Supplementary Data

Quality Control of Single-cell RNA-Seq by SinQC

Peng Jiang¹, James A. Thomson¹,²,³, and Ron Stewart¹

¹Morgridge Institute for Research, Madison, WI 53707, USA
²Department of Cell and Regenerative Biology, University of Wisconsin, Madison, WI 53706, USA
³Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, CA 93106, USA

Contents

1. scRNA-seq datasets 2
2. SinQC algorithm 2-4
3. SinQC software 4
4. Definition of high dropout rate cells 4
5. Supplementary results and discussion 5-9
   5.1 The main population cells in mixtures of cell types 5
   5.2 Comparison of SinQC with other scRNA-seq quality control (QC) methods 5-7
   5.3. Poor experimental batches are enriched in more technical artifacts 7
   5.4 The sensitivity and specificity of SinQC in mixed cell types 7-8
   5.5 The sensitivity and specificity of SinQC in different experimental batches 8
   5.6 Use with expression measures not from RSEM 8-9
6. Supplementary tables (Table S1-S4) 10-13
7. Supplementary figures (Figure S1-S11) 14-24
8. References 25
1. scRNA-seq datasets
Nine scRNA-seq datasets were used in this study to test the performance of SinQC. The first scRNA-seq dataset comes from a highly heterogeneous mixture of 11 cell types comprising 301 cells (Pollen, et al., 2014). The rest of eight scRNA-seq datasets include six experimental batches of our lab-generated human embryonic stem cell (ES) datasets (Leng, et al., 2015) and two mouse datasets (Islam, et al., 2011). The six batches of human ES datasets include three batches of H1 datasets (H1-Exp1 = 72 cells, H1-Exp2 = 81 cells, and H1-Exp3 = 75 cells) and three batches of fluorescent ubiquitination-based cell cycle indicator (FUCCI) sorted cell cycle known H1 cells (G1 = 91 cells, S = 80 cells, and G2 = 76 cells). The human single cells were prepared using the Fluidigm C1 system and sequencing was performed by the Illumina HiSeq 2500 system (Leng, et al., 2015). The two mouse datasets include 48 ES cells and 44 mouse embryonic fibroblasts (MEF cells) (Islam, et al., 2011). For human scRNA-seq samples, reads were mapped against the human transcriptome (RefGene v1.1.17). For mouse ES and MEF raw reads, we trimmed the first 12 nt, which contained the multiplexing barcode region and constant Cs region introduced via the sequencing protocol used (Islam, et al., 2011), and then the trimmed reads were mapped against the mouse transcriptome (Ensembl v1.75, protein coding genes). The mapping was performed with Bowtie 0.12.8 (Langmead, et al., 2009), allowing up to two mismatches and a maximum of 200 multiple hits. The expected counts (ECs) and transcripts per million (TPM) were estimated via RSEM 1.2.3 using default parameters (Li and Dewey, 2011).

2. SinQC algorithm
SinQC integrates both gene expression patterns and sample sequencing library qualities to detect technical artifacts. Given a batch of scRNA-seq data, SinQC uses gene expression patterns to classify cells as either gene expression outliers or cells of the main population. SinQC assumes that gene expression outliers contain both cells with real biological variation and technical artifacts, but cells of the main population, in general, are more likely to contain good quality cells. Thus, SinQC uses cells of the main population as controls to estimate data quality cutoffs and a corresponding false positive rate (FPR) (Figure S1A). SinQC determines data quality cutoffs by allowing a minimal fraction (user-defined) of cells of the main population to fail to pass these cutoffs. The technical artifacts are defined as gene expression outliers with poor data quality. Specifically, SinQC contains the following three steps:

Step 1: Classification of cells into gene expression outliers and cells of the main population
First, SinQC calculates a list of Spearman rank correlations comparing a given cell to the rest of the cells in the dataset (‘one-to-others’). Then, that given cell is removed and a list of pairwise Spearman rank correlations is calculated for the remaining cells (‘pairwise’). Finally, SinQC uses a one-sided Wilcoxon signed-rank test to assess whether the ‘one-to-others’ correlation is significantly lower than the set of ‘pairwise’ correlations. A similar procedure is also performed using Pearson product-moment correlations. Given two user-defined p-value cutoffs (one based on Spearman correlations and another based on Pearson correlations), SinQC classifies cells as either gene expression outliers or cells of the main population. The joint requirement using both correlations provides a more stringent criterion for defining gene expression outliers than either Spearman or Pearson alone. In this study, we define gene expression outliers as cells with p-values less than 0.001 in both Spearman and Pearson tests.

