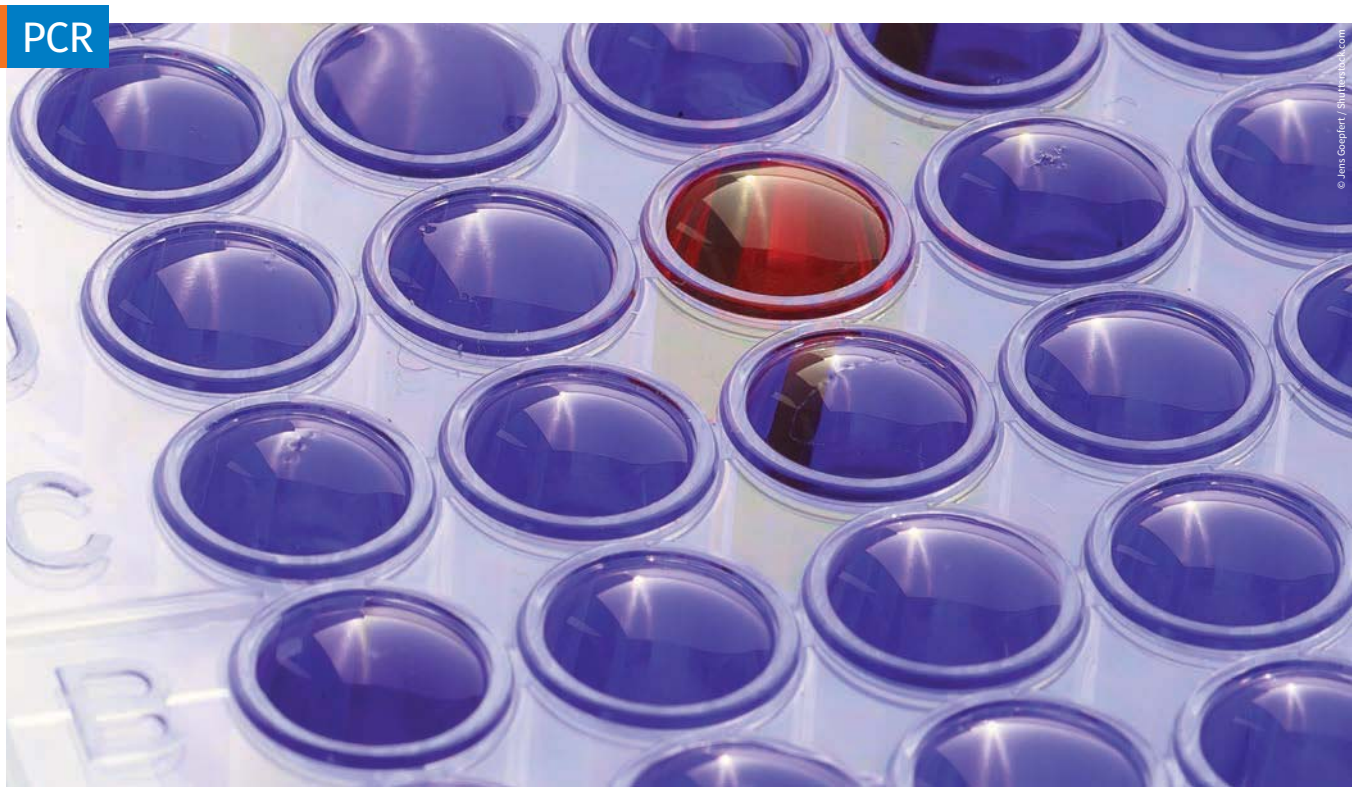

MIQE compliance in expression profiling and clinical biomarker discovery

Irmgard Riedmaier, Melanie Spornraft, Benedikt Kirchner and Michael W. Pfaffl
Technical University of Munich

