Mikael Kubista, Linda Strömbom, David Svec, Vendula Rusnakova & Anders Ståhlberg
TATAA Biocenter, Gothenburg, Sweden and the Institute of Biotechnology, CAS

High throughput single cell expression profiling:

Taking a closer look on biological response

Molecular analysis of tissue and in most cases also of bodily fluids is complicated because of tissue heterogeneity and the presence of many different cell types. Even cells of apparently the same type show substantial variation in gene expression under virtually identical conditions. When analysing classical samples based on tens of thousands of cells, this natural variability among cells is lost. With the advent of real-time quantitative polymerase chain reaction (qPCR), we have a most powerful tool to study diversity on the single cell level and can detect rare cells that are critical to treatment or survival.

In 1983, Kary Mullis conceptualised the polymerase chain reaction (PCR), for which he was awarded the 1993 Nobel Prize in chemistry. The idea was as simple as it was brilliant. Based on the natural ability of polymerases to copy nucleic acids in the presence of short complimentary oligonucleotides, Kary Mullis reasoned that using a heat stable polymerase, the reaction could be automated to perform multiple copying cycles by cycling temperature. At 95°C the strands of DNA separate, at 50-60°C primer oligonucleotides hybridise to the complimentary template strands and at 72°C a heat stable polymerase extends the primers copying the template molecule. Virtually any DNA molecule could be amplified multifold even from a single copy. PCR quickly became a key technology in genetic engineering and an important tool for infectious disease testing. However, PCR was at best semi-quantitative, since the amount of product generated depends on the amount of reagents added rather than on the initial number of template molecules. Bob Griffith, working with Russell Higuchi, serendipitously ran PCR in the presence of the DNA stain ethidium bromide, and found fluorescence increased as the dye bound to the DNA amplicons being formed[1]. The larger the initial number of template molecules, the sooner did the fluorescence signal develop. From the number of amplification cycles required to produce a threshold fluorescence signal the relative amount of initial template copies in two samples could be calculated. PCR became real-time quantitative PCR or qPCR for

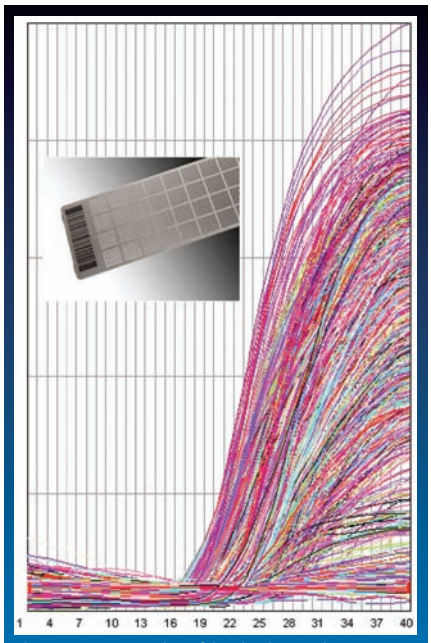

Figure 1 Example of high throughput qPCR measured in the OpenArray. 46 genes were assayed in 48 samples in a total of 2208 reactions. Insert: the OpenArray through-hole platform.

short (**Figure 1**, page 20). qPCR has the ultimate sensitivity to detect a single molecule, virtually infinite dynamic range, high reproducibility and excellent specificity, and it quickly became the preferred technique for quantitative analysis of nucleic acids. DNA could be analysed as is, while RNA had to be reverse transcribed (RT) to cDNA for subsequent PCR. The RT-qPCR was almost as reproducible and sensitive as qPCR[2]. Today, technical solutions are also available to quantify microRNAs and other non-coding RNAs using RT-qPCR[3] and proteins with immuno-qPCR[4].

