The added value of single-cell gene expression profiling

Anders Ståhlberg, Vendula Rusnakova and Mikael Kubista

Abstract

Cells are the basic unit of life and they have remarkable abilities to respond individually as well as in concert to internal and external stimuli in a specific manner. Studying complex tissues and whole organs requires understanding of cell heterogeneity and responses to stimuli at the single-cell level. In this review, we discuss the potential of single-cell gene expression profiling, focusing on data analysis and biological interpretation. We exemplify several aspects of the added value of single-cell analysis by comparing the same experimental data at both single-cell and cell population level. Data normalization and handling of missing data are two important steps in data analysis that are performed differently at single-cell level compared with cell population level. Furthermore, we discuss how single-cell gene expression data can be viewed and how subpopulations of cells can be identified and characterized.

Keywords: single-cell biology; single-cell analysis; RT–qPCR; single-cell transcriptomics; gene expression analysis

INTRODUCTION

The great value in studying cell diversity and biological processes at single-cell level has been recognized for long time, but the lack of sensitive and specific techniques to measure few molecules has been a major obstacle. During the past decade, refinement of established methods and also development of new technologies have opened up new possibilities to analyze single cells and even single molecules in highly reproducible manners. Today, most analytes including DNA, RNA and protein, as well as other biological parameters, such as cell morphology and action potential, can be measured with high precision at the single-cell level. Microscopy, fluorescence-activated cell sorting, real-time quantitative polymerase chain reaction (qPCR), mass spectroscopy, microarrays and next-generation sequencing are some commonly applied techniques to study individual cells. In general, one class of analytes is analyzed at a time with one of the methods. Recently, we developed a strategy to analyze DNAs, RNAs and proteins in the same single-cell using qPCR, opening up for multi-analyte correlation analysis in single cells [1]. For overview and comparison of different techniques, see recent reviews [2–5].

Early single-cell work focused on understanding the seemingly stochastic variations in mRNA and protein levels observed among individual cells. The variability was explained by transcriptional and translational bursts [6–8] and mathematical models, including the master equation, were developed to describe underlying biological processes [4,6]. However, other types of kinetics have been reported [9]. Mathematical modeling of transcriptional and translational kinetic at single-cell level is discussed elsewhere [4,6,9,10].

Currently, characterization of cell heterogeneity in mixed as well as in seemingly homogenous cell

Corresponding author. Anders Ståhlberg, Department of Pathology, Sahlgrenska Cancer Center, Sahlgrenska Academy at University of Gothenburg, Box 425, 40530 Gothenburg, Sweden. Tel: +46 31 7866735; E-mail: anders.stahlberg@gu.se

Anders Ståhlberg, PhD, is a principle investigator at the Sahlgrenska Cancer Center, University of Gothenburg, Gothenburg, Sweden. His primary research interest is to understand molecular mechanisms in tumor development and stem cell differentiation.
Vendula Rusnakova, MSc, is a PhD student at Charles University in Prague, Prague, Czech Republic. She is developing new approaches of single-cell gene expression profiling, including astrocyte characterization.
Mikael Kubista, PhD, is a head of the department of gene expression at the Institute of Biotechnology of the Czech Academy of Sciences, Prague, Czech Republic and the founder of TATAA Biocenter and MultiD Analyses (both Gothenburg, Sweden).
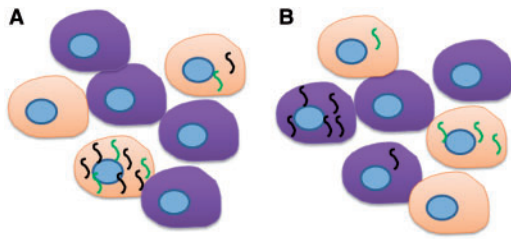
Figure I: Biomarker analysis at single-cell level. Single-cell analysis can be used to define genes with correlated expressions (**A**) or uncorrelated expressions (**B**) in specific cell types.

populations is the focus area of single-cell studies. Established cell types as well as previously unknown subpopulations have been identified and characterized in detail at single-cell level using qPCR [11–19]. Analysis at single-cell level eliminates many of the factors that confound gene expression profiling at cell population level. Correlation of transcript levels within individual cells has proven to be a most powerful identifier of cell types and distinct subpopulations of cells (Figure 1A and B) [11–13]. Another application of single-cell analysis is to define and characterize rare cell types such as stem cells and circulating tumor cells [20–22]. One challenge when studying rare cells is that a large number of cells must be screened and tested to find those cells that are relevant. Flow cytometry can handle large numbers of cells, but most other single-cell techniques do not offer the required throughput. An alternative approach is to enrich for the targeted cells before analysis, taking advantages of distinctive physical or biological properties of the rare cells [21,22]. Further development of technologies such as high-throughput droplet PCR opens for new possibilities for high-throughput screening of rare cells [23].

