Volume 59(Suppl.1):51-67, 2015
Acta Biologica Szegediensis
http://www.sci.u-szeged.hu/ABS

REVIEW

Microarray technology

Ágnes Zvara, Klára Kitajka, Nóra Faragó, László G. Puskás*

Institute of Genetics, Biological Research Center, Hungarian Academy of Sciences, Szeged, Hungary

ABSTRACT The normal functions of the cells are based on a strict and regulated expression of various genes. If this precise hierarchy of gene actions becomes unregulated or disturbed due to different genetic or environmental effects, the result will be abnormal cellular function that eventually could lead pathological alterations, including carcinogenic transformation or apoptosis. To understand the complex mechanisms and networks involved in biological processes and diseases, it is not enough to analyze isolated pathways, single gene functions or a single genetic event. A living organism has to be studied as a complex system and all genes involved in different biological processes need to be analyzed simultaneously: a systems biology approach should be applied. In the beginning of the 1990's years, a new, high throughput technology - called microarray technology – was developed to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Microarrays have dramatically accelerated many types of investigation since a microarray experiment can accomplish many genetic tests in parallel. This review summarizes some of aspects of the microarray technology, including sample preparations, application possibilities and data analysis.
Acta Biol Szeged 59(Suppl.1):51-67 (2015)

KEY WORDS

aCGH
DNA microarray
functional genomics
gene expression profile

Introduction

Until the end of the last century, studying gene function and regulation was restricted to examination of one or very few genes at a time. The different successful genome projects gave the opportunity to develop new, high throughput molecular methods for gene expression monitoring, mutation analysis of the whole genome (single nucleotide polymorphisms (SNP) analysis, array-based comparative genomic hybridization (CGH), protein expression and gene and protein interaction analysis. These novel functional molecular biology technologies are extraordinary tools to simultaneously monitor different mutations in the genome, monitoring all gene activities in one experiment, and analyze protein expression differences between diverse biological samples in a comparative way at different levels: genome, transcriptome or proteome. The basic role of functional biology is to identify new genes and gene functions, explore new regulatory networks involved in different cellular processes.

One of the most powerful and widespread high throughput methods is the microarray technology which has become an essential tool for a new discipline studying the expression of all genes in a genome simultaneously. This technology has been applied to a diverse range of studies such as for transcriptome analysis, detection and characterization of genetic variants (*e.g.*, SNP, copy number variants (CNVs)), studying DNA-protein interaction, and detecting genome methylation.

The main innovation of the microarray technology was the immobilization of different molecules (oligonucleotides, proteins, small drug like compounds) onto a solid and activated surface. These molecules are bound to these surfaces as a matrix in a well-defined order. This high density matrix arrangement of biologically active molecules is called microarray. In one spot of a microarray, high concentration of a given molecule is immobilized and can have specific interaction with its target. On one microarray thousands or millions of specific spots can be immobilized enabling the analysis of a full genome, transcriptome or proteome of an organism at a given time point.

Analysis of nucleic acids: hybridization techniques

The quantitative analysis of nucleic acids in an organism, tissue or cell is essential if we want to gather information about their roles, function and interactions. Almost all the techniques that could provide us these data based on one of the very important features of the DNA. Namely, the two complementary strands of the DNA - following the Watson-Crick base pairing rule - can complete and hybridize to each

Submitted March 8 2015; Accepted June 9, 2015
*Corresponding author. E-mail: pusi@brc.hu

other. Although many of the techniques used in nucleic acid research are not based on hybridization, this specific feature of nucleic acids is an essential base for emerging such high throughput technologies like microarray technology. The hybridization-based quantitative analysis of nucleic acids started with a discovery in 1965 by Gillespie and Spiegelman, when they described that not only the denatured DNA strands bound to nitrocellulose membrane but a single strand DNA could also bind to its immobilized complementary sequence (Gillespie et al. 1965). This observation became the basis of the detection and analysis of specific DNA sequences using labeled single strand DNA or RNA.

Southern in 1975 used ^{32}P isotope to label DNA molecules (radioactive labeling) in a special buffer solution. The first step in this method is the extraction of DNA or RNA from the sample tissues or cells followed by an enzymatic fragmentation (using precisely chosen restriction endonuclease in the case of DNA) and separation of the nucleic acid by gel electrophoresis. The separated nucleic acid fragments are then blotted and immobilized onto nitrocellulose or nylon membrane (usually baked at 80 °C for a couple of hours for nitrocellulose membranes, or using ultraviolet radiation in case of nylon membrane). The denatured and labeled single stranded DNA molecules find their complementary pairs on the nitrocellulose or nylon membrane showing the specific interaction as radioactive signal (Southern et al. 1975). The labeled single stranded nucleic acid is called probe, while the immobilized nucleic acid is called sample.

This approach was further improved by Kafatos and coworkers in 1979 (Kafatos et al. 1979). They spotted different DNA molecules onto nitrocellulose membrane in a well-defined order. Since they spotted many different unknown DNA molecules onto the surface, many samples could be tested in one step during the hybridization. Spots arranged on a solid surface in a well-defined order are called arrays. Using this so called *"dot-blot"* technique, the presence of an antibiotic resistance gene (labeled probe) can be detected in many different plants (their DNA spotted onto solid surface as samples) so for example, transgenic plant carrying a marker gene can be easily selected. The question that can be answered using dot blot technique is the presence of a homologue pair of a known nucleic acid sequence in a relatively wide range of unknown DNA population. This known nucleic acid fragment can be DNA or RNA molecule.

