Quality assessment of single-cell RNA sequencing data by coverage skewness analysis

Imad Abugessaisa[1], Shuhei Noguchi[1], Melissa Cardon[1], Akira Hasegawa[1], Kazuhide Watanabe[1], Masataka Takahashi[1], Harukazu Suzuki[1], Shintaro Katayama[2,3,4], Juha Kere[2,3,4,*], Takeya Kasukawa[1,5,*]

[1]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

[2]Folkhälsan Research Center, Helsinki, Finland

[3]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

[4]Stem Cells and Metabolism Research Program, University of Helsinki, Helsinki, Finland

[5]Institute for Protein Research, Osaka University, Suita, Osaka, Japan

Abstract

Analysis and interpretation of single-cell RNA sequencing (scRNA-seq) experiments are compromised with the presence of poor quality cells for many underlying reasons. For meaningful analyses, such poor quality cells should be excluded before detailed analysis. However, there exist no clear guidelines.

Here we present SkewedCID a novel quality-assessment method to identify poor quality single-cells in scRNA-seq experiments. The method relies on the measure of skewness of the gene coverage of each single cell as a quality measure. The gene coverage is characteristic of

each method, and different methods yield highly different coverage profiles. To validate the method, we investigated the impact of poor quality cells on downstream analysis and compared biological differences between good and poor quality cells.

In addition to skewness-based quality assessment, we developed models to measure the ratio of intergenic expression, suggesting genomic contamination, and foreign organism contamination of single-cell samples.

The method is robust and able to segregate poor quality cells from good ones and applicable to any type of scRNA-seq protocols. We tested our method in about 38,000 human and mouse cells generated by 15 scRNA-seq protocols. In addition to the quality method, our analysis brings new insights about the capability of scRNA-seq protocols in term of gene body coverage and the influence of the poor quality cells in scRNA-seq analysis.

Keywords

Single-cell RNA sequencing, single-cell data quality, intergenic expression, single-cell sample contamination

Introduction

Recent advances in scRNA-seq methods and protocols have enabled new discoveries and insights in the biology of cells[1, 2]. The method has been used to profile gene expression of individual cells under different biological conditions, and to identify new cell types and provide knowledge about different biological processes[3].

Data quality measures and quality-control (QC) methods aim to provide confidence in the quality of the data set and assure the robustness, reproducibility and high quality of any experimental study. In bulk RNA-seq experiments, different data quality measures are applied at consecutive experimental stages. At the early stage of experiment, the integrity of RNA (RIN value) is measured, the raw reads are evaluated (FASTQ) as well as the quality of the aligned reads (MAPQ), samples with low number of reads per sample are excluded, and genes with low expression value may be filtered.

In scRNA-seq no such standard measures are used, and the data quality may vary highly due to variation for biological reasons or experimental procedures. The variation of experimental procedures include cell capture methods, target of the sequencing protocols, and reaction failure, to name just a few.

Cell capture method might expose individual cells to stress and cause cell death. Cell capture site may contain debris due to broken cells, or contain multiple cells instead of a single cell. scRNA-seq protocols are designed to capture reads either at the end of the gene (5’or 3’ end) or the full gene body (entire transcript). The multitude of the scRNA-seq methods increases the complexity on the required quality assessment of the resulting data set. The failure or inadequate quality assessment might lead to the presence of poor quality cells (dead or live cells[4]) and thus

incorrect interpretation or compromised resolution, resulting from mis-clustering errors, propagation of specific cell type population, or poor sensitivity to detect differentially expressed genes (DEGs).

As reported in several publications[5], identification of poor quality cells is challenging since they might represent a large population of cells and may not be limited to dead cells. To identify poor quality cells experimentally (e.g., microscopic techniques or cell staining), is laborious and involve further manipulations possibly affecting the transcriptome. Computationally, several tools and methods have been developed to identify poor quality cells[4-11]. The first group of the methods classify the cells based on resulting sequence reads counts, number of expressed genes, gene expression patterns to detect outliers or cutoff value based on library size. The second group of computation methods use machine learning techniques (classifier-supervised learning) to classify cells based on their normalized expression profile and a training set. The training dataset is generated from experimentally classified scRNA-seq data. A full list of the current tools for single-cell QC is available at https://www.scrna-tools.org/tools?sort=name&cats=QualityControl.

In general, the above methods are based on existing approaches used in bulk RNA-seq QC and analysis, and they ignore the characteristics of scRNA-seq experiment, variation of methodology and the quality properties of individual cells compared to the bulk RNA-seq sample, resulting in limitations both in terms of the classification result or implementation[5]. Here we considered such limitations and introduced new approaches able to segregate poor quality cells from the good quality cells. SkewedCID enabled us to identify two classes of single cells that we refer to as good and skewed profile cells, respectively.

Results

A comparison of scRNA-seq protocols

In this study, we compare and analyze 15 scRNA-seq protocols using different approaches from the published studies[12, 13] In particular, we evaluate the capability and power of each protocol in terms of the full-length transcript coverage, variability in sequence depth, expression variation, ratio of intergenic expression and analysis of the unmapped sequence reads to the reference genome (**Fig. 1**). To perform this analysis, we collected and reprocessed 28 published human and mouse single-cell data sets comprising 37,993 single-cells generated by 15 scRNA-seq protocols as well as one new data set from the human MCF10A cell line.

Based on the target read capture strategy of the scRNA-seq method, 5 data sets measured gene expression at the 5'-end of the transcript (STRT, C1 CAGE and 10x Chromium 5'-end), 20 datasets measured gene expression of the full-length transcript (SMARTer, Smart-Seq, RamDA-seq, SUPer-Seq, Quartz-Seq, C1 single-cell mRNA-Seq, Smart-Seq2, TruSeq and Drop-Seq), and 4 studies measured gene expression at the 3'-end of the transcript (CEL-Seq, CEL-Seq2, and 10x Chromium 3'-end). The data sets covered tissues, primary cells and cell lines, and to make the results comparable, we designed the data sets as batch-matched cells and unmatched cells (**Table 1**). For the mouse batch-matched cells, we analyzed mES, CD4 T cells, fibroblasts and hematopoietic cells. For human batch-matched cells, we analyzed human embryo stem cells (hESCs). The mouse unmatched cells were adipocytes and PBMC cells, and the human unmatched cells were MCF10A cells, PBMC cells and HEK 293 and 3T3 cells (**Supplementary Table 1**).

Wide discrepancies in gene body coverage among scRNA-seq protocols

Gene body (full transcript length) coverage considers the distribution of the sequence tags over the entire transcripts. To analyze the gene body coverage of different protocols, we computed the gene body coverage for each single cell (**Methods**). Remarkably, the gene body coverage shows wide differences among the data-sets generated by different scRNA-seq protocols (**Fig. 2a and b**; **Panels a and b in Supplementary Figs. 1-5**). This indicates major variation and differences in scRNA-seq data sets produced by difference protocols.

We investigated the pattern of the gene body coverage for each single cell in individual data sets. The visualization of the gene body coverage profile revealed two patterns of gene body coverage. The first set of single-cells show well clustered gene coverage distribution according to the target sequence of the protocol. The second set of single-cells showed skewed gene body coverage distributions. The skewness in the distribution could be caused by several reasons.

In one type, there was bias towards the 3'-end of the gene body in case of the 5'-end sequence and full-length sequence protocols. The bias towards the 3'-end indicated by high coverage at the 3'-end of the gene body (**Fig. 2c)**. The tag-based sequencing of 5' or 3' ends methods [14-17] should have the peak coverage at either the 5' or 3' end of the gene with low/no coverage in the middle region of the gene body.

In the second type, there was high coverage in the middle of the gene for 5'-end and 3'-end sequence protocols (**Fig. 2c)**, in contrast to the full-length sequencing protocols.

In the third type, there was low coverage in the middle of the gene for full-length sequence protocols. This indicated by low coverage at mid-point of 5'-3'-end of gene body (**Fig. 2d)**.

The above variation (bias) and skewness in the gene body coverage among individual cells of the same data set reflect the success of each single-cell reaction. The variation in the gene body coverage of individual cell should be considered when analyzing scRNA-seq data.

Variation in gene expression and gene saturation among scRNA-seq protocols

Deep sequencing increases the statistical power to detect differentially expressed genes DEGs [18]. To analyze the variability in gene expression for each of the scRNA-seq protocols, we computed gene expression for each data set (**Methods**), and used normalized expression values. (**Fig. 2e and f**) and (**Panel c and d in Supplementary Figs. 1-4; and Supplementary Fig. 5c**) show variability in gene expression among single-cells of the same cell type generated by different protocols. The variability is more visible in the 5' end and 3' end sequence protocol compared to the full length sequence. The coefficient of variation over the mean shows higher CV/mean for the 5'-end and 3'-end sequence protocols, and lower CV/mean for the full length sequence protocols.

We used the mean of the normalized gene expression (μ FPKM) and compute the R^2 of the two data sets from the same batch-matched cell type (**Supplementary Fig. 6**). The figure illustrates the variability in the mean expression of scRNA-seq protocols; the two mES data set (E-MTAB-2600[19] and E-MTAB-2805[20]) produced by different labs using the SMARTer protocol showed strong correlation ($R^2 = 0.85$). Comparison of the mean expression of the mES data set generated by SMARTer (E-MTAB-2600)[19] and SUPer-Seq (GSE53386)[21] showed weak correlation ($R^2 = 0.61$). SMARTer has better mean expression correlation with full length sequence protocols compared to the 5'-end and 3'-end sequence protocols. The mES data set (GSE46980)[4] and (GSE29087)[22] were generated by two different versions of STRT[14]

protocols and showed weak correlation ($R^2 = 0.54$). In general, the mean gene expression values of data sets from the same cell type (mES) generated by different protocols were dissimilar. Similar patterns were illustrated using mouse CD4 T-cells, mouse fibroblasts, mouse hematopoietic cell and hESC (**Supplementary Figs. 7–10**). The high correlation between data sets demonstrated the similarity of the expression profiles, whereas poor correlation between data sets demonstrated the dissimilarity of the expression profiles.

