Methods 59 (2013) 3–9

Contents lists available at SciVerse ScienceDirect

Methods

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

Transcriptional biomarkers – High throughput screening, quantitative verification, and bioinformatical validation methods

Irmgard Riedmaier *, Michael W. Pfaffl

Physiology Weihenstephan, Technische University Munich, Weihenstephaner Berg 3, 85354 Freising, Germany
ZIEL Research Center for Nutrition and Food Sciences, Technische University Munich, Weihenstephaner Berg 3, 85354 Freising, Germany

ARTICLE INFO

Article history:
Available online 4 September 2012

Communicated by Kenneth Adolph

Keywords:
Transcriptional biomarkers
Omic technologies
Transcriptomics
Bioinformatics

ABSTRACT

Molecular biomarkers found their way into many research fields, especially in molecular medicine, medical diagnostics, disease prognosis, risk assessment but also in other areas like food safety. Different definitions for the term biomarker exist, but on the whole biomarkers are measureable biological molecules that are characteristic for a specific physiological status including drug intervention, normal or pathological processes. There are various examples for molecular biomarkers that are already successfully used in clinical diagnostics, especially as prognostic or diagnostic tool for diseases.

Molecular biomarkers can be identified on different molecular levels, namely the genome, the epigenome, the transcriptome, the proteome, the metabolome and the lipidome. With special "omic" technologies, nowadays often high throughput technologies, these molecular biomarkers can be identified and quantitatively measured.

This article describes the different molecular levels on which biomarker research is possible including some biomarker candidates that have already been identified. Hereby the transcriptomic approach will be described in detail including available high throughput methods, molecular levels, quantitative verification, and biostatistical requirements for transcriptional biomarker identification and validation.

1. Introduction

Identification of molecular biomarkers to distinguish physiological conditions or clinical stages is an emerging research field that has grown substantially during the last years. The main fields in which molecular biomarker research is performed are clinical diagnostics, risk assessment, and therapeutic areas, but also in other fields like food safety, where the request for biomarkers is coming into focus [1]. Within the National Institute of Health, a special "biomarkers definition working group" exists which defined the term biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [2]. Biomarkers are classified in few groups as FDA described below (FDA, qualification process for drug development tools, 2010; http://www.fda.gov):

A prognostic biomarker is a baseline patient or disease characteristic that categorizes individuals by degree of risk for disease occurrence or progression. Prognostic biomarkers informs about the natural history of the disorder in that particular patient in the absence of a therapeutic intervention.

A predictive biomarker is a baseline characteristic that categorizes individuals by their likelihood for response to a particular drug treatment. Such a predictive biomarker is used to identify whether a given individual is likely to respond to a treatment intervention in a particular way. It may predict a favorable response or an unfavorable response or adverse event.

A pharmacodynamic or activity biomarker is a dynamic assessment that shows that a biological response has occurred in an individual after having received a therapeutic intervention. These pharmacodynamic biomarkers may be treatment-specific or more broadly informative of disease response after intervention (FDA, qualification process for drug development tools, 2010 http://www.fda.gov) [3].

There are a number of levels on which molecular biomarkers can be identified, from the beginning of functional protein formation until the deposition of degradation products. Protein formation starts with the encoding of the amino acid sequence on genomic DNA. Epigenetics is an additional field that influences the generation, formation, and abundance of mRNA and later proteins by modifying genomic DNA. The versatile transcriptome with all its different components like mRNA, microRNA, short and long non-coding RNAs is the next level on which dynamic changes on

* Corresponding author. Address: Physiology Weihenstephan, TU Muenchen, Weihenstephaner Berg 3, 85354 Freising, Germany. Fax: +49 8161 714204.
E-mail address: irmgard.riedmaier@wzw.tum.de (I. Riedmaier).

1046-2023/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2012.08.012

the molecular level can occur. The functional proteome itself can be analyzed and at least the metabolites generated can act as potential biomarkers. Another new field in molecular biomarker discovery is the analysis of lipids and their metabolites.

Nowadays there are multiple laboratory methodologies available, enabling the analysis of all those putative molecular biomarkers in a high throughput manner. Those methods can be summarized to the "omic" technologies, namely, genomics, epigenomics, transcriptomics, proteomics, metabolomics, and lipidomics. The technological progress in all those "omic" fields allows the identification of molecular biomarkers in a high number of research areas.

Most research in the field of biomarkers is done in molecular medicine, where different applications of biomarkers can be distinguished. A major field for biomarkers is molecular diagnostics, to identify diseases as well as to monitor the progression of a disease [4]. Some simple examples are blood pressure or cholesterol levels as well as viral load e.g. in HIV diagnostics [2]. Those parameters can easily be determined with routine methods available in most diagnostic laboratories. Biomarkers can be used to predict the risk associated with a particular outcome after an event but also the progression of a disease, in terms of a survival rate or survival time. This enables more exact therapeutic decisions, and the reduction of healthcare costs. There are already DNA based molecular biomarkers available to predict the susceptibility of an individual to a disease e.g. by determining the genotype of specific genes [4]. A further field is risk assessment. Molecular biomarkers can help to associate particular risks or prediction to a particular sub-population. It is sometimes defined as "stratification" because it allows separating one sub-population into the total population with a remarkable biological or medical fact. A classical example of such application is the detection of Human Papilloma Virus (HPV) in women. As soon as the HPV infection persists it is predictive for a higher risk of developing precancerous lesions of the cervix, which can progress to invasive cervical cancer [5].

The interest on biomarker development is not restricted to medical diagnostics. Also in the field of food safety the identification of specific contaminations is coming into focus [1,6,7]. In all described fields, high throughput "omic" technologies are useful for the discovery and identification of molecular biomarkers.