The workflow of single-cell analysis requires delicate sample preparation, development of sensitive assays and often sophisticated data analysis. With time, as experience is accumulating, improved and simplified protocols for single-cell analysis will become available. In this review, we discuss important aspects of data pre-processing and how single-cell gene expression data can add value to our understanding of cell diversity and expression pathways controlling biological processes. To emphasize key aspects of single-cell gene expression profiling, we here analyze an experimental data set at both single-cell and cell population level by comparing the results.

SINGLE-CELL VERSUS CELL POPULATION ANALYSIS

To illustrate different aspects of single-cell gene expression profiling, we re-analyzed a previously published data set at both single-cell and cell population level [24]. The original data set consisted of the expression profiles for 41 genes measured in 303 individual astrocytes at four time points: 0, 3, 7 and 14 days after induced brain ischemia in mice, with day 0 being control [24]. Here, we focus our analysis on a smaller number of genes selected based on their expression features rather than biological relevance. The single-cell gene expression features we report for the selected genes are not new; they have been observed before [1,11–19].

Cell population data analysis follows the Minimum Information for Publication of Quantitative Real-Time PCR Experiments guidelines [25], while single-cell data analysis is performed as described [5,24]. The workflows for the two data pre-processing schemes are shown in Figure 2 [13,24]. Two steps differ fundamentally: missing data and normalization. It should be noted that the order of the data pre-processing steps in Figure 2 can be partly swapped, since some data processing is independent of the order. Furthermore, some steps will be canceled if the required information for that particular analysis is missing.

DATA PRE-PROCESSING: MISSING DATA

Missing data are common in scientific measurements and a classical problem in statistical analysis. The underlying cause of the missing data determines how they should be handled for downstream analysis. There are two fundamentally different reasons to miss data in qPCR: (i) the qPCR failed due to a technical error and (ii) the biological sample contained fewer target molecules than the workflow with all dilution steps, losses and imperfections could detect (limit of detection). Missing data due to technical failure should be replaced using information from (preferably) technical replicates or (less preferably) from similar samples using imputation [24]. However, missing values due to too few target molecules should be replaced by zeros in linear scale and by a cutoff value in logarithmic scale. Technical failures are handled early in the data pre-processing scheme, while transcript levels below limit of detection are
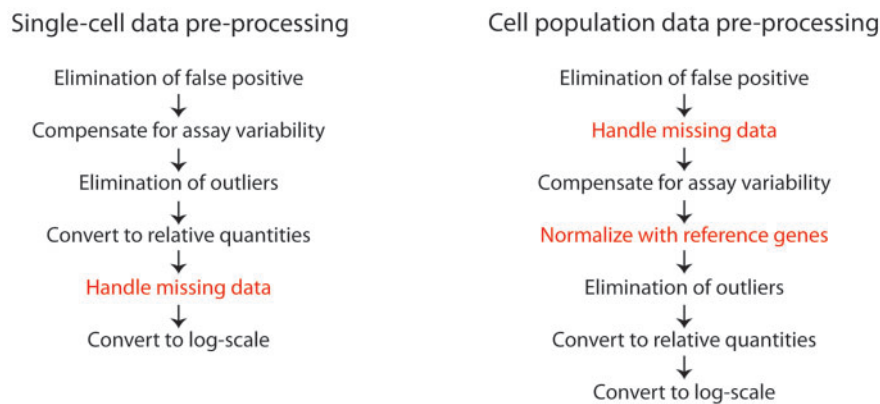
Single-cell data pre-processing

Elimination of false positive
↓
Compensate for assay variability
↓
Elimination of outliers
↓
Convert to relative quantities
↓
Handle missing data
↓
Convert to log-scale

Cell population data pre-processing

Elimination of false positive
↓
Handle missing data
↓
Compensate for assay variability
↓
Normalize with reference genes
↓
Elimination of outliers
↓
Convert to relative quantities
↓
Convert to log-scale

Figure 2: Simplified workflows for data pre-processing using qPCR. Handling of missing data and normalization are performed differently in single-cell and cell population analysis.

processed later in the workflow (Figure 2). How can one tell whether missing data are due to technical failures or expression levels below the limit of detection?

The transcriptome of typical mammalian cells consists of about 300 000 mRNA molecules represented by some 10 000 expressed genes [26]. Some genes are highly expressed (>100 transcripts per cell), but most genes have a mean expression of less than one molecule per cell. In the bacteria *Mycoplasma pneumoniae*, the mean mRNA count was determined to be as low as 0.04 molecules per cell [27]. The observed variability in expression level also scales to the average number of transcripts for respective gene [24]. If samples are analyzed in replicates, technical failures are usually readily recognized by inspecting the replicate measurements. However, single-cell profiling is usually performed without technical replicates to maximize the number of cells analyzed. In single-cell qPCR, extensively validated protocols and highly optimized assays are used, which reduces the risk for technical failures. A pragmatic approach to handle missing data in single-cell gene expression profiling is to assume that all missing data are due to too few molecules. In conclusion, high frequency of missing data is expected for most genes at single-cell level due to low mean expression levels and fluctuations over time (Figure 3).