In the *"reverse dot blot"* technique the sample and the probe are in reverse position. In this case the sample nucleic acid extracted from a biological sample is labeled with *e.g.,* isotope and the probes (many different sequences) are immobilized on a solid surface one by one. During the hybridization step the labeled sample nucleic acid finds its complementary probe on the surface. In this case, using the aforementioned antibiotic resistance example, the question to be answered is how many different resistance genes can be found in a given plant. Applying this technique, the presence of many genes, sequences can be tested in one experiment. The first relatively large reverse dot blot array was made by Saiki and coworkers in 1989 (Saiki et al. 1989). They bound synthetic oligonucleotides corresponding to different mutant alleles of the human HLA gene (human lymphocyte antigen) onto a nitrocellulose membrane. This miniature membrane was called DNA chip, or gene chip, although the high density microarrays used nowadays was developed only eight years later at Stanford University.

The *DNA chip technology* (or DNA microarray technology) is a reverse dot blot technique where the number of spotted oligonucleotides with known sequence can vary between several thousand to even a hundred thousand. Thus this technology is a hybridization based nucleic acid detection system, where thousands of different genes or nucleotide changes can be analyzed in one hybridization step.

Classification and preparation of DNA arrays, steps in a microarray experiment

The DNA arrays are usually classified according to the number of the different DNA molecules immobilized on their surfaces (Table 1). *Macroarrays* contain a few hundred to a maximum a thousand gene specific probes. The diameter of these spots is bigger than 300 micrometer. This technique was used mainly in experiments, where a small number of focused gene activities were tested or the presence of a small number of nucleic acid molecules wanted to be identified. The spotted probes are usually longer (>400 nt) cDNA molecules, or amplified and denatured PCR products. Since the sample nucleic acids are always radio-labeled, the quantity of the bound DNA is determined by autoradiography. The preparation of these

Table 1. Comparison of macro- and microarrays.

	Number of sports	Spot size (micrometer)	Immobilized nucleic acid type	Immobilized nucleic acid size (bp)	Surface	Printing
Macroarray	100-500	>300	cDNA, PCR product	>400	activated glass, nitrocellulose or nylon filter	usually manually
Microarray	1000-200000	80-300	cDNA, PCR product, oligonucleotide	>50	activated glass	spotted (printing robot, inkjet) or *in situ* synthesis

arrays does not require automated chip printer (printer robots) and sophisticated data analyzing system.

In contrast, *microarrays* have several thousand or a hundred thousand of spotted probes on their surface with a less than 300 μm in diameter. During microarray preparation, high precision and computer driven automated printing robot spots the DNA fragments onto the surface of chemically activated glass microscope slides (or in some cases a flexible nitrocellulose membrane can also be used). The printer head contains 4-48 precisely shaped pins with a thin capillary tunnel inside. All the pins suck the DNA solution by capillary action and put a micro spot (few nanoliter in volume) onto the surface. The diameter of the spot is determined by the capillary force and the surface tension. It usually means an approximately 80-300 micrometers in diameter and about 1-5 nanoliters solution regarding one spot. The microarrays contain longer cDNA fragments or shorter synthetic oligonucleotides in very high density. If electric current is applied, the ink-jet technology drops standard size spots onto the surface without actually touching the surface resulting uniform spot morphology. The reason this technology has not been spread widely for spotting pre-synthesized oligonucleotides is the high reagents necessity. However, when active precursors (phosphoramidites) are spotted, ink-jet technology results in highly precise deposition and high density microarrays.

Although the cDNA based arrays were more common in the early years, due to problems with annotation and clone identification, synthetic oligonucleotide (50–70-mer) platforms became more popular (Woo et al. 2004). Comparative studies showed that oligonucleotide arrays offer several advantages over cDNA platforms in terms of specificity, sensitivity, and reproducibility (Hughes et al. 2001). The glass surface in the case of cDNA fragment is usually treated with aminosilane or polylysine (amine groups on the surface) and ionic bond immobilizes the DNA onto the surface. In contrast, the synthetized oligonucleotides usually amino modified. When preparing oligonucleotide microarrays, there are active groups on the surface and the oligonucleotides are covalently bound onto the surface.

Besides the spotting techniques mentioned earlier (printer robots, ink-jet technology), oligonucleotides can be immobilized onto the surface by *in situ* oligonucleotide synthesis using special chemical solutions. Oligonucleotide based microarrays prepared by *in situ* synthesis are called *DNA chips* or *gene chips*. In these cases, the oligonucleotides are not prepared in advance and bound to the surface, but synthesized *in situ* onto the solid surface instead. The basic technology which was developed by the Affymetrix Inc. (www.affymetrix.com) is based on photolithography (Lockhart et al. 1996). The key element of the innovation was the development of a photosensitive protective group. The bound protective group inactivates the nucleotide, but after UV light exposure, the photosensitive group dissociates and the nucleotides become active. The whole surface of the chip is covered by a series of a computer designed and very precisely perforated masks. All these masks have a specific perforation pattern. After putting one mask on, UV light dissociates the protection groups in the affected spots. For example, if activated thymine (T) is added to the system with the appropriate coupling buffers, the first nucleotide will be T and the application of further perforated masks will determine the place of the next nucleotides. Since the newly bound nucleotides also contain the protective group, the process is continuous. To build a 25-mer oligonucleotide, 4 x 25 = 100 synthesis steps are necessary. The same chemistry is used for a more modern technology, where perforated masks are not necessary, because the photosensitive protection groups are dissociated by light directed with digitally controlled micro-mirror system (www.nimblegen.com). Using this technology, preparation of custom DNA chips is quicker and more economic.