Molecular diagnostics and biomarker discovery are gaining increasing attraction in clinical research. This includes all fields of diagnostics, such as risk assessment, disease prognosis, treatment prediction and drug application success control[1,2]. The detection of molecular clinical biomarkers is very widespread and can be developed on various molecular levels, like the genome, the epi-genome, the transcriptome, the proteome or the metabolome. Today, numerous high-throughput laboratory methods allow rapid and holistic screening for such marker candidates. Regardless of which molecular level is analysed, in order to detect biomarker candidates, high sample quality and a standardised and highly reproducible quantification workflow are prerequisites. This article describes an optimal and approved development strategy to discover and validate 'transcriptional biomarkers' in clinical diagnostics, which are in compliance with the recently developed MIQE guidelines[3]. We focus on the importance of sample quality, RNA integrity, available screening and quantification methods, and biostatistical tools for data interpretation.

The application of molecular biomarkers is a common research field in many different areas, including clinical diagnostics, therapeutic prediction, risk assessment and food safety. By applying molecular biomarkers, different physiological or pathophysiological conditions can be identified in patients, or stages of disease progress can be distinguished. There are numerous molecular levels on which such biomarkers can be determined: from the detection of DNA mutations or methylation patterns (genomics and epi-genomics), over-expression profiling of specific gene transcripts (transcriptomics), to the screening of functional proteins (proteomics) and the deposition of degradation products (metabolomics)[2,4,5].

The primary essential information for each protein is encoded in the genome, where epigenetic modifications have a great influence on the expression profile and expression rate of specific genes. This first step of building a functional protein is the transcription of the specific gene into the coding messenger RNA (mRNA). Beside mRNA, the

transcriptome also includes a huge variety of non-coding RNAs which have regulating functions on protein formation, like tRNAs, rRNAs, miRNAs, piRNAs and long non-coding RNAs. All these expressed transcripts are summarised in the transcriptome[6].

Compared to the technically very complex analysis of post-translational modified proteins or chemically closely related metabolites, today the analysis of the transcriptome is relatively fast and easy. Hence, the development and application of 'transcriptional biomarkers' for different purposes has risen tremendously in clinical diagnostics during the recent years[4,5].

> "In order to detect biomarker candidates, high sample quality and a standardised and highly reproducible quantification workflow are prerequisites"

The 'gold standard' method for the reliable and quantitative analysis of single gene transcripts is still reverse transcription polymerase chain reaction (RT-qPCR). However, if someone wants to perform a holistic screening for a high number of transcripts or even for all transcripts in a biological sample, recently developed next-generation sequencing (RNA-Seq) will be the method of choice. Since RNA-Seq is still relatively expensive to perform on any large sample set, preliminary screening on a subset of representative samples is conducted most of the time. To validate these candidates in all available biological samples, those transcripts can then be quantified and confirmed using RT-qPCR, and if such transcriptomic biomarkers form a valid and stable 'biomarker signature', they are suitable for the implementation in RT-qPCR routine analysis in clinical molecular diagnostic laboratories. It is also an advantage to apply RT-qPCR for molecular diagnostics, since it is fast to perform, relatively cheap, already established in almost all clinical laboratories and if applied according to the MIQE guidelines, valid and highly reproducible.

This article describes the technical requirements for the development of high quality transcriptomic biomarkers according to the MIQE guidelines[3], using RNA Seq and RT-qPCR.

Sampling and RNA quality assessment

An important consideration in biomarker research is the quality of the analysed samples and the nucleic acids contained within. It has already been shown multiple times that the integrity of all RNA transcripts has a tremendous effect on RT-qPCR performance, either for mRNA or microRNA quantification[7,8]. Obtaining high integer RNA starts with the quality of sampling, which needs to be fast and clean to avoid contamination with RNases. Storing tissue samples in formalin-fixed, paraffin-embedded, as it is mostly performed in clinical routine diagnostics, leads to a crosslinking and high degradation rate of RNA and therefore delivers limited or misleading biomarker information[9].