To investigate the variability in gene saturation for different scRNA-seq protocols, we used the Hanabi plot[23] (**Supplementary Figs. 11–17**). The figure demonstrates the detection power and variability in gene saturation for different protocols. The Hanabi plot considers the number of detected genes over the total counts. The 5'-end and 3'-end sequence protocols detected smaller numbers of genes with smaller total read counts compared to the full-length sequencing protocols.

QC method segregate poor and good quality single-cells

The results from the gene body coverage analysis discriminate two classes of cells, referred to as good single-cells and skewed single-cells, even in one dataset. Single-cell become skewed due to either technical failure during the sequencing or biological issue. To systematically classify the two classes of the cells, we developed an algorithm (SkewedCID). SkewedCID takes as input the gene body coverage of the scRNA-seq data set (**Fig. 3a**) (**Methods**). Systematically we applied the method for all data sets in this study (**Fig. 3b–d**) and (**Supplementary Figs. 18–23**). We removed single-cells with low numbers of mapped input reads (left charts of **Fig. 3b–d**) and (left panel of **Supplementary Figs. 10–15**). The remaining single-cells had high numbers of mapped

input reads. When applying our method, two distinct clusters of single-cells are visible (middle charts of **Fig. 3b–d**) and (middle charts of **Supplementary Figs. 18–23**).

Expression variation of housekeeping genes between skewed and good cells

To investigate the difference between the good and skewed cells in the resulting gene expression, we compared the normalized expression of the housekeeping genes of the good cells versus the skewed cells (right charts of **Fig. 3b–d**) and (right charts of **Supplementary Figs. 18–23**). The boxplot shows distinct differences in the variability in gene expression of the housekeeping genes (HKGs) between the two classes of cells with adjusted P-values < .001. The ratio of the skewed cells to good cells is different between different data set. As an example, the mES data set E-MTAB-2600[19] has a total number of single-cells (n=869), of which single-cells with low input mapped reads (n=19), good single-cells (n=765) and skewed single-cells (n=85). As another example, the data set GSE98664[24] with a total number of single-cells (n=364), there were no single-cells with low input mapped reads (n=0), good cells (n=338) and skewed cells (n=26).

Biological features of the skewed cells

To validate our QC methods, we used the mES data set GSE46980[4]. The authors of the data set classify the single-cells in their experiment as good quality cells (n=47), poor quality (n=40) and dead cells (n=9), totaling n=96 single-cells. We implemented SkewedCID on the live single-cells (n=87). The *t*-SNE plot (**Fig. 4a**) shows a clear distinction of the good and skewed cells. Our method reduced number of good cells from (n=47 to n=39) and increased poor quality from

n= 40 to n=48. This indicates that the standard QC procedures currently used in scRNA-seq analysis are inadequate to discover single-cells with poor quality.

To investigate the biological meaning of the good and skewed cells we used the microscopic image of the Fluidigm C1 chip provided by the authors in[4] and computed the cell size in pixels of all single-cells (**Methods**) and (**Supplementary Table 2**). The good cells have larger cell size compared to the skewed cells (**Fig. 4b**, bottom plot).

An additional analysis to investigate biological difference between good and skewed cells is the distribution of the cell-cycle phase in the data set. We assigned the cell-cycle phase for each single-cell based on their gene expression (**Methods**). The majority of the good cells were in the G2/M and M/G1 phase (n=26 of 39), bottom graph of **Fig. 4c**, suggesting that good cells pass the G2 checkpoint (G2/M) and the spindle checkpoint (M/G1). On the other hand, the majority of the skewed cells are in G1/S, S and G2 phase (n= 32 of 48), indicating that skewed cell reside around the S phase (Synthesis Phase) but do not pass to the Mitotic phase (chromosome separation), top panel of **Fig. 4c**.

We investigated the differences in coverage skewness between the good and skewed cells (**Fig. 4d**). The skewed cell possess high coverage skewness compared to the good cells.

Finally, we performed gene expression analysis between the good and skewed cells. The clustering of the top 100 most variable genes across cells illustrated in the heatmap (**Fig. 4e**) with good and skewed cells are clustered separately based on the gene expression of the top 100 genes. We performed gene set enrichment analysis (GSEA) on of the top 100 genes, we found the KRAS signaling DN pathway was enriched in the good cells (p-value < .001). The set of genes up-regulated by KRAS play roles in cell signaling[25].

Effect of skewed cells on downstream analysis

Since we noticed a great variability in the mean expression of the HKGs between good and skewed cells, we investigated the effect of skewed cells on the downstream analysis of the scRNA-seq experiment. The *t*-distributed stochastic neighbor embedding (*t*-SNE)[26], is a common dimension reduction technique in scRNA-seq analysis. *t*-SNE is usually performed after read count normalization[27]. We analyzed the impact of filtering skewed cells on *t*-SNE implementation using mES data set (**Fig. 5**). In the first data set E-MTAB-2600[19] generated by SMARTer protocol (**Fig. 5a**), the top panel shows the *t*-SNE plot of all single-cells clustered and colored by the growth factors used in the experiment, some of the single-cells are misplaced in the wrong cluster (mis-cluster). The *t*-SNE plot in the middle panel of (**Fig. 5a**), cluster and color the single-cells based on the classification of good and skewed cells. The majority of the skewed cells clustered together with few exceptions. The *t*-SNE plot in the bottom panel of (**Fig. 5a**) shows replotting the *t*-SNE after filtering the skewed cells. The plotting of the good cells only show distinct clustering of the cells based on the growth factors compared the *t*-SNE before the filtering of the skewed cells. This shows the impact of the skewed cells on the clustering of the single-cells.

The data set GSE98664[24] is a time-course analysis of mES development generated by RamDA-seq protocol (**Fig. 5b**). The top panel illustrates *t*-SNE with four clusters (four time-points). Each of the four clusters contains single-cells that do not belong to the same time-point (mis-clustered). The wrongly clustered single-cells are in fact skewed cells that stopped development but were mis-annotated by as developing cells. The middle *t*-SNE plot shows the clustering of the same data set based on good and skewed cells and the appearance of the skewed cells in two clusters (ES_12h & ES_24h). The bottom *t*-SNE illustrates re-clustering of the data

set after filtering the skewed cells; the plot shows clear improvement of the clustering result. Compared to the top *t*-SNE plot, the final *t*-SNE after filtering of the skewed cell removed the group of single-cells of time-point ES_72h from ES_12h and the ES_24h cells were removed from the cluster from ES_48h.

In the single-cell data set PRJDB5282[16], generated by C1 CAGE protocol (**Fig. 5c**), the skewed cells clustered separately on *t*-SNE plot (top of **Fig. 5c**). After filtering the skewed cells, *t*-SNE shows one cluster consisting of good cells.

All the above examples demonstrate the strong impact of skewed cells on the clustering results. The identification and filtering of the skewed cells is important to consider in any downstream analysis.

Ratio of intergenic expression

In scRNA-seq protocols, cDNA is obtained from the reverse transcription of RNA. This step is followed by amplification of cDNA by PCR or in vitro transcription before sequencing[28]. The amplification step is required due to the small amount of RNA found in an individual cell, and the workflow is prone to losses or biases[29]. To investigate the possibility of such problems resulting, e.g., from genomic DNA contamination, we developed a model to quantify the ratio of intergenic expression (**Methods**). For each cell, the model computes the ratio of intergenic expression. As a control, we considered matched cell type bulk RNA-Seq data set from ENCODE[30] (PloyRNA-Seq and Total RNA-Seq). As an example (**Fig. 6a**), the data set GSE68981[31] from mouse hematopoietic stem cells (HSCs) was analyzed with single-cell (C1-single-cell mRNA-Seq protocol) and bulk RNA-Seq used as control. As another example, we compared the human data sets GSE75748[32] (**Fig. 6b**), from human embryonic stem cells

(analyzed with single-cell SMARTer protocol) and bulk RNA-Seq. As demonstrated in the above examples, the ratio of intergenic expression is high in the scRNA-seq data compared to the bulk RNA-Seq.

Our analysis suggested that scRNA-seq data prone to high level of intergenic expression compared to bulk RNA-Seq. One possibility for such high read counts from intergenic regions is amplification of genomic DNA, as such signals are not observed in bulk RNA-seq data from same cell types. Our results suggest that single-cell data set should be evaluated for intergenic expression, possibly originating from genomic DNA amplification.

Annotation and classification of the sequence reads

One potential source of problems with scRNA-seq is microbial contamination (bacteria, archaea and viruses), we developed a workflow to annotate and classify unmapped sequence reads from scRNA-seq experiments (**Methods**). Sequence reads are annotated on three categories, single-mapped, multi-mapped, or unmapped on the reference genome. The majority of the sequence reads are annotated as mapped reads (**Fig. 6c, d**) (wheat color) with one exception of the data set E-MTAB-3346[33] (mouse thyme epithelial cells (mTECs) generated by Smart-Seq2 protocol) (**Fig. 6c**). The unmapped reads are annotated as positive sequencing control (phiX), or as reads belonging to unexpected organisms (bacteria, archaea, virus) (**Fig. 6c, d**). The dataset PRJEB8994[1] of gene expression during the first three days of human development (**Fig. 6d**) is the only data set that doesn't contain any reads annotated as unexpected organisms (bacteria, archaea, virus), possibly because the PRJEB8994 experiment was conducted in a sterile clinical environment.

A set of unmapped sequence reads were annotated as unexplained-unmapped (**Fig. 6c, d**) (gray color). These are reads that failed to find any match in the Kraken database. Notable, the two data sets GSE54695[34] and GSE70798[35] generated by CEL-Seq contain high numbers (> 50% of the raw reads) of unexplained-unmapped reads. A common source of the unmapped reads in scRNA-seq experiments might arise from the sequence linkers. Specific sequencing linkers are usually added to cDNA during library construction.