Step 2: Evaluation of RNA-seq library quality
SinQC uses three metrics to evaluate the scRNA-seq library quality for each sample: total number of mapped reads, mapping rate, and library complexity. The total number of mapped reads is the sum of mapped reads for all the genes. An extremely low number of mapped reads may affect the ability to characterize the transcriptome and could be due to either a low mapping rate or other technical issues introduced during sample prep or sequencing. The mapping rate is the total number of mapped reads divided by the read depth. Mapping rate can be effected by RNA degradation, contamination with genomic DNA, or other technical issues introduced during sample prep or sequencing. The library complexity is defined as the ratio of unique reads (the count of reads after removing duplicates) over the total number of all reads. This metric gives an indication of amplification biases.

For each cell, SinQC calculates a quantile score (QS) for each quality metric. Given a metric, the QS of a cell is defined as the number of other cells in the dataset with equal or lower values divided by the total number of cells. For example, if a cell has the twentieth highest mapping rate among a set of 80 cells, then the mapping rate QS for this particular cell is 0.75. A higher QS indicates better data quality.

SinQC combines these three data quality metrics for each cell into two meta-scores: Minimal Quantile Score (MQS) and Weighted Combined Quality Score (WCQS). MQS and WCQS provide measurements for different scenarios. MQS is defined as the minimal QS of the three quality metrics.

\[ MQS = \min \{ QS_i \} \quad i \in \{\text{mapped reads}, \text{mapping rate}, \text{library complexity}\} \]

MQS assumes that each of the three quality metrics is critical and that a deficiency in any of the three is a potential indicator of technical issues. Thus the ‘final quality’ of a cell depends on its lowest quality metric score.

In contrast, WCQS assumes that the importance of each quality metric may depend on specific experimental batches, protocols, and/or conditions. WCQS assumes that the importance of each quality metric for detecting technical artifacts is proportional to its ability to discriminate between gene expression outliers and cells of the main population. For example, given a batch of cells, if the mapping rate of a given batch of cells can perfectly discriminate between gene expression outliers and cells of the main population, then it is more likely that the mapping rate is a dominant player in detecting technical artifacts. In contrast, if a metric does not indicate differences between gene expression outliers and cells of the main population, then it should be removed from prediction of potential technical artifacts. WCQS calculates a weighted aggregation quality score for each sample defined as:

\[ WCQS = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i Z_i^2}} \]

where \( Z_i \) is the transformed Z-score of QS for data quality metric \( i \), according to

\[ Z_i = \Phi^{-1} (1 - P_i) \]

where \( \Phi \) is the standard normal cumulative distribution function, and \( P_i \) is the probability that quality metric \( i \) is lower in a given cell than in the rest of the cells. We estimate \( P_i \) is as

\[ P_i = 1 - QS_i \]
To avoid numerical error, we set maximal and minimal $Z_i$ as $+8.5$ and $-8.5$, which corresponds to $P_i < 10^{-16}$ and $P_i > (1 - 10^{-16})$, respectively. The $w_i$ is the weighting factor for data quality metric $i$ and is estimated according to the individual quality metric’s ability to discriminate between cells of the main population and gene expression outliers as

$$w_i = \begin{cases} 
\frac{AUC_i - 0.5}{0.5} & (AUC_i > 0.5) \\
0 & (AUC_i \leq 0.5)
\end{cases}$$

where $AUC_i$ is the area under the curve (AUC) of the receiver operating characteristic (ROC) curve for quality metric $i$. If a quality metric $i$ (e.g., mapping rate) can perfectly discriminate between cells of the main population and gene expression outliers, then $AUC_i = 1$ and thus $w_i = 1$. If the values of a quality metric $i$ are randomly distributed in cells of the main population and gene expression outliers, then the expected $AUC_i = 0.5$ and thus $w_i = 0$.

