentially confound analyses that take place across different scRNA-seq platforms. Moreover, these data show that ERCCs are sampled independently of endogenous mRNA, which casts doubt on the common assumption that ERCC spike-ins are less likely to be sampled in cells with more endogenous mRNA. ERCC spike-ins can therefore be used to measure technical noise which is orthogonal to biological variation.

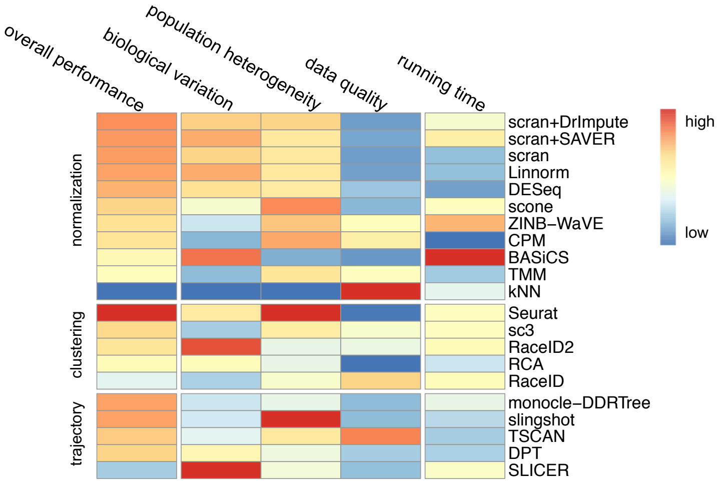


**Figure 6. Summary of results from methods comparisons using scRNA-mixology datasets.** Methods are ranked by overall performance in each category. The correlation between performance and data characteristics shows whether the result of a method is likely to be influenced by that characteristic. For example, the performance of *kNN smooth* changes a lot in the presence of variations in data quality, while *Seurat* has good performance in different situations. Results have been scaled and standardized to use the same colour scale.

To demonstrate the broad utility of these data, we performed systematic methods comparisons for 4 key tasks; normalization and imputation, clustering, trajectory analysis and data integration. The performance of methods varied across different datasets, with no clear winners in all situations, however, consistently satisfactory results were observed for *scran*, *Linnorm*, *DrImpute* and *SAVER* for normalization and imputation; *Seurat* and *SC3* for clustering; *Monocle2* and *Slingshot* for trajectory analysis and *MNN* for data integration. Interestingly, the various ensemble methods, which combine results from multiple algorithms in a bid to improve performance, did not always outperform individual methods. While the ensemble method *SC3* for clustering generally gave good performance, the *clusterExperiment* method was less reliable and the ensemble method for normalization, *scone* also gave mixed results on different datasets. Having multiple benchmark datasets with different numbers of cell groups and varying levels of biological noise allowed us to objectively assess performance with different data characteristics. As summarized in Figure 6, performance could be correlated to biological variation, population heterogeneity and data quality, with some methods more sensitive to these characteristics than others, indicating that they might best suit particular cases rather than general use. As an example, the trajectory methods *Monocle2* and *Slingshot* have similar overall performance, but *Slingshot* is more sensitive to the population structure than *Monocle2*, which means the latter may be better in more general situations. Data integration for scRNA-seq is still an emerging field, with limited systematic benchmarking performed to date. The varying complexity of our datasets provides an excellent resource for further methods development.

Our comparison is subject to a number of limitations such as the relatively small cell numbers and the linear mixture settings, which may not be a realistic model for developmental trajectories where regulatory gene expression may be non-linear and non-systemic. Also methods are mostly compared under default settings, which may not give optimal performance across all datasets.

Our benchmarking study serves as a demonstration of the different types of comparisons that can be performed using the comprehensive designs. Therefore, rather than performing an exhaustive comparison of all the methods available for every task, we have chosen to demonstrate a breadth of applications across key analysis tasks. These data can also be used to test the performance of different methods for data preprocessing (alignment, UMI deduplication, gene-level quantification), dimension reduction (PCA, t-SNE, UMAP) and differential expression analysis. They could also be used to explore any interactions between different methods (such as normalization and clustering) and performance, although the number of combinations, even with a few methods selected for each task would quickly become very large. Our benchmarking dataset will benefit future package developers as it allows new methods to be evaluated on the same standards, avoiding ambiguity caused by cherry-picking evaluation datasets. We hope that this study will reinvigorate interest in the important area of benchmark data generation and analysis, providing new insights into current best practice and guide the development of better scRNA-seq algorithms in the future to ensure the biological insights derived from single cell technology stand the test of time.

# Methods

## Study design

Three human lung adenocarcinoma cell lines HCC827, H1975 and H2228 were cultured separately and the same batch was processed in three different ways (Figure 1). Firstly, single cells from each cell line were mixed in equal proportions, with libraries generated using three different protocols: CEL-seq2, Drop-seq with Dolomite equipment and 10X Chromium.

Secondly, single cells were sorted from the three cell lines into 384-well plates, with an equal number of cells per well in different combinations. For most of the wells, we sorted 9 cells in total, with different combinations of three cell lines distributed in“pseudo trajectory” paths (Supplementary Figure 1B), where the major trajectory is similar to the RNA mixture design while the minor trajectory is the combination that only contains two cell lines instead of three, which is similar in design our previous study [15]. For the major trajectory, we also included the population control for each combination, which includes 90 cells in total (i.e a large sample) instead of 9, while maintaining the cell combinations from the different cell lines. Apart from the trajectory design, we also varied the cell numbers and qualities to study the data characteristics in these configurations. We include 9 replicates with 3 cells in total with one cell from each cell line, to simulate “small cells”. 20 cells with low integrity identified by PI staining. The 9-cell wells were sub-sampled after pooling to get single cell equivalents of RNA, with three replicates in 1/9 and one in 1/3. We applied different clean up ratios to the three replicates after library generation to induce batch effects of a purely technical nature and study how clean up affects the data.

Thirdly, the RNA were extracted in bulk for each cell line and the RNA was mixed in 7 different proportions and diluted to single cell equivalent amounts (Supplementary Figure 1)A). In total, there are 8 mixtures in the plate layout with 49 replicates of each mixture. The mix1 and mix2 samples have the same proportions of the three cell lines (H2228:H1975:HCC827 $\frac{1}{3} : \frac{1}{3} : \frac{1}{3}$) but were prepared separately in order to assess the variation introduced during the RNA dilution and mixture step. In addition to the RNA mixtures, we also designed a dilution series in the same plate to create variations in the amount of RNA added. The amounts ranged from 3.75pg to 30pg (Figure 2 A2) and were intended to simulate differences in cell size. In total, each mixture will have 4 different RNA starting amounts with replicate numbers per mixture of 6:14:14:14 for the 3.75:7.5:15:30 pg group respectively.

## Cell culture and mRNA extraction

The human lung adenocarcinoma cell lines H2228, H1975 and HCC827 were retrieved from ATCC (https://www.atcc.org/) and cultured in Roswell Park Memorial Institute (RPMI) 1640 medium with 10% fetal calf serum (FCS) and 1% Penicillin-Streptomycin. The cells were grown independently at 37°C with 5% carbon dioxide until near 100% confluency.

For the three cell lines, cells were dissociated into single cell suspensions in FACS buffer and sorted for the 9-cell-mixture and single cell experiment (see below for sorting strategy). The remaining cells were centrifuged and frozen at -80°C for later RNA extraction. RNA was extracted using a Qiagen RNA miniprep kit. The amount of RNA was quantified using both Invitrogen Qubit fluorometric quantitation and an Agilent 4200 bioanalyzer to get an accurate estimation. The extracted RNA was then diluted to 60 ng/$\mu$l and then mixed in different proportions, according to the study design. The different mixtures were further diluted to create an RNA series that ranged from 3.75pg to 30pg that was dispensed into CEL-seq2 and SORT-seq primer plates using a Nanodrop II dispenser. Prepared RNA mixture plates were sealed and immediately frozen upside down at -80°C.

