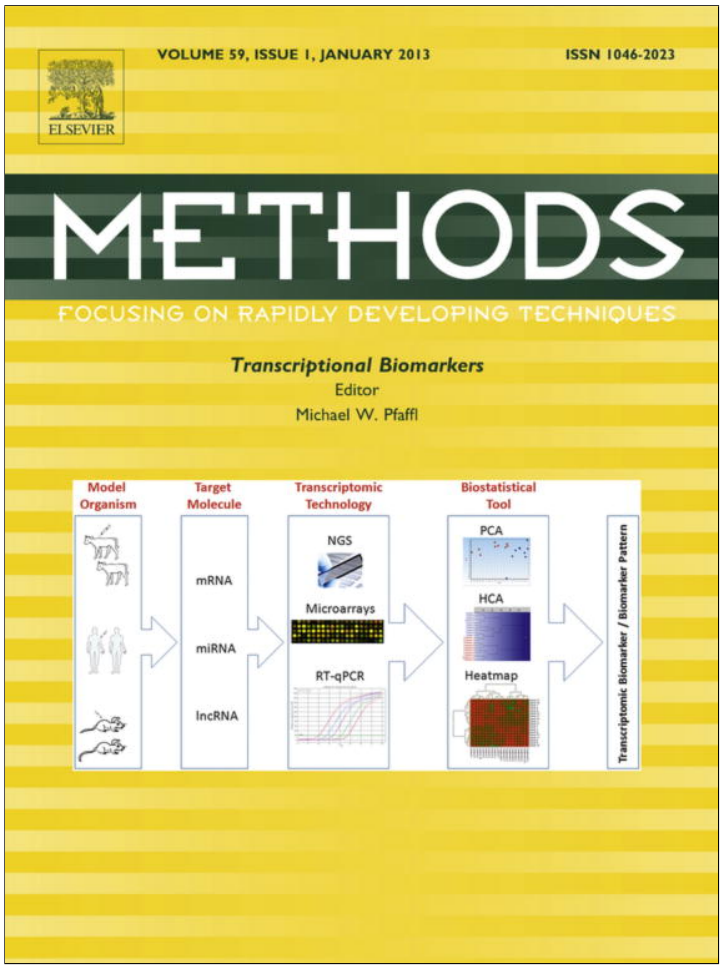


Methods 59 (2013) 80–88

Contents lists available at SciVerse ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

RT-qPCR work-flow for single-cell data analysis

Anders Ståhlberg [a,b,*], Vendula Rusnakova [c], Amin Forootan [b,d], Miroslava Anderova [e], Mikael Kubista [a,c,*]

[a] TATAA Biocenter, Gothenburg, Sweden
[b] Sahlgrenska Cancer Center, Department of Pathology, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden
[c] Institute of Biotechnology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[d] MultiD Analyses AB, Gothenburg, Sweden
[e] Department of Cellular Neurophysiology, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague, Czech Republic

ARTICLE INFO

Article history:
Available online 25 September 2012

Communicated by Michael W. Pfaffl

Keywords:
RT-qPCR
Single-cell data analysis
Single-cell biology
Data pre-processing
Missing data
Gene expression profiling

ABSTRACT

Individual cells represent the basic unit in tissues and organisms and are in many aspects unique in their properties. The introduction of new and sensitive techniques to study single-cells opens up new avenues to understand fundamental biological processes. Well established statistical tools and recommendations exist for gene expression data based on traditional cell population measurements. However, these work-flows are not suitable, and some steps are even inappropriate, to apply on single-cell data. Here, we present a simple and practical workflow for preprocessing of single-cell data generated by reverse transcription quantitative real-time PCR. The approach is demonstrated on a data set based on profiling of 41 genes in 303 single-cells. For some pre-processing steps we present options and also recommendations. In particular, we demonstrate and discuss different strategies for handling missing data and scaling data for downstream multivariate analysis. The aim of this workflow is provide guide to the rapidly growing community studying single-cells by means of reverse transcription quantitative real-time PCR profiling.

1. Introduction

In translational molecular research tissue heterogeneity is major complication. Tissues consist of several cell types that respond differently to stimuli. When studying the effects of environmental changes or responses to drugs only some of the cells respond and they may respond differently. Non-responsive cells only confound the measured signal and obscure analysis. The introduction of single-cell analysis has opened up for new possibilities to study tissue heterogeneity by detecting differences even among seemingly identical cells. Several strategies for single-cell analysis have been described and used to study various experimental systems [1–3]. Today, single-cell gene expression profiling using reverse transcription quantitative real-time PCR (RT-qPCR) is the most commonly used method to study individual cells. Single-cell RT-qPCR has been applied to many cell types including neurons [4], astrocytes [5,6], embryonic stem cells [7–9] and beta-cells [10].

Cell collection can be handled in high throughput using fluorescence activated cell sorting (FACS). Specific cells from body fluids and in vitro cultures can be enriched on the basis of surface markers using FACS. Individual cells can also be generated from most tissues by careful dissociation and then collected by FACS, but the context from which the cell is taken is lost during preparation [5,6]. Other means to extract individual cells are microaspiration [4,10] and laser capture microdissection [11,12]. Single-cells are then lysed and if possible no further purification or washing is performed. Purification-free lysis minimizes RNA losses [13,14]. Lysis is followed by reverse transcription, pre-amplification and finally qPCR [9,15–17]. If fewer than ten genes are analyzed and they are reasonably high expressed pre-amplification may not be needed [4–7]. Single-cell RT-qPCR measurements should be performed to the highest possible extent according to the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines [18].

When designing experiments the studied effect should be maximized relative to the confounding variation. Confounding variation has two main contributions: (1) Intersubject variation, caused by the natural biological heterogeneity among the studied subjects that give rise to different gene expression levels; (2) Technical variation, introduced by imprecision in the processing of the samples, comprising the steps of sampling, transport, storage, extraction, reverse transcription, pre-amplification, and qPCR. The confounding variation is reduced by using appropriate controls and references and by performing biological and technical replicates. The measured cycle of quantification (Cq) values are then

* Corresponding authors. Addresses: Sahlgrenska Cancer Center, Department of Pathology, Sahlgrenska Academy at University of Gothenburg, Box 425, 40530 Gothenburg, Sweden (A. Ståhlberg); TATAA Biocenter AB, Odinsgatan 28, 411 03 Gothenburg, Sweden (M. Kubista).

E-mail addresses: anders.stahlberg@neuro.gu.se (A. Ståhlberg), mikael.kubista@tataa.com (M. Kubista).

1046-2023/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2012.09.007

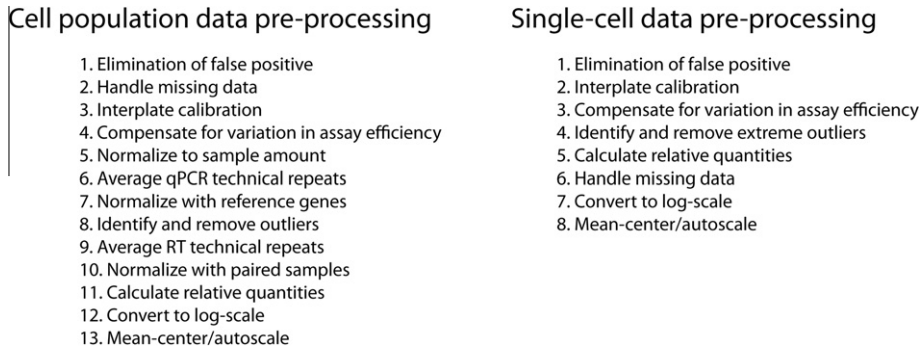


Fig. 1. Workflow for pre-processing of RT-qPCR data at cell population and single-cell level.

pre-processed, taking advantage of the controls and references used, to remove confounding variation ending up with as accurate Cq-values as possible for statistical analysis to extract biologically relevant information. Fig. 1 left shows the general qPCR data pre-processing workflow to reduce confounding variation and prepare data for statistical analysis for traditional samples based on large number of cells. The workflow lists the steps that may be relevant in appropriate order. In practice all pre-processing steps are not needed, since some steps cancel the effect of other steps. Some pre-processing steps considered routine when analyzing traditional samples are not only unsuitable but also inappropriate at single-cell level, since they would increase confounding variation. Other steps, hardly significant in cell population analysis are critical when analyzing individual cells. Here, we describe the steps relevant when pre-processing single-cell RT-qPCR expression data and present a convenient and robust workflow (Fig. 1, right).