**Step 3: Identification of technical artifacts**

We assume that good quality cells should pass particular MQS and WCQS cutoffs. SinQC uses cells of the main population as controls to determine these cutoffs. SinQC enumerates all possible combinatorial pairs of MQS and WCQS cutoffs in a given dataset, calculates the fraction of cells of the main population that pass both cutoffs of a pair, and then uses the remaining cells of the main population to estimate the corresponding false positive rate (FPR) for that pair (Figure 1A). Under a user-defined maximal FPR cutoff (e.g., FPR<5%), SinQC estimates the quality ‘bottom lines’ by requiring that at least ‘1-FPR’ fraction of cells of the main population pass both the MQS and WCQS cutoffs. If more than one pair of MQS and WCQS cutoffs results in the same FPR, then SinQC will choose the cutoff pair that maximizes the percentage of gene expression outliers failing to pass. Then SinQC applies these cutoffs to the gene expression outliers to identify technical artifacts. Technical artifacts are defined as gene expression outliers with poor data quality measurements.

**3. SinQC software**

SinQC is implemented in Python/R and can be run in any Linux/Unix environment. A SinQC input and output example is shown in Figure S1B. In addition to detecting technical artifacts, SinQC can also be used to generate general data quality information and gene expression patterns (e.g., PCA) for scRNA-seq data. The SinQC software and detailed user manual are available at http://www.morgridge.net/SinQC.html.

**4. Definition of high dropout rate cells**

For each cell, we calculate the total number of genes detected (TPM>1). Given an scRNA-seq dataset, we calculate the first quartile (Q1), the third quartile (Q3) as well as interquartile range 'IQR = Q3-Q1' of the number of genes detected among those cells. The high dropout rate cells are defined as the cells whose number of genes detected being less than ‘Q1 – 2 × IQR’.
5. Supplementary Results and Discussion

5.1 The main population cells in mixtures of cell types
We further investigated main population cells in mixed cell types. We applied SinQC to a mixed mouse ES (48 cells) and mouse MEF (44 cells) scRNA-seq dataset (Islam, et al., 2011). SinQC detects two gene expression outliers in ES cells and one gene expression outlier in MEF cells. The main population cells consist of 46 ES cells and 43 MEF cells (Figure S6 (A)). We then mixed the mouse ES cells and MEF cells in different ways to simulate datasets containing different proportions of subpopulations. As shown in Figure S6 (B), we split the mouse MEF cells into five portions and mixed each portion with all mouse ES (48 cells). After applying SinQC, the proportions of ES cells and MEF cells in main population cells are similar to the proportions of ES cells and MEF cells in their mixtures (ES (all) + 1/5 (MEF)). As shown in Figure S6 (C), we split the mouse ES cells into five portions and mixed each portion with all mouse MEF (44 cells). After applying SinQC, the proportions of ES cells and MEF cells in main population cells are similar to the proportions of ES cells and MEF cells in their mixtures (MEF (all) + 1/5 (ES)).

5.2 Comparison of SinQC with other scRNA-seq quality control (QC) methods
We compared SinQC with several alternative quality control (QC) methods for scRNA-seq data:

5.2.1 Using housekeeping genes to perform quality control (QC) for scRNA-seq
Several studies used housekeeping genes to perform quality control for scRNA-seq datasets. For example, cells not expressing housekeeping genes (e.g., Actb, Gapdh) or abnormally expressing them are filtered out (Ting, et al., 2014; Treutlein, et al., 2014). The assumption of methods in this category is that housekeeping genes are highly and consistently expressed, which is not necessarily true for single cells. A study using single-cell qPCR not only showed that the gene expression of housekeeping genes had high variation between individual cells but also that gene expression of housekeeping genes can even distinguish cell types (Oyolu, et al., 2012). Thus, a reliance on housekeeping genes to perform QC can result in removing cells with real biological variation.