## Cell sorting and single cell RNA sequencing

For CEL-seq2, single cells were flow sorted into chilled 384-well PCR plates containing 1.2$\mu$l of primer/ERCC mix using a BD FACSAria III flow cytometer. Sorted plates were sealed and immediately frozen upside down at -80°C. These plates, together with the RNA mixture plates, were taken from -80°C and processed using an adapted CEL-Seq2 protocol with the following variations. The second strand synthesis was performed using NEBNext Second Strand Synthesis module in a final reaction volume of 8 $\mu$l and NucleoMag NGS Clean-up and Size select magnetic beads were used for all DNA purification and size selection steps. For the 9-cell-mixture plates, clean up of the PCR product was performed with 2×0.7-0.9 bead/DNA ratio. For the single cell and RNA mixture plates, two different clean up ratios for the PCR product were used (0.8 followed by 0.9). The choice of clean up ratio was optimized from the QC results of the 9-cell-mixture data and the SORT-seq protocols.

The 9-cell-mixture plates were sorted according to the plate design. Each well contained 9 cells in total in different combinations, and was processed using our adapted CEL-seq2 protocol described above with variations in the pooling step. After the second strand synthesis, materials from the 9-cell-mixtures and 90-cell population controls were pooled separately into different tubes and the volumes were measured. Then for the 9-cell-mixture sample, 3×1/9 and 1×1/3 of the total pooled material were taken and these four samples were processed separately in the following step. At the PCR product clean up stage, the clean up ratios for the 3×1/9 samples were 0.7, 0.8 and 0.9 respectively, and 0.7 for the 1/3 9-cell-mixture sample and the 90-cell population controls.

The SORT-seq protocol is similar to CEL-seq2 but uses oil to prevent evaporation. This allows reductions in the reaction volume which can be dispensed using the Nanodrop II liquid handling platform (GC biotech). In summary, 2.0$\mu$l vapor-lock oil was added to each well of the plate, followed by 0.1$\mu$l of primer/ERCC mix. The reaction volume for RT and first strand synthesis are 0.075$\mu$l and 0.568$\mu$l respectively. The composition of the various mixes was the same as for CEL-seq2. The sample pooling was achieved by centrifuging the plates upside down into a container covered with parafilm and carefully separating the oil from the other materials. The PCR clean up ratio used for SORT-seq was 0.8

followed by 0.9. We experienced significant sample loss during sample pooling such that only 60% of the total volumes were recovered, which is lower compared to the CEL-seq2 protocol (90%).

For the 10X and Drop-seq protocols, cells were PI stained and 120,000 live cells were sorted for each cell line by FACS to acquire an accurate equal mixture of live cells from the three cell lines. This mixture was equally split into three parts, where one part was then processed by the Chromium 10X single cell platform using the manufacturer's (10X Genomics) protocol. The second part was processed by Dolomite Drop-seq with standard Drop-seq protocols [22]. The third part was sorted in a 384-well plate and processed using the standard CEL-seq2 protocol, with a PCR clean up ratio of 0.8 followed by 0.9. All samples, including Drop-seq, 10X and CEL/SORT-seq, were sequenced on an Illumina Nextseq 500.

## Data preprocessing and quality control

*scPipe* was used for data preprocessing and quality control to generate a UMI-deduped gene count matrix per dataset. In general all data was aligned using *Rsubread* [26] to the GRCh38 human genome and its associated annotations, with ERCC spike-in sequences added as special chromosomes. For 10X, we processed the 4,000 most enriched cell barcodes, with `comp=3` used in the function `scPipe::detect_outliers` for quality control to remove poor quality cells. For CEL-seq2 and SORT-seq, we used the known cell barcode sequences for cell barcode demultiplexing and `comp=2` was used in the function `scPipe::detect_outliers` for quality control. The biological variation for each dataset were represented by BCV, calculated by `edgeR::estimateGLMCommonDisp`, using the known population structure in the design matrix. For the single cell datasets, the population structure is the cell line identity, while for the mixture data, the population structure is the mixture combination. The background contamination was high for Drop-seq, so we first ran `scPipe::detect_outliers` with `comp=3` to remove outlier cells and then ran it again with `comp=2` to remove the background noise which consists of droplets that did not contain beads. The quality control metrics, including intron reads for each cell, were generated during cell barcode demultiplexing by the function `scPipe::detect_outliers`. Intron reads are defined as any read that map to the gene body but do not overlap an annotated exon.

## Analysis of ERCC spike-in counts

For the RNA mixture dataset, the $\log_2$ transformed total UMI counts, total ERCC spike-in counts and RNA amounts were used as input for linear mixed models. The function `lme4::lmer` was used to fit the model [2]. Model I was formulated with `UMI_count ~ mix + (1 | mRNA_amount)`, where `UMI_count` is the total UMI count for each cell, `mix` is the mixture number as a factor and `mRNA_amount` is the amount of RNA encoded as a factor. Model I was extended to include ERCC spike-in counts as a covariate in Models II and III. Model II incorporated spike-in counts as a fixed-effect term using the formula `UMI_count ~ mix + ERCC_count + (1 | mRNA_amount)`. In Model III, spike-in counts were included as a random-effect that can change according to different RNA amounts: `UMI_count ~ mix + (ERCC_count | mRNA_amount)`. Models II and III were compared against Model I using `anova`. Akaike information criterion (AIC) and $p$-values are given in Supplementary Table 3.

## Data normalization and imputation

The raw counts were used as input to each algorithm, and all methods were blind to the RNA mixture proportions and RNA amounts. To have a fair comparison, the normalized counts from algorithms such as *BASiCS* [50] and *SCnorm* [1] do not generate values on a log-scale were further $\log_2$ transformed after an offset of 1 was added to the counts. We used *edgeR* [38] to calculate count-per-million (CPM) and TMM (trimmed mean of $M$-values) values. The current version of *BASiCS* requires spike-in genes, so we didn't apply it to our datasets generated by 10X or Drop-seq which both lacked ERCC spike-ins. For *scone* [4], we used housekeeping genes obtained from the Single Cell Housekeeping Genes website [27] as negative controls, set the maximum number of unwanted variation components to 3 for the removal of unwanted variation method (RUV) and ignored QC metrics. For other methods, we used their default parameters.

The Pearson correlation coefficient was calculated using gene expression after normalization or imputation for samples with the same RNA mixture proportion and the same mRNA amount, as these samples are technical replicates and any differences in gene expression should be contributed by technical noise. We performed PCA using normalized counts and calculated the silhouette width on the first two PCs to assess whether normalization was able to preserve the known structure. For any clustering of $n$ samples (here a cluster refers to a particular mixture or a cell line), the silhouette width of sample $i$ is defined as

$$sil(i) \equiv \frac{b(i) - a(i)}{max(a(i), b(i))} \in [-1, 1] \tag{1}$$

where $a(i)$ denotes the average distance (Euclidean distance over the first two PCs of expression measures) between the $i$th sample and all other samples in the cluster to which $i$ belongs to, and $b(i)$ is calculated as below: for all other clusters $C$,

$$b(i) = min_C d(i, C) \tag{2}$$

where $d(i, C)$ denotes the average distance (the same as described above) of $i$ to all observations to $C$. Methods with better performance have higher silhouette width. The function `silhouette` from the package `cluster` [32] was used to calculate the silhouette width.