2. "Omic" technologies in biomarker research

Independent on which molecular level biomarkers will be analyzed – genome, epigenome, transcriptome, proteome, metabolome, and lipidome – there are different important factors that have to be recognized. The first step is to consider natural occurring biomarker variations within populations triggered by age, sex, species, race, breed, food or feed, immunological status or generally by the environment. Having "normal" or "control" individuals from a big population is very important to define if a certain change in the biomarker concentration is a real change or within the natural variation of the "normal" or "healthy" population. Another important fact is sample integrity and quality which has a tremendous effect on results regardless on which molecular level the markers are examined. Therefore constant sample integrity and the proper quality affirmation are necessary to identify reliable biomarkers. All in all, the more variables and quality control checks are considered the more valid the discovered biomarker will be [8,9].

The genetic information for the formation of functional proteins is encoded on the genomic DNA. This information is fixed after insemination. The analysis of the genetic code and the SNPs from any tissue source is useful to predict an individual's risk to get a specific disease. For example the predisposition for specific kinds of cancer can sometimes be determined by analyzing the sequence of known cancer specific genes. A new and interesting field in genomic diagnostics is the analysis of circulating DNA, which can be extracted from any body fluid, like blood, urine or epithelial swabs [10]. Recent studies showed that minor amounts of fetal DNA fragments can be found in the plasma of pregnant women. Those circulating DNAs allow maternal prenatal diagnostics in a non-invasive form and can be used to identify different genetic diseases in the unborn child, like Down Syndrome, Trisomie 13 or Trisomie 18 at a very early pregnancy phase [11–13].

Another upcoming field is the molecular analysis on the level of the epigenome, highly related to the analysis of the genome. Epigenetics can be defined as the study of heritable changes in gene expression that are not caused by changes in the sequence of genomic DNA, but by changes in DNA methylation or Histone modifications [14]. The expression of specific genes is dependent on the methylation status of its promoter region, whereat methylation of promoter regions leads to transcriptional repression and unmethylated promoters allow the transcription of the gene [15,16]. The methylation status of different promoter regions has already been proven to act as potential biomarker for different diseases. For example methylation analysis of Septin 9 and Vimentin can be used in the diagnosis of colorectal cancer [15,17–20]. Further, CDKN2A is a tumor suppressor which could be shown to be inactivated in lung cancer by promoter methylation [15,21] and SHOX2 analysis can help to distinguish between malignant lung cancer and other non cancerous lung diseases [15]. Overall genome methylation or single gene methylation analysis are a big challenge, because it has to be defined which genetic methylation level is "normal" and which one not, dependent on tissue type and developmental stage. From various studies we know that the methylation in the fetal and prenatal phase affects the development of the unborn and has major impact on the later adult physiology [14]. Interindividual differences in the methylation status of specific genes are present and there are also examples of genes whose promoter methylation increases with age [15] and though identification of stable epigenetic biomarkers will be difficult.

The next step in the formation process of new proteins is the transcription of coding genes leading to the formation of mRNAs. Therefore the analysis of gene expression is the first instance to analyze influences on the molecular level of target cells. Transcription of genes is a very dynamic process being able to adapt rapidly to external, environmental or physiological changes of target tissues, organs or cells [22]. Thus analyzing the expression of genes is a very potential way to identify biomarkers to describe the physiological status, a disease, the exposure to drugs, or other exogenous stimuli.

Nowadays it is known that only 1–2% of the human transcriptome encodes for proteins, thus is transcribed to mRNAs [23]. Other, non-coding RNAs are represented by ribosomal RNAs or transfer RNAs, whose function is already known [24]. But there are further new non-coding RNA families. Non-coding RNAs can be subdivided to small non-coding RNAs shorter than 200 nt and long non-coding RNAs with more than 200 nt [23]. Many of those non-coding RNAs show regulatory functions. The best known small RNA molecules are microRNAs (miRNAs), which are known to interact with mRNAs by complementary base pairing. Dependent on the degree of complementarity, binding of miRNA to the corresponding mRNA leads to inhibition of translation either by inducing degradation of mRNAs (100% complementarity) or by inhibiting translation without degradation (<100% complementarity) [25]. The analysis of miRNAs in biomarker discovery is an upcoming field and there are already several miRNA biomarkers available, e.g. for diabetes, liver disease or cancer [10,26–29]. The group of long non-coding RNAs (lncRNAs) are defined as transcripts with no open reading frame, thus non-protein coding [23] and are also known to have regulatory functions. The most popular lncRNA in humans is Xist, which mediates the inactivation of one X Chromosome in females [23,24], and H19, which is only expressed

from the maternal allele. It is highly expressed during embryonic development in most tissues of vertebrates and is down regulated in most tissues after birth [23,30–32]. H19 has also been shown to be associated with different kinds of cancer [23].

Biomarkers can also be detected on the level of the proteome. The most established and worldwide used proteomic biomarker is human choriongonadotropin (hCG) which is secreted by the placenta and therefore only present in blood or urine from pregnant women, and is thus the perfect marker for early pregnancy [1]. In the diagnosis of prostate cancer the prostate specific antigen (PSA) is used as early prognostic marker [33]. Detecting single proteins via immuno assays or chromatographical methods combined with mass spectrometry is easy once the assay is established. But the proteome is much more complex than for example the genome, because proteins are also characterized by multiple functional group attachments or interactions like protein–DNA or protein–protein [22]. Thus the analysis of the proteome in the biomarker field is much more complex than analyzing DNA or RNA.

