Unrevealing biological secrets by single cell expression profiling

During the last 30 years exceedingly sensitive and accurate technologies have been developed to measure and quantify amounts of nucleic acids and proteins. With techniques such as quantitative real-time PCR (qPCR)[1] and next-generation sequencing (NGS)[2] DNA, messenger RNAs (mRNA), micro RNAs (miRNA), long non-coding RNAs (lncRNA) and today even proteins can be measured in most types of samples allowing the comparison of environmental conditions, studies of diseases and monitoring of treatments.

By Dr Mikael Kubista and Dr Anders Ståhlberg

Despite these technological advances, understanding biological phenomena on molecular level and identifying useful biomarkers to support clinical decisions remain challenging because of large variation in data. Confounding technical variation may be substantial when (for profiling) unsuitably preserved samples are used or samples are handled improperly[3,4]. When suitable samples are tested using optimised and standardised procedures, technical reproducibility is high and confounding variation is predominantly biological. A methodology to estimate the variance contributions of the steps in an experimental protocol and to optimise the design has been described by Tichopad et al[5] and is today implemented in software such as GenEx[6]. When studying human samples intersubject variation due to the individuals being genetically different is substantial. On top of that is sample heterogeneity. Analysing different bits of the same tissue can show substantial variation. Homogenising the sample removes the influence of sampling, but will not necessarily improve the diagnostic value. The sampling variation is due to the presence of many different cell types that can have disparate functions, may react to diverse stimuli and have distinct responses to drugs. In the homogenised material the important cell type may be in minority and its response may go unnoticed due to the non-relevant and unimportant signals from the prevalent cells. Then there is dynamic variation in the RNA and protein levels in individual cells over time even in the absence of stimuli. Transcripts and proteins are usually produced in burst with a rapid rise in expression followed by slow decay[7,8]. The burst kinetics leads to highly skewed distribution of transcripts observed even among seemingly like cells, with most cells harbouring very few transcripts and the majority of the transcripts being present in only a small subset of the cells, which can be modelled with a log normal distribution (**Figure 1**)[9]. This variation in the number of transcripts among cells is usually the dominant contribution and is readily measured with high precision[5,10]. In traditional samples composed of thousands of cells, the cell-to-cell variation is lost by

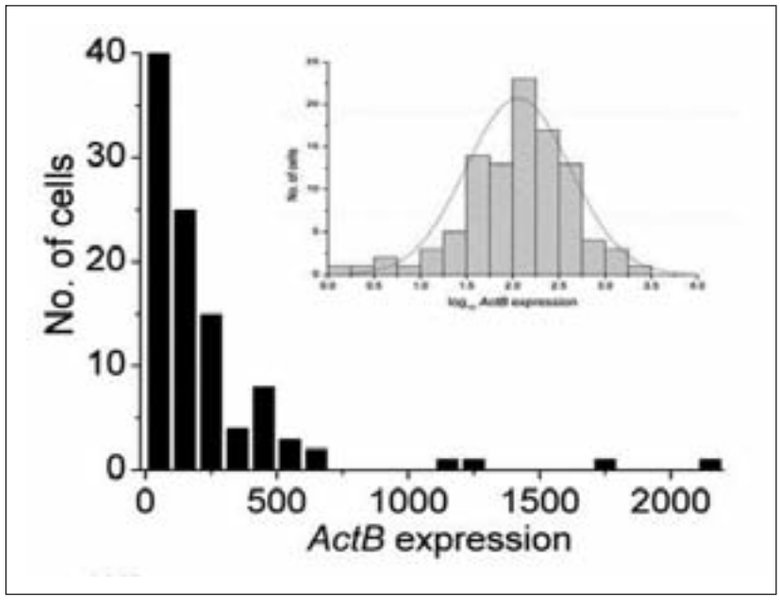

Figure 1: Distribution of β actin transcripts among seemingly β cells in culture; data shown in linear scale. Inset: the same data presented in logarithmic scale on the abscissa

the inherent averaging. As we shall see, the cell-to-cell variation is highly informative, providing an insight into the underlying biology of the studied tissues and their responses that cannot be measured by traditional means[11].