Figure 4 shows the binary (transcript detected/not detected) expression profiles for all the genes and time points at single-cell level for the experimental astrocyte data set. The 10 selected genes had overall intermediate to high expression levels, still large amounts of missing data are observed.
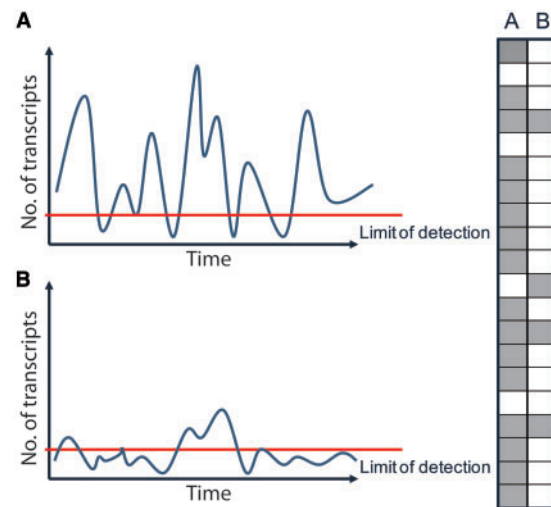
Figure 3: Missing data. Single-cell data sets have many missing data due to too few molecules in the measured sample. Total number of transcripts over time is shown for high (**A**) and low (**B**) abundant genes, respectively (left diagrams). The binary panel (right plot) shows the frequency of cells with detectable number (above limit of detection) of transcripts (gray). The limit of detection for high-performance assays is typically between one and five molecules.

DATA PRE-PROCESSING: NORMALIZATION

Comparison of classical samples containing many cells requires normalization to account for differences in the amount of material in the analyzed samples [25]. Several normalization strategies for qPCR data have been suggested. Most popular is to normalize with the expression of one or a few (two to five)
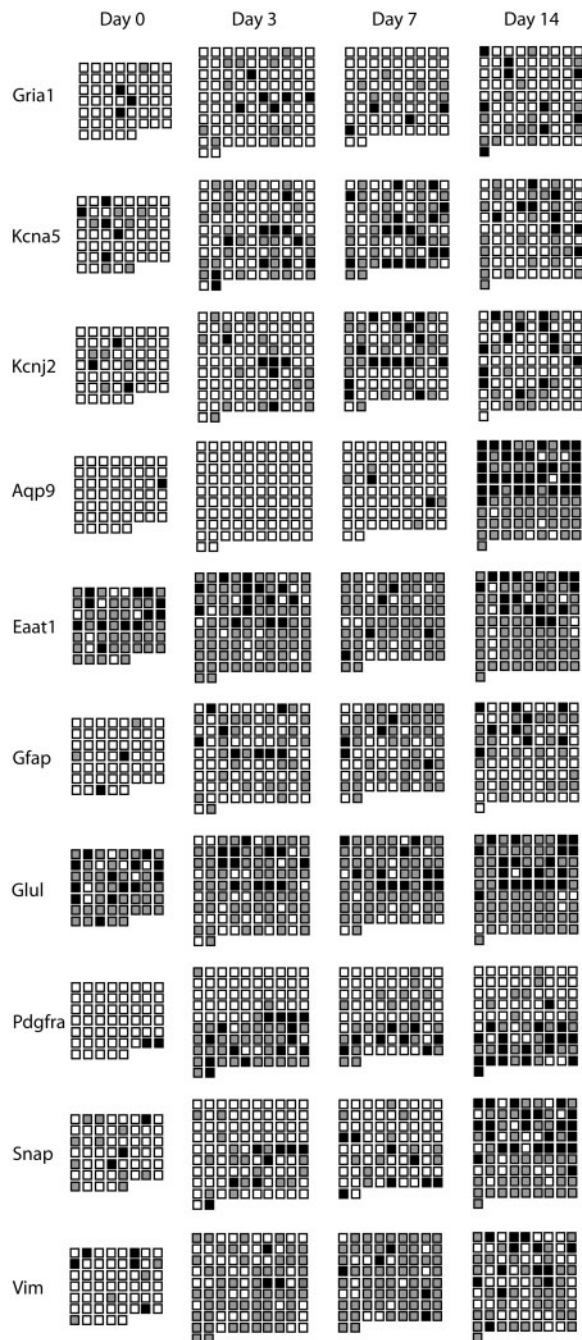
Figure 5: Cell population analysis. Relative gene expression profiles at four time points (days 0, 3, 7 and 14) are shown, with day 0 being the control (i.e. before brain ischemia). The relative gene expression level at day 0 was arbitrarily set to 1 for each gene. *Aqp9* expression was zero at day 3, but for presentation purposes it is set to 0.001.