Another "omic" technology that is getting more and more into the research focus is metabolomics, the analysis of metabolites whose occurrence is dependent on diseases, physiological status or external stimuli [1]. In former days single metabolites were analyzed using single metabolite assays. Due to technological developments, multiplexed metabolite measurements are possible using chromatography combined with mass spectrometry or NMR spectroscopy [1]. Promising biomarkers for different types of liver diseases are γ-glutamyl dipeptides. It could be shown that analyzing different levels of those peptides enables the discrimination of nine different types of liver diseases [34,35]. A further example is the analysis of L-valine, L-threonine, 3-hydroxybutyric acid, 1-deoxyglucose and glycine which enables the separation of colorectal cancer patients from healthy individuals [35]. Chen et al. [36] could show that 1-methyladenosine is a potential biomarker for early diagnosis of hepatocarcinoma, which would be of high interest, because late diagnosis leads to a high death rate of this disease. In the field of food residue analysis, the analysis of the metabolome is a new approach to detect the misuse of growth promoting agents in animals. A metabolite pattern could be identified by Courant et al. [37] which was useful to separate clenbuterol treated calves from untreated animals. One of those metabolites could be identified as creatine, which is a first potential marker for the abuse of clenbuterol in calves.

The lipidomics, a subgroup of metabolomics, presents a very new "omic" technology studying different lipids profiles and their metabolites in biological systems. As changes in lipid metabolism are related to different diseases and disease states, like in Alzheimer disease [38,39]. This technology is an upcoming and promising field for the identification of biomarkers.

As described, there are different molecular levels on which biomarkers can be investigated and determined. This article will focus on the analysis of the transcriptome describing available methods and examples for the establishment of transcriptional biomarkers for diseases and in the field of food safety.

3. Transcriptional biomarkers

3.1. Available methods to analyze the transcriptome

To analyze the expression of RNAs, different methods are available. These methods can be subdivided to targeted and untargeted methods. Targeted methods enable the analysis of the expression of single RNAs and untargeted methods are applied for a global screen and a huge amount of differentially expressed RNAs in one experiment [7]. Two main technologies for untargeted holistic transcriptome screening dominate the diagnostic field, namely gene expression microarrays and RNA-Sequencing.

Gene expression microarrays allow the analysis of the expression of RNA molecules whose sequence is already known. DNA fragments (probes) representing specific coding regions of a gene are immobilized on the surface [40]. After reverse transcription cDNA is labeled with a fluorescent dye and finally hybridized to the slide. The color coded cDNA binds to the probes via complementary base pairing. As there are multiple probes representing one gene immobilized on the slide, the amount of bound cDNAs can be determined by the intensity of the fluorescent signal. Specific software tools enable the analysis of the absolute or relative gene expression, depending on the applied microarray technology [40]. With microarray technology, genes whose sequence is known can be analyzed exon specific and down to various splice variants. But a high background level and the limited dynamic range of detection leads to less sensitivity and thus the expression of low abundant genes is difficult to determine [41].

A relatively new, more sensitive method is RNA-Sequencing (RNA-Seq), one application of next generation sequencing platforms. Thereby a cDNA library is produced from extracted RNA and all cDNA fragments are sequenced in a parallel high throughput manner. Using high sophisticated software tools, the generated number of short sequence reads will either be aligned to a reference genome or assembled *de novo* without having any genomic sequence available [41,42]. This holistic method is more sensitive than microarrays because it has no upper limit of quantification, shows a higher dynamic range of expression levels and has nearly no background signal. RNA-Seq can be applied for mRNAs or shorter microRNAs and is sensitive enough to enable the detection of "one single RNA molecule" [41].

To analyze a defined number of genes, targeted methods are useful. Northern Blot was the first method available for the targeted analysis of expressed genes. Within this method, RNA is separated via agarose gel electrophoresis and then transferred to a positive charged blotting membrane. Single stranded nucleic acids, either DNA or RNA, with the complementary sequence of the target RNA are labeled either with radioactive isotopes, chemiluminescent markers or with fluorescent dyes. With those hybridisation probes the target RNA can be identified [43].

If an exact RNA quantification in a biological sample is required, reverse transcription followed by quantitative polymerase-chain-reaction (RT-qPCR) will be the preferred method of choice. RT-qPCR is an advancement of original polymerase-chain-reaction (PCR), a method to amplify a defined unit of DNA using DNA polymerases. The method of RT-qPCR uses the action of fluorescent dyes to monitor the amplification course of the reaction [44]. Monitoring can either occur via non-specific fluorescent dyes, like SYBR Green I. Those dyes intercalate with newly generated double stranded DNA after amplification, which causes an increase of the fluorescence dye signal. Measuring the fluorescence after each PCR cycle allows the monitoring of the increase of the amount of DNA and specific analytical strategies enable the calculation of the amount of starting material [45,46]. As those dyes bind to all double stranded DNAs, differentiation of the target gene from other unspecifically amplified DNA strands or primer dimers is not possible. Another, more specific way is the use of labeled DNA probes, giving a fluorescent signal upon binding to the specific gene [45]. The use of those probes also allows the quantification of a panel of genes in one PCR reaction. Therefore different fluorescent dyes which emit light that is measurable at different wave lengths are used to label the DNA probes specific for different genes [47].

3.2. mRNA biomarkers

The search for mRNA biomarkers is already an established method in different life science fields. In pharmacogenomics it was successfully applied to establish treatment prediction with

specific drugs. Hereby the expression of drug sensitive and specific genes was analyzed to predict, if treatment with a specific drugs will be promising for the respective individual [22,48]. Using mRNA gene expression analysis is also helpful in the valid differentiation of types or stages of diseases. Thus different forms of heart disease, cancer or neuropsychiatric disorders can be distinguished by analyzing the expression of specific genes [22,48].