2. Description of method

2.1. Data set

A previously unpublished single-cell data set is used to illustrate the single-cell data analysis workflow. Expression of 41 genes was measured in 303 single astrocytes collected by FACS from mouse brains before (day 0) and after (day 3, 7 and 14) induced ischemia, using the GFAP/EGFP mouse model described elsewhere [6]. The cells were lysed, reverse transcribed, pre-amplified and analyzed with qPCR using the BioMark qPCR platform (Fluidigm) as described [5,13]. The detailed experimental protocol will be published elsewhere. The applied data set is representative for single-cells analyzed by RT-qPCR with respect to distribution of transcripts among cells, number of positive cells, and gene expression levels [5–10]. The pre-processing steps and the workflow is illustrated using GenEx (ver. 5.3, MultiD), but in principle any statistical software can be used.

2.2. Data arrangement

In most experiments groups are compared. Most common is to arrange data with Cq-values (the explained variables) in columns headed with experimental group labels (the explanatory variables). This layout provides easy overview of data, and basic statistics such as means and standard deviations (SD) are easily calculated (Fig. 2A). However, this arrangement of data is not practical for more advanced studies involving more than one factor, multiple markers, replicate measurement, multiplate measurements, etc. A more flexible layout is to arrange samples in rows and all variables in columns (Fig. 2B). Today, this is standard arrangement of data in most statistical software. The format is readily generalized to any number of markers and additional columns and rows can be added that specify the experimental design. In GenEx, the explanatory variables are referred to as classification columns and classification rows and have labels starting with #. In the example shown in Fig. 2C, #Repeat indexes qPCR technical replicates (samples with the same index are technical replicates on the qPCR level). These are expected to be highly similar and shall be averaged during data pre-processing. #Group indexes treatment groups that eventually shall be compared using a statistical test. Finally, the study is paired, meaning that each subject received both treatments and a sample was collected after each treatment. Paired study designs are more powerful, because the pairing eliminates much of the systematic subject variation between the compared groups.

2.3. Elimination of false positives

Amplification curves with atypical shapes are usually not processed correctly. Aberrant amplification curves also indicate sample specific problem, such as enzymatic inhibition, and should be removed from further analysis or reanalyzed. The quality of amplification curves is usually performed by visual inspection. For high throughput qPCR analysis, manual inspection is tedious and some automatic approaches based on kinetic analysis that indicate or remove suspicious data are available, but their reliability under various conditions still remains to be validated (www.labonnet.com and www.azurepcr.com) [19].

If reporter dyes are used and melting curve analysis has been performed it should be used to validate that amplification is specific. If aberrant PCR products are formed in addition to the expected product, data should be interpreted qualitatively only, since the Cq-value cannot be trusted. When probes are used melting curve analysis cannot be performed and any non-specific PCR products, which can influence the measured Cq-values by competing for reagents, will go unnoticed. In principle, a non-specific dye can be added into the probe based reaction to monitor formation of aberrant PCR products by melting curve analysis in a separate detection channel [20]. Tables S1A and B shows the data set before and after validation based on melting curve analysis.

2.4. Interplate calibration

Single-cell gene expression profiling experiments tend to be large scale because of the underlying lognormal variation of transcripts among the individual cells, which requires large number of cells to be analyzed to reach statistical significance [5,10,21]. Typically, 50 cells need to be profiled at each condition, and with current workflow 25–100 different transcripts are readily measured per cell. This leads to many thousands of reactions to be analyzed. Even with high throughput platforms multiple plates have to be run. In the processing of the measured raw data the qPCR

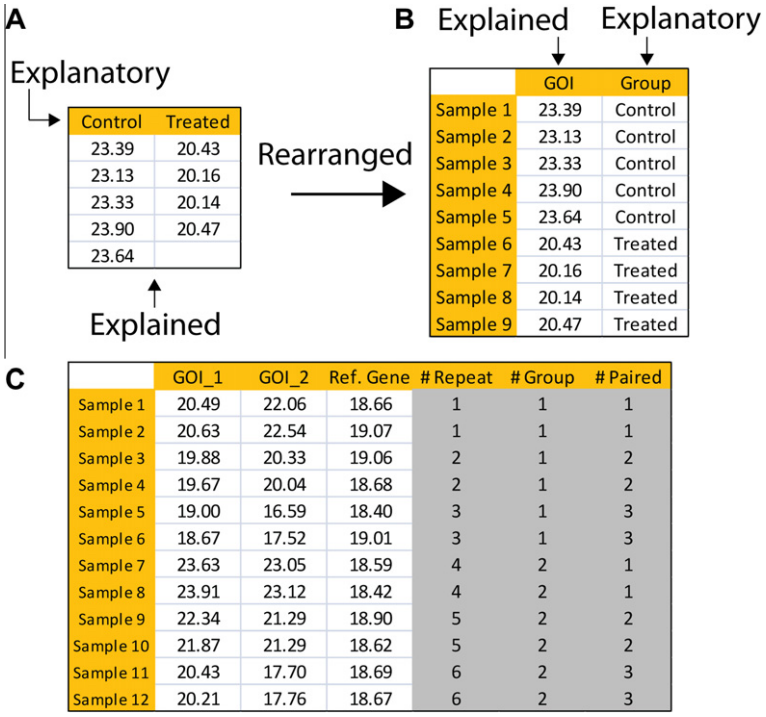


Fig. 2. Data arrangement. (A) Classical data arrangement with Cq-values (the explained variables) in columns headed with the experimental group labels (the explanatory variables). (B) Data arrangement with explained and explanatory variables in different columns is preferred for more advanced experimental designs. (C) Example of data arrangement with one column identifying samples, three columns with explanatory variables (# Repeat, # Group and # Paired) and three columns with explained variables (GOI_1, GOI_2 and Ref. Gene). GOI, Gene of interest; Ref. Gene, Reference gene.

instrument software subtracts baseline and reads out Cq-values as the crossing points of the amplification curves at threshold level or by calculating the maximum of the second derivative. The baseline corrections and the Cq estimations are performed independently for each run, which may introduce systematic variation. This bias is assay independent and can be compensated for running an interplate calibrator, which is a common sample assayed on all plates. Bias can also be introduced by variable assay performance over time. This can be differences in primer, probe, and mastermix batches or external factors such as the time to prepare the experiment. Table S1C shows the data set after interplate calibration. Indexes in the column labeled #plate indicates the independent qPCR runs and #IPC indicates the interplate calibrator. Global interplate calibration was applied to our data set. The average Cq-value of all genes for respective IPC ($Cq_{IPCAverage}$) was used to mean-center data between different runs using the equation:

$$Cq_{GOI}^{Calibtrated} = Cq_{GOI} - \frac{1}{m}\sum_{j=1}^{m}\left(Cq_{IPCAverage} - \frac{1}{n}\sum_{i=1}^{n}Cq_{IPCAverage}\right)$$

where $Cq_{GOI}^{Calibrated}$ and Cq_{GOI} are Cq-values after and before interplate calibration, respectively. m is the number of interplate calibrators in run m (qPCR replicates), and n is the total number of interplate calibrators ($m = 1$ and $n = 8$ in our data set, n = number of qPCR runs $*$ number of qPCR replicates). Calibrators can also be used to compensate for technical bias introduced during lysis, RT and pre-amplification [14,15,22,23].