A microarray experiment typically requires a (i) DNA microarray, (ii) labeled nucleic acids (RNA or DNA) from the tissue or cells of interest, (iii) proper hybridization environment, (iv) laser scanner to detect fluorescent signal, and (v) computer and software background (bioinformatics background) for data analysis and to properly interpret results. Some basic information about the DNA chips have already discussed in this review. Now we summarize some of the aspects of the other four key steps.

Labeling strategies, hybridization

The most wildly applied method to obtain labeled sample – if gene expression is in the focus of the study - is to convert the RNA to cDNA. During reverse transcription (RT), the whole RNA in the sample is reverse transcribed to cDNA with the help of reverse transcriptase enzyme. One option is to add fluorescent dye labeled nucleotide (usually Cy3 or Cy5 labeled dCTP or dTTP) to the RT reaction and the cDNA will be labeled directly. But in most cases, high amount of starting RNA is needed for this method. However, in many cases the amount of available biological sample (*e.g.*, FACS cells, laser micro-dissected tissue sample or in experimental systems where 1000-5000 cells are the subject of the investigation) is a limiting factor or the homogeneity of the sample is a serious issue. To override these problems, the starting material should be amplified, but it is crucial to keep the quantitative ratio presented in the original RNA population. The proper use of exponential (PCR) and linear (*in vitro* transcription, IVT) amplification can solve this problem (Puskás et al. 2002a, 2002b; Kitajka et al. 2002). PCR exponentially amplify the cDNA but there is a risk of distortion of the original quantity ratio. During IVT, the RT reaction is carried out by a specific primer containing T7 polymerase binding site so the double stranded cDNA can serve as template for the IVT reaction and the RNA population can be linearly amplified. In this case,

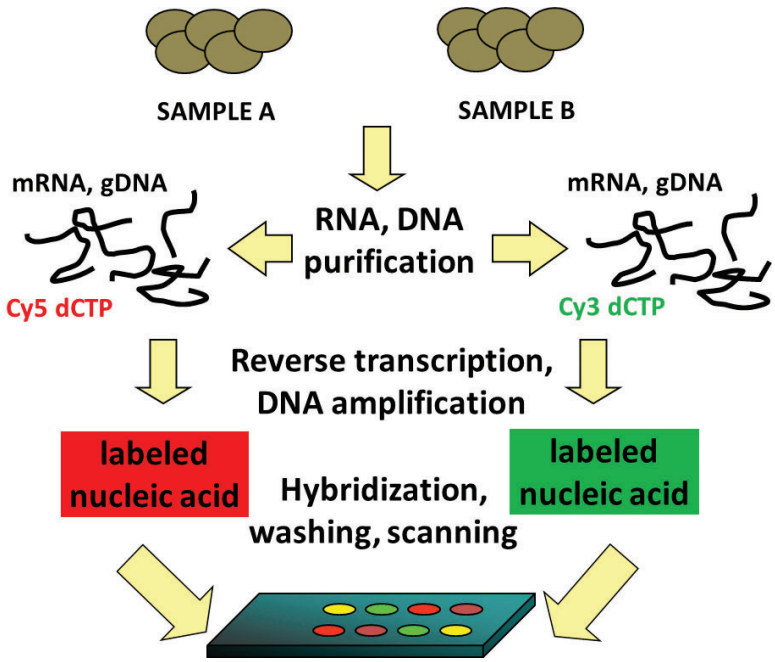

Figure 1. Diagram of a typical dual-color labeling and hybridization strategy. It is applicable both genomic DNA and mRNA labeling. Genome or transcriptome profile of treated/non-treated, healthy/diseased, two different tissues or organisms can be compared in one experiment. Genes active only or mainly in the *sample A* represented by red spots, genes expressed exclusively in *sample B* represented by green spots, while genes with about the same expression level in both samples are represented as yellow spots on the microarray. Violet spots show the non-expressed genes.

the fluorescent labeled nucleotide is part of the IVT reaction. Nowadays, the linear amplification methods are widely spread because of their better reproducibility.

It is very important to use the suitable fluorescent dyes for labeling. In a typical microarray experiment, the most frequently used fluorescent dyes are the cyanine based Cy3 and Cy5. Cy3 and Cy5 are synthetic dye belonging to polymethine group. Cy3 is fluorescent in green (~550 nm excitation, ~570 nm emission), while Cy5 is fluorescent in the red region (~650 excitation, 670 nm emission). These dyes should meet some important criteria, they should be spectrally well separated, fluoresce brightly when dry, which simplifies image acquisition, and should be incorporated with high specific activities with a variety of enzymes (Eisen and Brown 1999). There is an important disadvantage of Cy5 that should be considered during experiment design. It sometimes gives higher background levels on glass surfaces and is more sensitive to photobleaching than Cy3 (photobleaching is caused by intense light and occurs because the excited state of a molecule is generally much more chemically reactive than the ground state (van Hal et al. 2000). Different label incorporation efficiency and scanning artifacts may result in different Cy3- and Cy5 fluorescence intensities even when equal amounts of Cy3- and Cy5-labeled probes are present. Therefore it is essential to properly normalize the fluorescence signals, so these types of systematic errors can be kept to a minimum increasing microarray data quality (Bilban 2002).