Figure 4: Binary single-cell analysis. Single cells with (gray and black) and without (white) measurable transcript levels are indicated. The cells with highest expression levels that represent 70% of all transcripts of any given gene at each time point are indicated in black. $n_{Day0} = 53$, $n_{Day3} = 92$, $n_{Day7} = 74$ and $n_{Day14} = 82$.

reference genes that are constantly expressed in the compared samples.

To analyze the single-cell data set at cell population level, we summed the numbers of each transcript in every cell and each time point. Then, the
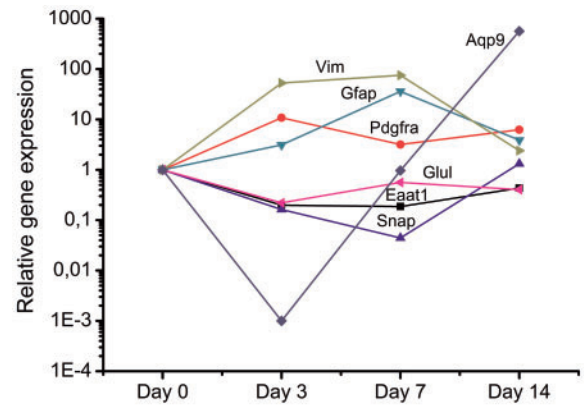
genes with most constant expression were identified for normalization of the data using analysis of variance as described in NormFinder [28]. The three genes *Gria1*, *Kcna5* and *Kcnj2* were found to be the best choice for normalization. Relative gene expression was calculated by normalizing to the geometric average expression of the three validated reference genes [29]. Figure 5 shows the normalized gene expression over time for the selected target genes *Aqp9*, *Eaat1*, *Gfap*, *Glul*, *Pdgfra*, *Snap* and *Vim*. The gene expression was arbitrarily set to 1 at day 0 for all genes (Figure 5). This choice of scaling has no effect on the downstream statistics, but provides an intuitive baseline for sample comparison [30]. This cell population analysis reveals that expression of *Gfap*, *Pdgfra* and *Vim* is transiently upregulated, while *Eaat1*, *Glul* and *Snap* is transiently downregulated, and expression of *Aqp9* is first downregulated and then strongly upregulated (Figure 5). Clearly, the cell population analysis has caught the essence of the overall biological response induced by the brain injury.