Mycoplasma contamination of cell cultures in the laboratory is common and special detection kits are available to ensure that the cell culture is free of such contamination. In bulk RNA-seq, it was reported that several large scale RNA-seq projects generate 9-20% of reads that do not map to the human reference genome[36-38].

Discussion

Advances in scRNA-seq have already impacted biology and medicine and will increasingly do so. It has enabled investigation of transcriptomic variation between individual cells, thereby enabling the discovery of new cell types, analyses of cellular response to stimulation, analyses of the nature and dynamics of cell differentiation and reprogramming, and study of transcriptional stochasticity. In spite of the technical advances, several challenges still remain and need to be understood for improved interpretation of the data. There is high variability in the performance of scRNA-seq protocols in terms of coverage, accuracy and specificity, impacting the quality of data generated by different scRNA-seq protocols[12, 13]. The variability among scRNA-seq protocols and the quality of the scRNA-seq data set might also impact global efforts to map transcription in human and mouse cells[39, 40].

Our analysis identified wide differences in patterns of gene coverage generated from same cell types by different scRNA-seq protocols. Based on the analysis of the gene body coverage, we identified two classes of single-cells in any data set: cells with method-specific distribution of gene body coverage, and cells with a skewed distribution. Each of the scRNA-seq protocols yields different gene body coverage. A skewed distribution with excess observed 3'-end bias could be attributed to technical or biological processes during any of the experimental steps, such as reaction failure, or cell death, triggering mRNA degradation. During embryo development, skewed distribution coverage suggests induction of maternal transcript degradation at early genome activation in the earliest developmental stages[41], as maternal transcript degradation takes place after the fertilization till day 3[1, 42] .

SkewedCID based on the observation of the two distribution patterns in the gene body coverage of scRNA-seq protocols and the identification of the good and skewed cells. The skewed cells might result from a technical failure of the reaction, or be either live or dead cells. The skewed live cells might potentially be quiescent satellite cells, a form of ($\mathbf{G_0}$ cell-cycle phase), cell enter $\mathbf{G_0}$ phase (resting) due to different environmental factors. The common feature between the identified skewed cells and the quiescent satellite cells is the low RNA content[43, 44].

We evaluated the impact of gene body coverage skewness on the expression of housekeeping genes and find significant differences in the expression of housekeeping genes between good cells and skewed cells. We further assessed the impact of filtering out the skewed cells from downstream analysis (clustering and differential gene expression analysis), and found that exclusion of skewed cells drastically changes *t*-SNE clustering results. Wrong clustering may lead to false discovery resulting from the skewed cells. The skewed cells show different

expression profile of the top expressed genes compared to the good cell, as seen in the heatmap. We also investigated the difference in biological features between skewed and good cells (cell-size and cell cycle). In contract to skewed cells, good cells are strongly correlated with cell size and cell cycle phase.

We recommend that skewed cells should be identified and excluded from any downstream analyses of scRNA-seq experiments.

The data set-to-data set similarity analyses in terms of gene expression profiles rather expectedly show weak expression correlation between data sets generated by different protocols. Our results show strong variation in sequence depth, detection power and gene saturation revealed by different scRNA-seq protocols. This result demonstrated challenges for the current efforts to computationally integrate heterogeneous scRNA-seq data sets generated by different protocols and labs[45-47].

We developed a model to estimate the intergenic expression in scRNA-seq, and observed high level of intergenic expression in single-cells compared to the control bulk data set. This might be caused by scRNA-seq contamination with genomic DNA reads, but other alternatives to explain this observation might also exist. Finally, we annotated and classify unmapped reads in order to find the source of contamination in scRNA-seq experiments.

In conclusion, our results demonstrated that a QC procedure to segregate good cells and skewed cells in scRNA-seq should be incorporated in any scRNA-seq experiment to avoid false interpretations of data. scRNA-seq experiments may suffer from biological and technical failures such as genomic DNA amplification and microbial contamination as possible sources.

Materials and Methods

Study design

Based on the objective of the scRNA-seq experiment, the protocols are divided in two categories; full-length sequence profiling or transcript end-tagging (5' or 3'). In full-length sequence protocols (SMARTer, Smart-Seq, SUPer-Seq, RamDA-seq, etc.) the sequence reads cover the entire gene body (5'to 3'-end) and quantify gene and transcript isoforms. The end-tagging based sequencing protocols (C1 CAGE, CEL-Seq, CEL-Seq2, STRT, 10x Chromium Single Cell 3' –end etc.) target one end of the transcript (5'-end or 3'-end) and are used to identify promoters (5' tagging) or give an estimate of transcript abundance. In our study design (**Fig. 1**). We analyzed data sets produced with 15 different protocols representing commonly used scRNA-seq methods of the above two categories.

Study data set

To perform fair comparison among different protocols, we used both batch-matched and unmatched mouse and human dataset of primary cells and cell lines (**Table 1**). In this study, we used three types of data sets: (1) we generated a data set for human MCF10A cells, (2) we reanalyzed published human and mouse scRNA-seq from International Nucleotide Sequence Database Collaboration (INSDC) data set, and (3) we reanalyzed published scRNA-seq data set for human and mouse from 10x Genomics data portal. The three types of the data sets are described below.

10× Genomics Chromium experiment of human MCF10A cells

10x Genomic Chromium data set was generated from the MCF10A cells (ATCC). The cells were grown in DMEM/F12(1:1) as described in [48]. RNA-Seq library was prepared using 10x Chromium Single Cell 3' –end Reagent Kits User Guide (v2 Chemistry). Libraries were sequenced using paired-end sequencing (26bp Read 1 and 98bp Read 2) with a single sample index (8bp) on the Illumina HiSeq 2500.

International Nucleotide Sequence Database Collaboration (INSDC) data set

The method for collecting and processing the raw read scRNA-seq from the public databases were illustrated in [49] and listed in **Supplementary Table 1**[1, 4, 15, 16, 19-22, 24, 31, 32, 34, 50-63]. In brief, published scRNA-seq were collected by searching PubMed for human and mouse scRNA-seq articles. Our strategy was to include different types of cells generated by different technology platforms. All single-cells in the database were annotated using different type of ontologies[64]. This strategy enabled us to cover a wide range of cell types and datasets generated by different platforms. We retrieved study accession number(s) of the original data deposited to International Nucleotide Sequence Database Collaboration (INSDC). The study accession numbers were used to retrieve sequence read files and metadata files from INSDC sites (DDBJ, EMBL-EBI, NCBI). To obtain FASTQ files, we implemented an automated program using the NCBI SRA Toolkit[65]. Metadata about each data set were collected as well. This metadata contains information about the cell type, protocol, sequence platform, single-cell isolation techniques, etc. We implemented automated script to retrieve data set metadata utilizing The Entrez Programming Utilities (E-utilities) from NCBI[66].

10x Genomics data set

We downloaded two data sets from 10x Genomics portal (https://support.10xgenomics.com/single-cell-vdj/datasets). The Human PBMCs of a healthy donor - 5' –end gene expression and cell surface protein (8,258 single-cell) BAM files were downloaded from http://cf.10xgenomics.com/samples/cell-vdj/3.0.0/vdj_v1_hs_pbmc2_5gex_protein/vdj_v1_hs_pbmc2_5gex_protein_web_summary.html The Mouse PBMCs from C57BL/6 mice - 5' gene expression (8,500 single-cell) BAM files were downloaded from http://cf.10xgenomics.com/samples/cell-vdj/3.0.0/vdj_v1_mm_c57bl6_pbmc_5gex/vdj_v1_mm_c57bl6_pbmc_5gex_web_summary.html

Data processing of the raw sequence reads

For the raw sequence data downloaded from INSDC, we run basic QC procedures to obtain quality assessment metrics of the raw sequence reads. The QC procedures includes testing of all FASTQ files with FastQC tool [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] to identify any quality issues. Additional QC procedures of the raw reads includes count of raw tags. The raw sequence reads were aligned to a recent reference genome build (GRCh38 (human) or GRCm38 (mouse) genome assembly). We used STAR software (version 2.5.1b)[67] with default settings and GENCODE gene annotations in the release v24 for human and the release

vM9 for mouse for all data set but not for (GSE98664) in which we used GENCODE vM22. Aligned reads in BAM file format together with the log files generated by STAR were used to obtain quality assessment metrics (total read count, number of uniquely mapped reads and assigned reads (mapped reads assigned to gene)). The mapping ratio, counts of mapped reads, unmapped and multi-mapped reads were summarized using SAMtools software[68]. For MCF10A data set, we used cell ranger2.1.1 for data processing. The data sets from 10x Genomics were processed with Cell Ranger version 3.1.0.

Reads count summarization and expression normalization

To obtain expression matrices, we quantified gene expression counts using featureCounts (in the Subread package Version 1.5.0-p1) [69]. The gene expression counts were normalized into transcripts per million reads (TPM) and fragments per kilobase million (FPKM) to generate a gene expression table for each study, according to the following standard formula.

TPM (gene-level expression) = mapped reads assigned to each gene) * 1,000,000 / mapped reads

FPKM (gene-level expression) = mapped reads assigned to each gene) * 1000 / (gene length (bp)) * 1,000,000 / mapped reads

Computation and visualization of the gene body coverage

To compute the gene body coverage for data set, we implemented[11] module geneBody_coverage.py. The module was used to check if reads coverage was uniform and if there was any 5'/3' end bias. The input for the module is indexed BAM files and gene model in BED format. All BAM files were sorted and indexed prior to this procedure. Gene models were

downloaded from http://rseqc.sourceforge.net/#download-gene-models-update-on-08-07-2014 . The gene model BED files for human hg38_Gencode_V28.bed.gz and for mouse mm10_Gencode_VM18.bed.gz were preprocessed to filter ribosomal RNA and transfer RNA. Since processing gene body coverage is a time-consuming task even in a high performance computer environment, data sets with large numbers of single-cells were split to smaller number of data sets to run small jobs in parallel. The result of the gene body coverage module is a vector of normalized values. The values cover the gene body from the 5'-end to the 3'-end scaled from 0-100 (positions). The value for each position range from (0-1), where 0 indicate now coverage and 1 indicate full coverage at the position on the gene body. The vector of the normalized values were post-processed in several steps for visualization and study the normality and skewness of the coverage (**Fig. 3a**).