Data analysis using *scran* (1.7.0), *DrImpute* (1.0) and *SCnorm* (1.1) was performed with default settings. For the RNA mixture data, *kNN smooth* (1.0) was run with $k = 8$ (increasing the number of $k$ significantly increases the correlation with technical replicates but destroys the biological information). We used $k = 2$ for *ZINB-WaVE* (1.1.5) normalization although varying the value of $k$ did not change our results. *BASiCS* (1.1.29) was run with 20,000 MCMC iterations, 1,000 warm up iterations and a thinning parameter of 10.

## Clustering

Our comparison of clustering methods used all mixture datasets apart from cellmix5 (which is the population control). To obtain truth for the single cell datasets, sc_CEL-seq2, sc_10x and sc_Drop-seq, we used *Demuxlet* [20], which exploits the genetic differences between the three different cell lines to determine the most likely identity of each cell. The predicted cell identities in each dataset corresponded largely to clusters seen when visualizing the data. Six methods, including *clusterExperiment* (1.99.2), *RaceID* (1.0), *RaceID2* (1.0), *RCA* (1.0), *SC3* (1.7.7) and *Seurat* (2.3.0) were compared. Each method is used as specified by the authors in its accompanying documentation. This includes any normalization and filtering steps. Furthermore, any parameters required were left as their defaults or chosen as described in the documentation.

In order to compare the performance of the clustering methods, we looked at two measures: entropy of cluster accuracy, $H_{accuracy}$, and cluster purity, $H_{purity}$. With $M$ and $N$ represent the cluster assignment generated from clustering methods and annotation (ground truth), we defined these measures as follows:

$$H_{accuracy} = -\frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} p(x_j)log(p(x_j))}{M} \tag{3}$$

$$H_{purity} = -\frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} p(x_j)log(p(x_j))}{N} \tag{4}$$

For the $H_{accuracy}$, $M$ denotes the cluster generated from a method, and $N_i$ is the real clusters in $i$th generated cluster. Similarly, in the $H_{purity}$ the $N$ denotes the real clusters while $M_i$ is the generated cluster for $i$th real cluster.

## Trajectory analysis

The comparison of trajectory analysis methods used all 9-cell mixture datasets (cellmix1 to cellmix4) and the RNA mixture dataset generated by the CEL-seq2 and Sort-seq protocols. For each dataset, the gene count matrix is normalized using the method in *scran* in cases where the method does not have an explicit normalization step. The top 1,000 most highly variable genes were selected using the `trendVar` and `decomposeVar` functions in *scran*, then 500 genes were randomly selected as the input features, repeated 10 times to assess the stability of the method. Five methods, including *Slingshot* (0.1.2), *Monocle2* (2.6.1), *SLICER* (0.2.0), *TSCAN* (1.16.0) and *DPT* (0.6.0) were compared on the above dataset. *Slingshot* requires the dimensionally reduced matrix and cluster assignment as input. Similar to the approach described in their paper, we used PCA (`scater::runPCA`) for dimensionality reduction and *k*-means clustering was performed on the first two PCs, then the first two PCs and the clustering results were used as input for *Slingshot*. DDR-Tree, a scalable reversed graph embedding algorithm, was used for *monocle2* for dimensionality reduction and tree construction. *SLICER* applies locally linear inference to extract features and reduce dimensions. To make it easier for comparison, the pure H2228 cells were selected as the root cells or root state when generating the trajectory and computing pseudotime. Then for the branching structure generated by each method, we searched for the best match to the two branches: H2228 to H1975 and H2228 to HCC827 and calculated the percentage of overlap of cells between the real path and the branch calculated by each method. Although we sampled genes to assess the robustness of the method, for the representative plot in Figure 4 we used the 1,000 most variable genes in the analysis.

## Data integration

Data integration methods main characteristics are described in Supplemental Table 4 using the R packages *zinbwave*(1.2.0), *scran*(1.8.2) for MNN, *Seurat*(2.3.4) for Diagonal Canonical Correlation Analysis (CCA) and *scMerge*(0.1.8). PCA and MINT analyses were performed using *mixOmics*(6.3.2) [41] and *Scanorama* (0.1) using the python library from Hie *et al.* (2018) [14].

We calculated the silhouette width to compare the clustering performance of the different methods to combine different protocols. In single cells and RNA mixtures the clusters are already defined based on either protocol (batch) information or cell line / RNA mixture information. Silhouette coefficients were calculated on the first two principal components from PCA for each method that output a data matrix (*MNN*, *ZINB-WaVe*, *scMerge* and *Scanorama*) or the first two resulting components for *MINT*. We excluded diagonal CCA from this analysis as the Euclidean distance that is calculated in the silhouette coefficient is not meaningful for t-SNE components. A high value for the batch clusters indicates that a strong protocol effect remains, whilst a low value for the cell group information indicates that the biological variation remains after the data integration process.

## Performance summary

Figure 6 summarizes the performance across all the evaluated methods and their correlation to data characteristics. For each method, a linear regression model is used to partition the variance of the performance to the data characteristics, where the data characteristics, such as biological variation and data quality, were used as covariates. Both the variance (sum of squares) and performance is $Z$-score scaled for better visualization on a heatmap.

## Data and code availability

These data are available under GEO SuperSeries GSE118767. A summary of the individual accession numbers is given in Supplementary Table 1. The processed `SingleCellExperiment` R objects, including all code used to perform the comparative analyses and generate the figures are available from https://github.com/LuyiTian/CellBench_data.

## Acknowledgements

We thank Dr Clare Weeden and Dr Marie-Liesse Asselin-Labat for providing the cell lines used in this study. We thank Mr Jaring Schreuder and Ms Dawn Lin for assistance in conducting experiments and Mr Isaac Virshup for assistance in the data integration process.

**Funding:** This work was supported by the National Health and Medical Research Council (NHMRC) Project Grants (GNT1143163 to MER, GNT1124812 to SHN and MER, GNT1062820 to SHN), Fellowship GNT1104924 to MER, GNT1087415 to KALC, the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (grant numbers 2018-182819 to MER, 2018-182885 to KALC), a Melbourne Research Scholarship to LT, Genomics Innovation Hub, Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS.

## Author contributions

LT designed, planned and performed experiments, conducted data analysis and wrote the manuscript. XD, SF, KALC and SS performed data analysis and wrote the manuscript. DAZ, TSW and AS performed experiments. SHN and MER designed the study. MER supervised the analysis and wrote the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

## References

1. R. Bacher, L. F. Chu, N. Leng, A. P. Gasch, J. A. Thomson, R. M. Stewart, M. Newton, and C. Kendziorski. SCnorm: Robust normalization of single-cell RNA-seq data. *Nature Methods*, 14(6):584–586, 2017.

2. D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 2015.

3. P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1098, 2013.

4. M. B. Cole, D. Risso, A. Wagner, D. DeTomaso, J. Ngai, E. Purdom, S. Dudoit, and N. Yosef. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *bioRxiv*, page 235382, 2017.

5. L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3):323–331, 2004.

6. A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, sep 2018.

7. S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 7(0):1297, aug 2018.

8. D. Grün, L. Kester, and A. Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014.

9. D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. Van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

10. D. Grün, M. J. Muraro, J. C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. van Es, E. Jansen, H. Clevers, E. J. de Koning, and A. van Oudenaarden. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, 19(2):266–277, 2016.

11. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016.

12. L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.

13. T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, and I. Yanai. CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):77, 2016.

14. B. L. Hie, B. Bryson, and B. Berger. Panoramic stitching of heterogeneous single-cell transcriptomic data. *bioRxiv*, page 371179, 2018.

15. A. Z. Holik, C. W. Law, R. Liu, Z. Wang, W. Wang, J. Ahn, M. L. Asselin-Labat, G. K. Smyth, and M. E. Ritchie. RNA-seq mixology: Designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Research*, 45(5):1–18, 2017.