As mentioned, statistical analysis identified *Gria1*, *Kcna5* and *Kcnj2* as optimal reference genes at cell population level. From single-cell analysis, we note that only very few cells host the majority of these transcripts at each time point (Figures 4 and 6). Furthermore, the *Gria1*, *Kcna5* and *Kcnj2* transcripts are found in different cells. Hence, if we would normalize expression of the target genes to that of the reference genes
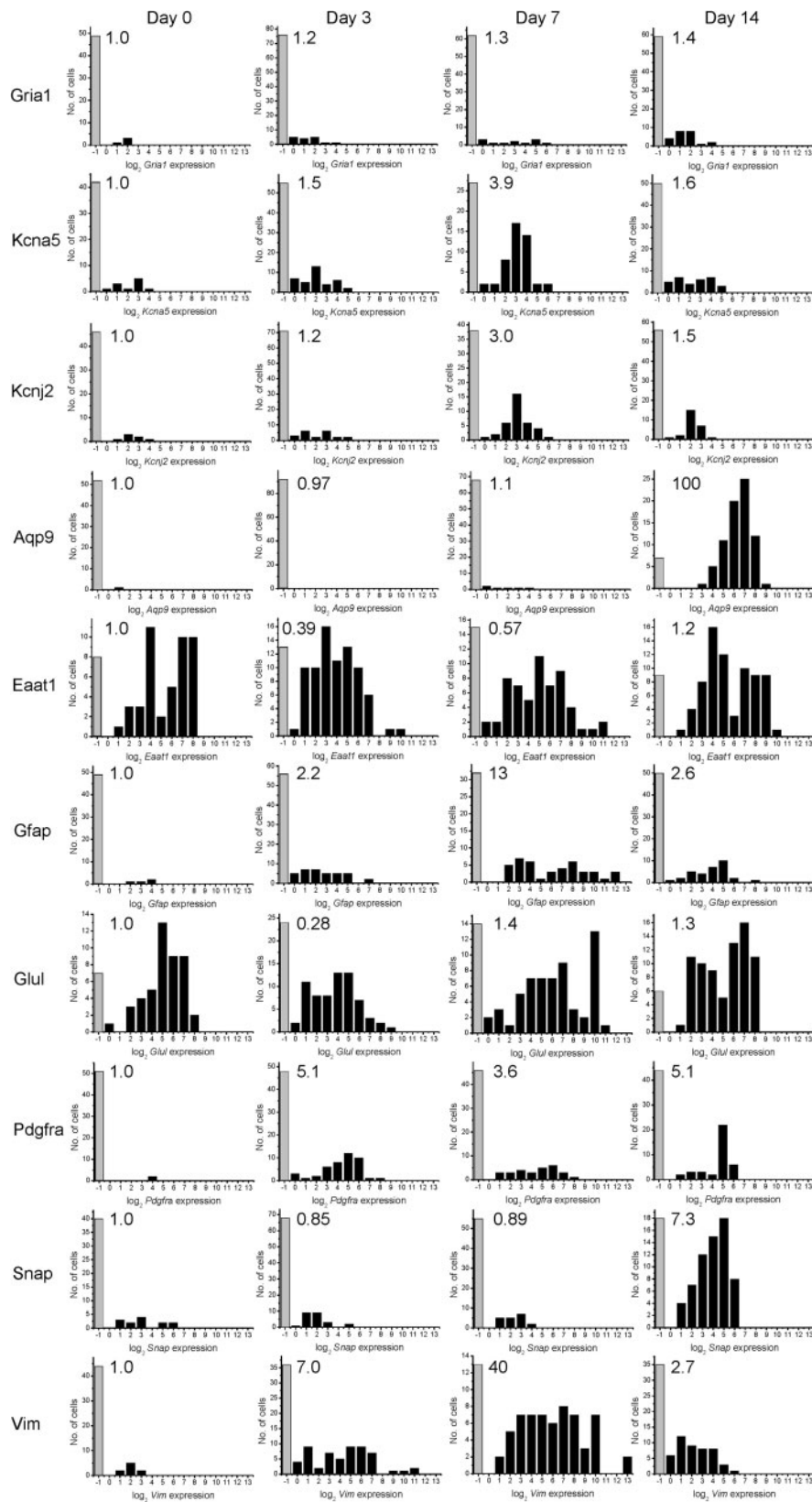
Figure 6: Distribution of transcripts among individual cells. Histograms showing relative gene expression levels (log$_2$-scale) for all genes at days 0, 3, 7 and 14. Day 0 is control (i.e. before brain ischemia). Cells without transcripts are indicated in gray. The typical gene expression of each gene and time point was calculated as geometric average by assigning missing data 0.5 transcripts, since calculations of geometric averages cannot include zeros. The relative mean expression was arbitrarily set to a value of 1 at day 0 for each gene.

separately in each individual cell, we would introduce very large unphysiological bias. The reason normalization to reference genes is applicable in cell population studies is due to the large number of cells analyzed, which averages out the large natural variation among individual cells. However, the averaging also eliminates much of the differences in expression contributed by subpopulations. There is also risk that the reference genes selected have different expression profiles in existing subpopulations; the Normfinder algorithm does not sense this.

Traditional cell population analysis is performed on thousands of cells, sometimes even many more. The interest to analyze samples containing fewer cells is, however, rapidly increasing. For example, fine needle aspirates collected frequently in minimal invasive diagnostics contain often only a few hundreds of cells for downstream analysis. In our data set, 52–92 cells were analyzed at each time point. One consequence of analyzing samples with so few cells is that estimated expression of low expressed genes is uncertain and results may become biased. For example, cell population analysis indicates that *Aqp9* is substantially downregulated at day 3, followed by dramatic upregulation at day 14 (Figure 5). Inspecting the single-cell data we find that only one cell expressed *Aqp9* at day 0 and none at day 3 (Figures 4 and 6). This vital biological information is lost in the cell population analysis, and the conclusion reached relied on expression in only one cell, corresponding to less than 1% of the cells (1 out of 145 cells expressed *Aqp9* during days 0–3). Single-cell data allow for more detailed and careful interpretation of the *Aqp9* expression data. One way to improve the accuracy of *Aqp9* expression analysis is to increase the amount of material analyzed. However, this is possible only to a certain extent. There is a maximum amount of material that can be reliably reversed transcribed and qPCR amplified in a single reaction. Furthermore, standard RNA extraction protocols purify total RNA or total mRNA and are not suitable to enrich rare transcripts. Consequently, quantification of low abundant transcripts will usually have poor precision when studying samples using common population analysis. In summary, single-cell analysis adds value by identifying the number of cells that actually express any given transcript.