When we compare two transcriptomes there are two options: using co-hybridization strategy with Cy5-labeled cDNA from the test sample and Cy3-labeled cDNA from the control sample (Fig. 1), or apply the same fluorescent dye for the comparable samples and hybridize them onto different DNA chip. As it is mentioned earlier, Cy3 and Cy5 dyes have different fluorescent properties which can be easily differentiated (non-overlapping emission and absorption spectra). In the first case labeled nucleic acid molecules compete with each other for the binding sites on the surface. The hybridization step between the labeled sample(s) and the immobilized probes carried out for 6-16 hours under specified conditions: usually at 50-65 ºC, in 10-200 µl of special buffer solution and in special metal or plastic chambers under glass coverslip. In most of the cases the chambers are put into a rotator in order to provide even dispersion. The hybridization temperature depends on the binding capacity of complementary nucleotides: in the case of 20-40 nt long oligonucleotides, lower temperature needs to be applied (37-45 °C), while longer complementary nucleotides need higher temperature (45-60 °C) to achieve higher stringency and more reliable data. Both the correct

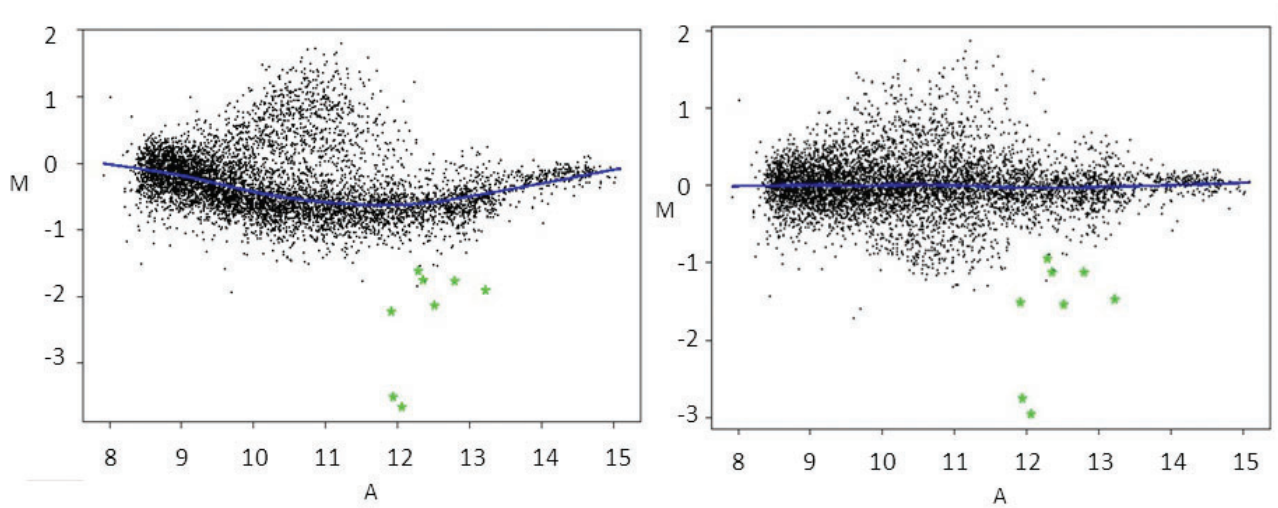

Figure 2. Spots before and after Lowess normalization. In low and high intensity range the intensity ratios are distorted due to the inaccurate background correction and spot saturation respectively. The regression based Lowess normalization correct this systematic error.

choice of hybridization temperature and the stringency of the washing step after the hybridization ensure us to get rid of signals from weaker or unspecific binding.

Laser scanner to detect fluorescent signal

The specific fluorescent intensities bound to the surface can be captured by a high resolution (usually 2-10 micron) fluorescent laser scanner. The scanner generates 1-3 pictures files according to the number of applied dyes and experiment strategy. The analysis of the spots is made by special software. In every spot, the dye intensities correlate to the quantity of the bound and hybridized dye. Usually DNA miroarray laser scanners possess two lasers: one excites Cy3, the other excites Cy5 fluorescent dyes. Thus, the fluorescent intensity in case of a gene-expression experiment, is directly proportional to the given gene product, *i.e.* the RNA molecules, which is directly proportional to the activity of the gene. During the analyses, many correction factors should be applied including: the intensity of the local background, the uniformity of the spot, any spotting problems, outlier spots, normalization, etc. Usually the special analysis softwares incorporate all the algorithms need to be applied to get reliable data.