16. M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, jun 2018.

17. L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

18. Z. Ji and H. Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117–e117, jul 2016.

19. L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, sep 2011.

20. H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, R. E. Gate, S. Mostafavi, A. Marson, N. Zaitlen, L. A. Criswell, and C. J. Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, dec 2018.

21. V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hemberg. SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, 2017.

22. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, may 2015.

23. I.-Y. Kwak, W. Gong, N. Koyano-Nakagawa, and D. Garry. DrImpute: Imputing dropout events in single cell RNA sequencing data. *bioRxiv*, 19(1):181479, dec 2017.

24. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, aug 2018.

25. H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan, M. Wong, P. J. Choi, L. J. Wee, A. M. Hillmer, I. B. Tan, P. Robson, and S. Prabhakar. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49(5):708–718, 2017.

26. Y. Liao, G. K. Smyth, and W. Shi. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10), 2013.

27. Y. Lin, S. Ghazanfar, D. Strbenac, A. Wang, E. Patrick, T. Speed, J. Yang, and P. Yang. Housekeeping genes, revisited at the single-cell level. *bioRxiv*, page 229815, 2017.

28. Y. Lin, S. Ghazanfar, K. Wang, J. A. Gagnon-Bartsch, K. K. Lo, X. Su, Z.-G. Han, J. T. Ormerod, T. P. Speed, P. Yang, et al. scmerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo-replication. *bioRxiv*, page 393280, 2018.

29. M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, Dec 2014.

30. A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, 2016.

31. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

32. M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2018. R package version 2.0.7-1 — For new features, see the 'Changelog' file (in the package source).

33. M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. de Koning, and A. van Oudenaarden. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394, 2016.

34. S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann. zUMIs -A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, (June):1–9, 2017.

35. E. Purdom and D. Risso. *clusterExperiment: Compare Clusterings for Single-Cell Sequencing*, 2017. R package version 1.4.0.

36. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, 2017.

37. D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J. P. Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284, dec 2018.

38. M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

39. M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3):R25, Mar 2010.

40. F. Rohart, A. Eslami, N. Matigian, S. Bougeard, and K.-A. Lê Cao. Mint: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC bioinformatics*, 18(1):128, 2017.

41. F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao. mixomics: An r package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):1–19, 11 2017.

42. W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, page 276907, 2018.

43. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

44. Sequencing Quality Control Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9):903–914, sep 2014.

45. C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, feb 2018.

46. K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, dec 2018.

47. V. Svensson, K. N. Natarajan, L. H. Ly, R. J. Miragaia, C. Labalette, I. C. Macaulay, A. Cvejic, and S. A. Teichmann. Power analysis of single-cell rnA-sequencing experiments. *Nature Methods*, 14(4):381–387, 2017.

48. L. Tian, S. Su, X. Dong, D. Amann-Zalcenstein, C. Biben, A. Seidi, D. J. Hilton, S. H. Naik, and M. E. Ritchie. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput Biol*, 2018.

49. P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7(September 2016):39921, jan 2017.

50. C. A. Vallejos, J. C. Marioni, and S. Richardson. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology*, 11(6):e1004333, 2015.

51. L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

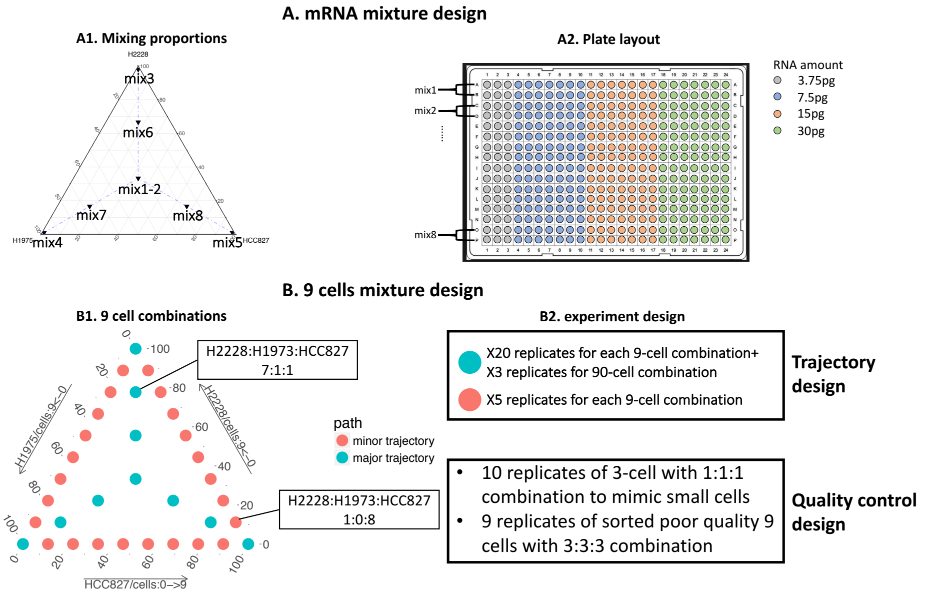
52. F. Wagner, Y. Yan, and I. Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*, page 217737, 2018.

53. J. D. Welch, A. J. Hartemink, and J. F. Prins. SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, 17(1):047845, 2016.

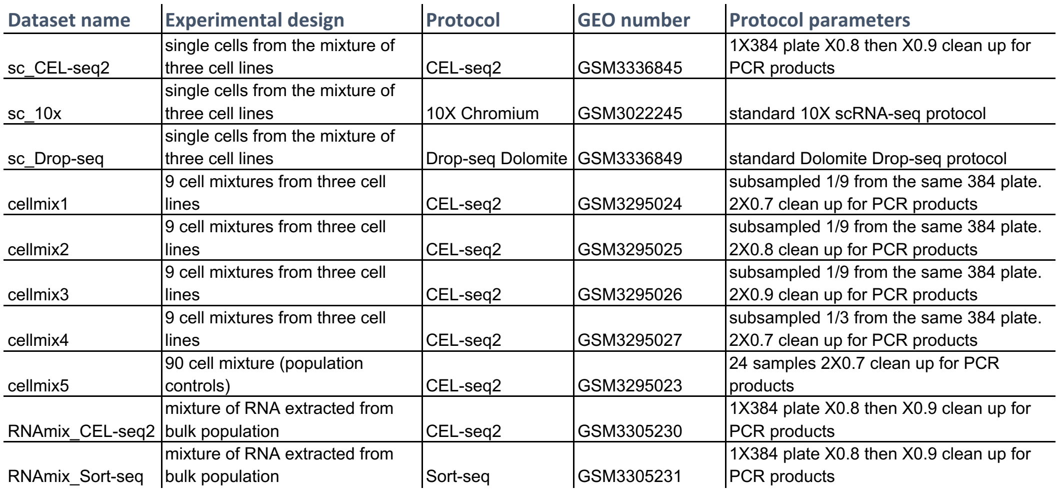
54. S. H. Yip, P. C. Sham, and J. Wang. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings in Bioinformatics*, (October 2017):1–7, 2018.

55. S. H. Yip, P. Wang, J. P. A. Kocher, P. C. Sham, and J. Wang. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic acids research*, 45(22):e179, 2017.