Already in the very first qPCR single cell profiling study, which measured expression of five genes in individual mouse β cells, it was found that the majority of these genes showed uncorrelated expression; transcripts of the different genes were most abundant in different cells[8]. Only one gene pair showed correlated expression, which on the contrary was highly correlated, signifying the transcripts were produced at the same time in the cells. These were the two insulin genes present in rodents. Despite being located in different chromosomes their transcriptional bursts are evidently synchronised. The genes have high sequence similarity, since Ins1 was once formed as accidental retrotransposition of partially processed Ins2 mRNA. They should share many transcription regulatory factors and therefore are likely to be transcribed in a common transcription factory[12]. Since this first publication the field of single cell profiling has exploded, very much spurred by the development of reagents that simplify workflow and high throughput qPCR instruments that reduce running cost and hands-on time (**Figure 2**). Individual cells are collected typically by fluorescence activated cell sorting (FACS), which is most convenient and suitable for high throughput analysis. However, the sorting conditions may stress cells affecting the expression of some genes. Alternative methods to collect cells are microaspiration and laser microdissection. The cells are sorted directly into a lysis buffer compatible with downstream reverse transcription[13], which eliminates losses due to washing[14]. The cDNA is then preamplified, by performing multiplex PCR a limited number of cycles, to increase target concentrations[15]. This step is critical, since in the subsequent parallel qPCR, the number of target molecules per reaction chamber should be some 20 minimum to avoid serious confounding sampling ambiguity[19]. For mainstream applications Fluidigm has developed the microfluidic C1 Single-Cell Auto Prep System that automatises the steps from single cell collection (though without sorting) to the production of preamplified cDNA[16]. Next step is high throughput qPCR, which is performed in nanolitre volumes to keep reagent consumption down[17]. The very large amount of measured data is then transferred to software such as GenEx in a convenient workflow using instrument specific wizards for qPCR data pre-processing and subsequent multivariate analysis[18]. The most informative genes are analysed and used as basis for pathway analysis[19]. The three leading solution providers Life Technologies, Fluidigm and Roche have all aligned with MultiD Analysis and Ingenuity systems to secure robust and reliable workflow from single cell collection, preanalysis, data collection, data mining and pathway analysis.

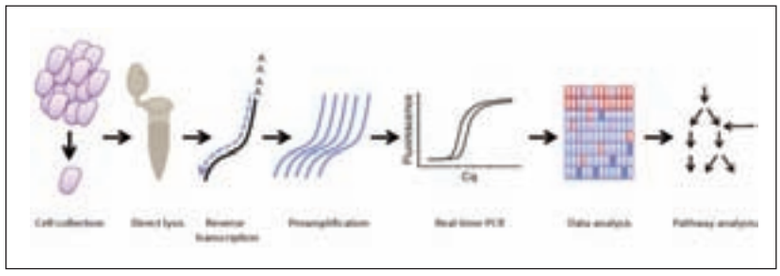

Figure 2: High throughput single cell workflow. Individual cells are collected typically by fluorescence activated cell sorting (FACS), microaspiration or laser microdissection, lysed using reagents that are compatible with downstream processing steps and avoid losses due to washing, optimised reverse transcription, preamplification, which is multiplex PCR performed limited number of cycles, to increase target concentrations, high throughput qPCR, qPCR data pre-processing and multivariate analysis, and pathway analysis

During the last 10 years pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins[20]. Traditionally, pathway analysis is applied on global expression data on classical samples composed of large numbers of