Computation of intergenic expression

To quantify the possibility of genomic DNA contamination (e.g. due to the PCR amplifications of the starting material of genomic DNA), we modelled the following formula:

Possibility of genomic contamination (%) = (((total number of mapped reads) - (the total number of reads that are assigned to a gene feature in GENCODE annotation)) / ((total number of mapped reads))) * 100.

The possibility of genomic DNA contamination was computed for each single-cell and summarized per the data set.

Analysis and annotation of the unmapped reads

In our pipeline for aligning raw reads to the reference genome, we kept log files and unmapped reads. We investigated the source and ration of the unmapped reads from each FASTQ file. Our workflow to analyse the unmapped reads started with filtering of ribosomal RNA and artificial reads from the BAM file of the unmapped reads using TagDust tool[70]. Next we filtered multi-mapped reads to get true unmapped reads and finally, we performed blast search on the most frequent unmapped reads [71]. The remaining reads were screened for microbial contamination sequences. To screen for such organisms, we utilized metagenomics tools: sequana[72] and Kraken [73]. The Kraken tool provided 8GB database of complete bacterial, archaeal and viral genomes (https://ccb.jhu.edu/software/kraken/dl/minikraken_20171019_8GB.tgz).

Cell size estimation of the data set GSE46980

The data set GSE46980 of mESs [4] was generated by the STRT protocol and provided full annotation of quality status of the single-cells (n=96). The authors classify each single-cell as either dead (depleted before cell capture by the flow-cell) or live cell. The live cells were further classified as either low quality cells or good quality cells (see [4] for details on how the annotation was performed). We used this data set to compare our QC method of good and skewed cells. Additionally, the data set provided a microscopic image of the Fluidigm C1 chip. In the microscopic image, each Fluidigm C1 chip (a 96-well plate) was imaged after cell capture and a grid of thumbnails was generated for each chip. To verify some of the morphological phenotypes of the good and skewed cells, we estimated morphological properties of the cells

based on the microscopic image, cell size, areas, circularity, skewness roundness and solidity that were calculated using the ImageJ tool [74].

Cell cycle phase prediction

To further evaluate some of the biological phenotypes of the good and skewed single-cells, we predicted the cell-cycle phase, the cell-cycle phase predicted computationally based on the expression profile of the single-cell[75]. We obtained the predefined human cell-cycle marker set provided in [76]. As for the mouse cell-cycle markers, the orthologous mouse genes of the human cell-cycle gene marker were obtained **Supplementary Table 3**. The cell-cycle phase predictor [75] assign any of (S, G1/S, G2, M/G2 , G2/M) phase to each single-cell.

Statistical tests, boxplots and plotting tools

Unless otherwise indicated, all p-values were obtained with two-sided t-test. In all boxplots, center lines indicate median values, box heights indicate the inter-quartile range of data. The ggplot2 library from R software version 3.5.1 (2018-07-02) was used for plotting of all plots and figures.

Data availability

The raw read data listed in (**Supplementary Table 1**) are available from INSDEC sites. We developed SCPortalen, a single-cell database in which we deposited all results from this study at http://single-cell.clst.riken.jp/

Acknowledgements

This work was supported by research grants for the RIKEN Center for Life Science Technologies, RIKEN Center for Integrative Medical Sciences and RIKEN Open Life Science Platform project from MEXT, Japan. SK and JK were supported in part by Knut and Alice Wallenberg Foundation (KAW2015.0096) (Sweden), Swedish Research Council, Jane and Aatos Erkko Foundation (Finland), and Sigrid Jusélius Foundation (Finland). This work was initiated when JK was a Japan Society for the Promotion of Science Fellow at RIKEN Center for Integrative Medical Sciences, Yokohama.

References

[1] Töhönen V, Katayama S, Vesterlund L, Jouhilahti EM, Sheikhi M, Madissoon E, et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun. 2015;6:8207.

[2] Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570:332-7.

[3] Giladi A, Amit I. Single-Cell Genomics: A Stepping Stone for Future Immunology Discoveries. Cell. 2018;172:14-21.

[4] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11:163-6.

[5] Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17:29.

[6] Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371-5.

[7] Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10:1093-5.

[8] Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500:593-7.

[9] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381-6.

[10] DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28:1530-2.

[11] Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184-5.

[12] Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65:631-43.e4.

[13] Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14:381-7.

[14] Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. Nat Protoc. 2012;7:813-28.

[15] Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17:77.

[16] Kouno T, Moody J, Kwon AT-J, Shibayama Y, Kato S, Huang Y, et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nature Communications. 2019;10:360.

[17] Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666-73.

[18] Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30:301-4.

[19] Kim JK, Kolodziejczyk AA, Ilicic T, Illicic T, Teichmann SA, Marioni JC. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun. 2015;6:8687.

[20] Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33:155-60.

[21] Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015;16:148.

[22] Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21:1160-7.

[23] Haberle V, Forrest AR, Hayashizaki Y, Carninci P, Lenhard B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 2015;43:e51.

[24] Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9:619.

[25] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739-40.

[26] ***Maaten Lvd***, ***Hinton G***. Visualizing Data using t-SNE. Journal of Machine Learning Research2008. p. 2579--605.

[27] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50:96.

[28] Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20:257-72.

[29] Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. Nat Methods. 2014;11:22-4.

[30] Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794-D801.

[31] Yang J, Tanaka Y, Seay M, Li Z, Jin J, Garmire LX, et al. Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors. Nucleic Acids Res. 2017;45:1281-96.

[32] Chu LF, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. 2016;17:173.

[33] Brennecke P, Reyes A, Pinto S, Rattay K, Nguyen M, Küchler R, et al. Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. Nat Immunol. 2015;16:933-41.

[34] Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11:637-40.

[35] Meredith M, Zemmour D, Mathis D, Benoist C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. Nat Immunol. 2015;16:942-9.

[36] Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014;32:915-25.

[37] Consortium G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648-60.

[38] Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods. 2017;14:135-9.

[39] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017;6.

[40] Consortium TM, coordination O, coordination L, processing Oca, sequencing Lpa, analysis Cd, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562:367-72.

[41] La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560:494-8.

[42] Dobson AT, Raja R, Abeyta MJ, Taylor T, Shen S, Haqq C, et al. The unique transcriptome through day 3 of human preimplantation development. Hum Mol Genet. 2004;13:1461-70.

[43] Fukada S, Uezumi A, Ikemoto M, Masuda S, Segawa M, Tanimura N, et al. Molecular signature of quiescent satellite cells in adult skeletal muscle. Stem Cells. 2007;25:2448-59.

[44] Hüttmann A, Liu SL, Boyd AW, Li CL. Functional heterogeneity within rhodamine123(lo) Hoechst33342(lo/sp) primitive hemopoietic stem cells revealed by pyronin Y. Exp Hematol. 2001;29:1109-16.

[45] Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019;37:685-91.

[46] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, III, et al. Comprehensive Integration of Single-Cell Data. Cell.

[47] Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat Methods. 2019;16:695-8.

[48] Watanabe K, Panchy N, Noguchi S, Suzuki H, Hong T. Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition. NPJ Syst Biol Appl. 2019;5:21.

[49] Abugessaisa I, Noguchi S, Böttcher M, Hasegawa A, Kouno T, Kato S, et al. SCPortalen: human and mouse single-cell centric database. Nucleic Acids Res. 2018;46:D781-D7.

[50] Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14:R31.

[51] Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, et al. Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep. 2014;7:1130-42.

[52] Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods. 2016;13:329-32.

[53] Proserpio V, Piccolo A, Haim-Vilmovsky L, Kar G, Lönnberg T, Svensson V, et al. Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. Genome Biol. 2016;17:103.

[54] Zhang X, Xu C, Yosef N. Simulating multiple faceted variability in single cell RNA sequencing. Nat Commun. 2019;10:2611.

[55] Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SA, et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature. 2016;534:391-5.

[56] Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343:193-6.

[57] Reinius B, Mold JE, Ramsköld D, Deng Q, Johnsson P, Michaëlsson J, et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet. 2016;48:1430-5.

[58] Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 2015;25:1860-72.

[59] Grover A, Sanjuan-Pla A, Thongjuea S, Carrelha J, Giustacchini A, Gambardella A, et al. Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. Nat Commun. 2016;7:11075.

[60] Dueck H, Khaladkar M, Kim TK, Spaethling JM, Francis C, Suresh S, et al. Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. Genome Biol. 2015;16:122.

[61] Leng N, Chu LF, Barry C, Li Y, Choi J, Li X, et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nat Methods. 2015;12:947-50.

[62] Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016;165:1012-26.

[63] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161:1202-14.

[64] Abugessaisa IEA, Sivertun A. Ontological approach to modeling information systems. The Fourth International Conference onComputer and Information Technology, 2004 CIT '042004. p. 1122-7.

[65] Leinonen R, Akhtar R, Birney E, Bonfield J, Bower L, Corbett M, et al. Improvements to services at the European Nucleotide Archive. Nucleic Acids Res. 2010;38:D39-45.

[66] NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2017;45:D12-D7.

[67] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.

[68] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.

[69] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923-30.

[70] Lassmann T, Hayashizaki Y, Daub CO. TagDust--a program to eliminate artifacts from next generation sequencing data. Bioinformatics. 2009;25:2839-40.

[71] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403-10.

[72] Cokelaer T, Desvillechabrol D, Legendre R, Cardon M. 'Sequana': a Set of Snakemake NGS pipelines. Journal of Open Source Software. 2017.

[73] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.

[74] Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9:671-5.

[75] Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.

[76] Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell. 2002;13:1977-2000.

Table Legends

Table 1: Sample design and tested single-cell RNA sequence protocols.

Figure Legends

Figure 1: SkewedCID workflow.

Flow diagram summarizing the Workflow for skewness-based quality assessment of single-cell RNA-Seq experiment. Data sets are collected from three sources, in-house, public data, and 10xGenomics web resource. The figure illustrates the computation analysis tasks, and the result from each task.