The search for gene expression biomarkers is also entering the field of food safety and residue analysis in food, especially the analysis of growth promoting agents. Hereby physiological changes, caused by treatment with anabolic agents are detected on the level of the transcriptome and those differentially expressed genes should act as first biomarker candidates [1,6]. There are different publications available dealing with the analysis of the transcriptome after treatment with anabolic agents. Gene expression was analyzed in reproductive organs like testis, uterus, ovary or vaginal epithelial cells, in muscle tissue, liver and blood following administration of different anabolic substances [49–53]. Up to now, no single marker gene could be identified for a single substance or an anabolic substance cocktail in any organ or tissue [6]. In most studies, a number of genes whose expression was influenced by treatment could be identified. Hence the identification of a biomarker pattern consisting of various expressed genes will be more promising than finding stand alone single markers. Some authors deal with the use of biostatistical tools for pattern recognition, e.g. principal components analysis or cluster analysis to visualize separation between treated and untreated individuals [49–53]. Hence the identification of biomarker patterns for the identification of illegally treated individuals seems to be very promising.

3.3. miRNA biomarkers

Instead of the classical analysis of mRNA the quantitative analysis of miRNA is more and more used for biomarker establishment. miRNAs are small non-coding RNA molecules with about 20–22 nucleotides which are involved in post-transcriptional processing of mRNA. In this way they are able to regulate physiological pathways and metabolic processes [54] and therefore impact the entire cellular physiology, organ development, and tissue differentiation. Most miRNAs are known to be expressed in a physiological-, tissue-, and disease-specific manner [25]. Due to their short length they are less sensitive to RNase exposure and hence are more stable than the longer mRNA with an average length of 2 kb [25]. It is already proven that miRNAs have the potential in the diagnosis of specific types of cancer. For example tissue derived from gastrointestinal cancer can be differentiated from non-gastrointestinal cancer tissue by analyzing specific miRNA profiles [26]. As also described for mRNA, the miRNA profile characterization gives insights in the progression of specific diseases or the response to a given therapeutic approach [55,56].

The expression of miRNAs cannot only be measured in tissue or cell culture samples, they are also present in body fluids, like urine, blood or even milk [57–60]. Some of those circulating miRNAs are already known to be specific disease markers, especially for different forms of cancer [61–63]. It has been shown, that miR-141 is a potential plasma marker for prostate cancer [61]. There exists the hypothesis that tumour cells secrete micro vesicles containing miRNAs into the blood stream and therefore those circulating miRNAs are very potential biomarkers in the field of cancer diagnostics [58].

As already described for mRNAs, miRNAs are also entering the field of food safety and residue analysis. In a first study, the expression of a panel of miRNAs was analyzed in bovine liver after the administration of growth promoters (trenbolone acetate plus estradiol) using RT-qPCR arrays. Various miRNAs could be shown to be highly regulated (up-regulated: miR-34a, miR-181c, miR-20a, and miR-15a; down-regulated: miR-29c, miR-130a, and miR-103) in the liver of treated animals. Combining those results with those already obtained for mRNA expression in the same biological samples, treated animals could be separated from untreated individuals using PCA [64]. This study shows that the combined expression analysis of mRNA and miRNA and their parallel data analysis is a potential approach for the identification of biomarker patterns for the treatment with growth promoters.

3.4. lncRNA biomarkers

Non-coding RNAs with a length of more than 200 nt belong to the group of long non-coding RNAs (lncRNAs). In biomarker research the group of lncRNAs is coming into focus, especially in cancer research. Due to its regulatory functions, different potential lncRNA biomarker candidates are already available.

One of the first identified lncRNAs, H19, is a biomarker for tumors of the esophagus, liver, bladder, colon, and for metastases in the liver. A loss of methylation in its promoter region leads to a strong up-regulation of this lncRNA, indicating the presence of tumor tissue [23,65–67]. Another marker lncRNA is HOTAIR, which is a prognostic marker and a marker of cancer invasiveness as it is up to 2000-fold up regulated in primary and metastatic breast cancer tissue compared to normal breast tissue. High levels of HOTAIR are correlated with metastasis and a poor survival rate for the patient [23,68]. A further candidate is MEG 3, which is expressed in a high number of tissues in humans whereat it is highly expressed in the brain and the pituitary gland [23,69,70]. In various brain cancer types, MEG3 is not detectable at all and in several human cancer cell lines MEG3 has shown to suppress cell growth, indicating a role as tumor suppressor [23,70].

Those examples indicate that the role of RNAs is not only to be an intermediate state between genes and proteins but also show regulatory functions giving them the potential to act as transcriptional biomarkers in various diagnostic and research fields.

4. Biostatistical tools for biomarker identification

Valid biomarkers hold promise for the increasing success rates of clinical trials. But the biomarker discovery requires an intensive search across a broad spectrum of molecular data [71]. The identification of single biomarkers on the mRNA or the miRNA level is not possible in most pathological disorders. In such cases a set of multiple biomarkers must be present to distinguish between specific disease types, disease states or applied treatments [72]. The first important question is how to deal with these high-throughput data sources to get the desired information? One approach is the integrative data analysis of multiple biomarker levels, herein regulated genes and their characteristic dynamic regulation on various transcriptomic levels (mRNA, miRNA, and lncRNA). This could help to generate an integrative gene expression pattern [72,73]. Hence the second important question is how to get the right transcribed marker set out of this integrative pattern and finally find the specific transcriptional biomarker pattern that is visible at a glance? The best way seems to be the construction of clusters where all elements (individuals or patients) have similar characteristics. The data integration has made strides in developing management and analysis tools for structured biological data, but best practices are still evolving for the integration of high-throughput data with less structured data from experimental or clinical studies [71]. To achieve the goal, different multivariate analysis methods are available, which are used for biomarker selection and validation, namely hierarchical cluster analysis (HCA) and principal components analysis (PCA) (Fig. 1). With such bioinformatical tools the visualization of togetherness, treatment groups or expression patterns, in a two or three dimensional graph are possible [1,74].