2.5. Assay efficiency correction

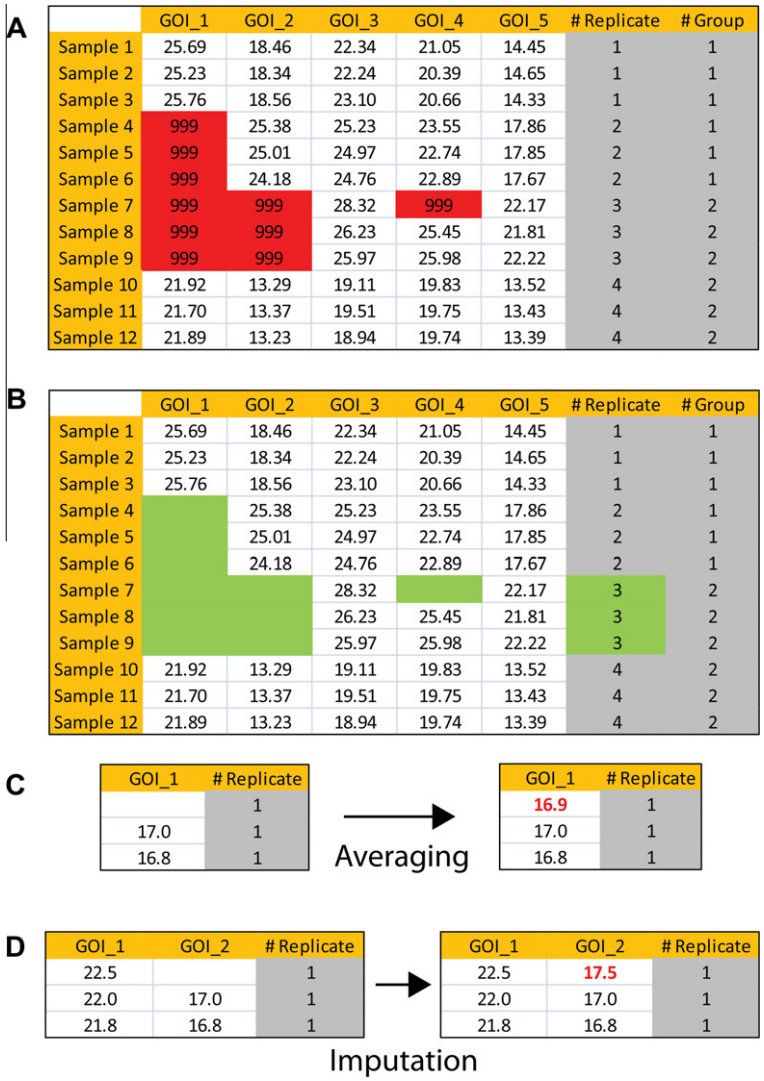
PCR efficiency is assay dependent and influenced by factors such as amplicon length and sequence, mastermix composition, and temperature profile. PCR efficiencies can be estimated from standard curves [18,24] and used to correct the measured Cq-values using the equation:

$$Cq_{E=100\%} = Cq_{E}\frac{\log(1+E)}{\log(2)}$$

where Cq_E is the uncorrected Cq value and E ($0 \leqslant E \leqslant 1$) is the PCR efficiency. The importance of the correction depends on application. Most single-cell data are analyzed without additional normalization and autoscaled by genes (see below). Data scaled this way are independent of PCR efficiency and therefore not affected by the correction. In practice, there may be small effect of PCR efficiency introduced by the handling of off-scale and missing data (see below), since it influences which cells are above the off-scale threshold Cq. Figure S1 shows that the effect of PCR efficiency correction in a Principal Component Analysis (PCA) plot is negligible. Figure S2 shows how the calculated distribution of transcripts among cells is affected by the PCR efficiency correction. Comparing mean expression levels between groups, using e.g. *t*-test or non-parametric tests, the calculated p-values are not affected by the PCR efficiency correction, while the difference between the mean expressions is affected. Assuming 100% PCR efficiency when the true efficiency is lower also underestimates the relative quantities (RQ) and the variability in transcript levels between individual cells. Differences in pre-amplification efficiencies between genes affect the data in a similar way, i.e., negligible effect on analyses performed with autoscaled data and on calculated *p*-values. No correction for any variation in PCR efficiency among genes during pre-amplification is performed here.

2.6. Missing and off scale data

Analyzing small samples there will be missing data (reactions that do not give rise to any Cq-value) and off-scale data (reactions that give Cq-values too high to be trusted). How missing data appear in the data file depends on the qPCR instrument used. Some instruments leave them blank, other indicate them with "NAN" (= Not A Number), while some assign an extreme value to them,



	GOI_1	GOI_2	GOI_3	GOI_4	GOI_5	# Replicate	# Group
Sample 1	25.69	18.46	22.34	21.05	14.45	1	1
Sample 2	25.23	18.34	22.24	20.39	14.65	1	1
Sample 3	25.76	18.56	23.10	20.66	14.33	1	1
Sample 4	999	25.38	25.23	23.55	17.86	2	1
Sample 5	999	25.01	24.97	22.74	17.85	2	1
Sample 6	999	24.18	24.76	22.89	17.67	2	1
Sample 7	999	999	28.32	999	22.17	3	2
Sample 8	999	999	26.23	25.45	21.81	3	2
Sample 9	999	999	25.97	25.98	22.22	3	2
Sample 10	21.92	13.29	19.11	19.83	13.52	4	2
Sample 11	21.70	13.37	19.51	19.75	13.43	4	2
Sample 12	21.89	13.23	18.94	19.74	13.39	4	2

	GOI_1	GOI_2	GOI_3	GOI_4	GOI_5	# Replicate	# Group
Sample 1	25.69	18.46	22.34	21.05	14.45	1	1
Sample 2	25.23	18.34	22.24	20.39	14.65	1	1
Sample 3	25.76	18.56	23.10	20.66	14.33	1	1
Sample 4		25.38	25.23	23.55	17.86	2	1
Sample 5		25.01	24.97	22.74	17.85	2	1
Sample 6		24.18	24.76	22.89	17.67	2	1
Sample 7			28.32		22.17	3	2
Sample 8			26.23	25.45	21.81	3	2
Sample 9			25.97	25.98	22.22	3	2
Sample 10	21.92	13.29	19.11	19.83	13.52	4	2
Sample 11	21.70	13.37	19.51	19.75	13.43	4	2
Sample 12	21.89	13.23	18.94	19.74	13.39	4	2

GOI_1	# Replicate
	1
17.0	1
16.8	1

GOI_1	# Replicate
16.9	1
17.0	1
16.8	1

GOI_1	GOI_2	# Replicate
22.5		1
22.0	17.0	1
21.8	16.8	1

GOI_1	GOI_2	# Replicate
22.5	17.5	1
22.0	17.0	1
21.8	16.8	1

Fig. 3. Missing data. (A) Data can be missing due to technical failure or to too low gene expression level. (B) Any extreme Cq-values (e.g. 999) reported by the instrument software should be eliminated. (C and D) Missing data can be replaced using information from technical replicates, either by the average of the successful measurements or by imputation, which also considers systematic differences between replicate samples based on information from other genes.

such as 999. In particular, the latter has to be amended in the pre-processing of qPCR data, since an extreme value will have profound influence on any downstream parametric statistical analysis and the conclusions reached will be dominated by the extreme values rather than on those measurements that produce reliable data. In any workflow extreme Cq-values assigned by the instrument software should be removed or replaced with "NAN" (Fig. 3A).