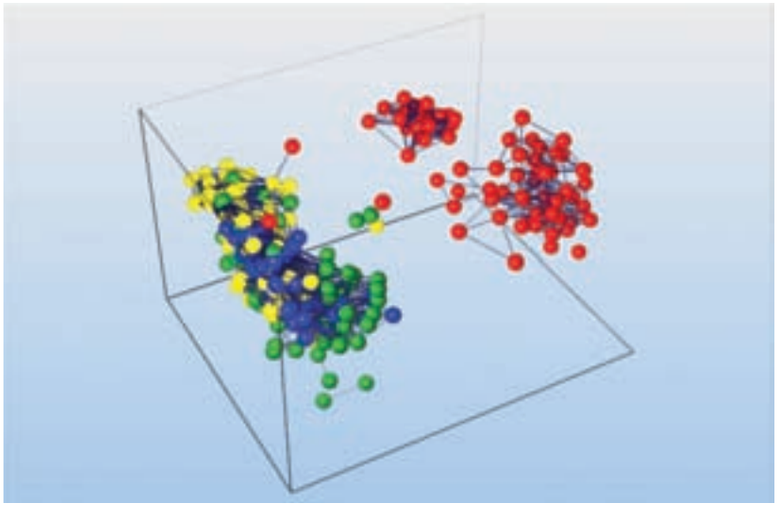

Figure 3: Clustering of astrocytes collected after brain trauma in mice in a three-dimensional PCA plot. The astrocytes were collected before trauma (blue), three days (yellow), seven days (green), and 14 days (red) after trauma. The data have been filtered for variance using dynamic PCA in GenEx (MultiD Analyses)

cells. The cell types present show different behaviour leading to complex patterns, often with thousands of genes showing altered expression. Disintegrating the sample into single cells separates their responses and distinct profiles characteristic of each cell type are obtained. This great improvement is a consequence of the single cell being the studied unit, but also to the burst kinetics. When exposed to an environmental change that affects several expression pathways, it is possible distinct pathways are triggered in different cells. These factors dramatically reduce the complexity of the observed expression patterns, making it much easier to elucidate the underlying expression networks and pathways. In addition to the classical multivariate analysis, correlation analysis is most powerful and frequently reveals distinct groups of genes with correlated expressions[21,22]. The networks can be further analysed calculating partial correlations to separate true direct dependence from apparent indirect dependences. The difference between direct and indirect correlation can be understood by considering some weather statistics collected in Sweden and reported few years ago in a newspaper. The paper noted that ice cream sales rose on hot summer days. The paper also noted that swimming incidents increased with temperature, and concluded eating ice cream was dangerous. Of course there was correlation between ice cream consumption and swimming incidents, but this was accidental caused by the two observations having a common trigger. Such indirect non-relevant effects can be removed by analysing partial correlations. One study using this approach is activation of astrocytes in response to brain trauma in the form of focal cerebral ischaemia using mice models[20,23]. These astrocytes express green fluorescent protein under the control of the astrocyte marker Glial Fibrillary Acidic Protein (GFAP) and can be sorted by FACS from brains of mice sacrificed at different time points after trauma for single cell profiling. The cells can then be classified based on their global expression patterns, typically after filtration to remove the least affected genes that in the analysis would only contribute with noise (**Figure 3**). Visualised in a three-dimensional principal component analysis (PCA) plot, where the axes reflect combined expressions of gene that maximise separation, we see clear differences between astrocytes collected before trauma (blue), and at three (yellow), seven (green), and 14 days (red) after trauma. The changes reflect activation of the astrocytes in response to the trauma and also reveal that the active population prevalent after 14 days is heterogeneous being composed of two distinct subgroups (two red clusters). Genes important for the activation are identified by the PCA, as well as genes differentially expressed between the two subgroups of the reactive astrocytes. Correlation between the genes' expressions on the single cell level defines networks of genes that produce transcripts at the same time in the