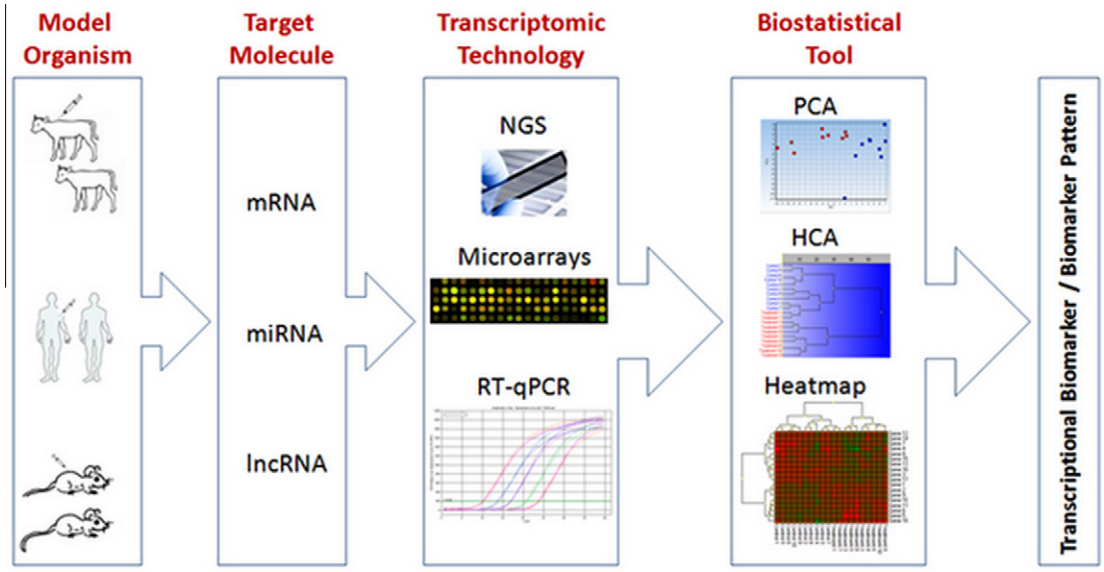

Fig. 1. Workflow for the identification of transcriptomic biomarkers.

Some recently published studies used PCA and HCA successfully to visualize a separation between treated and untreated individuals by evaluating significantly regulated genes [49–53,64].

4.1. Hierarchical cluster analysis

The preferred used method for the visualization of treatment groups and/or expression patterns is hierarchical clustering. The real advantage of hierarchical clustering compared to the direct visualization methods is that a high dimensionality of the data set is reduced to a convenient two-dimensional representation of subject similarities [75,76]. Thereby multiple measured gene transcripts and numerous biological individuals in various treatment groups can be analyzed in an easy way. HCA is able to classify biological samples according to their expression profile into different clusters, or more precisely, the partitioning of a data set into subsets. The goal of clustering is to create subsets that share common trait that is a matchable expression pattern. The more comparable the expression pattern of the analyzed samples is, the closer the position in the cluster. Hierarchical clustering can be performed either for the expressed gene transcripts or for the other dimensions represented by biological samples or treatment groups. HCA uses distance measure to identify pairs of individuals showing high similarity based on the gene expression pattern. Within many steps, those with the highest similarity are merged in a cluster, and then the process is repeated. The result of the analysis is a tree dendrogram displaying the distances between the individuals based on the similarity of the transcribed biomarkers [76,77]. Using hierarchical clustering a tree dendrogram can either be designed for the measured genes (in all biological samples) or for the samples (based on all measured expressed genes). Using a heatmap analysis these two classification approaches can be fused, resulting in a two-dimensional, mostly color-coded description of the whole experimental matrix (genes × samples). It displays in a very convenient way all samples versus gene expression where each tile is colored with a different intensity according to all available data.

4.2. Principal components analysis

A further useful biostatistical visualization method to group expression data and to validate transcriptional biomarkers is principal components analysis (PCA). Comparable to HCA, the PCA is a mathematical procedure that converts a highly complex multidimensional data set into a lower number of variables called principal components (PC) [45,77]. Mostly a two or three dimensional graphical data output is favored. The classification of the genes is based on unscaled gene expression data (mostly Cq values) or the overall fold change of the gene expression magnitudes [78]. Each analyzed individual will be represented by one spot in the graphical output file. PCA has effectively been employed to visualize a treatment pattern in veterinary science in bovines [49–53,64]. Applying the integrative approach of analyzing transcribed biomarkers on multiple layers, herein for mRNA and miRNA, a more significant expression pattern and a better separation between the treatment groups could be achieved. A clear separation of the two treatment groups indicates that PCA is a good tool for pattern recognition in gene expression biomarker research.

The advantage of PCA in comparison to HCA methods is obvious. PCA allows a much clearer recognition and more precise differentiation of the treatment groups, because the commonalities in gene expression pattern are visualized by the symbol interspaces in two dimensions.

4.3. Software tools

Various software tools are available to perform highly sophisticated multi-dimensional gene expression data analysis, including HCA and PCA, to identify and validate transcribed biomarkers. On the one hand stand-alone software can be purchased or freeware packages are available on the internet.

The "Genex" software package (MultiD, Gothenburg, Sweden) offers a lot of applications mainly dedicated to real-time PCR data analysis, to identify and validate transcribed biomarkers on the mRNA and miRNA level. Genex further supports the correct qPCR data analysis in a MIQE compliant way [76,79] (http://Genex.-gene-quantification.info).

The "Genevestigator" software tool (Nebion, Zürich, Switzerland) aims to detect specific expression patterns in a multi-dimensional expression space. On the basis of a huge number of expression data and physiological conditions processed from thousands of Affymetrix microarrays. The intuitive interface allows obtaining lists of potential biomarker candidate genes that can then be validated using further implemented Genevestigator analysis tools or manually in future experimental trials. Genevestigator

provides several clustering tools for array analysis or meta-profiles, while the similarity expression pattern is measured across arrays or physiological conditions [8,9]. In addition a new advanced biclustering method allows identifying groups of genes that have similar profiles in a subset of conditions, irrespective of their profile similarity in the other physiological conditions. Recent studies have shown that biclustering performs better than methods that require similarity over all conditions [8,9] (http://www.genevestigator.com).