Bioinformatics

One of the advantages of the microarray approach compared to other classic high throughput methods is the close proximity of the immobilized probes on the surface. This fact opens the door to a high numbers of uniform measurements under the exact same circumstances. Thus, results from different spots can be more comparable and the conclusions are more accurate. Besides, local distortions can easily change the intensities and cause low rate error. Since a typical microarray contains 10 or 20 thousands spots, this law rate error still can generate hundreds of false negative or false positive results. In order to eliminate these errors, different softwares and statistical methods should be used during the data analysis. In different phases of an experiment, numerous errors can affect the results and the final output such as: the applied technology, laboratory protocols or human error. Theoretically, measurement errors and biological variance can be separated from each other. Applying different normalization methods the measurement errors can be minimalized. There are two parts of the measurement error: first is the systematic error causing more or less standard measurement deviation (*e.g.*, Cy3/Cy5 efficiency distortions) and the second one the random deviation. By doing several repeats of an experiment the measurement errors can be reduced to the minimum. The global normalization methods are suitable for scaling the expression vectors. Two commonly used types of normalization exist: (i) one is based on the total intensity and (ii) the second is the regression based Lowess corrected median normalization (Berger et al. 2004). In the first case, all the spot intensities are divided by the sum of the total intensities on the array. This method is based on the assumption, that the RNA quantity is constant under the analyzed conditions. Lowess normalization assumes that the dye bias appears to be dependent on spot intensity. In low and high intensity range the intensity ratios are distorted due to the inaccurate background correction and spot saturation respectively. The regression based methods like Lowess normalization are suitable for elimination these systematic errors (Fig. 2). In order to get reliable data, comprehensive statistical analysis has

also to be performed. Typically both biological and technical (dye-swap) replica experiments are carried out to gain data for statistical analysis. To find genes with significant expression differences between two comparable biological samples (*e.g*., diseased versus control samples) Student's t-test can be used, while comparing more than two different conditions (*e.g*., drug concentration, age, sex, control) ANOVA should be applied. In microarray data analysis, *p-values* derived from the mentioned statistical tests have to be adjusted by multiple testing corrections to correct for occurrence of false positives. False positives are the genes that are found to be statistically different between conditions, but in fact they are not. A typical microarray experiment measures several thousand genes simultaneously across different conditions. When testing for potential differential expression across those conditions, each gene is considered independently from one another. In other words, a *t-test* or ANOVA is performed on each gene separately. The incidence of false positives (or genes falsely called differentially expressed when they are not) is proportional to the number of tests performed and to the critical significance level (p-value cutoff). Four types of multiple testing corrections are usually used: (i) Bonferroni, (ii) Bonferroni Step-down (Holm) (Holm et al. 1979), (iii) Westfall and Young Permutation (Westfall et al. 1993), (iv) Benjamini and Hochberg False Discovery Rate. The methods are listed in order of their stringency, with the Bonferroni being the most stringent, and the Benjamini and Hochberg FDR being the least stringent. The more stringent a multiple testing correction is, the less false positive genes are allowed. The trade-off of a stringent multiple testing correction is that the rate of false negatives (genes that are called non-significant when they are) is very high (Benjamini et al. 1995).

After the several statistical corrections the differently expressed genes can be further analyzed by the most commonly used hierarchical clustering method. This is the most popular method for comprehensive gene expression data analysis. During clustering, different samples are grouped together into clusters based on similarities in their gene expression patterns and are connected by a series of branches (clustering tree or dendrogram). Experiments with similar expression profiles can also be grouped together using the same method. In Fig. 3, different thyroid diseases are clustered based on their gene expression pattern that obtained by microarray experiments. Thyroid carcinomas (two malignant and one benign) are clearly grouped together while the other diseased samples (autoimmune and hormonal diseases) are separated from that cluster. Gene expression based clustering methods are very important in those cases where no simple methods available to separate a disease group from another or from a control group. Typical example is the outcome prediction of different chemotherapy. In many cancer cases, the treatment is effective in one patient or patient group but totally ineffective for another. Clustering based on differences in gene expression fingerprints can be generated which could predict the efficacy of the chemotherapy. During clustering, usually marker genes or gene groups are looked for that defines a sample cluster.

To get overall functional information about the significant differentially expressed gene set, Gene Ontology (GO) analysis and functional gene clustering are routinely used. To reveal significantly enriched biological functions and pathways, DAVID bioinformatics system and database (Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov) (Sherman et al. 2007) or other third party software like GeneSpring (Agilent) or IPA (Ingenuity Pathway Analysis) can be used. These analyses are suitable for identifying significantly enriched biological themes, particularly GO terms related to multiple genes, discover enriched functional-related gene groups in a particular gene set comparing to a background set of genes. In order to gain information about the possible common regulatory elements of differentially expressed genes, analysis of proximal promoter and distant regulatory element of genes with altered expression can also be done. The online system called DiRE (Distant Regulatory Elements), based on the Enhancer Identification (EI) method and determines the chromosomal location and functional characteristics of DIREs. It enables to analyze complex cooperative activity of different regulatory elements like proximal promoters and distant regulatory elements such as enhancers, repressors, and silencers. The Regulatory Elements (RE) are ranked by their importance and occurrence in the input co-expressed gene set (Pennacchio et al. 2007; Gotea et al. 2008). To explore possible connections/interactions between proteins coded by significantly altered genes, Ingenuity Pathway analysis platform and STRING interaction database online tool (http://string-db.org/) are very useful. These databases link proteins based on genomic context, experimental evidence, co-expression and data from other databases such as PubMed, MINT, KEGG, BIND and BioGRID.