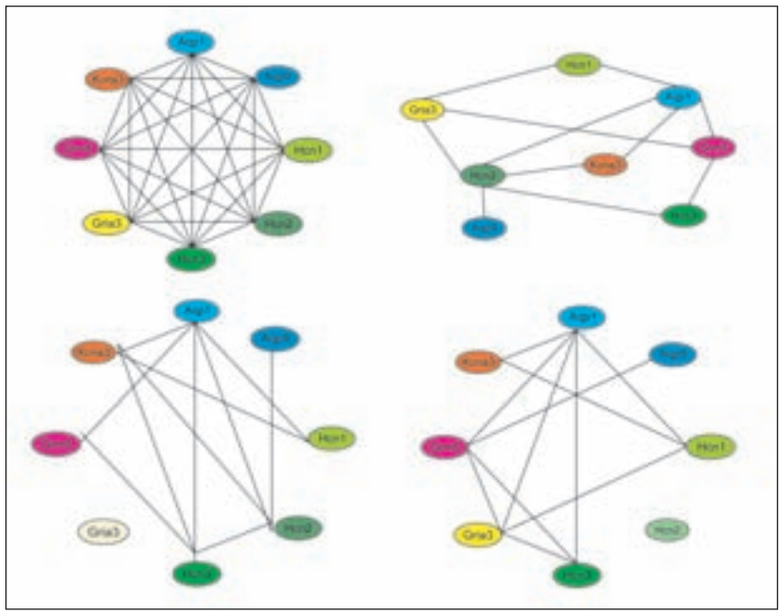

Figure 4: Networks showing dependences among genes expressed simultaneously in the same astrocyte cells during their activation in response to trauma

same cell, suggesting their simultaneous presence may be biologically significant (**Figure 4**).

Proteins can also be measured using qPCR. Two techniques dominate, both being based on the simultaneous binding of two specific antibodies to a common protein. In the original Proximity Ligation Assay (PLA) two antibodies, each tagged with an oligonucleotide, bind to a common protein target (**Figure 5, top**)[24]. An adapter oligonucleotide complementary to the oligonucleotide ends is used as seal to catalyse ligation producing a long DNA template. The amount of long DNA formed is quantitatively measured by qPCR, which reflects the initial amount of the target protein. The technique is today available through Life Technologies as the Taqman protein assays[25]. In the related Proximity Extension Assay (PEA) the tethered oligonucleotides overlap and can be extended without adapter (**Figure 5, bottom**). The PLA and PEA can be combined with preamplification for the parallel analysis of up to 92 proteins in high throughput format[26]. Recently, protocol was developed to split a single cell sample for the simultaneous analysis of DNA by qPCR, mRNA, miRNA, and lncRNA by RT-qPCR, and protein by PLA-qPCR[27]. This was the first example of multianalyte single cell profiling and demonstrated correlation of the levels of related DNA, RNA and protein. In the future this approach will be most valuable to study intricate interactions between the different types of biomolecules occurring in the cell.

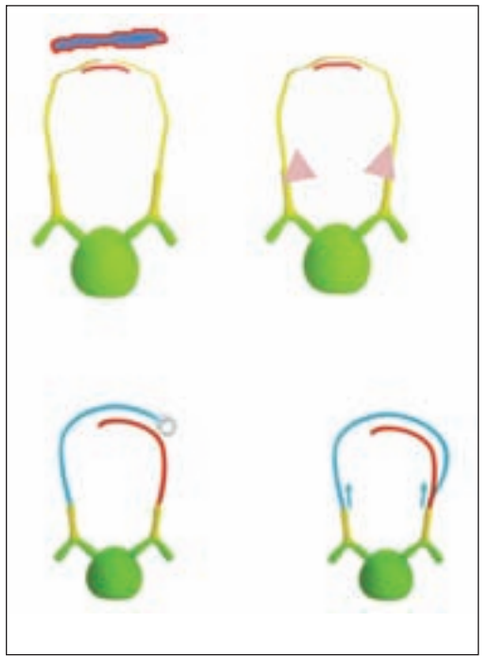

Figure 5: Proximity ligation (PLA, top) and proximity extension (PEA, bottom) assays to detect protein targets. Two antibodies, each tagged with an oligonucleotide, bind to the same protein. The oligonucleotides are either ligated to create a linear DNA template that can be PCR amplified (PLA), or the oligonucleotides overlap and can be extended to produce a double-stranded DNA template (PEA). The amount of DNA template is quantified by qPCR, which reflects the amount of the targeted protein that was present