To further investigate the feasibility of using housekeeping genes to perform quality control for scRNA-seq datasets, we calculated the gene expression levels (TPMs) for two housekeeping genes (Actb and Gapdh) in a mouse scRNA-seq dataset (Islam, et al., 2011), containing 48 mouse ES cells and 44 mouse MEF cells. As shown in Figure S7, Gapdh is significantly higher expressed in ES cells than in MEF cells (P = 5.6e-06, 1-sided Wilcoxon rank sum test) while Actb is significantly lower expressed in ES cells than in MEF cells (P < 2.2e-16, 1-sided Wilcoxon rank sum test). This suggests that it is infeasible to use housekeeping genes to perform QC for scRNA-seq datasets.

We applied SinQC to the same dataset and detected two technical artifacts (‘ESC_46’ and ‘ESC_32’) (FPR < 5%). Those two cells have extremely low mapping rates: 3.49% and 4.35% for ‘ESC_46’ and ‘ESC_32’, respectively, suggesting that they are real technical artifacts.

5.2.2 Using median gene expression values or the number of genes detected (TPM > 1) to perform quality control (QC)
Low data quality (e.g., low mapping rate) can result in fewer number of genes detected or low median gene expression values. However, the number of genes detected (TPM > 1) can also be biologically related. The number of genes detected vary depending on the quality of a particular library and cell types (Kharchenko, et al., 2014). We calculated the number of genes detected in a highly heterogeneous scRNA-seq dataset containing 301 cells (mixture of 11 different cell types) (Pollen, et al., 2014). As shown in Figure S8, the number of genes detected is highly cell type dependent, suggesting using the number of gene detected to identify technical artifacts will result in substantial bias.

For this highly heterogeneous scRNA-seq dataset, the technical artifacts detected by SinQC overall have fewer genes detected if compared with QC pass cells. But this does not mean that the cells with fewer genes detected are technical artifacts. For example, we investigate whether the high dropout rate cells (cells with significantly fewer number of genes detected in a given scRNA-seq dataset, Supplementary Data: Definition of high dropout rate cells) are technical artifacts in the mixture of 11 different cell types. There are two high dropout rate cells in total (‘Cell.GW16_11’ and ‘Cell.GW16_8’), but neither of them is detected as a technical artifact by SinQC (Table S3). As shown in Table S4, these two high dropout rate cells have very good data quality indeed. For example, ‘Cell.GW16_11’ has a mapping rate of 73%, which is among the highest mapping rate (percentile = 99.34%) among all the cells, and all three quality metrics (mapped reads, mapping rate and library complexity) are high for both ‘Cell.GW16_11’ and ‘Cell.GW16_8’ (Table S4), suggesting these two cells are not technical artifacts and SinQC correctly predict them as QC pass cells.

We further investigate whether the high dropout rate cells in eight “low-heterogeneity datasets”, which contain one cell type for each dataset, are technical artifacts. We applied SinQC to each of these low-heterogeneity datasets. Interestingly, all of the high dropout rate cells in “low-heterogeneity datasets” are detected as technical artifacts (Table S2).

The number of genes detected is determined by both of the data quality and cell types. For “low-heterogeneity datasets” (one cell type), the cells with significantly fewer number of genes detected are more likely to be technical artifacts (Table S2). But for highly heterogeneous datasets (mixture of cell types), fewer number of genes detected could be related to both technical issues and real biological variation (Table S4).

5.2.3 Using ‘genes detected and/or mapping rate’ to perform quality control (QC)

The basic idea of using ‘genes detected and/or mapping rate’ (Kumar, et al., 2014) to perform QC is that the fewer number of genes detected could be due to technical issues or biological heterogeneity or both. But if a cell with fewer genes detected is also associated with low mapping rate (mapping rate is technical related), the cell might be more likely to be a technical artifact. Of the approaches discussed, this approach is the most conceptually similar to SinQC. However, SinQC has two advantages:

First, the determination of the mapping rate cutoff and the number of genes detected cutoff is very difficult and arbitrary. SinQC assumes that gene expression outliers contain both cells with real biological variation and technical artifacts, but the rest of the cells (main population cells) in general, are more likely to contain good quality cells. Thus, SinQC uses cells of the main population as controls to estimate data quality cutoffs that can maximize the probability that
the technical artifacts are correctly detected while also minimizing the false positives.
Second, in addition to mapping rate, SinQC also takes other library quality metrics into consideration (e.g., library complexity).