Storing samples in stabilising solution, e.g., PAXgene blood or tissue (PreAnalytix), LeucoLock (Life Technologies), or RNAlater (Ambion), or snap freezing in liquid nitrogen would be the preferred sampling and storage strategy to obtain high quality RNA. It has been shown that the RNA integrity number, which is determined by Agilent Technologies'Agilent 2100 Bioanalyzer, directly correlates with the resulting Cq value and hence with the quantification result. The better the RNA integrity and the higher the RIN value, the more specific the RNA or microRNA molecules that can be quantified in the respective sample[7,8].

Strategy for transcriptomic biomarker discovery

Screening by next generation sequencing and validation by RT-qPCR

There are two main strategies for biomarker detection for a specific physiological status, disease or treatment. On the one hand, there is the so called 'targeted approach', whereby a limited number of well-known and established biomarker candidates that are potentially influenced by the examined condition are quantified. Regarding the analysis of the transcriptome, RT-qPCR will be the first method of choice. Depending on how many biomarker candidates shall be analysed, single gene assays or qPCR arrays can be applied[5].

On the other hand, there is the so called 'untargeted approach', whereby all transcript sequences of a sample will be screened and therefore quantified. This allows for the detection of new and unknown biomarker candidates that would not be recognised by the researcher using the targeted approach. Some years ago, microarray technology was the 'gold standard' for the broad screening of differentially expressed genes. Since next-generation sequencing (NGS) technology, with its array of applications like RNA-Seq or small RNA-Seq, has become more and more popular, it has displaced microarray analysis[10]. Applying NGS, a holistic and very sensitive quantification of all different RNA species is possible. Compared to microarray analysis, it has no upper limit of quantification, nearly no background signal and a higher dynamic range in the expression levels[11].

If a stable and treatment-specific biomarker signature should be established, a high number of biological samples is required to exclude markers that are also dependent on generic factors, like age, gender, stress and nutrition. As screening methods like RNA Seq are still very expensive, the untargeted approach is usually applied only in a small subset of samples. After analysing the data for potential biomarker candidates, they can be validated via single RT-qPCR assays in all available samples[11,12]. Validating small RNA-Seq data using

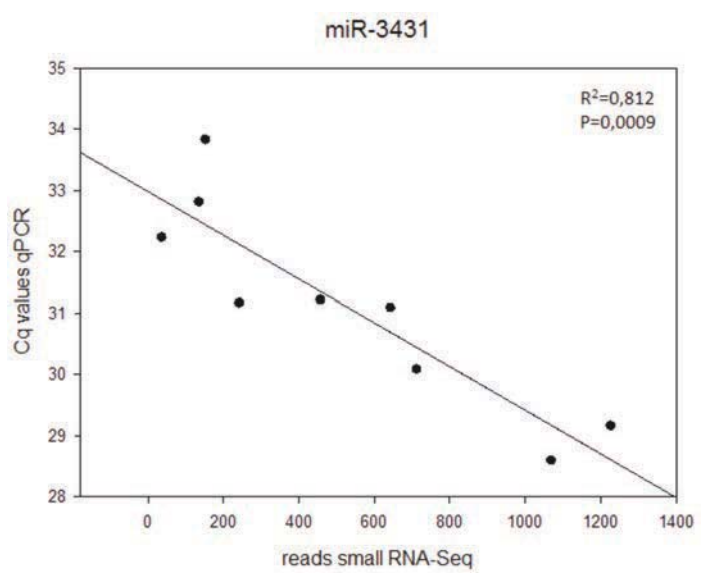

Figure 1: Correlation analysis of miR-3431 expression between results obtained from small RNA-Seq analysis and single miRNA target RT-qPCR analysis

single target RT-qPCR assays showed that the obtained read count correlated well with the Cq values from qPCR (**Figure 1**; page 34).

Biostatistical tools for multivariate data analysis

The physiological answer to a treatment, drug application or disease is usually a complex regulation cascade, whereby the expression of multiple mRNAs and their connected small RNAs are impacted. For the most part, a meaningful expression pattern of significantly regulated gene transcripts is the outcome of transcriptomic biomarker research[13,14]. But to obtain the intended 'unique biomarker signature', highly advanced bioinformatical tools for data visualisation, data comparison, data grouping or treatment cohort separation must be applied. The goal is the visual and statistical separation of the treated 'abnormal', or patho-physiological, from the 'normal' physiological status. So-called multivariate data analysis tools like hierarchical cluster analysis (HCA), heatmaps or principal components analysis (PCA) can be ideally used for this purpose[13].