Before handling missing data it helps realizing there are two distinctly different reasons a reaction does not produce a Cq-value: (1) the reaction chamber contained template molecules but the PCR failed (2) the reaction did not produce any product because there was no template in that particular reaction chamber. These two cases should be handled differently. If technical replicates are available (Fig. 3B), missing data due to reaction failure can be restored based on replicate information. The simplest approach is to restore them by the average of the successfully recorded replicates (Fig. 3C). A more advanced and accurate approach is imputation. Imputation accounts also for any systematic differences between replicates based on information from other assays (Fig. 3D). Imputation shall only be applied to sets that have common replicates. Typically, imputation is used for technical replicates; if a biological replicate is missing it is usually better to exclude it. If several levels of technical replicates are available, imputation can be repeated at each level. For single-cell expression data technical replicates are usually not available, since it is more cost efficient to analyze more cells than to collect replicate data. Should technical (qPCR) replicates be available they are averaged at this stage of pre-processing.

Very high Cq-values reported by the instrument software are often not reliable even when the correct PCR product is formed, since they correspond to very few molecules or even fractions of a molecule and have become large due to various problems such as failed amplification of single molecule targets in the initial cycles, delayed amplification due to competing reactions, or inhibition. They may also be false positives (primer-dimers and other aberrant products). These Cq-values are not useful and should be discarded. A pragmatic approach is to delete all Cq-values above a certain threshold. In principle, Cq-values higher than the Cq-value expected for a reaction starting with a single template molecule ($Cq_{N=1}$) should not be trusted. $Cq_{N=1}$ depends on instrument, PCR efficiency, applied detection chemistry, and instrument settings. On classical microtiter plate based instruments $Cq_{N=1}$ is usually around 35–37 cycles. On high throughput platforms using smaller reaction volumes and more sensitive

optics $Cq_{N=1}$ is some 10 cycles lower. $Cq_{N=1}$ is assay dependent and can in principle be determined as the intercept of a qPCR standard curve with known concentrations. However, this is not practical, since it requires preparing purified starting material (preferably cDNA) of known concentration for every target and should be run in a matrix similar to that of the single-cell experiments. Pragmatic is to delete all Cq-values above a certain threshold, Cq_{Cutoff}, which is the same for all the assays. A reasonable Cq_{Cutoff} can be chosen by inspecting control charts (Fig. 4). The first control chart shows the variation of Cq-values across samples for the different assays (Fig. 4A). In this chart cells that have overall very few transcripts (all Cq-values are high) are identified (Table S1D) and may be considered failed or at least anomalous. Furthermore, poor performing assays are identified by lines that never or rarely reach below Cq_{Cutoff}. A second control chart presents the measured Cq-values of each assay in a box and whiskers plot (Fig. 4B). Box and whiskers plots are expected to be symmetric because of the underlying lognormal distribution of transcripts among individual cells. Asymmetry may indicate that the number of copies of the particular transcript is too low to be reliably detected giving rise to truncation at high Cq. However, distributions of transcripts can have many features and genes should in general not be eliminated due to their expression characteristics. Tentative outliers are identified using standard statistical tests and are indicated with symbols. A third control chart shows the fractions of missing Cq-values, off scale Cq-values ($Cq > Cq_{Cutoff}$) and valid Cq-values ($Cq \leqslant Cq_{Cutoff}$) for each gene (Fig. 4C). This plot identifies assays that are not contributing with meaningful information. The control charts can be calculated and inspected for different Cq_{Cutoff}-values, guiding the selection of an appropriate Cq_{Cutoff} for further analysis.

Most single-cell expression data analyses can be performed directly on Cq-values, but results become more intuitive and suited for visualization if expression values are converted to relative copy numbers. RQ can be calculated assuming $Cq_{cutoff} = Cq_{N=1}$:

$$\text{Relative quantities of cDNA molecules (RQ)} = 2^{Cq_{cuttoff} - Cq}$$

However, data should be expressed in $\log_2$-scale, since the underlying distribution of transcripts among cells is lognormal. A practical approach is to set all missing data to -1 in $\log_2$-scale, corresponding to 0.5 molecule in RQ. For the data in our example we applied $Cq_{Cutoff} = 27$ (Table S1E) and converted data to RQ using the equation above (Table S1F). Remaining missing data were assigned RQ = 0.5 (Cq Off scale = 27 + 1, Table 1F). The rational of assigning $Cq_{Cutoff} + 1$ to the missing data is that Cq_{Cutoff} is global representation of $Cq_{N=1}$, which is the Cq expected for a single template molecule. Hence, $Cq_{Cutoff} + 1$ represents a concentration lower than a single molecule, in fact, half that concentration. Of course, we cannot have 0.5 molecules in a test tube, but due to sampling ambiguity we cannot conclude a test sample is negative when we obtain a negative PCR analyzing an aliquot. We only know that the sample contains fewer molecules than we are able to detect with the current protocol. Even with a perfect PCR sampling ambiguity, which can be modeled by the Poisson distribution, results in a limit of detection (LOD) of 3 molecules at 95% confidence, i.e., the volume fraction analyzed shall in average contain 3 molecules for a reaction to give positive PCR in 95% of the cases [18]. For real samples LOD is typically higher because of losses in sample processing, RT and imperfect PCR. Analyzing experimental data, a more conservative approach is to set Cq_{Cutoff} to a lower value (25–26 cycles on a high throughput instrument) with a larger offset to the missing data. For example, assigning $Cq_{Cutoff} + 4$ to the missing data is equivalent to assigning a concentration to those samples that is 1/16 of LOD. The effect of the offset is to weight the importance of not detecting a transcript in a cell. We have found that a good strategy is to use a small offset when analyzing homogeneous cohort of cells that are expected to express the same genes, and