RT-qPCR profiling soon became the preferred platform to validate expression markers preselected by microarray screening that signify disease state, therapeutic response or predict survival. Based on those markers, companies develop diagnostic, theranostic and prognostic tests. First out was Genomic Health with onco*type* DX for breast cancer[5]. The test is

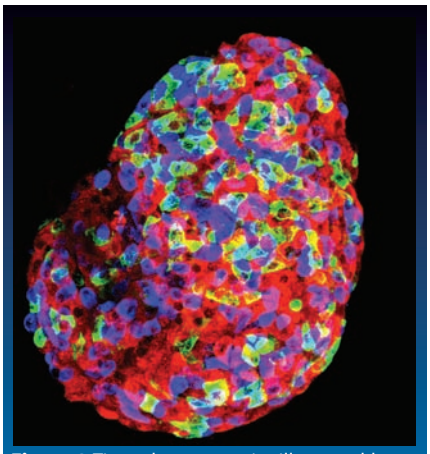

Figure 2 Tissue heterogeneity illustrated by an islet of Langerhans. Confocal image of an isolated islet immunolabelled for insulin (red), glucagon (green) somatostatin (blue). Reproduced from Sven Göpel's doctoral thesis[17]

based on the expression of 21 genes to predict the benefit of chemotherapy and 10-year distant recurrence in women with early-stage breast cancer to support adjuvant treatment decisions[6]. Onco*type* DX was the first test to achieve approval from regulatory authorities and obtain insurance coverage in the US. Recently, Genomic Health launched also a prognostic test for colorectal cancer. XDx, Inc., has developed the FDA-cleared RT-qPCR AlloMap Molecular Expression Testing, which provides transplant physicians with a tool to estimate the probability of acute cellular rejection for post-cardiac transplant patients[7]. Exagen develops qPCR based tests for diagnosis of Irritable Bowel Syndrome, Inflammatory Bowel Disease and the differentiation of Ulcerative Colitis from Crohn's Disease. CardioDx has a test based on the expression of 23 genes to classify coronary artery disease in patients with chest pain. In Europe Ipsogen develops qPCR tests for blood cancer[8], TcLand Expression develops qPCR tests for companion diagnostics in Immune Mediated Disorders and biomarkers in Solid Organ Transplantation[9] and Diagenic has qPCR tests for Alzheimer's disease, breast cancer and a test for Parkinson's disease under development[10]. These tests are all very promising and they set the stage for companion diagnostics, but they all suffer from limited predictive power due to sample complexity. These tests are performed on biopsy or blood samples composed of tens of thousands of cells of many different types, most of which may not be affected by the disease, and among those that are affected, a minority may be determining the outcome.

To address the complexity of sample heterogeneity, our team developed single cell expression profiling[11]. Using qPCR, we measure the number of transcripts in individual cells with very high precision and sensitivity. In our first paper, we studied expression in insulin secreting β cells from the islet of Langerhans, which is highly heterogeneous (**Figure 2**). Surprisingly, we also found large and skewed variation also in the transcript levels among the β cells. The majority of cells contained only some transcripts, while a few cells were loaded with the assayed mRNA. After some modelling, we

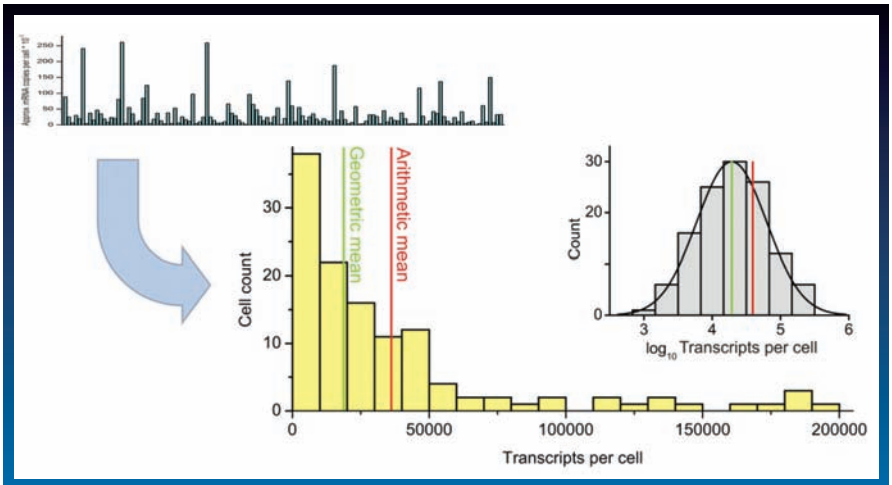

Figure 3 Top left: Insulin I expression in individual beta cells. Bottom left: Expression levels arranged into a histogram with linear scale. Arithmetic and geometric means are indicated. The distribution is highly skewed towards low expression. Right: Expression levels arranged into a histogram with logarithmic scale. Geometric and arithmetic means are indicated. The data are fitted to a normal distribution.