56. L. Zappia and A. Oshlack. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience*, 7(7):giy083, 2018.

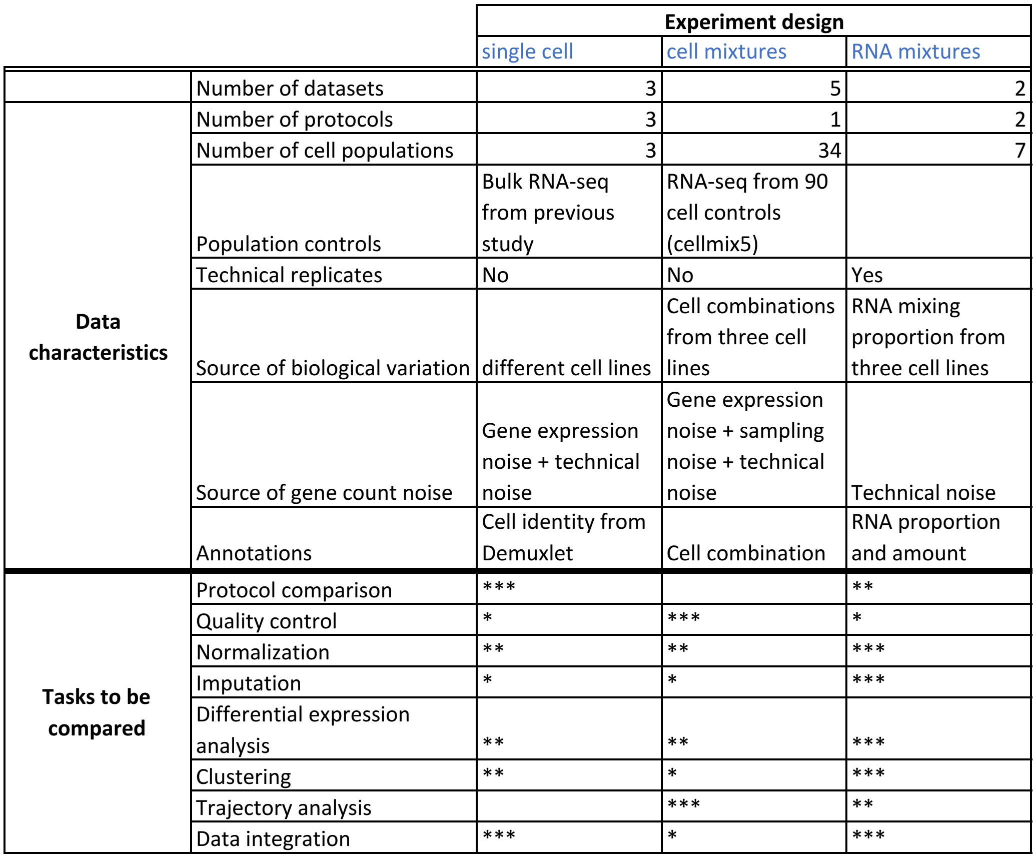


**Supplementary Figure 1. An overview of the RNA mixture and 9-cell mixture designs from the benchmark study.** (**A**) Mixing RNA extracted from bulk samples to get 8 mixtures with different proportions of RNA from 3 cell lines (**A1**), with different amounts of RNA for each ranging from 3.75pg to 30pg (**A2**). (**B**) Cell mixtures, with 9 cells in total for each well in various combinations from the 3 cell lines (**B1**). The number of replicates for each combination varies, as does the number of low quality control samples included (**B2**).



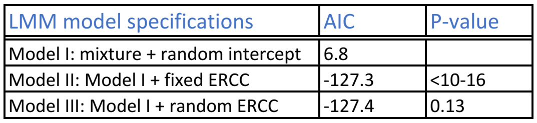
| Dataset name | Experimental design | Protocol | GEO number | Protocol parameters |
|---|---|---|---|---|
| sc_CEL-seq2 | single cells from the mixture of three cell lines | CEL-seq2 | GSM3336845 | 1X384 plate X0.8 then X0.9 clean up for PCR products |
| sc_10x | single cells from the mixture of three cell lines | 10X Chromium | GSM3022245 | standard 10X scRNA-seq protocol |
| sc_Drop-seq | single cells from the mixture of three cell lines | Drop-seq Dolomite | GSM3336849 | standard Dolomite Drop-seq protocol |
| cellmix1 | 9 cell mixtures from three cell lines | CEL-seq2 | GSM3295024 | subsampled 1/9 from the same 384 plate. 2X0.7 clean up for PCR products |
| cellmix2 | 9 cell mixtures from three cell lines | CEL-seq2 | GSM3295025 | subsampled 1/9 from the same 384 plate. 2X0.8 clean up for PCR products |
| cellmix3 | 9 cell mixtures from three cell lines | CEL-seq2 | GSM3295026 | subsampled 1/9 from the same 384 plate. 2X0.9 clean up for PCR products |
| cellmix4 | 9 cell mixtures from three cell lines | CEL-seq2 | GSM3295027 | subsampled 1/3 from the same 384 plate. 2X0.7 clean up for PCR products |
| cellmix5 | 90 cell mixture (population controls) | CEL-seq2 | GSM3295023 | 24 samples 2X0.7 clean up for PCR products |
| RNAmix_CEL-seq2 | mixture of RNA extracted from bulk population | CEL-seq2 | GSM3305230 | 1X384 plate X0.8 then X0.9 clean up for PCR products |
| RNAmix_Sort-seq | mixture of RNA extracted from bulk population | Sort-seq | GSM3305231 | 1X384 plate X0.8 then X0.9 clean up for PCR products |

**Supplementary Table 1. Summary of the benchmarking datasets generated.** Information on the 3 experimental designs employed, the single cell protocols used, the GEO accession numbers and parameters applied when generating cDNA libraries is listed.



| | | Experiment design | | |
|---|---|---|---|---|
| | | single cell | cell mixtures | RNA mixtures |
| | Number of datasets | 3 | 5 | 2 |
| Data characteristics | Number of protocols | 3 | 1 | 2 |
| | Number of cell populations | 3 | 34 | 7 |
| | Population controls | Bulk RNA-seq from previous study | RNA-seq from 90 cell controls (cellmix5) | |
| | Technical replicates | No | No | Yes |
| | Source of biological variation | different cell lines | Cell combinations from three cell lines | RNA mixing proportion from three cell lines |
| | Source of gene count noise | Gene expression noise + technical noise | Gene expression noise + sampling noise + technical noise | Technical noise |
| | Annotations | Cell identity from Demuxlet | Cell combination | RNA proportion and amount |
| Tasks to be compared | Protocol comparison | *** | | ** |
| | Quality control | * | *** | * |
| | Normalization | ** | ** | *** |
| | Imputation | * | * | *** |
| | Differential expression analysis | ** | ** | *** |
| | Clustering | ** | * | *** |
| | Trajectory analysis | | *** | ** |
| | Data integration | *** | * | *** |

**Supplementary Table 2. Summary of the data characteristics and data analysis tasks that can be compared by each experimental design.** The suitability of each experimental design to benchmark specific tasks is indicated by the scale $* < ** < ***$ i.e. the RNA mixture datasets include a dilution series which induces different dropout levels, making it an ideal dataset for comparing imputation methods.