Application of microarrays

Gene expression monitoring, basics and general considerations

A primary goal of an expression profiling study is to characterize genes that expressed differentially in two experimental groups (Szűcs et al. 2010). The array-based gene expression analysis based upon a comparison of expression patterns between two samples. Total mRNA from both samples are first extracted, purified and then converted to cDNA (Murphy et al. 2002). The two cDNAs are then labeled with different fluorescent dyes and are hybridized onto probes spotted in high density onto the surface of the glass surface. The hybridization signals and intensities are then analyzed and the differentially expressed genes are selected in the two analyzed

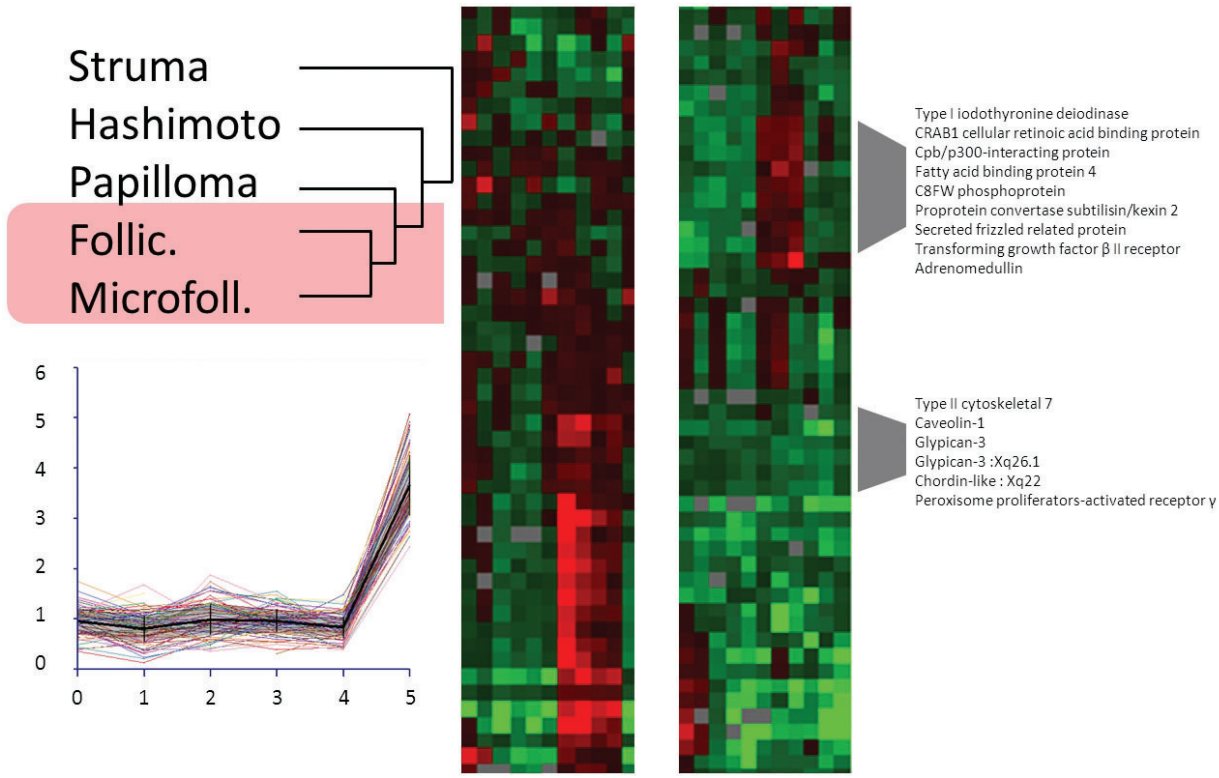

Figure 3. Clustering of different thyroid carcinomas by gene expression pattern. Thyroid carcinomas (two malignant and one benign (transparent red box)) are clearly grouped together while the other diseased samples (autoimmune and hormonal diseases) are separated from that cluster.

samples. These differences in gene expression are causes or consequences of a given disease state, show reactions to a drug treatment or to environmental changes, reflect altered gene function, elucidate genetic regulation and biological pathways underlying specific physiological conditions, refer to hosts' response to pathogenic infections and can be a first step in disease diagnosis and drug discovery (Murphy et al. 2002). These molecular fingerprints can also serve as markers of a given disease and help subgrouping disease clusters according the effectiveness of treatment. Microarray technology allows the changes in cells at the gene expression level induced by various effects (*e.g.*, pharmaceutical treatment, pathological processes) to be traced, new biochemical markers and genes responsible for pathological phenotype to be discovered, drug effects to be followed and the treatment to be optimized (Csont et al. 2007; Erdi et al. 2012; Szatmári et al. 2014). The differences in gene expression of the treated and untreated cells or tissues provide information about the regulation of the enzymatic pathways influenced by drugs, about the enzymes, transporters playing a role in drug resistance. Identification of gene expression patterns may provide vital information for understanding the pathological processes and contribute to diagnostic decisions and therapies tailored to individual patient.