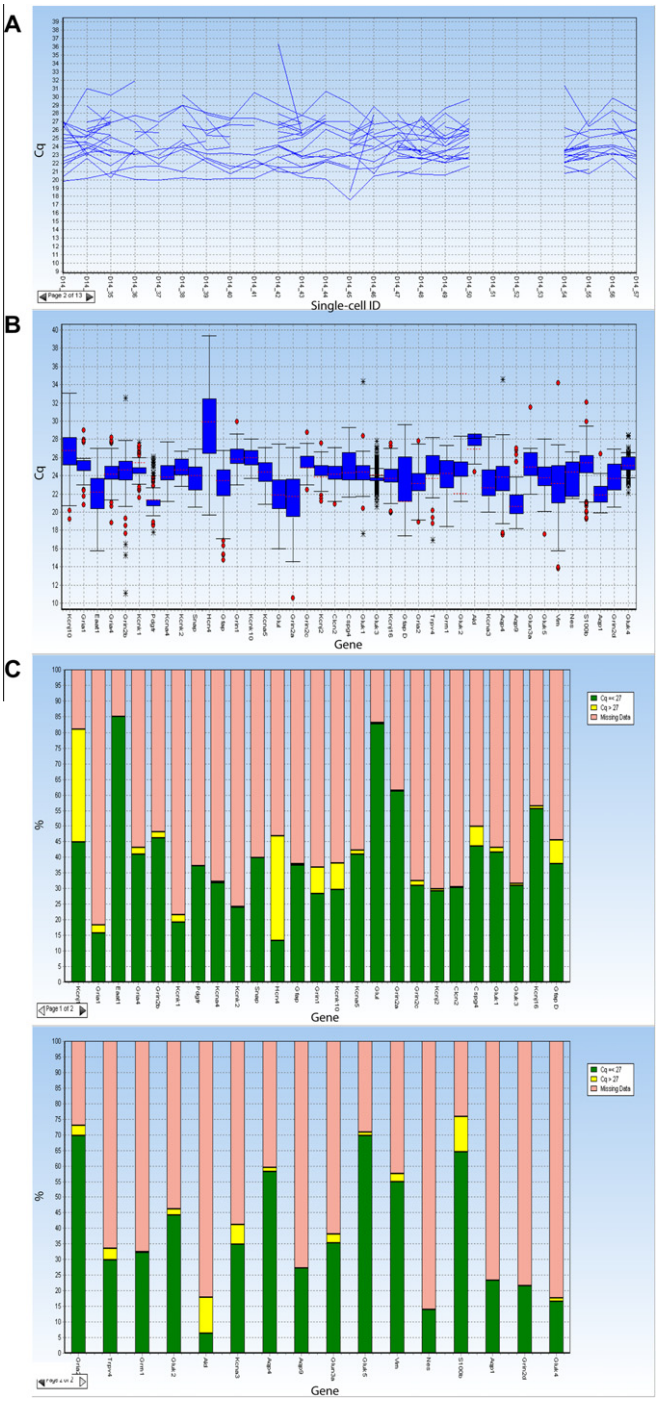


Fig. 4. Quality control charts. (A) Variation of Cq-values across samples for the different assays. In this chart anomalous single-cells that generated few reliable data (high rate of missing data and high Cq-values for all genes) are identified. Here, two cells (D14_51 and D14_53) showed no expression what so ever and should be eliminated from further analysis. (B) Box and whisker plot provides an overview of the spread of the genes' expressions. Potential outliers are indicated with circles and extreme outliers with stars. Potential and extreme outliers have expression levels outside 1.5∗IQR and 3∗IQR, respectively (IQR = Quartile 3–1). Quartile 1 and 3 represent the bottom and top of the box and are the 25th and 75th percentile of the gene expression values. Box and whisker plots with many potential outliers indicate genes (*Kcnk1*, *Gluk3* and *Gluk4*) with skewed distributions of transcripts among the individual cells. (C) Gene quality graph. The percentages of single-cells with Cq below and above Cq_{cutoff} and with missing data are shown. Genes with many Cq-values larger than Cq_{cutoff} (*Kcnj10* and *Hcn4*) have either very low expression and/or the assays are performing poorly.

a large offset when objective is to distinguish between cell types that exclusively express some markers. Figure S3 shows the effect of changing $Cq_{cutoff} = 27$ to $Cq_{cutoff} = 25$ on PCA, and Figure S4 shows the effect of offset when assigning different Cq-values to the missing data. For this particular data set the Cq_{Cutoff}-value and the offset have insignificant effect on the analysis result. This is usually the case. Still, we recommend users to test the effects of Cq_{Cutoff} and offset when analyzing new data to validate the robustness of their handling of missing data. One should also be cautious when key classification genes are expressed at low level with high Cq-values. It is then important to set Cq_{Cutoff} at a value that distinguishes between positive and negative cells with respect to this marker. In some cases it may be necessary to treat key classifiers separately, using an assay specific $Cq_{cutoff.}$ Genes that may be biased due to the choice of Cq_{cutoff} are identified in the control charts (Figs. 4B and C). The rationale of handling missing data after, instead of before, the transformation to RQ is that RQ of the Cq_{cutoff}-value (and not of $Cq_{cutoff} + 1$) shall be assigned arbitrary expression of 1 (zero in log_2-scale). The RQs after handling missing data and RQs in log_2-scale are shown in Tables S1G and S1H, respectively. Conversion of relative quantities to the number of cDNA molecules or to the number of mRNA molecules requires calibration with standards [9,14,25]. Missing data due to few target molecules in the reaction vessel is in single-cell studies with good assays more common than reaction failure. Missing data can therefore be replaced with Cq_{cutoff} + offset instead of imputation or averaging even when qPCR replicates are available. In practice the two options to handle missing data will produce very similar results.

2.7. Basic statistics and distributions

To get an overview of the data we calculate some basic statics for all the genes studied, including the number (or fraction) of cells expressing each gene, and the mean and standard deviation (SD) of all the genes' expressions. Mean and SD are calculated on data in logarithmic scale, since the underlying distribution is lognormal, but can be converted to linear scale for presentation. The fraction of cells that express a particular gene correlates well with the gene's mean expression (Fig. 5A and B, Spearman correlation = 0.78, $P < 0.01$). The mean expression is calculated as the arithmetic average of the relative quantities expressed in logarithmic scale (log_2). This is equivalent to calculating the geometric average of the data in linear scale, and reflects the medium number of transcripts per cell, which is the number of transcripts expected in the typical cell of the population. From Fig. 5C we calculated that the median number of molecules per reaction well was 3.1 ($RQ_{median} = 3.1$ assuming $Cq_{cutoff} = Cq_{N=1}$). To convert this to the median number of mRNA molecules per cell one must consider RT, pre-amplification, and qPCR efficiencies as well as all dilution steps. The SD of the measured Cq-values reflects variability, which is dominated by the heterogeneity of the cells. Usually SD scales with the average expression of the gene (Fig. 5D-E, spearman correlation = 0.98, $P < 0.01$). This correlation is quite linear, which is equivalent to the relative standard deviation (SD/mean, also known as the coefficient of variation, CV) being independent of concentration. For our data CV was roughly 100% (mean and SD are equal). A transcript that has deviant CV should be suspected to have differently regulated transcription or anomalous stability.

It is good idea to visualize the distributions of the different transcripts among the cells either as frequency histograms or violin plots (Fig. 6). From experience we know that under most conditions the distribution of a transcript among homogenous cells can be fitted with a lognormal distribution [5,8–10,14]. Recent theoretical considerations suggest that a Poisson-Beta distribution may be more appropriate to model single-cell expression data [26]. However, the two distributions have similar features and are both suitable for analysis of single-cell data with parametric statistics. Deviant distributions, in particular highly skewed and bimodal (two maxima) distributions, indicate that several cell types are present and/or that the cells respond differently to stimuli [5,10]. In our example the distribution of *Vim* transcripts among the cells changes over time. In the reference material (day 0) most cells had low/no *Vim* expression. Upon injury *Vim* was upregulated and reached maximum expression at day 7. At day 14 *Vim* was downregulated again, but did not return all the way to the basal expression level of the healthy brain.