Can one go beyond the single cell? Indeed one can. The oocyte from the African claw frog *Xenopus laevis* is a colossal single cell that can be mounted in a cryostat. Furthermore its two hemispheres, known as the animal and vegetal pole, have different shades allowing the oocytes to be oriented. The mounted, oriented oocyte can be sliced across the animal-vegetal axis and the RNA extracted, reverse transcribed and quantified. Three distinct intracellular gradients of transcripts are observed in the *Xenopus laevis* oocyte (**Figure 6**). One group of transcripts is located near the cell's centre, somewhat offset towards the animal pole (green). A second group of transcripts is polarised towards the vegetal pole (pink) and a third group of transcripts is found exclusively in the most vegetal segments (red). Clearly, the distribution of transcripts within the oocyte is asymmetric and gene dependent. When the oocyte divides, the first two cleavages are along the animal vegetal-axis at 90° apart and do not introduce asymmetry from the animal-vegetal gradient of transcripts among the blastomeres. The third cleavage axis is across the animal-vegetal axis and introduces asymmetry. Notably, the animal-vegetal asymmetry that manifests first at the 8-cell stage, was already present in the oocytes.

The emerging single cell profiling platforms offer new insights into cell biology expected to lead to novel discoveries and even challenge some dogmas. Particularly exciting will be the new possibilities to characterise cell types and study their differentiation and proliferation. Whereas no two cells are identical in appearance, generalised patterns of morphology and biochemistry suggest that all cells conform to a relatively limited number of patterns referred to as cell types. The tens of trillions (10^{13}) of cells in a human body are often said to be made up of 210 cell types subdivided into 20 categories assembled in 1989 based primarily on function[28]. A more recent classification suggests there are 411 cell types[29]. However, a precise and unambiguous definition of cell type is notoriously difficult[30]. Environmental conditions,