A further free bioinformatical approach to discover and validate expressed biomarkers is to use R programming language, which is summarized in the "Bioconductor" project database (http://www.bioconductor.org). Bioconductor is an open source, open development software project to provide tools for the analysis and comprehension of high-throughput data [80]. There are multiple R packages and meta-data packages available, which provide the analysis of various data sources, on the genomic, epigenomic, and transcriptomic level. The broad goal is to provide widespread access to a full range of powerful statistical and graphical methods for data analysis. In our particular purpose to analyze transcribed biomarkers the following packages are interesting. For quantitative real-time PCR data analysis and normalization a bundle of free projects are available, e.g. "HTqPCR", "qpcrNorm", "SLqPCR", or "ddCT" (summarized in "The qpcR library – Analysis of real-time PCR data using R" – http://www.dr-spiess.de/qpcR.html) [81]. Further specialized packages for multi-dimensional expression analysis, PCA, HCA or biomarker discovery are available in the database, e.g. "BioMark" or "optBiomarker" project (http://www.bioconductor.org/help/search/index.html?q=biomarker).

Comparing the described software packages, "Genex" and "Genevestigator" are working on a windows based environment and are therefore more intuitive and user friendly. Advantage of the "Bioconductor" packages is that they are freeware, but expect an advanced operator who is able to handle and modify the often very complex text based input script lines. In "Bioconductor" the graphical output of the results is limited and very rudiment in appearance.

5. Conclusions

The demand for the development of molecular biomarkers in a high number of diagnostic and prognostic fields has grown during the last years. The discovery is based on changes on different molecular levels due to treatment with drugs, changes in the physiological status, disease state or other pathological processes. The development of new high throughput methods for the analysis of biological molecules has enabled screening for those changes on different molecular levels. Detecting single specific biomarkers like hCG is only possible in exceptional cases. Today unique biomarker patterns are used for valid identification. Since the transcription of genes is a very dynamic process and being able to adapt rapidly to environmental, physiological or pathological changes, the transcriptome is preferable used for the identification of transcribed biomarkers. To validate the identified transcribed biomarker set with highly significance, the application of bioinformatical tools is necessary. Using biostatistical methods for dimensionality reduction and pattern recognition, a clear separation between "normal" and "treated" or "diseased" can be achieved and thereby confirm the identified transcriptional biomarkers. The workflow for transcriptomic biomarker research is summarized in Fig. 1.