5.3. Poor experimental batches are enriched in more technical artifacts
If a single-cell RNA-seq experiment contains hundreds or thousands of cells, it is likely that they are processed in several experimental batches (e.g., single cell preparation and RNA sequencing). Thus some experiments will have overall better or worse data quality than others. For our lab-generated human embryonic stem cell (ES) datasets (Leng, et al., 2015), we processed them in three different experimental batches (H1-Exp1 = 72 cells, H1-Exp2 = 81 cells, and H1-Exp3 = 75 cells). As shown in Figure S9 (A-B), the first experimental batch has overall lower data quality (e.g., mapping rate and library complexity) than experiment 2 and 3. We applied SinQC on those pooled three batches of cells and detected 15 technical artifacts in total (Figure S9 (C)). Among these 15 technical artifacts, 12 of them come from Experiment 1. As shown in Figure S9 (D), technical artifacts identified by SinQC are significantly enriched in Experiment 1 (P = 9.34e-05, Fisher’s Exact Test, Experiment 1 versus Experiment 2 + Experiment 3), which is expected.

5.4 The sensitivity and specificity of SinQC in mixed cell types
SinQC assumes that gene expression outliers contain both technical artifacts and biological variant cells, but cells of the main population, in general, are more likely to contain good quality cells. Thus, SinQC uses cells of the main population as controls to estimate data quality score cutoffs and a corresponding false positive rate (FPR). However, given a FPR, it is a challenge to estimate the corresponding false negative rate (technical artifacts that are missed by SinQC), due to that scRNA-seq has no ‘ground-truth’ for ‘bad samples’.

Sensitivity (also called the true positive rate) is the proportion of positives (‘technical artifacts’) that are correctly identified. Specificity (also called the true negative rate) measures the proportion of negatives (‘good quality single cells’) that are correctly identified. Since scRNA-seq has no ‘ground-truth’ for ‘good samples’ and ‘bad samples’, it is a challenge to estimate these two measurements directly. To further compare the sensitivity and specificity of SinQC in high-heterogeneity and low-heterogeneity datasets, we do not calculate the sensitivity and the specificity directly. Instead, we applied SinQC to datasets with mixture of different portions of cell types, and compared the overlap of technical artifacts detected among them. For example, using a mouse scRNA-seq dataset (48 ES cells and 44 MEF cells) (Islam, et al., 2011), we mixed the cells into three different categories: high-heterogeneity (48 ES cells + 44 MEF cells), medium-heterogeneity (‘ES cells (all) + 1/5 (MEF) cells’ and ‘MEF cells (all) + 1/5 (ES) cells’), and low heterogeneity ((48 ES cells) and (44 MEF cells), separately). As shown in Figure S5, SinQC detects two technical artifacts (ESC_46 and ESC_32) in the high-heterogeneity dataset (48 ES cells + 44 MEF cells). These two technical artifacts can also be robustly detected either in medium-heterogeneity datasets or low heterogeneity datasets (Figure S5). However, if we run SinQC on each individual ES (48 cells) or MEF (44 cells) dataset separately, we can detect more artifacts, compared to running SinQC on pooled mixture datasets (48 ES cells + 44 MEF cells). Therefore, we conclude that SinQC increases specificity at the cost of dropping sensitivity when the extent of heterogeneity in a dataset is high.
In highly heterogeneous cell populations, detecting technical artifacts carries a higher risk of dropping real biological variation cells. The increased specificity and decreased sensitivity of SinQC for highly heterogeneous cell populations is a good feature that can minimize the false positives. Therefore, we do not recommend relaxing the FPR (false positive rate) threshold for highly heterogeneous samples.

Although a direct estimation of sensitivity and specificity to produce an ROC curve for scRNA-seq datasets is infeasible due to no ‘ground-truth’ of ‘bad’ or ‘good’ samples, we further investigated the relationship between the portion of main population cells and the portion of gene expression outliers being classified as ‘low data quality cells’ given different data quality score cutoffs. We applied SinQC to a more highly heterogeneous scRNA-seq dataset, which contains 301 cells consisting of 11 cell types (Pollen, et al., 2014). The relationship between the proportion of main population cells (false positive rate) and the proportion of gene expression outliers that are below different quality score cutoffs are shown in Figure S10. Given the same quality score cutoffs, the main population cells exhibit much lower proportions that fail to pass these cutoffs if compared with gene expression outliers. For example, if the quality score cutoffs are set so that 20% of the main population cells are below them, ~ 80% of gene expression outliers will be below these quality score cutoffs. This further suggest that gene expression outliers contain both cells with real biological variation and technical artifacts, but the main population cells in general, are more likely to contain good quality cells.