Figure 2: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using mES.

9 datasets from mouse embryonic stem cells. **(a)** Distribution of the mapped reads (tags) across the genes. Each panel shows gene body coverage percentile per dataset. The x-axis represents the gene body from 5' end to 3' end scaled from 0-100, and the y-axis gene coverage (0-1). Each line represents a single cell. **(b)** Mean calculated for bin size = 10. (**c**-**d**) Skewed distribution of the gene body coverage. **(c)** & (**d)** show the bias towards the 3'-end of the gene body (magenta dashed box) and low coverage in the middle of the gene body (blue dashed box). (STRT as 5'-end sequence protocol) shows the bias towards the 3'-end of the gene body (magenta dashed box) and high coverage in the middle of the gene body (green dashed box). **(e)** The variability in gene expression plot. The X-axis is the mean of the normalized gene expression (FPKM), Y-

Axis is the coefficient of variation. CV/mean correlates the sequence depths and variability in gene expression among different protocols. Figure **(f)**, grouping of the smooth lines from **(e)**.

Figure 3: Classification of the good and skewed coverage distribution cells.

Overall description and steps for the QC methods; **(a)** The figure illustrates the method to discriminate skewed cells with skewed coverage distribution. **(b–d)** application of the QC methods in 3 different data sets. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good cells with normal gene coverage and skewed cells with skewed coverage distribution. Right chart**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells.). We applied our method to other dataset (**Supplementary Figs.** 18–23)

Figure 4: Validation of QC method.

We used an existing experimentally validated dataset to validate the QC methods. The data set GSE46980. **(a)** *t*-SNE illustrating the clustering of single-cells. **(b)** Line and point graph with error bars representing the standard error of the mean of the single-cell size. Upper panel shows that cell size of the single-cells as annotated by the data set authors (dead, low quality and good quality cells). The bottom plot shows that cell size of the cells as annotated after applying our QC methods (Skewed cells and Good cells). **(c)**, distribution of predicted cell-cycle phases among skewed cells (top) and good cells (bottom). **(d)** Coverage skewness comparison between

good and skewed cells. **(e)** Heatmap showing the top 100 differentially expressed genes in the data set. Columns are cell category (good and skewed) and rows are gene names.

Figure 5: Effect skewed cells on downstream analysis.

To evaluate the effect of the skewed cells when performing downstream analysis, t-SNE generated for several datasets. **(a)** mES single-cell treated with three types of growth factors. On the top is the t-SNE before the classification and colored by the growth factor used, in the middle a t-SNE of the skewed and good cells, on the bottom t-SNE after removing the skewed cells. **(b)** mES with four development time-points. Top t-SNE of the dataset before the classification. The cluster in the bottom show the majority of the cells at 12 h, with few cell from 72 h. In the middle t-SNE we observed that the skewed cells are cells from 72h. The bottom t-SNE show better clustering of the single-cell per development time-point after filtering the skewed cells. **(c)** t-SNE perfectly discriminate good and skewed cells.

Figure 6: Ratio of intergenic expression and annotation of the reads.

(a–b) High level of intergenic expression in scRNA-seq **(a)** mouse and **(b)** human data compared to bulk. Box plots are grouped by protocol name. **(c–d)** Classification of reads annotations per dataset for **(c)** mouse and **(d)** human data. Mapped reads are separated between uniquely mapped and multimapped on the reference genome. Unmapped reads are annotated in categories according to the explanation of unmapping, namely tagdust filtered, positive sequencing control (phiX), contamination by a foreign organism (archea, bacteria, virus). Unmapped reads that couldn't be explained are presented in gray.

Supplementary Table Legends

Supplementary Table 1: List of the data set used in the analysis.

Supplementary Table 2: Analysis of the microscopic image for the data set GSE46982.

Supplementary Table 3: List of the cell cycle marker genes for mouse.

Supplementary Figure Legends

Supplementary Figure 1: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using mouse CD4 T cells.

4 datasets from mouse CD4 T cells. **(a)** Distribution of the mapped reads (tags) across the genes. Each panel shows gene body coverage percentile per dataset. The x-axis represents the gene body from 5' end to 3' end scaled from 0-100, and the y-axis gene coverage (0-1). Each line represents a single cell. **(b)** Mean calculated for bin size = 10. **(c)** The variability in gene expression plot. The X-axis is the mean of the normalized gene expression (FPKM), Y-Axis is the coefficient of variation. CV/mean correlates the sequence depths and variability in gene expression among different protocols. **(d)** Grouping of the smooth lines from **(c)**.

Supplementary Figure 2: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using mouse fibroblast

4 datasets from mouse fibroblast. **(a)** Distribution of the mapped reads (tags) across the genes. Each panel shows gene body coverage percentile per dataset. The x-axis represents the gene body from 5' end to 3' end scaled from 0-100, and the y-axis gene coverage (0-1). Each line represents a single cell. **(b)** Mean calculated for bin size = 10. **(c)** The variability in gene

expression plot. The X-axis is the mean of the normalized gene expression (FPKM), Y-Axis is the coefficient of variation. CV/mean correlates the sequence depths and variability in gene expression among different protocols. **(d)** Grouping of the smooth lines from **(c)**.

Supplementary Figure 3: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using mouse hematopoietic cells

3 datasets from mouse hematopietic cells. **(a)** Distribution of the mapped reads (tags) across the genes. Each panel shows gene body coverage percentile per dataset. The x-axis represents the gene body from 5' end to 3' end scaled from 0-100, and the y-axis gene coverage (0-1). Each line represents a single cell. **(b)** Mean calculated for bin size = 10. **(c)** The variability in gene expression plot. The X-axis is the mean of the normalized gene expression (FPKM), Y-Axis is the coefficient of variation. CV/mean correlates the sequence depths and variability in gene expression among different protocols. **(d)** Grouping of the smooth lines from **(c)**.

Supplementary Figure 4: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using human embryo.

4 datasets from mouse human embryo cells. **(a)** Distribution of the mapped reads (tags) across the genes. Each panel shows gene body coverage percentile per dataset. The x-axis represents the gene body from 5' end to 3' end scaled from 0-100, and the y-axis gene coverage (0-1). Each line represents a single cell. **(b)** Mean calculated for bin size = 10. **(c)** The variability in gene expression plot. The X-axis is the mean of the normalized gene expression (FPKM), Y-Axis is the coefficient of variation. CV/mean correlates the sequence depths and variability in gene expression among different protocols. **(d)** Grouping of the smooth lines from **(c)**.

Supplementary Figure 5: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using human and mouse un-matched single-cells.

5 different human and mouse datasets. **(a)** Distribution of the mapped reads (tags) across the genes. Each panel shows gene body coverage percentile per dataset. The x-axis represents the gene body from 5' end to 3' end scaled from 0-100, and the y-axis gene coverage (0-1). Each line represents a single cell. **(b)** Mean calculated for bin size = 10. **(c)** The variability in gene expression plot. The X-axis is the mean of the normalized gene expression (FPKM), Y-Axis is the coefficient of variation. CV/mean correlates the sequence depths and variability in gene expression among different protocols.

Supplementary Figure 6: Dataset-to-dataset similarity of the mean expression: mouse ES cells.

Scatterplot of the mean value of the gene expression (FPKM) to compare the variability of gene expression in mES data set. Each thumbnail illustrates two datasets generated from same cell type. The dataset are either generated by same or different protocols, and some cases shows dataset generated by same protocols from different labs.

The figure illustrated that gene expression of dataset from cell type, generated by different protocols are dissimilar in the average expression.

Supplementary Figure 7: Dataset-to-dataset similarity of the mean expression: mouse CD4 T cells.

Scatterplot of the mean value of the gene expression (FPKM) to compare the variability of gene expression in mouse CD4 T cells data set. Each thumbnail illustrates two datasets generated from same cell type. The dataset are either generated by same or different protocols, and some cases shows dataset generated by same protocols from different labs.

Supplementary Figure 8: Dataset-to-dataset similarity of the mean expression: mouse fibroblast cells.

Scatterplot of the mean value of the gene expression (FPKM) to compare the variability of gene expression in mouse fibroblast cells data set. Each thumbnail illustrates two datasets generated from same cell type. The dataset are either generated by same or different protocols, and some cases shows dataset generated by same protocols from different labs.

Supplementary Figure 9: Dataset-to-dataset similarity of the mean expression: mouse hematopoietic cells.

Scatterplot of the mean value of the gene expression (FPKM) to compare the variability of gene expression in mouse hematopoietic cells data set. Each thumbnail illustrates two datasets generated from same cell type. The dataset are either generated by same or different protocols, and some cases shows dataset generated by same protocols from different labs.

Supplementary Figure 10: Dataset-to-dataset similarity of the mean expression: human embryo.

Scatterplot of the mean value of the gene expression (FPKM) to compare the variability of gene expression in human embryo cells data set. Each thumbnail illustrates two datasets generated

from same cell type. The dataset are either generated by same or different protocols, and some cases shows dataset generated by same protocols from different labs.

Supplementary Figure 11: Hanabi plot Gene saturation using mouse ES cells.

The x-axis shows total counts, and y-axis shows the number of detected genes in 9 mouse ES cell data set.

Supplementary Figure 12: Hanabi plot Gene saturation using mouse CD4 T cell.

The x-axis shows total counts, and y-axis shows the number of detected genes in 4 mouse CD4 T cells data set.

Supplementary Figure 13: Hanabi plot Gene saturation using mouse fibroblast cells.

The x-axis shows total counts, and y-axis shows the number of detected genes in 4 mouse fibroblast cells data set.

.

Supplementary Figure 14: Hanabi plot Gene saturation using mouse hematopoietic cells

The x-axis shows total counts, and y-axis shows the number of detected genes in 3 mouse hematopoietic cells data set.

Supplementary Figure 15: Hanabi plot Gene saturation using mouse PBMC

The x-axis shows total counts, and y-axis shows the number of detected genes in 2 mouse PBMC cells data set.