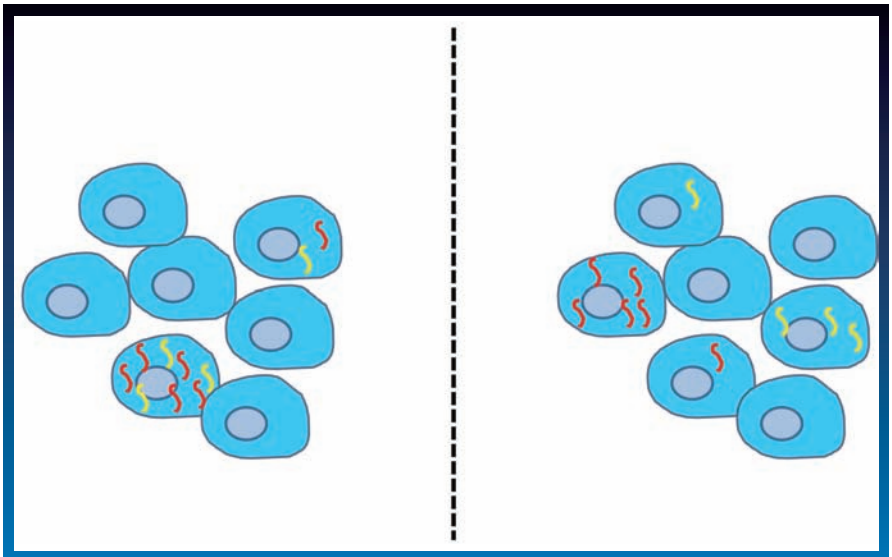

Figure 4 Illustrating correlation. Left: Red and yellow transcripts correlate on the single cell level. Right: Red and yellow transcripts correlate on sample level.

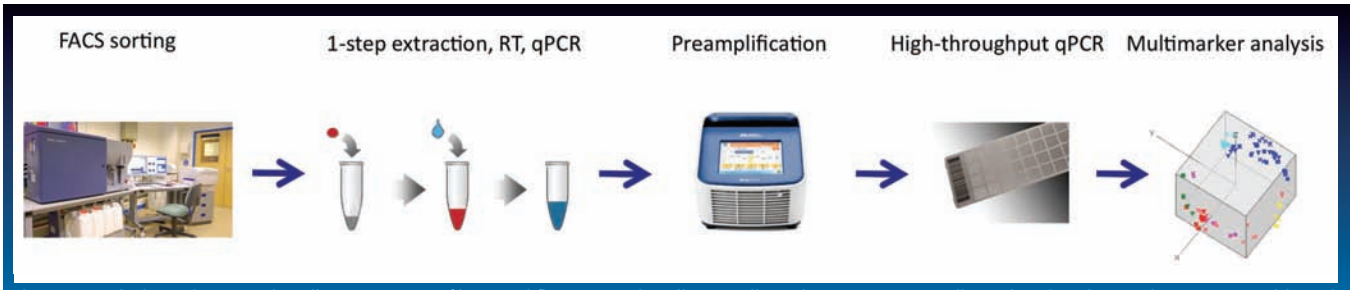

Figure 5 High throughput single cell expression profiling workflow. 1) Single cells are collected using FACS; 2) Cells are lyzed under conditions compatible with RT-qPCR; 3) RNA is reverse transcribed into cDNA and pre-amplified; 4) cDNA is assayed in parallel singleplex qPCR for large number of targets in a high throughput platform; 5) Genes and cells are classified using multivariate statistics.