HCA is a frequently-used tool for the two-dimensional illustration of multiple parameters available for a sample. Within HCA analysis, the expression profile of different samples is divided into subgroups, with the goal to create subsets that share as many common characteristics (transcriptional biomarkers) as possible. The more common the expression profile of two biological samples appears, the nearer they are positioned in the created cluster. Clustering is performed in many repeating cycles within the probands and the best biomarker similarities are combined in one cluster. Finally, this leads to a tree-shaped dendrogram (**Figure 2A**) that displays the distance between the samples based on their individual expression profile of the applied parameters[13,15].

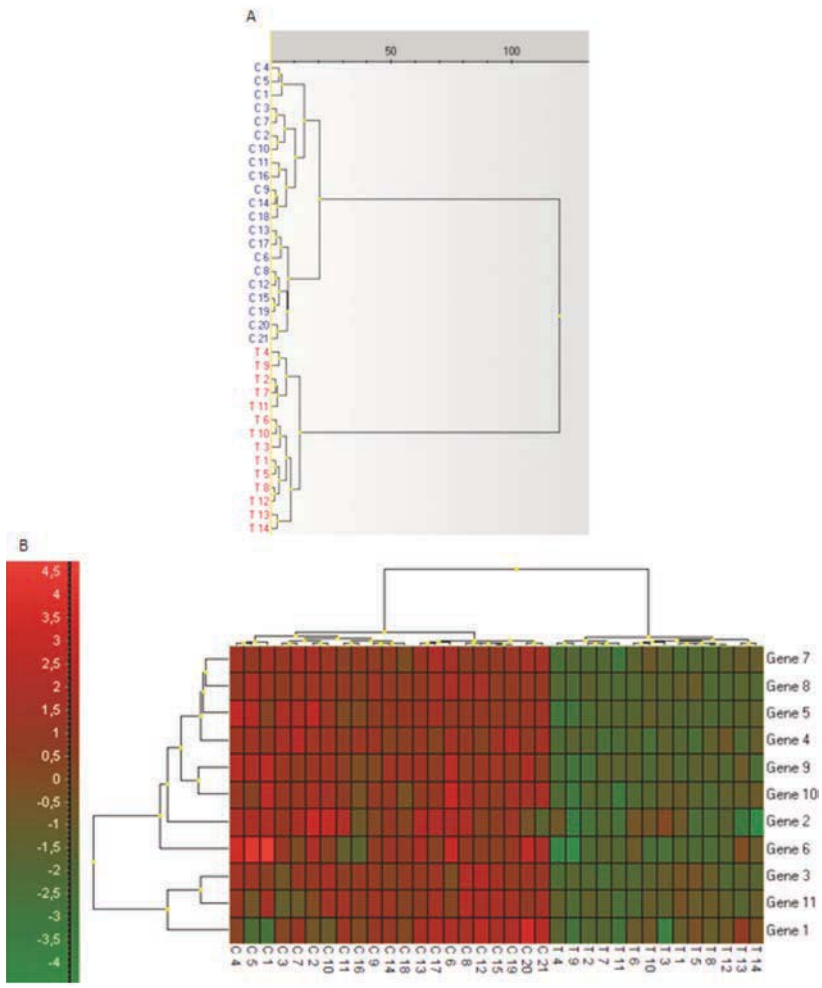

Figure 2: HCA (A) and heatmap (B) analysis of a biomarker research trial with 21 controls and 14 treatment samples. Herein 11 independent gene transcripts were integrated in the analysis applying Genex software[15] (MultiD, Sweden)