The intuitive way to present single-cell gene expression data is per cell. Importantly, single-cell data should not be normalized using reference genes, since no gene has ever been reported to have constant and stable expression among individual cells. Instead, very high variability in specific mRNA levels among cells is seen due to transcriptional kinetics (Figure 3) [6,7]. One option could be to perform global normalization, i.e. to normalize to the mean expression of all genes. But this is complicated since we are not analyzing all the genes with qPCR and those that are studied are not randomly selected. The studied genes rather reflect specific biological processes that are induced or stimulated to change. Normalization to the mean average of the studied genes is then likely to introduce bias. Furthermore, data comparison between studies that use slightly different markers is also biased, since the 'global' mean is not truly global, but is the mean of those genes that actually are measured. Finally, single-cell profiling studies have many missing data, and there is no good way to handle those in global normalization. In brief, methods to normalize cell population data are not suitable for single-cell data. We have found that the most useful and natural way to present and report single-cell gene expression data is per cell.

TRANSCRIPTS ARE LOGNORMALLY DISTRIBUTED AMONG INDIVIDUAL CELLS

Cells that contributed to 70% of the total number of a particular transcript were in minority at population level in our experimental data set (indicated as black in Figure 4). For example, only 6 out of 92 cells contributed with 70% of all the *Gria1* transcripts at day 3. The reason few cells contribute with majority of the transcripts for any particular gene is the highly skewed distribution of transcripts normally observed among individual cells (Figure 6). Gene expression data from single cells can usually be adequately modeled with a lognormal distribution, although recent advances in theoretically biology suggest that the underlying distribution is more complex [4,6]. A consequence of lognormal distribution is that the geometric average reflects the median cell, rather than the normal arithmetic average that is obtained by dividing the total number of transcripts with the number of cells analyzed (Figure 6). The geometric average can only be determined from single-cell studies, not from studies of cell populations.

TRANSCRIPT DISTRIBUTIONS ARE INFORMATIVE AT SINGLE-CELL LEVEL

Figure 6 illustrates some gene expression responses to the induced brain trauma at single-cell level in our experimental data set. *Aqp9* and *Snap* are strongly activated in most cells at day 14 after having hardly been expressed in any cells at earlier time points. *Gfap* and *Vim* are transiently expressed with a peak at day 7, when *Gfap* and *Vim* expression is also observed in most cells. Deeper inspection of the data reveals that *Gfap* and *Vim* are expressed in the same cells, suggesting that their activations are synchronized (Figure 4 and Table 1). The *Gfap* and *Vim* expression profiles at single-cell level are in concordance with the cell population analysis. In contrast, *Aqp9* and *Snap* are expressed in few cells only during days 0–7 and while single-cell profiling reveals the correct picture, the expression analyzed at population level is stochastic, since the amount of transcripts in the only or very few expressing cells present vary.

Heavily skewed and even bimodal distributions of transcripts among single cells are observed for *Eaat1* and *Glul* and indicate presence of multiple cell types, or subpopulations of cells that perhaps are cells in different reversibly interchangeable states (Figure 6). The distribution of *Eaat1* transcripts is bimodal at day 0 and at day 14, while cells with high transcript levels are missing days 3–7. *Glul* shows similar expression pattern and its expression also correlates with *Eaat1* expression (Table 1). These data suggest that one cell subpopulation characterized by high *Eaat1* and *Glul* expression present initially disappears at day 3 to reappear later. From our data, we cannot tell whether expressions of *Eaat1* and *Glul* are

uniformly downregulated in this group of cells or whether the subpopulation does not exist at days 3–7, since we are not measuring the same cells over time.

Pdgfra appears upregulated similarly to *Gfap* and *Vim*, when considering only expression levels (Figures 5 and 6). But, inspecting Figure 4 and Table 1, we see that *Pdgfra* is expressed in different cells than the other genes, in particular in cells that do not express *Glul*. This example shows that analyzing population data only gives part of the picture; analyzing correlations between gene expressions among the single cells also provides information about cell subpopulations. This is particularly interesting when cells are exposed to stimuli, and can reveal regulatory gene expression networks [13,15,17].

IDENTIFICATION AND CHARACTERIZATION OF DEFINED SUBPOPULATIONS

Principal component analysis, hierarchical clustering and self-organizing maps (SOMs) are algorithms that can be used to classify cells based on their expression profiles, dividing them into cell states or subpopulations [11–13,15]. However, the subpopulations will not become meaningful until they can be biologically interpreted.