There are three major groups of tools for studying gene expression at the transcript level: (i) hybridization-based techniques, such as Northern blotting, subtractive hybridization, DNA microarrays or macroarrays, (ii) PCR-based techniques such as differential display and RDA (representational difference analysis) (iii) Sequence-based techniques such as SAGE (serial analysis of gene expression) and new generation sequencing protocols (NGS) (Kozian et al. 1999)

Northern blot analysis was the first technique that made the identification of different mRNA in a given sample possible. Radiolabeled RNA or DNA probes hybridized to RNA fragments separated by gel electrophoresis and immobilized onto nylon filter (Alwine et al. 1977). This technique is rarely used now, however in some specific projects (*e.g.*, miRNA analysis) can be a useful method for detecting transcripts or confirming expression data obtained with other experiments although, in this case, big amount of starting material (usually 10 μg RNA) is needed to get reliable result. Using phosphoimaging, the labeled bands on the membrane can be quantified.

Subtractive hybridization was the first technique that enabled identification of differentially expressed genes. Subtrac-

tive cDNA libraries are generated by hybridizing an mRNA pool of one origin to an mRNA pool of a different origin (Hedrick et al. 1984). Transcripts with no complementary strand are then used for cDNA library construction. Despite the fact that numerous genes were successfully identified with this method, it had serious disadvantages: (i) only small fraction of gene expression differences can be successfully discovered, (ii) it requires large amount of RNA sample, and (iii) quite laborious and time-consuming.

In 1992, a new PCR-based method called differential display PCR (DD-PCR) has been emerged. This technique is a one-tube method to compare differentially expressed genes systematically (Liang et al. 1992). RNA from two different biological samples is amplified by PCR after a reverse transcription (RT), and the generated fragments that reflect the expression pattern of the given sample, are separated by denaturing gel electrophoresis. Differentially expressed genes could be isolated from the gel, sequenced and identified. Numerous studies have been published, which - despite the serious disadvantages (maintain the quantitative correlation after RT and PCR reactions, repeatability, and the elimination of false positive signals) - applied this method successfully (Prashar et al. 1996; Vogeli-Lange et al. 1996). However, this method was almost entirely replaced by new generation sequencing technologies.

Serial analysis of gene expression (SAGE) technique uses a sequence-based strategy that allows parallel analysis of a large number of transcripts (Velculescu et al. 1995). The method is based on two principles: (i) a short, 8-9 base pair long nucleotide sequence tag contains enough information for the identification of the transcript, (ii) concatenation of these short tags allows the efficient analysis of transcripts in a serial manner by the sequencing of multiple tags within a single clone. Results obtained with this technique allow the determination of significant quantitative relationship between mRNA populations derived from various experimental procedures. This method is much more sensitive in detecting low copy number transcripts. The aforementioned techniques are material-intensive and time consuming. For these reason, efforts had been undertaken to develop methods for high-throughput screening which are based on new generation sequencing technologies.

DNA microarray technology offers the possibility of high-throughput systematic analysis of the transcriptome in one experiment. The most informative and probably the most important application of DNA chips is the parallel study of gene expression from different biological samples focusing on the functionally active parts of the genome (DeRisi et al. 1996; Bittner et al. 2000; Ernst et al. 2002; Stremmel et al. 2002). DNA microarrays with sets of cDNA fragments or gene specific oligonucleotide on their surfaces can be used to obtain a molecular fingerprint of gene expression of cells in a given time point in a comparative way (Blohm et al. 2001; Koppler et al. 2002; Lin et al. 2002). Gene expression monitoring means quantification of RNA molecules in different conditions of cells or tissue grafts. mRNA population of a cell or tissue in a given condition is called transcriptome. Differences between transcriptomes due to any genetic or environmental factors (*e.g.*, treated cells versus non-treated, or diseased state versus control, transgenic plant versus wild type) is the most interesting question in biological systems and can be answered by one DNA chip experiment. The activity (corresponds to mRNA quantity) of nearly 22000 human genes can be monitored in one hybridization step using microarray technology.

The simplest application of the microarrays is the tissue specific gene expression analysis. If total RNA extracted from a tissue sample is labeled and hybridized onto the surface of a DNA microarray, active genes in a given time point can be monitored by the analysis and quantification of fluorescent spots. For this type of experiment only one fluorescent labeling and one hybridization step are needed. More informative experiments can be conducted, where two fluorescent dyes are applied. These analyses focus on the differences in gene expression patterns and can be divided to two different groups: (i) comparison of expression patterns of two different tissues or biological sample (ii) comparison of the same tissue or biological sample but at different time point, after different treatment or in a different developmental stage.

In the first case, as an example experiment, we can compare the gene activity patterns in leaves and flowers of a plant. If the chosen plant in this experiment is *Arabidopsis*, we need *Arabidopsis* specific DNA microarray to analyze specific gene expression profiles. RNA extracted from the two different tissues are labeled either Cy5 (red) or Cy3 (green) dye, mixed and hybridized onto the *Arabidopsis* specific DNA microarray. On the scanner image red, green and yellow spots will represent the relative RNA quantities. The red spots will refer to genes active exclusively or mainly in the flower, while green spots show the genes active only or mainly in the leaf. The yellow spots refer to genes with very similar expression level in both tissues. It is understandable even from this very simple hypothetical experiment that new and comprehensive information about gene activities cannot be gained by any other traditional techniques. Genes active only in the flower (red spots on the microarray) determine the color, shape, size, scent of the flower, while genes responsible for photosynthesis should be looked for among genes expressed exclusively in the leaf (green spots).