Supplementary Figure 16: Hanabi plot Gene saturation using human embryo.

The x-axis shows total counts, and y-axis shows the number of detected genes in 4 human embryo cells data set.

Supplementary Figure 17: Hanabi plot Gene saturation using human MCF10A, PBMC and HEK & 3T3

The x-axis shows total counts, and y-axis shows the number of detected genes in in different human cells data set.

.

Supplementary Figure 18: Classification of the good and skewed coverage distribution cells: mouse ES cells.

Application of the QC methods. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good cells with normal gene coverage and skewed cells with skewed coverage distribution. Right chart**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells).

Supplementary Figure 19: Classification of the good and skewed coverage distribution cells: mouse CD4 T cells.

Application of the QC methods. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good

cells with normal gene coverage and skewed cells with skewed coverage distribution. Right chart**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells).

Supplementary Figure 20: Classification of the good and skewed coverage distribution cells: mouse fibroblast.

Application of the QC methods. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good cells with normal gene coverage and skewed cells with skewed coverage distribution. Right chart**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells).

Supplementary Figure 21: Classification of the good and skewed coverage distribution cells: mouse hematopoietic cells

Application of the QC methods. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good cells with normal gene coverage and skewed cells with skewed coverage distribution. Right panel**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells).

Supplementary Figure 22: Classification of the good and skewed coverage distribution cells: human embryo.

Application of the QC methods. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good cells with normal gene coverage and skewed cells with skewed coverage distribution. Right chart**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells).

Supplementary Figure 23: Classification of the good and skewed coverage distribution cells: human MCF10A and HEK & 3T3 mix.

Application of the QC methods. Left chart filter cells with low-input reads, middle chart perform trimmed clustering on the coverage matrix of the cell with high read count. A proportion of the most outlying observations is trimmed (the skewed cells). This results into two sets of cell: Good cells with normal gene coverage and skewed cells with skewed coverage distribution. Right chart**,** the classification of cells was validated with the housekeeping genes (boxplot of the expression of the house keeping genes of the good vs. skewed cells).

Figure 1: Workflow for comparison of scRNA-seq protocols

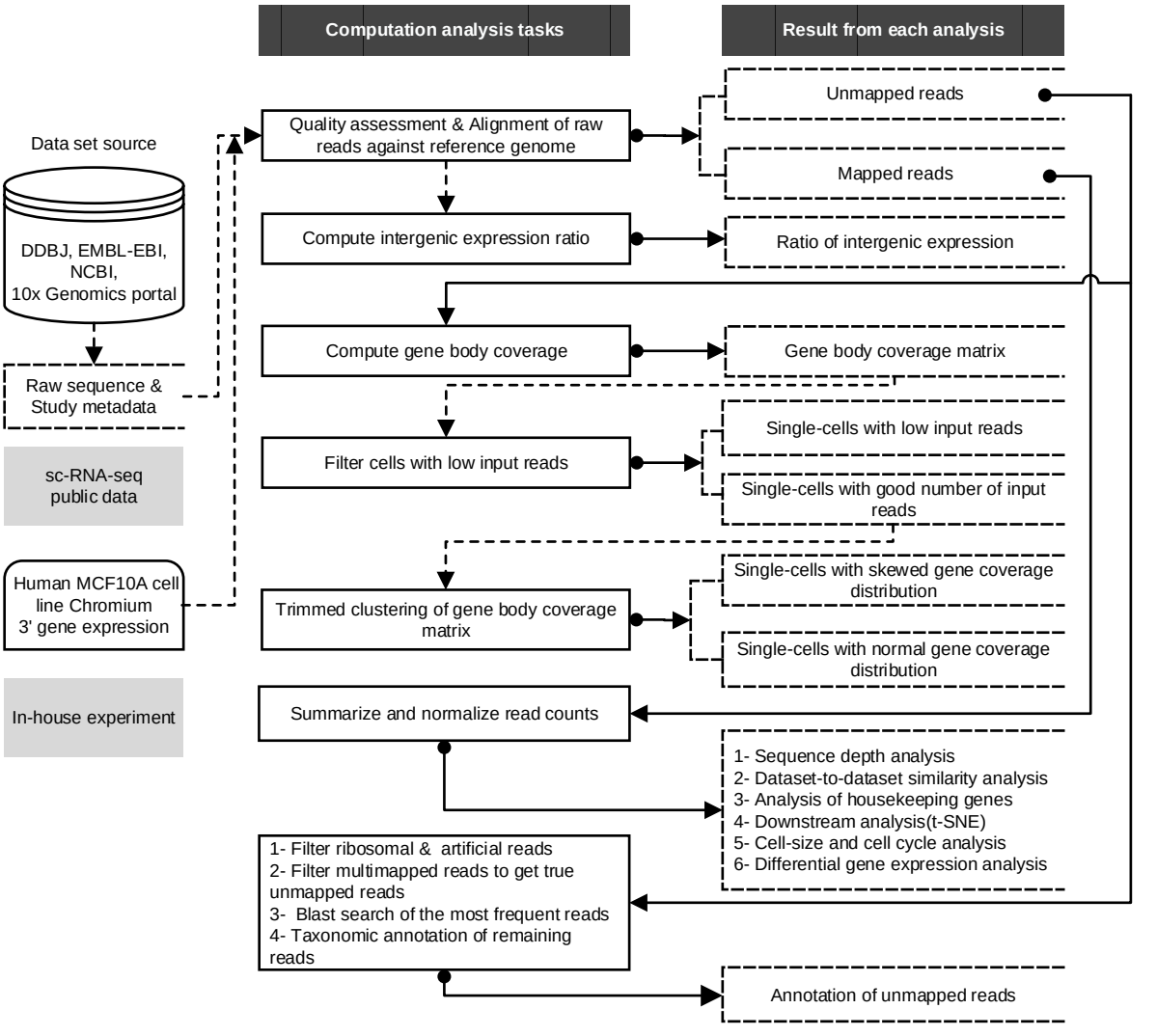
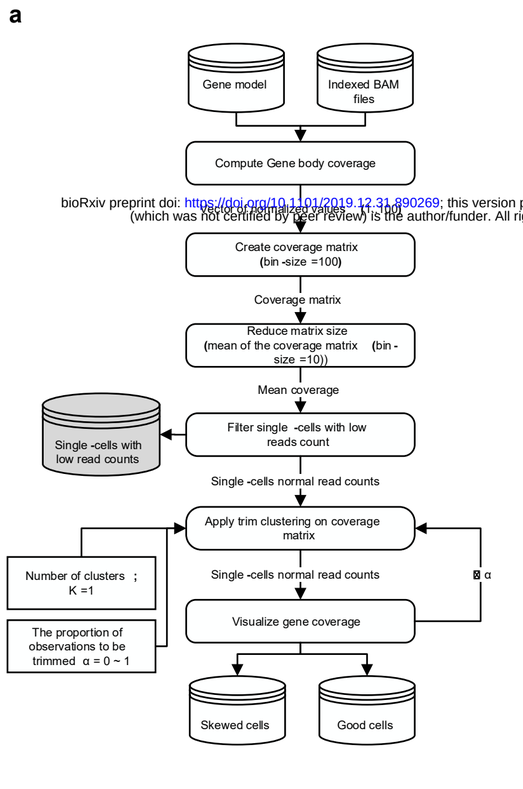
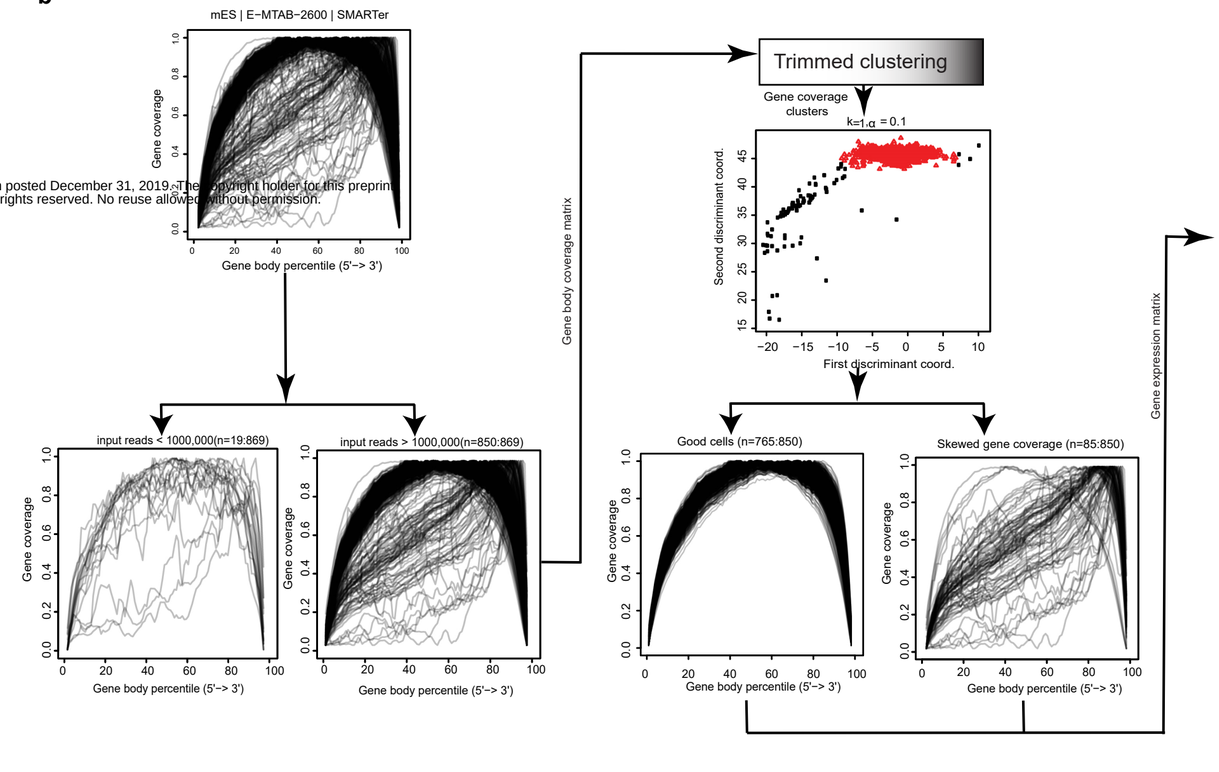
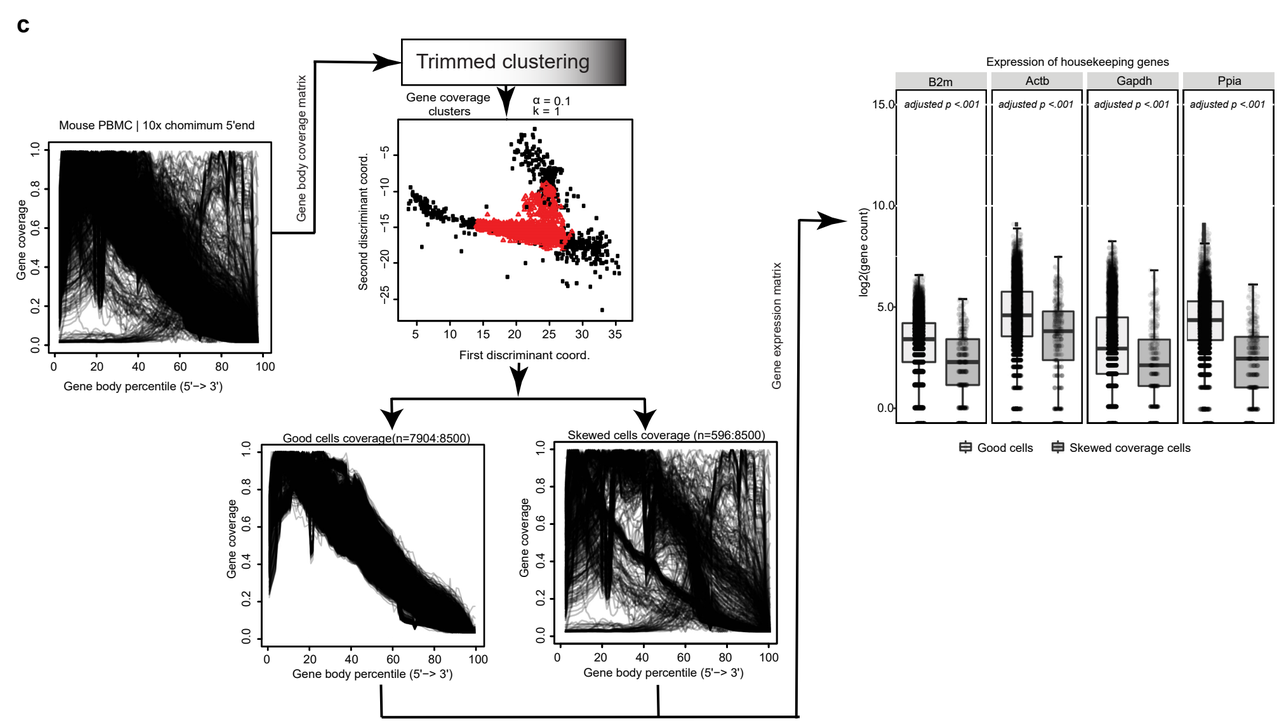