found the variation is consistent with a log normal distribution, i.e., a Gaussian distribution when the transcript levels are presented in logarithmic scale (**Figure 3**, page 22). One consequence of log normal distribution is that the average number of transcripts per cell calculated from a classical experiment performed on many cells does not reflect the activity of the typical or median cell in the population. Rather, the number of transcripts in the typical cell is given by the geometric average, which can only be calculated from single cell studies. In our pioneer work, we measured the expression of five genes per cell and found that four of the genes were expressed in different cells. Only for two genes, Ins1 and Ins2, did transcript levels correlate, suggesting a mechanism regulates their co-expression. With qPCR, we measure only static distribution of transcripts. Dynamic measurements using fluorescence in situ hybridisation have shown that transcription in eukaryotic cells occurs in bursts[12]. Hence, cells with many copies of a particular transcript are those that recently had a transcriptional burst for the corresponding gene, and those genes that have similar transcript levels should have their transcriptional bursts correlated. Notably, Ins1 and Ins2 are located in different chromosomes suggesting rather sophisticated mechanisms regulate transcriptional bursting. From a biological perspective, it makes sense that transcripts needed in equal amount for production of ternary proteins or perhaps part of the same regulatory pathway are co-regulated. Hence, correlation of transcript levels on the single cell level is a very strong indication that the genes are involved in the same biological processes. As we shall see, the correlation is a fingerprint of the particular cell type. Notably, correlation of transcription levels among classical samples is much less informative, since genes may be responding independently to a particular condition or stimuli applied (**Figure 4**, page 22).

In our pioneer work, we measured the expression of five genes per cell, using laborious and costly protocols that required many experimental steps. During the following years, our centre, in collaborations with Roche, Life Technologies and MultiD Analysis, made important progress in streamlining the process of single cell expression profiling that greatly increased the throughput, multiplex capacity, flexibility and robustness (**Figure 5**, page 23). Fluorescence activated cell sorting (FACS) became the preferred method to collect cells in suspension, although we also use manual picking using a micro manipulator under a microscope, which allows cells to be pre-selected based on surface antigens using immunomagnetic techniques. At our centre in the Czech Republic, we are running the COHERTA study in collaboration with Roche and AdnaGen[13], studying expression markers including HER2 in circulating tumour cells (CTC) enriched from peripheral venous blood. CTC markers are expected to be superior for treatment decisions, since they reflect the state of the metastatic process, to markers from the primary tumour that is surgically removed. Sorted cells are lysed using novel extraction reagents such as the CelluLyser (TATAA), Real-time Ready Cell Lysis kit (Roche), or SingleCell to CT (Life Technologies). These are compatible with reverse transcription and PCR, allowing the cell to be processed with minimum material losses. To measure the expression of more than some 5 – 10 genes per cell, the material has to be pre-amplified, because of the aliquoting required for parallel singleplex qPCR analyses. Pre-amplification for qPCR is usually based on multiplex PCR, wherein all targeted transcripts are amplified for a limited number of cycles, such that reagents are not consumed to a significant level and bias is avoided. Life Technologies have developed the TaqMan PreAmp master mix for this and TATAA offers custom optimisation of pre-amplification for single cell analysis. Next, the samples are analysed in a high throughput qPCR system. At TATAA, we have the OpenArray from Applied Biosystems and the BIOMARK from Fluidigm. A high throughput system is a must to keep costs down, because the number of reactions needed to obtain good statistics are of the order of thousands (**Figure 1**, page 20). Finally, data is analysed. Classical approaches based on normalisation with reference genes are not meaningful, because the transcript levels of genes in general do not correlate on the single cell level. Instead, data are mean centred or autoscaled and classified using multivariate methods such as Principal Component Analysis (PCA), the Self-Organised Map (SOM), and Hierarchical Clustering. The GenEx software from MultiD is excellent for this purpose. It has readers for the high throughput qPCR platforms, and single cell data are readily pre-processed and clustered to reveal correlations between genes and similarities among the cells[14]. Recently, we used single cell expression profiling to reveal the presence of novel astrocyte subtypes based on correlation between transcript levels[15]. The subtypes differ mainly in the expression of Vimentin, and are clearly distinguished when considering expression of multiple genes, even though neither subtype exclusively expresses any single marker (**Figure 6**). Based on correlations of transcript levels, we could design interaction maps between the most important genes[15].