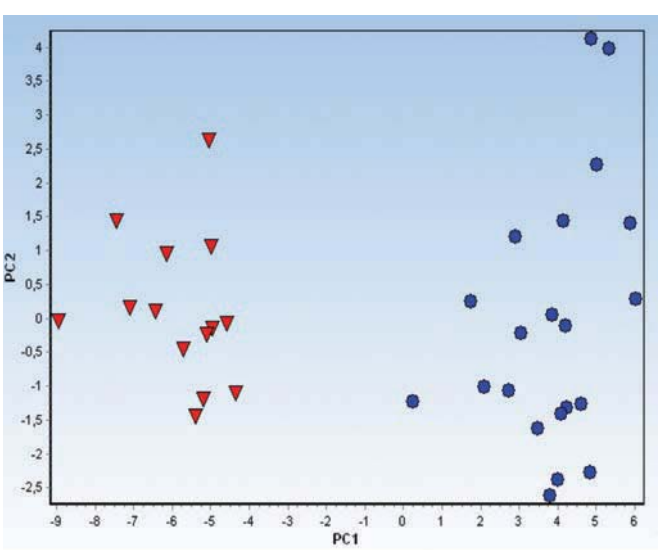

Figure 3: PCA analysis of a biomarker research trial with 21 untreated controls (blue circles) and 14 treated probands (red triangles). Herein, 11 independent gene transcripts were integrated in the analysis applying Genex software[15] (MultiD, Sweden)

A special application of HCA is the creation of a two-dimensional heatmap (**Figure 2B**). Clustering can be performed in parallel for the measured parameters, e.g., quantified transcripts and investigated samples/individuals. Using a heatmap, those two parameters can be combined in one plot, resulting in a colour-coded presentation of the complete experimental matrix (applied parameters vs. samples).

PCA is a further biostatistical method for visualising the affiliation of a sample to a group based on the similarity of specific parameters. PCA is also based on a statistical procedure, able to convert big multi-dimensional data sets into two-, or three-dimensional variables called principal components[13,16,17]. Using PCA in transcriptomic biomarker discovery, the classification of samples is mostly based on the expression values obtained from NGS or high-throughput RT-qPCR expression profiling experiments, whereat each sample/individual is represented by one data point on the PCA graph (**Figure 3**).

Summary and conclusion

The use of transcriptomic biomarkers has already entered different fields of clinical research. Obtaining reliable and reproducible results is most important in developing valid biomarker signatures[2,5]. High sample quality and high RNA integrity is a first essential step to

reach this goal. For the detection and expression profiling of transcriptomic biomarkers, different general quantification strategies are available. Either the expression of a set of predefined genes is quantified using RT-qPCR or a holistic approach is taken to monitor the whole transcriptome of a biological sample, applying RNA-Seq[10]. Regardless of which of those strategies is followed, the result is ideally a set of candidate genes, whose expression is changed. To get the intended information out of such a biomarker set, multivariate data analysis tools, like HCA, PCA or heatmaps are very helpful. In the research field of establishing new sensitive detection methods for drug abuse in veterinary molecular diagnostics, those biostatistical methods have already been successfully employed[18].

Michael W. Pfaffl studied Agriculture with a focus on Animal Science in 1986 at the Technical University of Munich (TUM). His second TUM university degree, in Biotechnology, was performed in parallel with his PhD. In 1997, he obtained his PhD in Molecular Physiology, in the field of molecular muscle and growth physiology at TUM, at the Chair of Physiology. In June 2003, he completed his Habilitation (Dr. habil.) at Center of Life and Food Sciences Weihenstephan with the title *'Livestock transcriptomics: Quantitative mRNA analytics in molecular endocrinology and mammary gland physiology'*. In early 2010 he became Professor of Molecular Physiology at the TUM in Freising. Today, he has reached the Principal Investigator status at the Institute of Physiology and is one of the leading scientists concerning RT-qPCR technology and its data analysis in mRNA and small-RNA expression profiling. In 2004 he founded, together with Sylvia Pfaffl (MBA), the Biotec Marketing, Communication and Consulting company bioMCC. Contact Michael at: Michael.Pfaffl@wzw.tum.de.