5.5 The sensitivity and specificity of SinQC in different experimental batches
Applying SinQC to three pooled batches of human H1 single cell RNA-seq data, we can detect 15 technical artifacts (H1-Exp1=12, H1-Exp2=3, H1-Exp3=0) (Figure S9). However, if we run SinQC batch by batch, we detect not only these 15 artifacts but also 11 additional ones (Table S1) (H1-Exp1=12, H1-Exp2=11, H1-Exp3=3). This suggests that SinQC is more sensitive if run batch by batch. This is because pooling batches will increase the diversity of the population being studied owing to batch effects in scRNA-seq data. Since SinQC uses relative measurements to determine data quality cutoffs, the increased diversity in pooled batches will relax the absolute data quality cutoffs thus allowing more gene expression outliers to pass these cutoffs.

5.6 Use with expression measures not from RSEM
To obtain mapped reads and gene expression values (TPMs), SinQC will read all “*.genes.results” files, which are generated by RSEM (Li and Dewey, 2011). For users who do not use RSEM, they can make a customized RSEM file (“*.genes.results”) to run SinQC. These customized RSEM files can be made from the output of any other tools used to make RNA-seq expression estimations (such as cufflinks (Trapnell, et al., 2012)) that produce gene expression information (e.g., TPM, RPKM, and FPKM) and mapped read counts. A SinQC accepted customized RSEM file for each cell contains seven fields, as shown in Figure S11.

SinQC accepted customized RSEM file:
The file should contain a header and seven tab-delimited fields.
   (1) **Gene information**: gene symbols or gene IDs. We recommend gene symbols or gene IDs should not contain space. E.g., GSTP1 (valid name), GSTP_1 (valid name), GSTP.1 (valid name), GSTP 1 (invalid name).
   (2) **ID**: gene ID
(3) **Tag1**: place holder (SinQC will ignore this field, but need to fill it with any string characters, e.g., ‘XXX’)

(4) **Tag2**: place holder (SinQC will ignore this field, but need to fill it with any string characters, e.g., ‘XXX’)

(5) **Counts**: mapped read counts

(6) **Gene expression values**: TPMs or RPKMs or any user calculated gene expression values

(7) **Tag 3**: place holder (SinQC will ignore this field, but need to fill it with any string characters, e.g., ‘XXX’)

SinQC will read fields (1) (2) (5) (6) and ignore fields (3) (4) (7).
### 6. Supplementary Tables

**Table S1.** The number of gene detected (TPM>1) of QC pass cells vs. technical artifacts detected by SinQC in eight low-heterogeneity scRNA-seq datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total Cells</th>
<th>Artifacts</th>
<th>PASS (QC)</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ave. Genes Detected</td>
<td>95% CI b</td>
</tr>
<tr>
<td>Human H1-Exp1 (72</td>
<td>12)</td>
<td>8653</td>
<td>(8482,8824)</td>
<td>3737</td>
</tr>
<tr>
<td>Human H1-Exp2 (81</td>
<td>11)</td>
<td>8268</td>
<td>(8120,8452)</td>
<td>5486</td>
</tr>
<tr>
<td>Human H1-Exp3 (75</td>
<td>3)</td>
<td>7678</td>
<td>(7532,7824)</td>
<td>5361</td>
</tr>
<tr>
<td>Human G1 (91</td>
<td>2)</td>
<td>8008</td>
<td>(7898,8118)</td>
<td>7711</td>
</tr>
<tr>
<td>Human S (80</td>
<td>17)</td>
<td>8978</td>
<td>(8835,9121)</td>
<td>7834</td>
</tr>
<tr>
<td>Human G2 (76</td>
<td>4)</td>
<td>8716</td>
<td>(8598,8834)</td>
<td>7268</td>
</tr>
<tr>
<td>Mouse ES (48</td>
<td>4)</td>
<td>2596</td>
<td>(2260,2932)</td>
<td>1172</td>
</tr>
<tr>
<td>Mouse MEF (44</td>
<td>1)</td>
<td>5409</td>
<td>(5024,5794)</td>
<td>1489</td>
</tr>
</tbody>
</table>