Figure 2: Gene coverage skewness and variation in expression among single-cell RNA-Seq protocols using mES

Figure 3 : Classification of the good and skewed coverage distribution cells







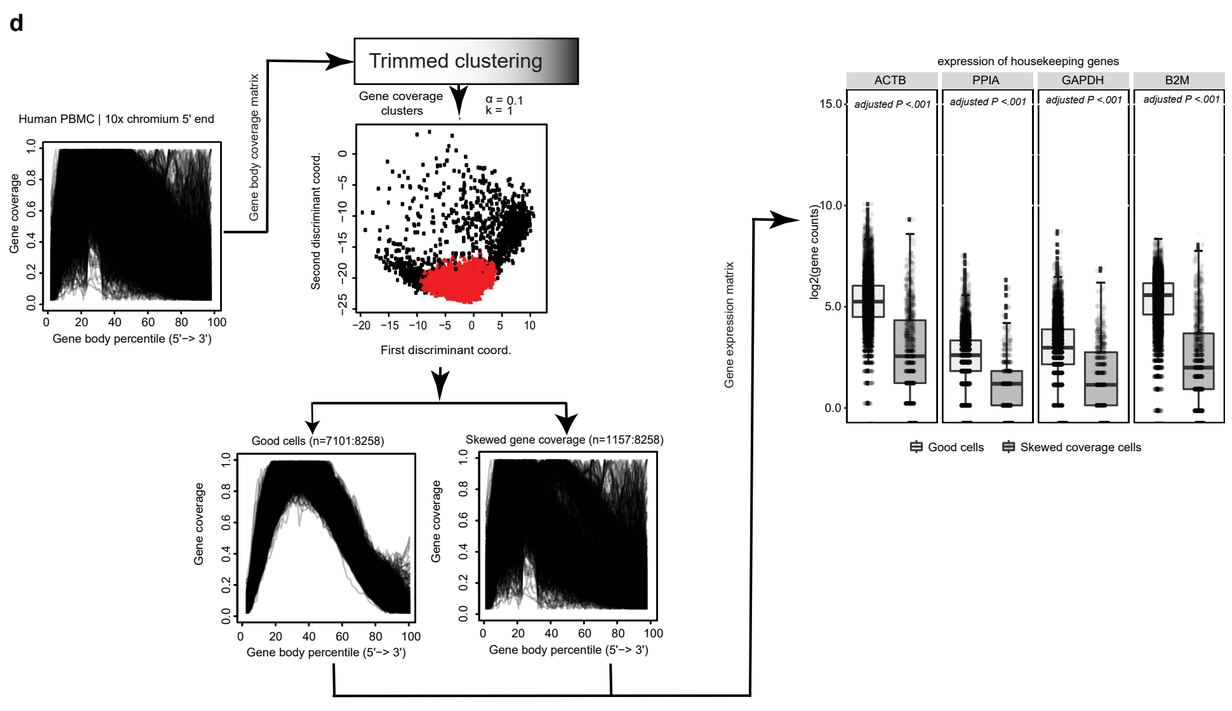
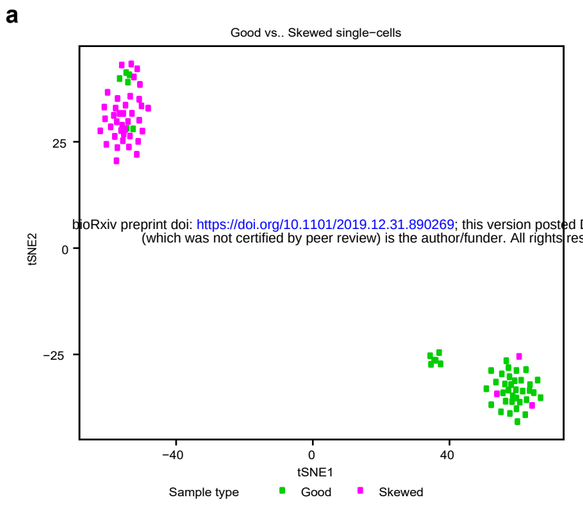
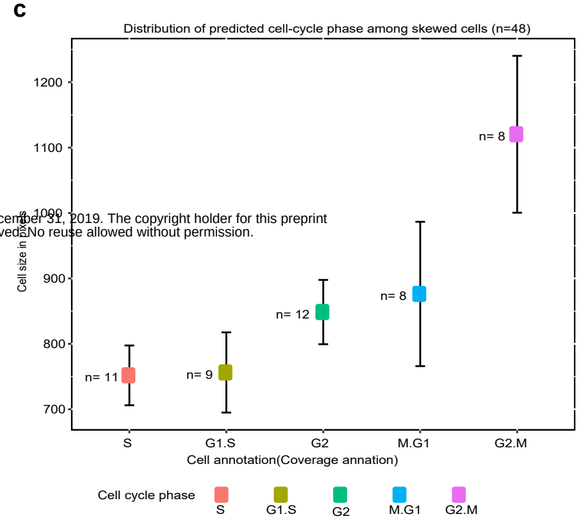
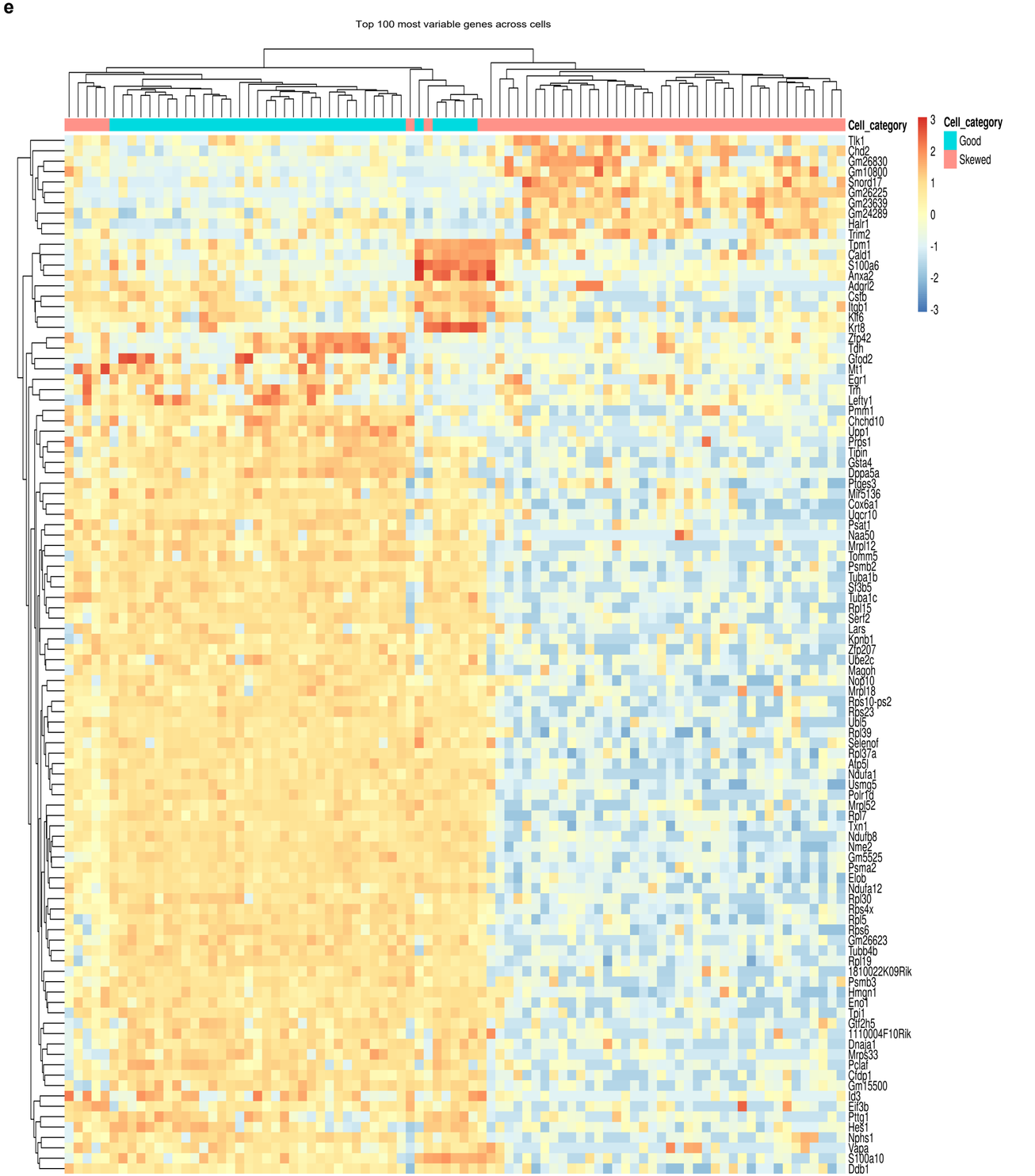
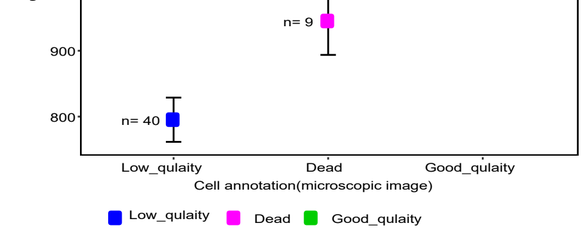