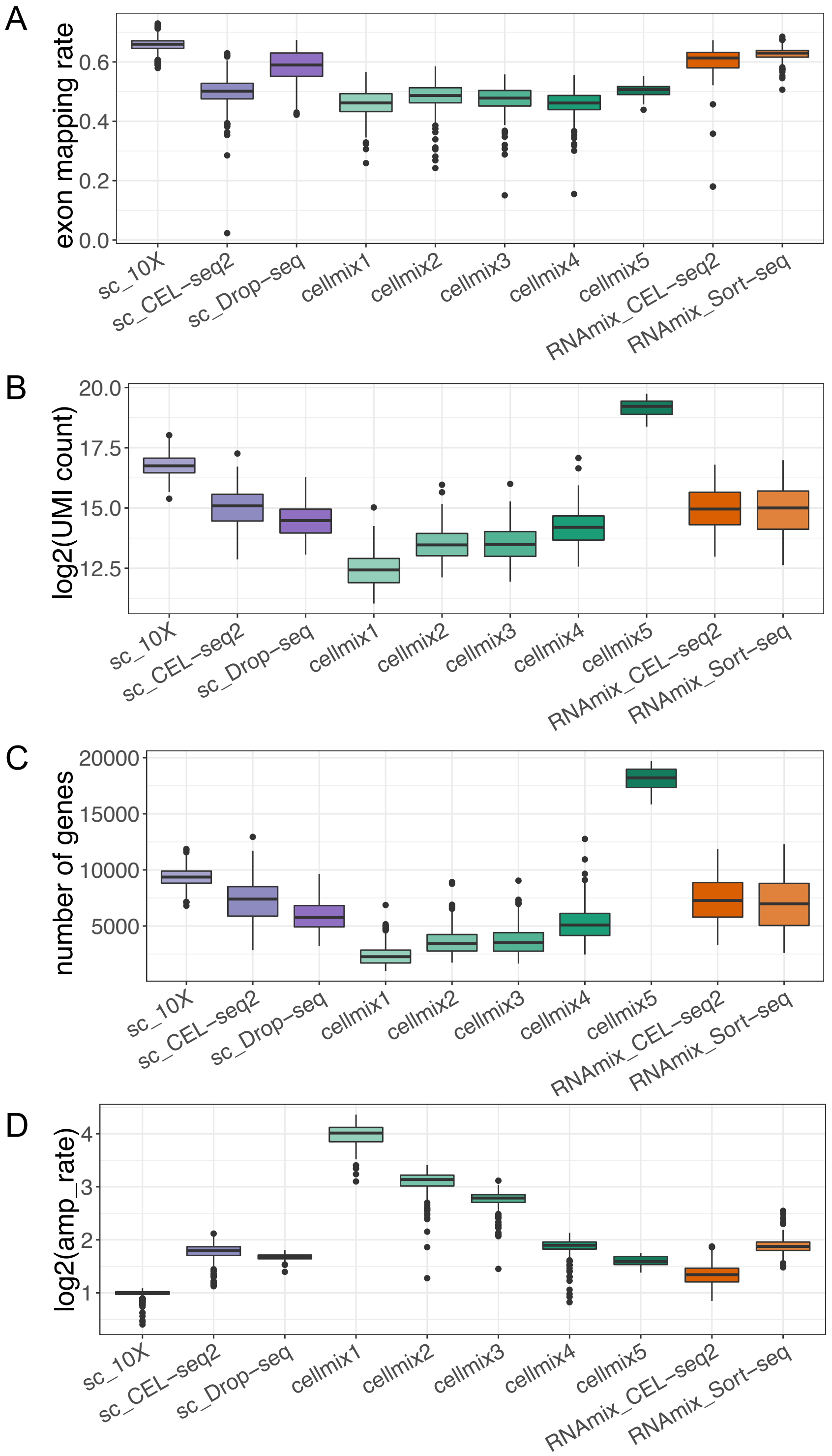


| LMM model specifications | AIC | P-value |
|---|---|---|
| Model I: mixture + random intercept | 6.8 | |
| Model II: Model I + fixed ERCC | -127.3 | <10-16 |
| Model III: Model I + random ERCC | -127.4 | 0.13 |

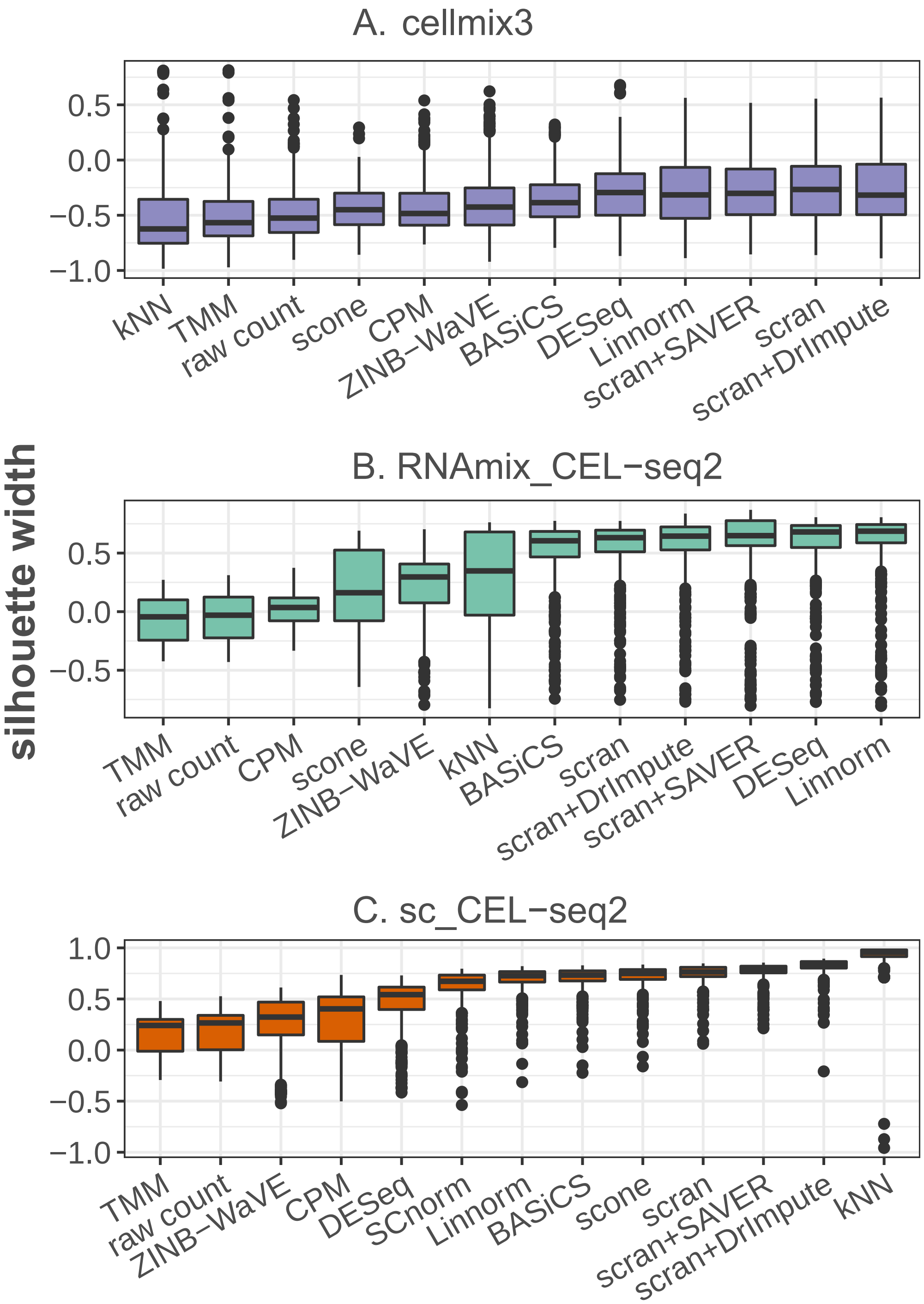
**Supplementary Table 3. Summary of the ANOVA model comparison.** Table of results from the model comparison of 3 linear mixed models, with the AIC and $p$-value which tests whether the coefficients added explain additional variation relative to the simpler model. Model I assumes a random intercept for different RNA amounts, while model II adds a fixed effect for the ERCC spike-ins (i.e. constant effect for different RNA amounts) and model III assumes both the coefficient and the intercept are random for different RNA amounts.

**Supplementary Table 4. Summary of integrative methods used to combine data from different protocols and scNRA-seq studies**. Methods can be classified into batch effect correction - where a batch-corrected data matrix is output, adjustment where the batch effect is accounted for in the model and dimension reduction methods where components or factors summarizing the batch-corrected data are output. Their hyperparameters are listed (*italic* indicates default parameters). HVG stands for Hyper-Variable Genes. SEG stands for Stably Expressed Genes.