**a** SinQC: FPR < 5%

**b** CI: The 95% confidence interval (CI) is estimated by $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$
Table S2. The high dropout rate cells (HDRCs) in eight low-heterogeneity scRNA-seq datasets are identified as artifacts by SinQC.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HDRCs</th>
<th>Identified as Artifacts (FPR&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human H1-Exp1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Human H1-Exp2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Human H1-Exp3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Human G1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Human S</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Human G2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mouse ES</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mouse MEF</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*The HDRCs are defined as the cells with the number of genes detected (TPM>1) less that 'Q1–2 × IQR' (Supplementary Data: Definition of High dropout rate cells).
**Table S3.** The high dropout rate cells (HDRCs) in a highly heterogeneous mixture populations and identified as artifacts by SinQC.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HDRCs(^{a})</th>
<th>Identified as Artifacts (FPR&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture of 11 cell types</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>(Pollen, et al., 2014) (301 cells)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\) The HDRCs are defined as the cells with the number of genes detected (TPM>1) less that 'Q1–2 × IQR' (Supplementary Data: Definition of high dropout rate cells).
Table S4. Data quality of high dropout rate cells (HDRCs) in a highly heterogeneous mixture population (Pollen, et al., 2014).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene Detected (TPM &gt; 1)</th>
<th>Total Reads</th>
<th>Mapped Reads (Percentile)</th>
<th>Mapping Rate (Percentile)</th>
<th>Library Quality (Percentile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell.GW16_11</td>
<td>2303</td>
<td>8295554</td>
<td>6087193 (97.34%)</td>
<td>0.73 (99.34%)</td>
<td>0.99 (87.71%)</td>
</tr>
<tr>
<td>Cell.GW16_8</td>
<td>2427</td>
<td>8229623</td>
<td>5625746 (96.68%)</td>
<td>0.68 (95.02%)</td>
<td>0.98 (77.41%)</td>
</tr>
</tbody>
</table>

‘Cell.GW16_11’ and ‘Cell.GW16_8’ are identified as high dropout rate cells (HDRCs) in dataset (Pollen, et al., 2014). The dataset contains 301 cells which consisting of 11 cell types. The HDRCs are defined as the cells with number of genes detected (TPM>1) less that ‘Q1–2 × IQR’ (Methods). Both ‘Cell.GW16_11’ and ‘Cell.GW16_8’ are QC pass cells based on SinQC (FPR < 5%).
7. Supplementary Figures

A  Gene Expression Patterns

- MPCs: Cells with good data quality
- GEOS: Cells with poor data quality

- FPR = +
- Artifact Rate = ▲▲
- GEO with good data quality
- Technical Artifact

Data Quality Map

B  Input and Output Examples of SinQC Software

- Raw Reads Folder
- RSEM Results Folder
- Parameter Options:
  - PValueCutoff--Distinct--Spearman
  - PValueCutoff--Distinct--Pearson
  - Max_FPR