References

[1] I. Riedmaier, C. Becker, M.W. Pfaffl, H.H. Meyer, J. Chromatogr. A 1216 (2009) 8192–8199.
[2] S.E. Ilyin, S.M. Belkowski, C.R. Plata-Salaman, Trends Biotechnol. 22 (2004) 411–416.
[3] W. Tang, Z. Hu, H. Muallem, Pharmacogenomics Personalized Med. 4 (2011) 95–107.
[4] R. Mayeux, NeuroRx 1 (2004) 182–188.
[5] E.E. Moore, J.D. Wark, J.L. Hopper, B. Erbas, S.M. Garland, Twin. Res. Hum. Genet. 15 (2012) 79–86.
[6] I. Riedmaier, M.W. Pfaffl, H.H. Meyer, Drug Test. Anal. 3 (2011) 676–681.
[7] G. Pinel, S. Weigel, Trends Anal. Chem. 29 (2010) 1269–1280.
[8] P. Zimmermann, O. Laule, J. Schmitz, T. Hruz, S. Bleuler, W. Gruissem, Mol. Plant 1 (2008) 851–857.
[9] T. Hruz, O. Laule, G. Szabo, F. Wessendorp, S. Bleuler, L. Oertle, P. Widmayer, W. Gruissem, P. Zimmermann, Adv. Bioinf. 2008 (2008) 420747.
[10] H. Schwarzenbach, D.S. Hoon, K. Pantel, Nat. Rev. Cancer 11 (2011) 426–437.
[11] J.A. Canick, E.M. Kloza, G.M. Lambert-Messerlian, J.E. Haddow, M. Ehrich, B.D. van den, A.T. Bombard, C. Deciu, G.E. Palomaki, Prenat. Diagn. (2012) 1–5.
[12] J.A. Canick, G.E. Palomaki, J. Med. Screen. 19 (2012) 57–59.
[13] G.E. Palomaki, C. Deciu, E.M. Kloza, G.M. Lambert-Messerlian, J.E. Haddow, L.M. Neveux, M. Ehrich, Genet. Med. 14 (2012) 296–305.
[14] J. Qiu, Nature 441 (2006) 143–145.
[15] T. Mikeska, C. Bock, H. Do, A. Dobrovic, Expert. Rev. Mol. Diagn. 12 (2012) 473–487.
[16] J.G. Herman, S.B. Baylin, N. Engl. J. Med. 349 (2003) 2042–2054.
[17] T. Devos, R. Tetzner, F. Model, G. Weiss, M. Schuster, J. Distler, K.V. Steiger, R. Grutzmann, C. Pilarsky, J.K. Habermann, P.R. Fleshner, B.M. Oubre, R. Day, A.Z. Sledziewski, C. Lofton-Day, Clin. Chem. 55 (2009) 1337–1346.
[18] M. Tanzer, B. Balluff, J. Distler, K. Hale, A. Leodolter, C. Rocken, B. Molnar, R. Schmid, C. Lofton-Day, T. Schuster, M.P. Ebert, PLoS ONE 5 (2010) e9061.
[19] J.D. Warren, W. Xiong, A.M. Bunker, C.P. Vaughn, L.V. Furtado, W.L. Roberts, J.C. Fang, W.S. Samowitz, K.A. Heichman, BMC Med. 9 (2011) 133.
[20] W.D. Chen, Z.J. Han, J. Skoletsky, J. Olson, J. Sah, L. Myeroff, P. Platzer, S. Lu, D. Dawson, J. Willis, T.P. Pretlow, J. Lutterbaugh, L. Kasturi, J.K. Willson, J.S. Rao, A. Shuber, S.D. Markowitz, J. Natl Cancer Inst. 97 (2005) 1124–1132.
[21] A. Merlo, J.G. Herman, L. Mao, D.J. Lee, E. Gabrielson, P.C. Burger, S.B. Baylin, D. Sidransky, Nat. Med. 1 (1995) 686–692.
[22] A.K. Sandvik, B.K. Alsberg, K.G. Norsett, F. Yadetie, H.L. Waldum, A. Laegreid, Clin. Chim. Acta 363 (2006) 157–164.
[23] E.A. Gibb, C.J. Brown, W.L. Lam, Mol. Cancer 10 (2011) 38.
[24] C.P. Ponting, P.L. Oliver, W. Reik, Cell 136 (2009) 629–641.
[25] C. Becker, A. Hammerle-Fickinger, I. Riedmaier, M.W. Pfaffl, Methods 50 (2010) 237–243.
[26] J. Lu, G. Getz, E.A. Miska, Nature 435 (2005) 834–838.
[27] A. Esquela-Kerscher, F.J. Slack, Nat. Rev. Cancer 6 (2006) 259–269.
[28] A.K. Pandey, P. Agarwal, K. Kaur, M. Datta, Cell. Physiol. Biochem. 23 (2009) 221–232.
[29] T.A. Farazi, J.I. Spitzer, P. Morozov, T. Tuschl, J. Pathol. 223 (2011) 102–115.
[30] J.C. Castle, C.D. Armour, M. Lower, D. Haynor, M. Biery, H. Bouzek, R. Chen, S. Jackson, J.M. Johnson, C.A. Rohl, C.K. Raymond, PLoS ONE 5 (2010) e11779.
[31] F. Poirier, C.T. Chan, P.M. Timmons, E.J. Robertson, M.J. Evans, P.W. Rigby, Development 113 (1991) 1105–1114.
[32] O. Lustig, I. Ariel, J. Ilan, E. Lev-Lehman, Mol. Reprod. Dev. 38 (1994) 239–246.
[33] M. Djulbegovic, M.M. Neuberger, P. Dahm, N. Engl. J. Med. 366 (2012) 2228–2229.
[34] T. Soga, M. Sugimoto, M. Honma, M. Mori, K. Igarashi, K. Kashikura, S. Ikeda, A. Hirayama, T. Yamamoto, H. Yoshida, M. Otsuka, S. Tsuji, Y. Yatomi, T. Sakuragawa, H. Watanabe, K. Nihei, T. Saito, S. Kawata, H. Suzuki, M. Tomita, M. Suematsu, J. Hepatol. 55 (2011) 896–905.
[35] A. Zhang, H. Sun, X. Wang, Anal. Bioanal. Chem. (2012).
[36] F. Chen, J. Xue, L. Zhou, S. Wu, Z. Chen, Anal. Bioanal. Chem. 401 (2011) 1899–1904.
[37] F. Courant, G. Pinel, E. Bichon, F. Monteau, J.P. Antignac, B.B. Le, Analyst 134 (2009) 1637–1646.
[38] R.B. Chan, T.G. Oliveira, E.P. Cortes, L.S. Honig, K.E. Duff, S.A. Small, M.R. Wenk, G. Shui, P.G. Di, J. Biol. Chem. 287 (2012) 2678–2688.
[39] M.R. Wenk, Cell 143 (2010) 888–895.
[40] K.M. Kurian, C.J. Watson, A.H. Wyllie, J. Pathol. 187 (1999) 267–271.
[41] Z. Wang, M. Gerstein, M. Snyder, Nat. Rev. Genet. 10 (2009) 57–63.
[42] V. Costa, C. Angelini, I. De Feis, A. Ciccodicola, J. Biomed. Biotechnol. 2010 (2010) 853916.
[43] J.C. Alwine, D.J. Kemp, G.R. Stark, Proc. Natl. Acad. Sci. USA 74 (1977) 5350–5354.
[44] R. Higuchi, G. Dollinger, P.S. Walsh, R. Griffith, Nat. Biotechnol. 10 (1992) 413–417.
[45] M. Kubista, J.M. Andrade, M. Bengtsson, A. Forootan, J. Jonak, K. Lind, R. Sindelka, R. Sjoback, B. Sjogreen, L. Strombom, A. Stahlberg, N. Zoric, Mol. Aspects Med. 27 (2006) 95–125.
[46] C.J. Smith, A.M. Osborn, FEMS Microbiol. Ecol. 67 (2009) 6–20.
[47] R.N. Gunson, S. Bennett, A. Maclean, W.F. Carman, J. Clin. Virol. 43 (2008) 372–375.
[48] D. Seo, G.S. Ginsburg, Curr. Opin. Chem. Biol. 9 (2005) 381–386.
[49] C. Becker, I. Riedmaier, M. Reiter, A. Tichopad, M.J. Groot, A.A. Stolker, M.W. Pfaffl, M.F. Nielen, H.H. Meyer, J. Steroid Biochem. Mol. Biol. 125 (2011) 192–201.
[50] C. Becker, I. Riedmaier, M. Reiter, A. Tichopad, M.W. Pfaffl, H.D. Meyer Heinrich, Horm. Mol. Biol. Clin. Invest. 2 (2011) 257–265.
[51] I. Riedmaier, A. Tichopad, M. Reiter, M.W. Pfaffl, H.H. Meyer, Anal. Chim. Acta 638 (2009) 106–113.
[52] I. Riedmaier, M. Reiter, A. Tichopad, M.W. Pfaffl, H.H. Meyer, Exp. Clin. Endocrinol. Diabetes 119 (2011) 86–94.