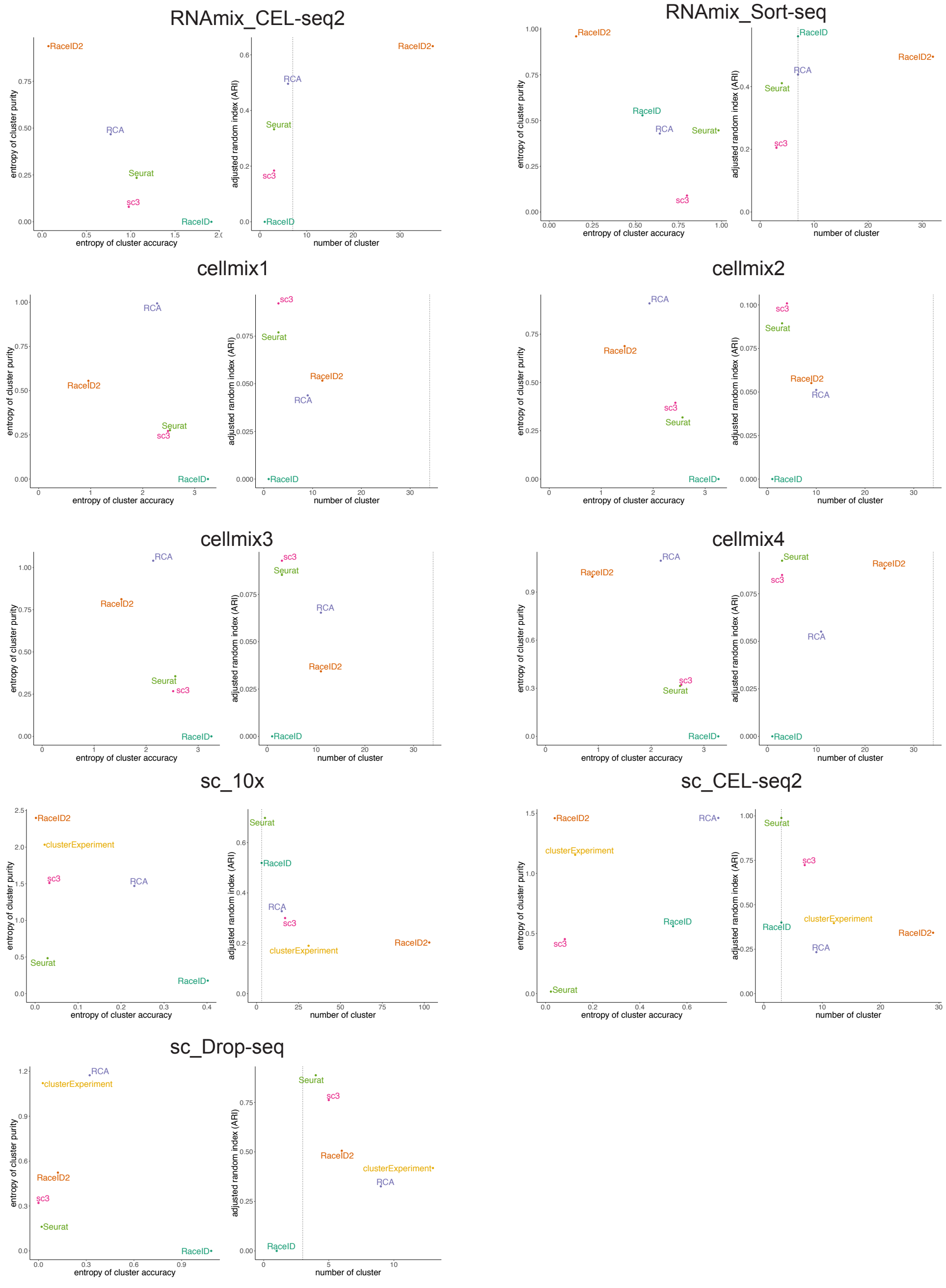
| Method | Correct | Adjust | Dim. reduction | # genes | Main parameters | Ref |
|---|---|---|---|---|---|---|
| MNN | ✓ | | | HVG | - *Number of nearest neighbors*<br>- *Bandwidth of smoothing kernel* | [12] |
| MINT supervised | | | ✓ | all genes | - Number of components<br>- If gene selection: Number of genes to select | [40] |
| ZINB-WaVe | ✓ | ✓ | | HVG | - Number of factors | [37] |
| diagonal CCA | | ✓ | ✓ | HVG | - Number of components<br>- Reference dataset<br>- *If multiCCA: number of iterations* | [43] |
| scMerge unsupervised | ✓ | | | all genes<br>(+ SEG) | - Number of K-means clusters<br>- *Number of factors*<br>- *Ratio of pseudo replicates*<br>- *Distance metric* | [28] |
| Scanorama | ✓ | | | HVG | - *Number of HVG*<br>- *Number of nearest neighbors (NN)*<br>- *Choice of approximate kNN*<br>- *Gaussian kernel function parameter* | [14] |



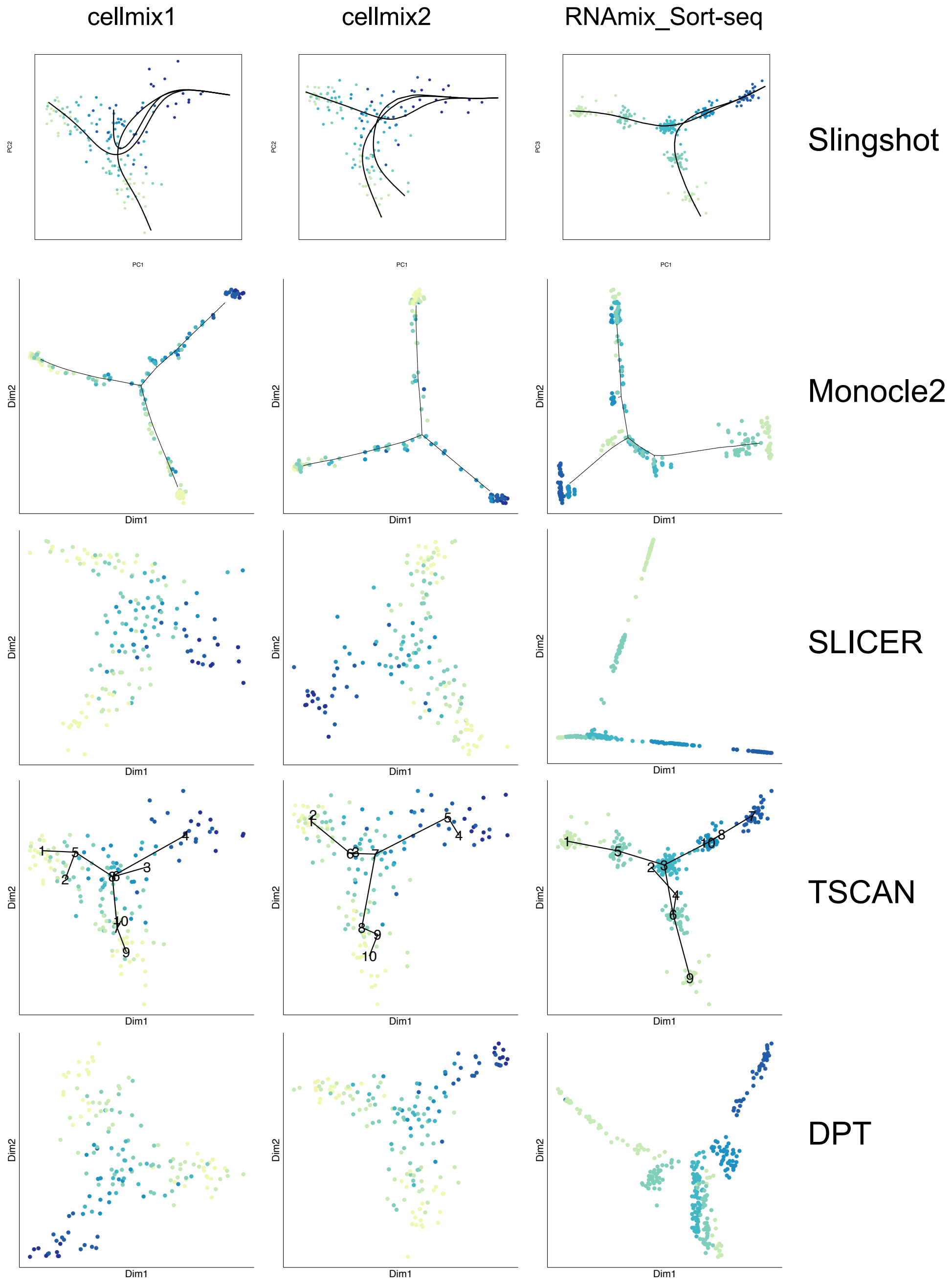
**Supplementary Figure 2. Box plots of quality control metrics for the samples from each benchmarking dataset.** (**A**) The percentage of reads that map to exons. (**B**) The total number of counts per cell after UMI deduplication. (**C**) The number of genes detected (genes with a count of a least 1) in each cell. (**D**) The amplification rate, which is defined by the ratio between the reads mapping to exons and the UMI counts after UMI deduplication. This measure reflects the library complexity.