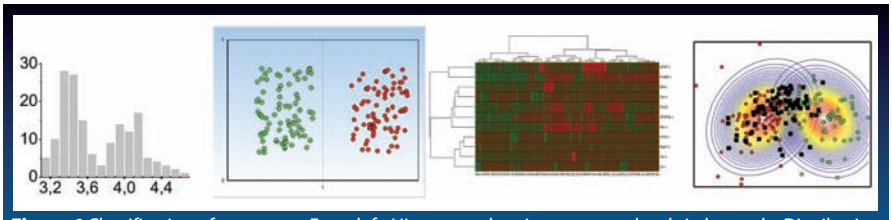

Figure 6 Classification of astrocytes. From left: Histogram showing astrocyte levels in log scale. Distribution is bimodal (two maxima) signifying the presence of two cell subtypes; Based on the expression of eleven genes the cells were classified into two groups using SOM; Heat map and hierarchical clustering separating the cells into two groups; PCA showing the cells colored based on the SOM and hierarchical classifications. The PCA is overlaid with cells dissociated from neurospheres generated from P4 brains (black symbols), showing these cells are similar to the astrocyte subtype that has low vimentin expression.

Single cell expression profiling is expected to become a key platform to identify new drug targets and to study the mechanism of action of lead compounds. In molecular diagnostics, single cell expression profiling will be important for theranostics and prognostics of complex diseases, in particular for cancer, where cells with stem cell properties and also cells resistant to treatment can be revealed. Particularly interesting is expression profiling of disseminated tumour cells, which is the focus of this year's qPCR symposium taking place June 13-17 in Prague[16].

References

1. R. Higushi. Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. Bio/Technology 11, 1026 – 1030 (1993)
2. Properties of the Reverse Transcriptase Reaction in mRNA Quantification, A. Stålberg, J. Håkansson, X. Xian, H. Semb, M. Kubista, Clinical Chemistry, 2004, 50:3, 509-515
3. http://www.exiqon.com; http://www.appliedbiosystems.com
4. Development and evaluation of three real-time immuno-PCR assemblages for quantification of PSA, K.Lind, M. Kubista, J. Immun. Meth. 304 (2005) 107-116.
5. http://www.genomichealth.com
6. Paik S, Shak S, Tang G et al.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27), 2817-2826 (2004).
7. http://www.xdx.com/
8. http://www.ipsogen.com/
9. http://www.tcland-expression.com/
10. http://www.diagenic.com/
11. M. Bengtsson, A. Stålberg, P. Rorsman, M. Kubista. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Research (2005) 1388-1392
12. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. and Tyagi, S. (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol., 4, e309.
13. http://www.adnagen.com/
14. http://www.multid.se/
15. Anders Ståhlberg, Daniel Andersson, Johan Aurelius, Maryam Faiz, Marcela Pekna, Mikael Kubista and Milos Pekny. Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations Nucleic Acids Research, 2010, 1–12, doi:10.1093/nar/gkq1182.
16. http://www.qpcrsymposium.eu/
17. http://www.lu.se/o.o.i.s?id=12588&postid=41379

Professor Mikael Kubista was among the pioneers in qPCR. Starting in 1991, his laboratory developed dyes and probes and founded LightUp Technologies as Europe's first company focusing on qPCR based molecular diagnostics. In 2001, Kubista founded the TATAA Biocenters (www.tataa.com), as Europe's leading qPCR service provider and main organiser of hands-on training in molecular diagnostics. TATAA also organises the annual qPCR symposia in Europe (www.qpcrsymposium.eu) and USA (www.qpcrsymposium.com). TATAA is member of the European effort SPIDIA (www.spidia.eu) aiming to standardise the pre-analytical steps in molecular diagnostics. Under Kubista's leadership TATAA pioneered single cell expression profiling. Working as advisor for Unesco Kubista introduced qPCR in Africa and in the Middle East, and he also co-founded MultiD Analyses (www.multid.se) that develops market leading GenEx software for qPCR data mining.