2.8. Mean-centering and auto-scaling

Expression data can be analyzed one gene at a time using traditional univariate statistics, such as the *t*-test, Anova and regression. These tests assume data are normal distributed, which is typically satisfied for single-cell gene expression when expressed in logarithmic scale. Univariate statistical methods can be applied directly on the measured data or on calculated logarithmic expression quantities. They usually work well on single-cell data, but suffer from ambiguity if many genes are analyzed, since the false positive rate increases and many genes will appear differentially expressed due to chance only. An alternative, or rather a complementary approach, is to analyze the data using multivariate statistical methods. These methods take the collective response of all the genes into account. Popular multivariate statistical methods suitable for qPCR expression profiling include PCA, Hierarchical clustering, and the Self organizing map [27]. When using multivariate methods we must consider the expression levels of all the genes, since if no scaling is applied the more expressed genes will have higher weights in the analysis and will dominate the result. This is usually not desired. Rather, it is preferred to give all the genes same weights in the analysis, making them equally important in the data processing. This is accomplished by autoscaling the data, which is done by calculating *z*-score for each gene by subtracting its mean expression and divide by its SD [27,28]. Thus, a *z*-score of 2 indicates that a gene in a particular sample is overexpressed by two SDs relative to its mean expression in all the samples. Autoscaling is preferred when mining single-cell qPCR expression profiling data. There are few cases, though, when autoscaling is not suitable. When the panel includes genes that are not responsive to the studied conditions and their expression among the samples show only random variation, autoscaling will only increase the noise these genes contribute to the data set. If the number of non-responsive genes is large the data quality may be compromised by autoscaling. An option is then to only mean-center the data. Mean-centering is performed by subtracting the average expression of each gene, but not dividing with the SD. Mean-centered data have zero average expression for all the genes, but their levels still vary. In practice one often starts analyzing mean-centered data, identifying the responsive genes, and then inactivating those non-responsive for further analysis, which is performed on autoscaled data. Variable selection tools such as dynamic PCA are very useful for identification of the most responsive and relevant genes. Sometimes it is also of interest to analyze only a subset for samples. If data are mean-centered or autoscaled and samples are removed from the data set, the original scaling is not valid and has to be repeated. Software such as GenEx handles this automatically, but it has to be remembered if analyzing data manually or using spreadsheet program such as Microsoft Excel. If too many data are missing one should also use mean-centering and autoscaling with care, since the offset chosen may have pronounced effect on the scaling. This can be tested by reanalyzing data using a different offset as described above. One should also test the influence of genes with low expression levels and many missing data on the multivariate analysis result by reversibly inactivating them. Still, an option, if

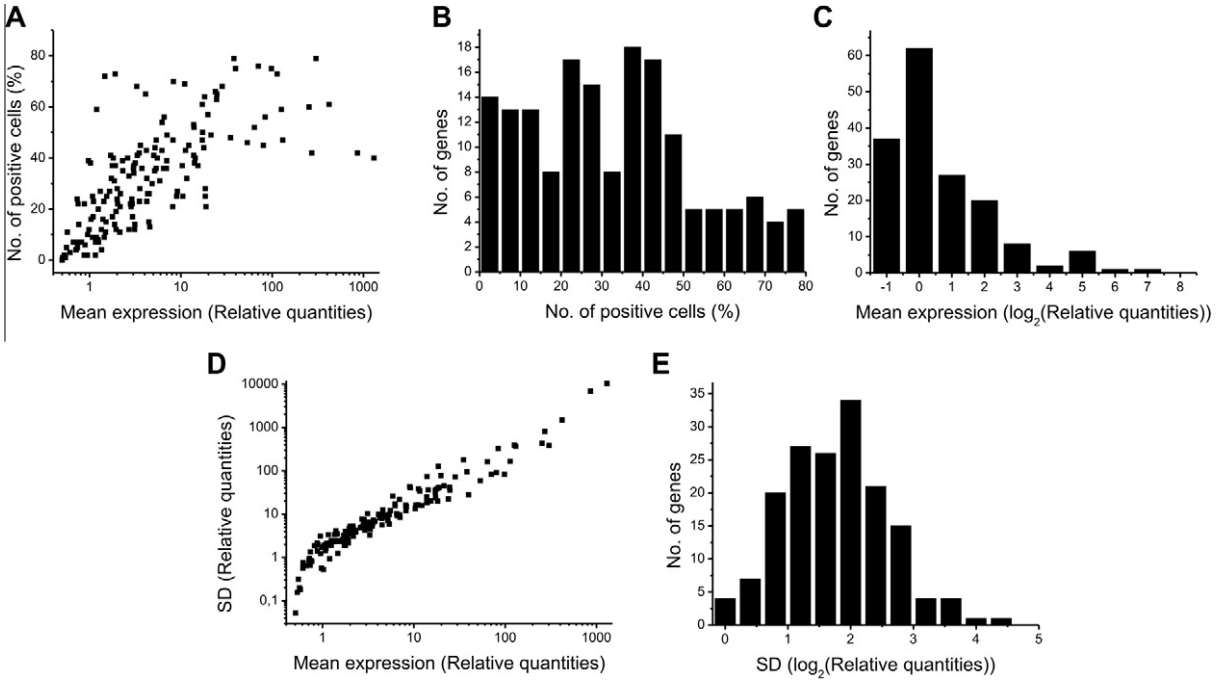


Fig. 5. Features of gene expression profiling at single-cell level. (A) Correlation between the number of cells containing a gene transcript and the mean expression of the same gene. (B) Number of genes versus percentages of positive expressing cells. (C) Diagram showing the frequency of expression levels among the genes. (D) Correlation between SD and mean expression (data in logarithmic scale). (E) Variation of expression levels among the cells presented as frequency distribution of the standard deviations of the genes' Cq-values. Genes' expressions at different time points after injury were treated as independent variables (N_{tot} = 164).

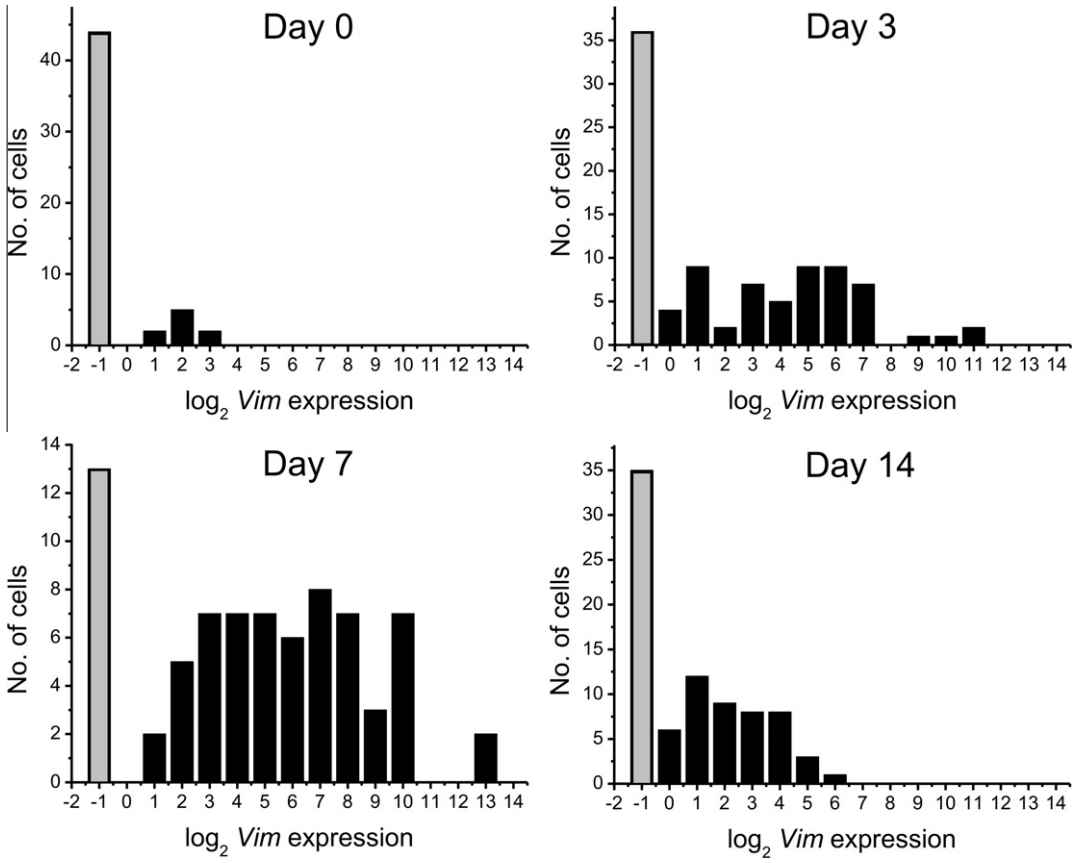


Fig. 6. Distribution of Vim transcripts among single-cells at different time points after injury. Frequency of cells with different expression levels of *Vim* before (day 0) and after injury (day 3, 7, and 14). The gray bars indicate the number of cells with no detectable expression of *Vim*.