We have shown that SOM can be used to divide genes and/or samples into subpopulations based on their expression profiles [13,29]. Here, we use SOM to classify the cells at day 14 into two subpopulations (Figure 7A). To interpret the biological relevance of the two subpopulations, we calculated the expression of all genes in the typical cell of the two

Table 1: Spearman correlation coefficients for all single cells

	Grial	Kcna5	Kcnj2	Aqp9	Eaatl	Gfap	Glul	Pdgfra	Snap	Vim
Grial	1									
Kcna5	0.10	1								
Kcnj2	**0.16**	**0.37**	1							
Aqp9	0.14	−0.02	−0.02	1						
Eaatl	0.12	0.04	0.01	**0.21**	1					
Gfap	0.13	**0.22**	0.11	0.05	**0.25**	1				
Glul	0.10	**0.15**	**0.20**	**0.21**	**0.53**	**0.23**	1			
Pdgfra	**0.17**	**0.19**	0.08	0.03	−0.08	0.06	**−0.34**	1		
Snap	**0.28**	0.10	0.10	**0.53**	0.10	0.08	0.09	**0.16**	1	
Vim	0.14	**0.28**	0.15	−0.10	0.12	**0.42**	0.06	**0.33**	−0.02	1

Notes: Bold indicates ≥99% significance. A correlation coefficient of 1 reflects perfect correlation, −1 reflects perfect anti-correlation and 0 is absence of correlation.
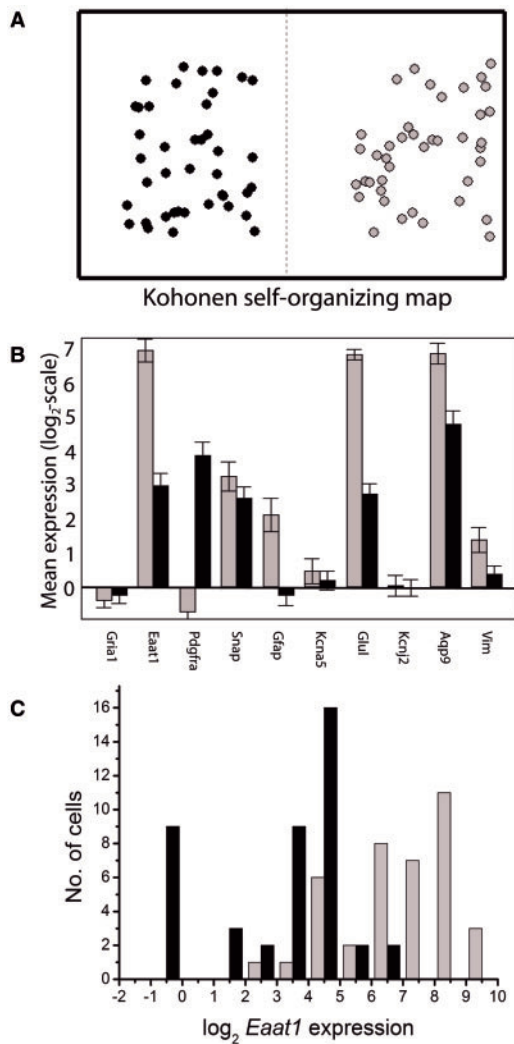
Figure 7: Classification and characterization of subpopulations. (**A**) Two subpopulations of cells (gray and black) were defined using Kohonen SOM at day 14. (**B**) Characterization of the typical cell in each subpopulation. (**C**) Distributions of *Eaat1* transcripts show that the originally observed bimodality in *Eaat1* expression levels (Figure 6) can be accounted for by the two subpopulations identified using SOM. Parameters for the SOM were 2×1 map, 0.1 learning rate, two neighbors and 10 000 iterations. The observed clustering was reducible and independent of parameter settings.

subpopulations (Figure 7B). *Aqp9*, *Eaat1*, *Gfap*, *Glul* and *Vim* are upregulated and *Pdgfr* is downregulated in one subpopulation (gray group) relative to the other subpopulation (black group). The gene expression profile we find for the astrocytes in the subpopulation indicated as gray in Figure 7B has expression features commonly attributed to activated astrocytes. In Figure 7C, we plot the distributions of *Eaat1* transcripts among cells for the two subpopulations

separately. As expected, the bimodal pattern of *Eaat1* seen in Figure 6 is resolved by the classification into groups by means of SOM. Closer inspection of Figures 4 and 6 and Table 1 shows that the SOM clustering picks up the major trend at day 14. If we instead use SOM to generate three subpopulations, the third group reveals no new feature relative to the first two groups. The third group shows characteristics of an intermediate state between the two subpopulations (Figure 7). These cells may be in transition between cell states, perhaps astrocytes becoming activated [13]. In conclusion, we present a strategy to identify subpopulations of cells using SOM and show how these subpopulations can be characterized.

FUTURE PROSPECTS
Techniques are rapidly developing for single-cell applications, and the importance of single-cell analysis is realized in many fields of research and diagnostics. Most studies of individual cells are still explorative, but with the improvement of protocols, robust methods will soon be available that will introduce these methods also into diagnostic and prognostic applications.