Output

Figure S1. Illustration of SinQC framework. (A) Basic principles of SinQC to detect technical artifacts. Cells can be separated out based on gene expression patterns into gene expression outliers (GEOS) and cells of the main population (MPCs). Using the two data quality metrics Minimal Quantile Score (MQS) and Weighted Combined Quality Score (WCQS), which combine three data quality metrics (mapped reads, mapping rate and library complexity), cells can be separated into good and poor quality cells. A false positive rate (FPR) is estimated by the fraction of cells of the main population that fail to pass the MQS and WCQC cutoff pair. Under a user-defined maximal FPR allowed, SinQC determines MQS and WCQC cutoffs and then defines technical artifacts as gene expression outliers that also have poor data quality. (B) Input and output examples of SinQC software. See the SinQC user manual (http://www.morgridge.net/SinQC.html) for a detailed description of inputs, outputs, and parameters.
Figure S2. Applying SinQC to a highly heterogeneous dataset containing a mixture of 11 cell types. (A-C) Statistical significance of data quality metrics (A: Mapped Reads, B: Mapping rate, C: Library complexity) lower in technical artifacts detected by SinQC (FPR < 5%) compared with cell that pass QC pass. The p-values are calculated by a 1-sided Wilcoxon rank sum test.
**Figure S3.** The number of genes detected in artifacts (FPR < 5%) versus in QC pass cells. The scRNA-seq contains 11 mixed cell types. The p-value is calculated by 1-sided Wilcoxon rank sum test.
Figure S4. Statistical significance of data quality metrics being lower in technical artifacts detected by SinQC (FPR < 5%) when compared to cells that pass QC. The p-values are calculated by a 1-sided Wilcoxon rank sum test.
Figure S5. Applying SinQC to datasets with mixtures of different portions of cell types. The symbol shapes indicate the cell types (circle: mouse ES, triangle: mouse MEF). All: Combining mouse ES and MEF datasets together and ran SinQC together. ES: Applying SinQC to ES cells only. MEF: Applying SinQC to MEF cells only. ‘ES + 1/5 (MEF)’: Randomly splitting MEF dataset into five portions, and combine one portion with ES full dataset. ‘1/5 (ES) + MEF’: Randomly splitting ES dataset into five portions, and combine one portion with MEF full dataset. The technical artifacts detected by SinQC (FDR < 5%) are highlighted in red.
**Figure S6.** The main population cells in mixed cell types (mouse ES and MEF cells) (Islam, et al., 2011). (A) Applying SinQC to the mixed mouse ES (48 cells) and MEF (44 cells). SinQC detects two gene expression outliers in ES cells and one gene expression outlier in MEF cells. The main population cells consist of 46 ES cells and 43 MEF cells. (B) Simulation of total cells consisting 1/5 MEF cells and all ES cells. The MEF cells are split into five portions. The mixture datasets consist each portion of MEF cells with all ES cells. (C) Simulation of total cells consisting 1/5 ES cells and all MEF cells. The ES cells are split into five portions. The mixture datasets consist each portion of ES cells with all MEF cells.
**Figure S7.** The gene expression levels (TPMs) of two housekeeping genes: Gapdh (A) and Actb (B) in each cell. (A) Gapdh is significantly higher expressed in mouse ES cells than in mouse MEF cells (P = 5.6e-06, 1-sided Wilcoxon rank sum test). (B) Actb is significantly lower expressed in mouse ES cells than in mouse MEF cells (P < 2.2e-16, 1-sided Wilcoxon rank sum test).
Figure S8. The number of genes detected is cell type dependent. The dataset contains a mixture of 11 cell types (Pollen, et al., 2014). (A) The number of genes detected for each cell type (all cells). (B) The number of genes detected for each cell type (QC pass cells). The technical artifacts, which are detected by SinQC (FPR < 5%), are excluded from calculation.
Figure S9. Applying SinQC to three pooled batches of human H1 single cell RNA-seq data. (A) The Experiment 1 (72 cells) have overall lower mapping rate than Experiment 2 (81 cells) and Experiment 3 (75 cells). (B) The Experiment 1 (72 cells) have overall lower library complexity than Experiment 2 (81 cells) and Experiment 3 (75 cells). (C) Technical artifacts identified by SinQC (FPR < 5%) in pooled scRNA-seq batches (Experiment 1 + Experiment 2 + Experiment 3) are visualized in PCA analysis. The PCA analysis is performed based on TPM values. (D) Technical artifacts identified by SinQC are significantly enriched in Experiment 1 (P = 9.34e-05, Fisher’s Exact Test, Experiment 1 versus Experiment 2 + Experiment 3).
Figure S10. The proportion of main population cells and the proportion of gene expression outliers that are below different quality score cutoffs. The dataset contains 301 cells which consisting of 11 cell types (Pollen, et al., 2014).
Figure S11. SinQC accepts customized RSEM format for use with expression measures not generated by RSEM. To obtain mapped reads and gene expression values (e.g., TPMs) for each cell, SinQC will read the “*.genes.results” file for each cell. Each “*.genes.results” file contains seven tab-delimited fields. The fields which are used by SinQC are highlighted in red color.
8. References