there are many cell specific genes and the objective is to distinguish cell types, is to analyze the data in binary mode; i.e., considering only whether a gene is expressed or not in a cell. Mathematically this is equivalent to using a very large (infinite) offset when handling missing data. The offset used when handling missing data is actually setting the weight to the importance of a transcript not being present in a cell. Fig. 7 shows PCA with mean-centered (Fig. 7A) and autoscaled (Fig. 7B) data. Figure S5 shows the PCA loadings calculated with the two scaling options and the impacts of the genes. Only minor differences caused by the scaling are observed in the PCA scores and loadings for our data (Figs. 7 and S5).

Traditional autoscaling and mean-centering is applied to variables, which in qPCR are the genes, with the objective of giving them the same weight in the analyses [27,28]. For single-cell expression profiling one may also apply autoscaling or mean-centering to the samples. For each cell the average expression of all the genes is subtracted and optionally the data are divided by the standard deviation of the genes' expressions in that cell. Mean-centering the data per cell is particularly interesting, since it is equivalent to global normalization or at least truncated global normalization, since we only measure the expression of a fraction of all the genes. Such normalization will account for cell specific variations in extraction, RT yield, and pre-amplification/PCR efficiencies. Autoscaling data per cell standardizes the expression data to a common scale; for example, a cell type may be characterized by having certain genes expressed at least 2 SDs above the mean expression of all the other genes studied in that cell. Such scaling removes, for example, the influence of total expression level per cell and it allows comparison of single-cell with few cells data.

2.9. Preprocessing steps not suitable for single-cell analysis

As indicated by the workflow in Fig. 1 some steps of classical qPCR data pre-processing are not used on data from single cells. Traditional RT-qPCR data are almost always normalized to compensate for differences in the amount of sample material analyzed. Most common is to normalize with endogenous reference genes [18]. Their expression is not only expected to be proportional to the amount of material, but it also reflects any loss in the handling and processing of the samples. However, transcription in individual cells occurs in bursts and under normal conditions no gene has constant steady state level of transcripts [29]. Since the transcriptional bursts of most genes are uncorrelated one cannot use common reference genes for normalization. In principle global normalization based on the average expression of all the transcripts can be performed (or mean center data per sample, see above), but the robustness of this approach is still to be proven. Complications include the rather small number of genes typically analyzed per cell and the ambivalence in the handling of missing data. Another option may be to normalize to a highly abundant or degenerative transcript, such as the Alu repeat in human. This approach, however, is complicated by the genomic DNA background present.

The accuracy of a RT-qPCR measurement is proportional to the number of molecules analyzed and if the number becomes very small accuracy is compromised due to sampling (Poisson) ambiguity. Instead of dividing a sample in duplicates that are analyzed separately and the results averaged, it is better to measure the single whole sample with twice the transcript amount [14]. Generally all samples should be processed keeping all dilution steps, including lysis, RT, pre-amplification and qPCR, to a minimum. As a rule of thumb the sample volume analyzed in a reaction vessel by qPCR should contain a minimum of 10–25 target molecules, which will contribute with sampling noise (SD) of 0.5–0.3 cycles. Notably, instead of aliquoting a cDNA sample with few molecules for separate singleplex qPCRs, it is better to first pre-amplify the cDNA. Even though preamplification introduces some variation it is less than the sampling error introduced when aliquoting a highly diluted

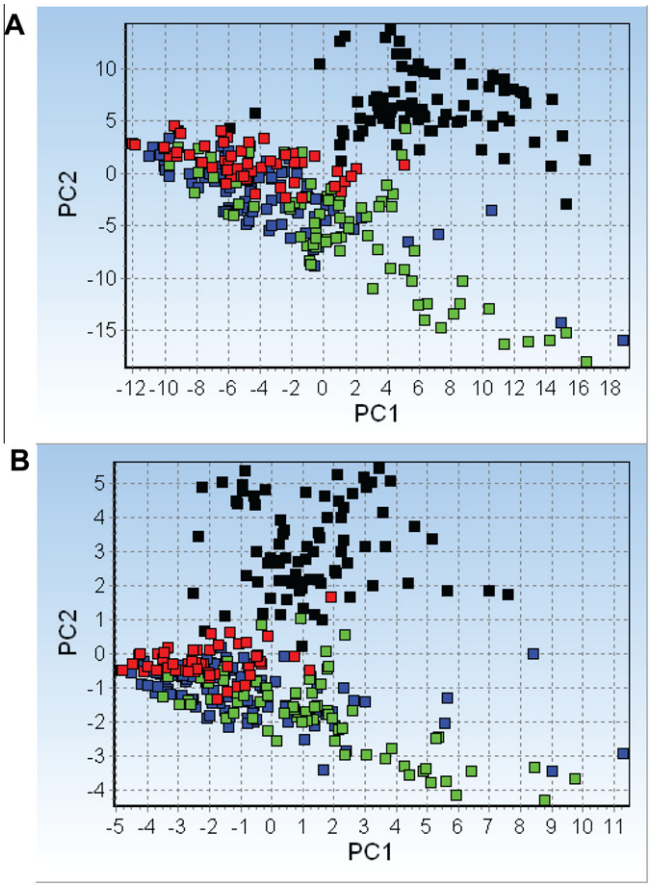


Fig. 7. Principal component analysis of single-cells based on their expression profiles. PCA based on mean-centered (A) and autoscaled (B) data. Loadings for principal component analysis are shown in Figure S5. Data points represent individual astrocytes collected from mouse brains before (day 0, red) and after (day 3, green, day 7, blue, and day 14, black) ischemia.

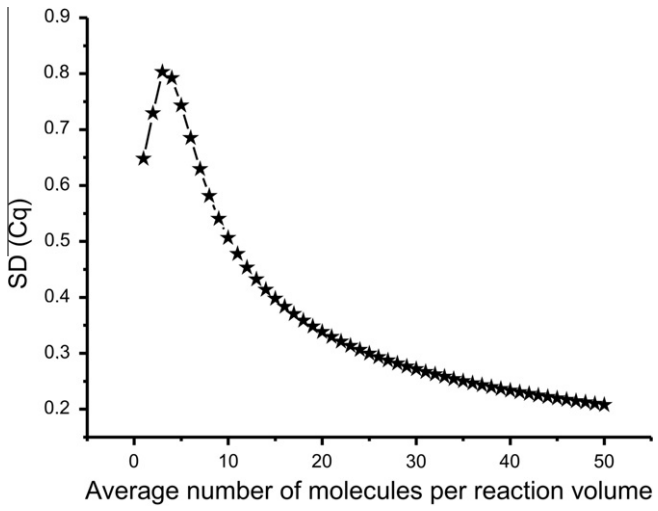


Fig. 8. Variation due to sampling ambiguity. Standard deviation of Cq-values of replicate samples caused by sampling ambiguity calculated by the Poisson distribution. The SD increases rapidly when the number of molecules decreases. The reason maximum SD is obtained at about 3 molecules is that only positive samples are considered.

sample (Fig. 8) and [15,22,23]. After pre-amplification the amount of transcripts is usually high enough to allow technical qPCR replicates to be measured, as a reassurance if a qPCR fails. From a cost-performance and statistical perspective, though, it is usually better to analyze larger number of single-cells than performing technical replicates. Of course, in the validation of assays, including controls, standard curves and LOD, technical replicates should be measured [18].