Key Points

- Single-cell analysis is an emerging concept in molecular biology and diagnostics.
- Comparison between cell population and single-cell gene expression profiling highlights the advantages of single-cell analysis.
- Handling of missing data and sample normalization are two data pre-processing steps that are performed fundamentally different in single-cell analysis compared with cell population analysis.
- Biomarker analysis including correlation studies at single-cell level opens up new possibilities to understand biological processes with refined resolution.
- Single-cell analysis can be used to identify and characterize novel subpopulations of cells.

References

1. Ståhlberg A, Thomsen C, Ruff D, Aman P. Quantitative PCR analysis o f DNA, RNAs, and proteins in the same single cell. *Clin Chem* 2012;**58**:1682–91.

2. Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. *Annu Rev Genet* 2011;**45**:431–5.

3. Wu M, Singh AK. Single-cell protein analysis. *Curr Opin Biotechnol* 2012;**23**:83–8.

4. Larson DR, Singer RH, Zenklusen D. A single molecule view of gene expression. *Trends Cell Biol* 2009;**19**:630–7.

5. Wang D, Bodovitz S. Single-cell analysis: the new frontier in 'omics'. *Trends Biotechnol* 2010;**28**:281–90.

6. Raj A, Peskin CS, Tranchina D, et al. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 2006;**4**:e309.

7. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. *Curr Biol* 2006;**16**:1018–25.

8. Yu J, Xiao J, Ren X, et al. Probing gene expression in live cells, one protein molecule at a time. *Science* 2006;**311**: 1600–3.

9. Zenklusen D, Larson DR, Singer RH. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* 2008;**15**:1263–71.

10. Paulsson J. Summing up the noise in gene networks. *Nature* 2004;**427**:415–9.

11. Guo G, Huss M, Tong GQ, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyte. *Dev Cell* 2010;**18**:675–85.

12. Dalerba P, Kalisky T, Sahoo D, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011;**29**:1120–7.

13. Ståhlberg A, Andersson D, Aurelius J, et al. Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic Acids Res* 2011;**39**:e24.

14. Narsinh KH, Sun N, Sanchez-Freire V, et al. Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *J Clin Invest* 2011;**121**: 1217–21.

15. Norrman K, Strömbeck A, Semb H, Ståhlberg A. Distinct gene expression signatures in human embryonic stem cells differentiated towards definitive endoderm at single-cell level. *Methods* 2013;**59**:59–70.

16. Liss B, Franz O, Sewing S, et al. Tuning pacemaker frequency of individual dopaminergic neurons by Kv4.3L and KChip3.1 transcription. *EMBO J* 2001;**20**:5715–24.

17. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 2005;**15**:1388–92.

18. Warren L, Bryder D, Weissman IL, Quake SR. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc Natl Acad Sci USA* 2006;**103**:17807–12.

19. Gibson JD, Jakuba CM, Boucher N, et al. Single-cell transcript analysis of human embryonic stem cells. *Integr Biol* 2009;**1**:540–51.

20. Powell AA, Talsaz AH, Zhang H, et al. Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines. *PLoS One* 2012;**7**: e33788.

21. Pantel K, Brakenhoff RH, Brandt K. Detection, clinical relevance and specific biological properties of disseminating tumor cells. *Nat Rev Cancer* 2008;**8**:329–40.

22. Alix-Panabieras C, Schwarzenbach H, Pantel K. Circulating tumor cells and circulating tumor DNA. *Annu Rev Med* 2012;**63**:199–215.

23. Markey AL, Mohr S, Day PJ. High-throughput droplet PCR. *Methods* 2010;**50**:277–81.

24. Ståhlberg A, Rusnakova V, Forootan A, et al. RT-qPCR work-flow for single-cell data analysis. *Methods* 2013;**59**: 80–8.

25. Bustin SA, Benes V, Garson JA, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 2009;**55**:611–22.

26. Velculescu VE, Madden SL, Zhang L, et al. Analysis of human transcriptomes. *Nat Genet* 1999;**23**:387–8.

27. Maier T, Schmidt A, Güell M, et al. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol Syst Biol* 2011;**7**:511.

28. Andersen CL, Jensen JL, Orntoft TF, et al. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 2004;**64**:5245–50.

29. Ståhlberg A, Elbing K, Andrade-Garda JM, et al. Multiway real-time PCR gene expression profiling in yeast *Saccharomyces cerevisiae* reveals altered transcriptional response of ADH-genes to glucose stimuli. *BMC Genomics* 2008;**6**: 170.

30. Kubista M, Andrade JM, Bengtsson M, et al. The real-time polymerase chain reaction. *Mol Aspects Med* 2006;**27**: 95–125.