References

References

1. Hulka BS (1990) Overview of biological markers. In: Biological markers in epidemiology (Hulka BS, Griffith JD, Wilcosky TC, eds), pp 3–15. New York: Oxford University Press
2. Atkinson AJ (2001) NCI-FDA Biomarkers Definitions Working Group; Biomarkers and surrogate endpoints: preferred definitions and conceptual framework; Clin. Pharmaco. Ther. 69: 89–95
3. Bustin, SA, Benes, V, Garson, JA, Hellemans, J, Huggett, J, Kubista, M, Mueller, R, Nolan, T, Pfaffl, MW, Shipley, GL, Vandesompele, J, Wittwer, CT. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem., 2009, 55
4. Sewall CH, Bell DA, Clark GC, Tritscher AM, Tully DB, Vanden Heuvel J, Lucier GW (1995) Induced gene transcription: implications for biomarkers. Clin Chem. 12(2): 1829-1834
5. Riedmaier, I. and Pfaffl, MW. Transcriptional biomarkers – high throughput screening, quantitative verification, and bioinformatical validation methods. Methods, 2013, 59
6. Kiss T.: Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO J. 2001, 20(14): 3617-3622
7. Becker, C, Hammerle-Fickinger, A, Riedmaier, I, Pfaffl, MW. mRNA and microRNA quality control for RT-qPCR analysis. Methods, 2010, 50
8. Fleige, S and Pfaffl, MW. RNA integrity and the effect on the real-time qRT-PCR performance. Mol. Aspects Med., 2006, 27
9. Kalmar, A, Wichmann, B, Galamb, O, Spisak, S, Toth, K, Leiszter, K, Tulassay, Z, Molnar, B. Gene expression analysis of normal and colorectal cancer tissue samples from fresh frozen and matched formalin-fixed, paraffin-embedded (FFPE) specimens after manual and automated RNA isolation. Methods, 2013, 59
10. Wang, Z, Gerstein, M, Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 2009, 10
11. Riedmaier, I, Benes, V, Blake, J, Bretschneider, N, Zinser, C, Becker, C, Meyer, HH, Pfaffl, MW. RNA-sequencing as useful screening tool in the combat against the misuse of anabolic agents. Anal. Chem., 2012, 84
12. Riedmaier, I, Becker, C, Pfaffl, MW, Meyer, HH. The use of omic technologies for biomarker development to trace functions of anabolic agents. J. Chromatogr. A, 2009, 1216
13. Bergkvist, A, Rusnakova, V, Sindelka, R, Garda, JM, Sjogreen, B, Lindh, D, Forootan, A, Kubista, M. Gene expression profiling – Clusters of possibilities. Methods, 2010, 50
14. Kubista, M, Andrade, JM, Bengtsson, M, Forootan, A, Jonak, J, Lind, K, Sindelka, R, Sjoback, R, Sjogreen, B, Strombom, L, Stahlberg, A, Zoric, N. The real-time polymerase chain reaction. Mol. Aspects Med., 2006, 27
15. GenEx qPCR data analysis software version 5.0 (MultiD, Gothenburg, Sweden)
16. Lee, G, Rodriguez, C, Madabhushi, A. Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. IEEE/ACM. Trans. Comput. Biol. Bioinform., 2008, 5
17. Beyene, J, Tritchler, D, Bull, SB, Cartier, KC, Jonasdottir, G, Kraja, AT, Li, N, Nock, NL, Parkhomenko, E, Rao, JS, Stein, CM, Sutradhar, R, Waaijenborg, S, Wang, KS, Wang, Y, Wolkow, P. Multivariate analysis of complex gene expression and clinical phenotypes with genetic marker data. Genet. Epidemiol., 2007, 31 Suppl 1
18. Riedmaier, I, Pfaffl, MW, Meyer, HH. The physiological way: monitoring RNA expression changes as new approach to combat illegal growth promoter application. Drug Test. Anal., 2012, 4 Suppl 1