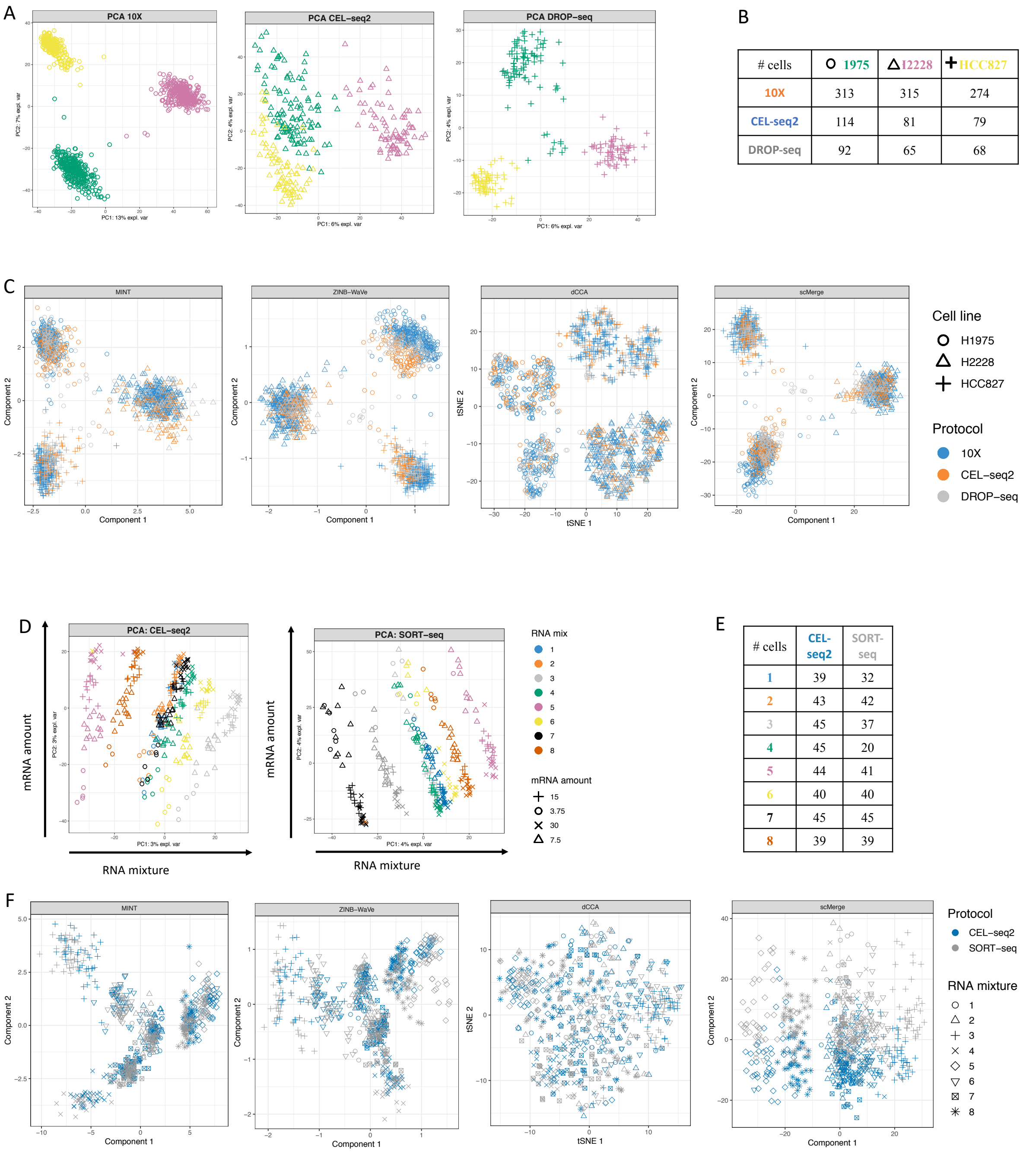
**Supplementary Figure 3. Box plots of silhouette widths for different normalization and imputation methods for 3 benchmarking datasets.** Silhouette widths calculated using the known biological groups after different normalization or imputation methods, using the cellmix3 (**A**), RNAmix_CEL-seq2 (**B**) and sc_CEL-seq2 (**C**) datasets. Methods are ordered by the average silhouette width.



**Supplementary Figure 4. Comparisons of clustering methods using clustering entropy and ARI.** The comparison of clustering methods in all datasets using the entropy of clustering accuracy and purity as detailed in the Methods and ARI, with dashed lines indicating the actual number of clusters.



**Supplementary Figure 5. Visualization of results from all trajectory methods evaluated in our study.** Results for cellmix1, cellmix2 and the RNAmix_Sort-seq analyses are shown. The dimension reduction method chosen for each method was as follows: PCA for *Slingshot* and *TSCAN*, DDR tree for *Monocle2*, diffusion map for *DPT* and LLE for *SLICER*.



| # cells | ○ 1975 | △ I2228 | + HCC827 |
|---|---|---|---|
| 10X | 313 | 315 | 274 |
| CEL-seq2 | 114 | 81 | 79 |
| DROP-seq | 92 | 65 | 68 |

| # cells | CEL-seq2 | SORT-seq |
|---|---|---|
| 1 | 39 | 32 |
| 2 | 43 | 42 |
| 3 | 45 | 37 |
| 4 | 45 | 20 |
| 5 | 44 | 41 |
| 6 | 40 | 40 |
| 7 | 45 | 45 |
| 8 | 39 | 39 |

**Supplementary Figure 6. Data integration results for single cell and RNA mixture datasets.** The top panels (**A-C**) present the analysis for the single cell datasets, while the bottom panels (**D-F**) show results for the RNA mixtures. **A, D** PCA sample plots of each protocol individually where colours indicate the known cell types / RNA mixture. **B, E** The number of cells per protocol and known cell groups. **C, F** Outputs from the additional integrative methods are represented into a reduced dimensional space either intrinsic from the method (MINT, ZINB-WaVe) or using t-SNE (diagonal CCA) or PCA where colours indicate protocol information.