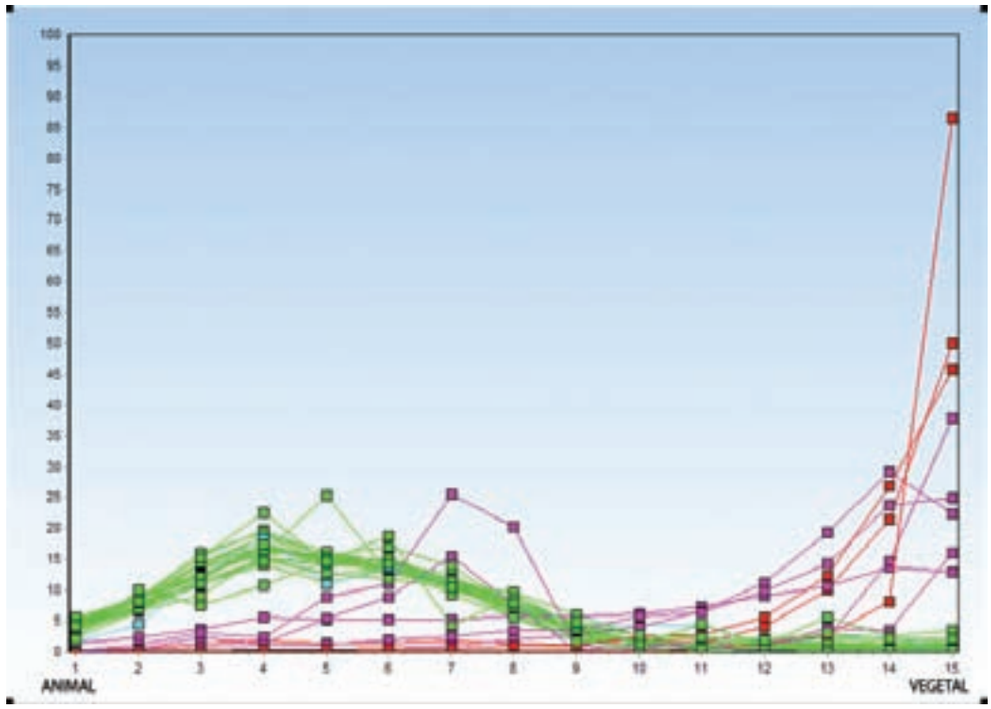

Figure 6: Intracellular profiles of transcripts in *Xenopus laevis* oocyte. Majority of transcripts are located in the cell centre somewhat polarised towards the animal pole (green), a second group of transcripts is polarised towards the vegetal pole (pink), and a third group of transcripts is found exclusively in the most vegetal segments (red)

external stimuli, number and nature of neighbouring cells, signals from remote cells through hormones, exosomes and other signalling substances, access to nutrients, oxygen and other vital substances, removal of waste products, phase of cell cycle, accumulated somatic mutations, integrated viruses, transposons, epigenetic alterations, any rearrangements and perhaps even its age and generation will affect a cell's molecular activities. Some may lead to virtually irreversible differentiation, while others may lead to reversible or even temporal changes only. Single cell profiling is expected to shed light on these processes, perhaps by identifying cell type specific expression networks, helping to reach a definition on cell type and defining the molecular events that make a change irreversible. DDW

Dr Mikael Kubista *is founder and CEO of the TATAA Biocenter (www.tataa.com). He was one of the pioneers contributing to the development of quantitative real-time PCR (qPCR) and coauthored the MIQE guidelines. Together with Dr Anders Ståhlberg, Kubista introduced qPCR for single cell expression profiling and he also developed qPCR tomography for intracellular expression profiling. Kubista's team at TATAA Biocenter has developed many of the reagents for high throughput single cell expression profiling and quality control currently in use.*

Dr Anders Ståhlberg *is co-founder of the TATAA Biocenter and is working as principal investigator at the University of Gothenburg. His main interest is tumour biology and stem cells, with focus on understanding cell hierarchy and cell transition processes. Together with Dr Mikael Kubista's team at TATAA, he has developed several strategies to study single cells, including approaches to analyse many sample types as well as different analytes including RNA, DNA and proteins from the same single-cell.*

References

1 Kubista, M, Andrade, JM, Bengtsson, M, Forootan, A, Jonak, J, Lind, K, Sindelka, R, Sjöback, R, Sjögreen, B, Strömbom, L, Ståhlberg, A, Zoric, N. The Real-Time Polymerase Chain Reaction, Molecular Aspects of Medicine (2006) 27, 95-125.

2 Metzker, ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46. doi: 10.1038/nrg2626. Epub 2009 Dec 8.

3 Pazzagli, M, Malentacchi, F, Simi, L, Orlando, C, Wyrich, R, Günther, K, Hartmann, CC, Verderio, P, Pizzamiglio, S, Ciniselli, CM, Tichopad, A, Kubista, M, Gelmini, S. SPIDIA-RNA: First external quality assessment for the pre-analytical phase of blood samples used for RNA based analyses. Methods 59, 20-31 (2013).

4 Kubista, Mikael, Björkman, Jens, Svec, David and Sjöback, Robert. RNA quality matters. European Pharmaceutical Reviews Vol 17, Issue 6, 2012.

5 Tichopad, A, Kitchen, R, Riedmaier, I, Becker, C, Stahlberg, A, Kubista, M. Design and Optimization of Reverse-Transcription Quantitative PCR Experiments. Clinical Chemistry 55:101816–1823 (2009).

6 www.multid.se.