[53] J.C. Rijk, A.A. Peijnenburg, P.J. Hendriksen, J.M. Van Hende, M.J. Groot, M.W. Nielen, BMC Vet. Res. 6 (2010) 44.
[54] R.C. Lee, R.L. Feinbaum, V. Ambros, Cell 75 (1993) 843–854.
[55] V. Benes, M. Castoldi, Methods 50 (2010) 244–249.
[56] H.M. Heneghan, N. Miller, M.J. Kerin, Curr. Opin. Pharmacol. 10 (2010) 543–550.
[57] O.F. Laterza, L. Lim, P.W. Garrett-Engele, K. Vlasakova, N. Muniappa, W.K. Tanaka, J.M. Johnson, J.F. Sina, T.L. Fare, F.D. Sistare, W.E. Glaab, Clin. Chem. 55 (2009) 1977–1983.
[58] N. Kosaka, H. Iguchi, T. Ochiya, Cancer Sci. 101 (2010) 2087–2092.
[59] N. Kosaka, H. Izumi, K. Sekine, T. Ochiya, Silence 1 (2010) 7.
[60] E.M. Kroh, R.K. Parkin, P.S. Mitchell, M. Tewari, Methods 50 (2010) 298–301.
[61] P.S. Mitchell, R.K. Parkin, E.M. Kroh, B.R. Fritz, S.K. Wyman, E.L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K.C. O'Briant, A. Allen, D.W. Lin, N. Urban, C.W. Drescher, B.S. Knudsen, D.L. Stirewalt, R. Gentleman, R.L. Vessella, P.S. Nelson, D.B. Martin, M. Tewari, Proc. Natl. Acad. Sci. USA 105 (2008) 10513–10518.
[62] C.H. Lawrie, S. Gal, H.M. Dunlop, B. Pushkaran, A.P. Liggins, K. Pulford, A.H. Banham, F. Pezzella, J. Boultwood, J.S. Wainscoat, C.S. Hatton, A.L. Harris, Br. J. Haematol. 141 (2008) 672–675.
[63] E.K. Ng, W.W. Chong, H. Jin, E.K. Lam, V.Y. Shin, J. Yu, T.C. Poon, S.S. Ng, J.J. Sung, Gut 58 (2009) 1375–1381.
[64] C. Becker, I. Riedmaier, M. Reiter, A. Tichopad, M.W. Pfaffl, H.D. Meyer Heinrich, Analyst 136 (2011) 1024–1029.
[65] K. Hibi, H. Nakamura, A. Hirai, Y. Fujikake, Y. Kasai, S. Akiyama, K. Ito, H. Takagi, Cancer Res. 56 (1996) 480–482.
[66] Y. Fellig, I. Ariel, P. Ohana, P. Schachter, I. Sinelnikov, T. Birman, S. Ayesh, T. Schneider, J. Clin. Pathol. 58 (2005) 1064–1068.
[67] I.J. Matouk, N. De Groot, S. Mezan, S. Ayesh, R. bu-lail, A. Hochberg, E. Galun, PLoS ONE 2 (2007) e845.
[68] R.A. Gupta, N. Shah, K.C. Wang, J. Kim, H.M. Horlings, D.J. Wong, M.C. Tsai, T. Hung, P. Argani, J.L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R.B. West, Nature 464 (2010) 1071–1076.
[69] N. Miyoshi, H. Wagatsuma, S. Wakana, T. Shiroishi, M. Nomura, K. Aisaka, T. Kohda, M.A. Surani, T. Kaneko-Ishino, F. Ishino, Genes Cells 5 (2000) 211–220.
[70] X. Zhang, Y. Zhou, K.R. Mehta, D.C. Danila, S. Scolavino, S.R. Johnson, A. Klibanski, J. Clin. Endocrinol. Metab. 88 (2003) 5119–5126.
[71] M.D. Sorani, W.A. Ortmann, E.P. Bierwagen, T.W. Behrens, Drug Discov. Today 15 (2010) 741–748.
[72] A.A. de, I. Kubisch, G. Breier, G. Stamminger, N. Fersis, A. Eichler, S. Kaul, U. Stolzel, Oncology 82 (2012) 3–10.
[73] T.J. Molloy, L.A. Devriese, H.H. Helgason, A.J. Bosma, M. Hauptmann, E.E. Voest, J.H. Schellens, Br. J. Cancer 104 (2011) 1913–1919.
[74] T.H. Thai, D.P. Calado, S. Casola, K.M. Ansel, C. Xiao, Y. Xue, A. Murphy, D. Frendewey, D. Valenzuela, J.L. Kutok, M. Schmidt-Supprian, N. Rajewsky, G. Yancopoulos, A. Rao, K. Rajewsky, Science 316 (2007) 604–608.
[75] G. Lee, C. Rodriguez, A. Madabhushi, IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (2008) 368–384.
[76] A. Bergkvist, V. Rusnakova, R. Sindelka, J.M. Garda, B. Sjogreen, D. Lindh, A. Forootan, M. Kubista, Methods 50 (2010) 323–335.
[77] J. Beyene, D. Tritchler, S.B. Bull, K.C. Cartier, G. Jonasdottir, A.T. Kraja, N. Li, N.L. Nock, E. Parkhomenko, J.S. Rao, C.M. Stein, R. Sutradhar, S. Waaijenborg, K.S. Wang, Y. Wang, P. Wolkow, Genet. Epidemiol. 31 (Suppl. 1) (2007) S103–S109.
[78] J. Vandesompele, M. Kubista, M.W. Pfaffl, Reference Gene Validation Software for Improved Normalization, 2009.
[79] 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.
[80] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, J. Zhang, Genome Biol. 5 (2004) R80.
[81] C. Ritz, A.N. Spiess, Bioinformatics 24 (2008) 1549–1551.