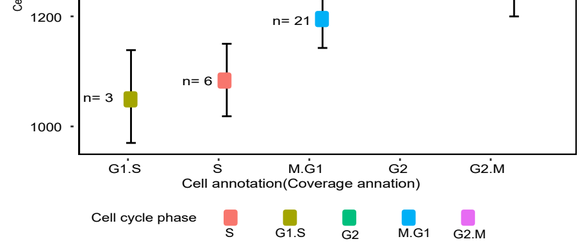
Figure 4: Validation of QC method using GSE46980

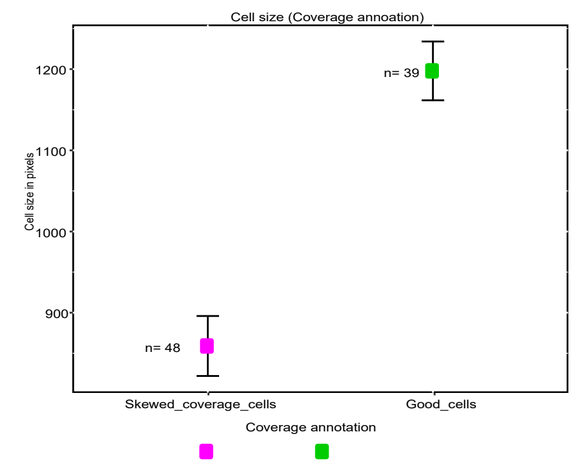












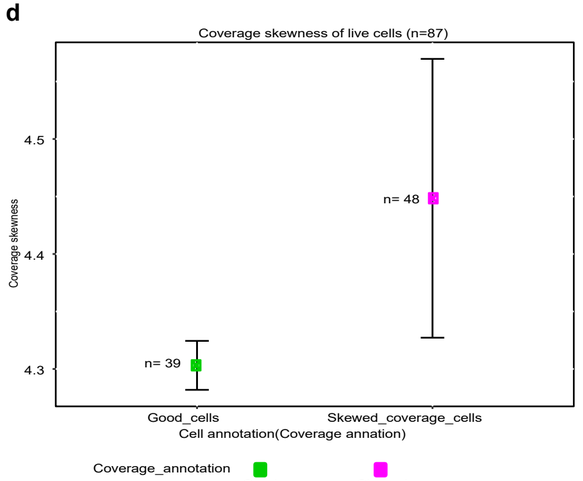


Figure 5: Effect of skewed cells on downstream analysis

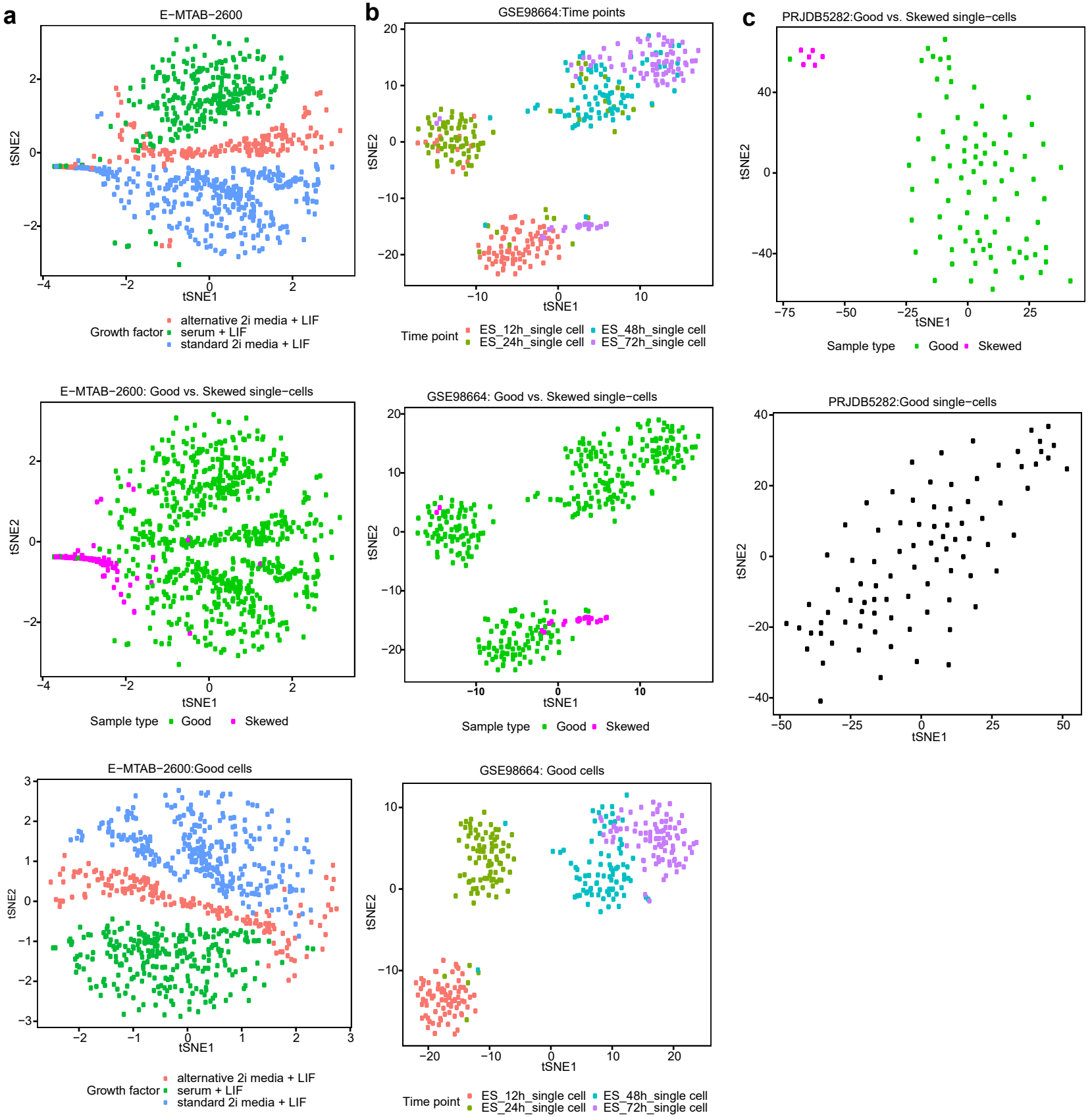


Figure 6: Ratio of intergenic expression and annotation of the reads

Batch-matched human and mouse cells		
scRNA-seq library protocol	**Species**	**Cell type**
CEL-seq	Mouse	mES
CEL-seq2	Mouse	Fibroblast
C1 CAGE	Mouse	mES
C1 single-cell mRNA-seq	Mouse	CD4 T cell & Hematopoietic
Quartz-seq	Mouse	mES
RamDA-seq	Mouse	mES
SMARTer	Mouse	mES, CD4 T cell, Fibroblast & Hematopoietic
Smart-seq	Mouse	CD4 T cell & Fibroblast
Smart-seq2	Mouse	Fibroblast
STRT	Mouse	mES
SUPer-seq	Mouse	mES
SMARTer	Human	Embryo
Smart-seq2	Human	Embryo
STRT	Human	Embryo
Un-Batch-matched human and mouse cells		
Chromium 5' end	Mouse	PBMC
TruSeq	Mouse	Adipocyte
Chromium 5' end	Human	PBMC
Drop-seq	Human	HEK & 3T3
Chromium 3' end	Human	MCF10A