7 Chubb, JR, Trcek, T, Shenoy, SM, Singer, RH (2006). Transcriptional pulsing of a developmental gene. Current biology : CB 16 (10): 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 Bengtsson, M, Ståhlberg, A, Rorsman, P and Kubista, M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Research 15, 1388-1392 (2005).

10 Livak, KJ, Wills, QF, Tipping, AJ, Datta, K, Mittal, R, Goldson, AJ, Sexton, DW, Holmes, CC. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. Methods. 2013 Jan;59 (1):71-9, 2012.

11 Ståhlberg, Anders, Rusnakova, Vendula and Kubista, Mikael. The added value of single-cell gene expression profiling. Briefings in Functional Genomics (2013) doi: 10.1093/bfgp/elt001 First published online: Feb 7, 2013.

12 Rieder, Dietmar, Trajanoski, Zlatko and McNally., James G. Transcription factories. Front Genet. 2012; 3: 221.

13 Ståhlberg, Anders, Håkansson, Joakim, Xian, Xiaojie, Semb, Henrik and Kubista, Mikael. Properties of the Reverse Transcription Reaction in mRNA Quantification. Clinical Chemistry 50:3, 509-515, 2004.

14 Cellulyser (www.tataa.com), RealTime ready Cell Lysis (www.roche-applied-science.com), Single-cell-to-CT (www.invitrogen.com).

15 GrandMaster PreAmp (www.tataa.com), Taqman PreAmp (www.invitrogen.com), RealTime ready cDNA Pre-Amp Master (www.roche-applied-science.com).

16 http://www.fluidigm.com/c1system.html.

17 Biomark (www.fluidigm.com), OpenArray/QuantStudio (www.lifetechnologies.com).

18 GenEx (www.multid.se).

19 iReport (www.ingenuity.com).

20 Khatri, Purvesh, Sirota, Marina, Butte, Atul J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol. 2012 February; 8(2): e1002375.

21 Benesova, Jana, Rusnakova, Vendula, Honsa, Pavel, Pivonkova, Helena, Dzamba, David, Kubista, Mikael. MiRNAoslava Anderova. Distinct Expression/Function of Potassium and Chloride Channels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice. PLoS ONE, 2012, 7(1), 1-13.

22 Ståhlberg, Anders, Rusnakova, Vendula, Forootan, Amin, Anderova, MiRNAoslava, Kubista, Mikael. RT-qPCR work-flow for single-cell data analysis. Methods 59, 80-88 (2013).

23 Rusnakova, Vendula, Honsa, Pavel, Dzamba, David, Ståhlberg, Anders, Kubista, Mikael, Anderova, MiRNAoslava. Heterogeneity of Astrocytes: From Development to Injury – Single Cell Gene Expression. PLOS ONE, in press.

24 Fredriksson, S, Gullberg, M, Jarvius, J, Olsson, C, Pietras, K, Gústafsdóttir, SM, Ostman, A, Landegren, U. Protein detection using proximity-dependent DNA ligation assays. Nat Biotechnol. 2002 May;20(5):473-7.

25 http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/protein-expression/understanding-taqman-chemistry-based-protein-assays.html.

26 http://www.tataa.com/services/descriptions/protein-profiling/.

27 Ståhlberg, Anders, Thomsen, Christer, Ruff, David and Åman, Pierre. Quantitative PCR Analysis of DNA, RNAs, and Proteins in the Same Single Cell. Clinical Chemistry 58:12 (2012).

28 Alberts, B, Bray, D, Lewis, J, Raff, M, Roberts, K and Watson, JD (1989). Molecular Biology of the Cell, 2nd Edn. Garland Publishing, Inc, New York.

29 Vickaryous, MK, Hall, BK. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. Biol Rev Camb Philos Soc. 2006 Aug;81(3):425-55. Epub 2006 Jun 22.

30 Valentine, JW (2002). Cell-type number and complexity. In Encyclopedia of Evolution, Vol. 1 (ed. M. Pagel), pp. 144–146. Oxford University Press, Oxford.