2.10. Classification algorithms

Analysis of data and data mining is usually hypothesis driven and therefore application dependent and the different methods available have been discussed elsewhere [5,27,28]. Here we only use PCA to visualize some of the important features of the pre-processing using an example data set collected on astrocytes harvested at different time points after brain injury in mice. PCA clearly reveals changes in gene expression profiles at the single-cell level occurring over time, reflecting heterogeneity and cell transformation induced by the injury. Other powerful methods to analyze and classify individual cells include hierarchical clustering, and the self organizing map [5,27,28].

3. Concluding remarks

Gene expression profiling using RT-qPCR took a major leap forward by the publication of the MIQE guidelines, which help users analyzing classical samples [18]. For single-cell profiling some aspects (e.g., reference gene normalization) of the standard workflow are not valid or applicable, while other options (e.g., mean-centering and autoscaling along samples) are almost compulsory. Here we present a complete workflow for the pre-processing of single-cell RT-qPCR data that is robust and prepares the data in a meaningful way for further analysis and mining. Using an example data set we also present several characteristics of single-cell expression data. For some pre-processing steps we present options and also recommendations based on the experience gathered so far. The aim of this workflow is to provide guidelines to the rapidly growing community studying single-cells by means of RT-qPCR profiling. To further stimulate exchange and experiences in the single-cell expression field we have made our example data available in supplement.

4. Disclosure

A.S., V.R., A.F. and M.K. declare stock ownership in TATAA Biocenter AB. A.F. and M.K. also declare stock ownership in MultiD.

Acknowledgements

This work was partly supported by grants from Assar Gabrielssons Research Foundation, Johan Jansson Foundation for Cancer Research, Socialstyrelsen, Swedish Society for Medical Research, The Swedish Research Council (A.S. 521-2011-2367), EMBO Short Term Fellowship (V.R.), ESF Functional Genomics Short Visit Grant (A.S. and V.R.), BioCARE National Strategic Research Program at University of Gothenburg, Wilhelm and Martina Lundgren Foundation for Scientific Research, the grant of Czech Ministry of Education ME10052, the research project AV0Z50520701, and GACR GA P303/10/1338.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2012.09.007.

References

[1] T. Kalisky, P. Blainey, S.R. Quake, Annu. Rev. Genet. 45 (2011) 431–435.
[2] M. Wu, A.K. Singh, Curr. Opin. Biotechnol. 23 (2012) 83–88.
[3] D.R. Larson, R.H. Singer, D. Zenklusen, Trends Cell Biol. 19 (2009) 630–637.
[4] B. Liss, O. Franz, S. Sewing, R. Bruns, H. Neuhoff, J. Roeper, EMBO J. 20 (2001) 5715–5724.
[5] A. Ståhlberg, D. Andersson, J. Aurelius, M. Faiz, M. Pekna, M. Kubista, M. Pekny, Nucleic Acids Res. 39 (2011) e24.
[6] J. Benesova, V. Rusnakova, P. Honsa, H. Pivonkova, D. Dzamba, M. Kubista, M. Anderova, PLoS One 7 (2012) e29725.
[7] A. Ståhlberg, M. Bengtsson, M. Hemberg, H. Semb, Clin. Chem. 55 (2009) 2162–2170.
[8] K.H. Narsinh, N. Sun, V. Sanchez-Freire, A.S. Lee, P. Almeida, S. Hu, T. Jan, K.D. Wilson, D. Leong, J. Rosenberg, et al., J. Clin. Invest. 121 (2011) 1217–1221.
[9] K. Norrman, A. Strömbeck, H. Semb, A. Ståhlberg, Distinct gene expression signatures in human embryonic stem cells differentiated towards definitive endoderm at single-cell level, Methods 59 (2012) 59–70.
[10] M. Bengtsson, A. Ståhlberg, P. Rorsman, M. Kubista, Genome Res. 15 (2005) 1388–1392.
[11] F. Kamme, R. Salunga, J. Yu, D.T. Tran, J. Zhu, L. Luo, A. Bittner, H.Q. Guo, N. Miller, J. Wan, et al., J. Neurosci. 23 (2003) 3607–3615.
[12] J. Gründemann, F. Schlaudraff, O. Haeckel, B. Liss, Nucleic Acids Res. 36 (2008) e38.
[13] A. Ståhlberg, M. Bengtsson, Methods 50 (2010) 282–288.
[14] M. Bengtsson, M. Hemberg, P. Rorsman, A. Ståhlberg, BMC Mol. Biol. 9 (2008) 63.
[15] B.C. Fox, A.S. Devonshire, M.O. Baradez, D. Marshall, C.A. Foy, Anal. Biochem. 472 (2012) 178–186.
[16] P. Dalerba, T. Kalisky, D. Sahoo, P.S. Rajendran, M.E. Rothenberg, A.A. Leyrat, S. Sim, J. Okamoto, D.M. Johnston, D. Qian, et al., Nat. Biotechnol. 29 (2011) 1120–1127.
[17] G. Guo, M. Huss, G.Q. Tong, C. Wang, L. Li Sun, N.D. Clarke, P. Robson, Dev. Cell 18 (2010) 675–685.
[18] S.A. Bustin, V. Benes, J.A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G.L. Shipley, J. Vandesompele, C.T. Wittwer, Clin. Chem. 55 (2009) 611–622.
[19] T. Bar, M. Kubista, A. Tichopad, Nucleic Acids Res. 40 (2012) 1395–1406.
[20] K. Lind, A. Ståhlberg, N. Zoric, M. Kubista, Biotechniques 40 (2006) 315–319.
[21] A. Ståhlberg, M. Kubista, P. Åman, Expert Rev. Mol. Diagn. 11 (2011) 735–740.
[22] A.S. Devonshire, R. Elaswarapu, C.A. Foy, BMC Geomics 12 (2011) 118.
[23] M. Reiter, B. Kirchner, H. Müller, C. Holzhauer, W. Mann, M.W. Pfaffl, Nucleic Acids Res. 39 (2011) e124.
[24] A. Ståhlberg, P. Åman, B. Ridell, P. Mostad, M. Kubista, Clin. Chem. 49 (2003) 51–59.
[25] A. Ståhlberg, M. Kubista, M. Pfaffl, Clin. Chem. 50 (2004) 1678–1680.
[26] A. Raj, C.S. Peskin, D. Tranchina, D.Y. Vargas, S. Tyagi, PLoS Biol. 4 (2006) e309.
[27] A. Bergkvist, V. Rusnakova, R. Sindelka, J.M. Garda, B. Sjögreen, D. Lindh, A. Forootan, M. Kubista, Methods 50 (2010) 323–335.
[28] A. Ståhlberg, K. Elbing, J. Andrade-Garda, B. Sjögreen, A. Forootan, M. Kubista, BMC Genomics 9 (2008) 170.
[29] A. Raj, A. van Oudenaarden, Cell 135 (2008) 216–226.