Comparison of the same tissue or biological sample at different time point or after different treatment is the other mostly applied experiment types. The “clearest system” for this type of experiments is when different states of cell cultures induced by any physical or chemical stress are analyzed. In spite of the advantageous homogeneity of the biological sample (in most cases contains only one type of cell), one of

the major disadvantage of this system that it does not reflect the complexity of a living organism. Thus, complex diseases cannot be modeled properly in an *in vitro* system. Analyzing a complex system (*e.g.*, post mortem human samples, surgical tissue biopsies) however needs careful sample preparation, the more homogenous and more precisely dissected the samples are the more accurate the result will be. Non-homogeneity in the samples can be misleading since results will reflect not only the disease state but also the gene expression differences within the tissue sample deriving from different cell type composition. Pooling the samples is a good strategy to reduce this type of error to the minimum. Pooling samples is very effective for example when samples from organisms with different genetic backgrounds are analyzed and common gene expression markers are looked for (Zvara et al. 2005; Virok et al. 2011).

In order to predict a disease outcome, exact subtyping of the samples is crucial. One of the most common cancer types in human is the malignant neoplastic changes in the skin. In spite of big efforts, no accepted histopathological or immunohistochemical prediction markers defining subset of melanoma according the outcome were found until the development of high throughput molecular methods. DNA microarray technology had a huge impact on molecular classification of melanomas. Bittner et al. (2000) discovered a subset of melanoma classified by mathematical analysis of gene expression in a series of samples. They showed that melanoma is a useful model to identify genes critical for aspects of the metastatic process including tumor cell motility.

Identification of new molecular markers based on gene expression profiling provides vital information for possible future drug development projects or for tailored or personalized medicine. In case of a complex neuropsychiatric disorder, like schizophrenia, the current diagnosis is based on complex clinical symptoms existing for more than six months. In these cases, the use of easily detectable peripheral molecular markers could substantially help to improve and speed up the diagnosis the disorders. Screening the peripheral blood lymphocytes (PBL) of 13 drug-naive/drug-free schizophrenic patients, Zvara et al. (2005) found two marker genes (dopamine receptor D2 and the inwardly rectifying potassium channel Kir2.3) to be overexpressed in a microarray experiment. The increased mRNA levels were confirmed either TaqMan or SybrGreen based quantitative real-time PCR (QRT-PCR). The use of molecular markers provides more rapid and precise opportunity to diagnose and predict an outcome or a drug response of a certain disease. It can help to find the optimal medication for the patients.

Microarray technology also provides a good possibility to follow a precisely regulated biological process in a comprehensive way. As an example, high density cDNA microarray was used to get a global picture of gene expression profiles during pear fruit development and climacteric ripening. These are complex processes involving major changes in fruit metabolism (Fischer et al. 1991). Biochemical processes occur in a well-defined order under the control of a large set of ripening-specific genes leading to changes in texture, pigmentation, taste and aroma. Fonseca et al. (2004) analyzed a series of fruit samples at different time points of ripening process and compare it to fruit that failed to ripe (FR) due to precocious harvesting. They found different transcripts correspond to kinases and phosphatases were induced specifically during early developmental stages of pear fruit. While another set of genes (transcripts encoding for cell wall modifications, and pigment and aroma biosynthesis) were activated at the onset of the climacteric period when fruit softening rates also increased. Some transcripts putatively involved in defense response, oxidative stress, primary and secondary metabolism, signaling and transcription regulation were also detected. Better understanding the ripening process in fleshy fruit is a prerequisite for improving fruit quality and storage potential.

Environmental changes (biotic or abiotic, internal or external) significantly alter gene expression profile of a biological system. By tracking those changes, we can reveal the molecular and biochemical background, and mechanism of action of a drug or a biologically active molecule. Kitajka et al. (2004) studied brain gene-expression changes in response to different polyunsaturated fatty acids (PUFA) -enriched diets in rats with high-density microarrays. PUFAs are well-known essential structural components of the central nervous with a role of controlling learning and memory. In aged rats fed throughout life with PUFA-enriched diets, transthyretin, alpha-synuclein, and calmodulins were found to have altered expression. These genes play important role in synaptic plasticity and learning. It was also shown that omega-3-deficient diets during the perinatal period cause altered gene function in the offspring throughout their lifetime. They concluded that PUFA-enriched diets lead to significant changes in expression of several genes in the central nervous tissue, and these effects appear to be mainly independent of their effects on membrane composition.

Detection of chromosomal alterations with oligonucleotide arrays

Genetic research has been revolutionized with the development of high-throughput genome analysis tools which allow simultaneous analysis of many genomic regions. Chromosomal alterations in the cancer genome are quite frequent. These differences include single nucleotide polymorphisms (SNPs one nucleotide mismatch) and copy number variations (CNVs) described as genomic fragments ranged in size from 100 base pairs to 1 kilobase or more (The 1000 Genomes Project Consortium 2010; Boone et al. 2010). CNVs usually result from structural genomic alterations such as a deletion

(loss), duplication (gain), an insertion (usually a gain) or unbalanced translocations/inversions that may lead to either loss or gain of sequences near the breakpoints (Feuk et al. 2006). Abnormal copy numbers of these regions have been implicated with several diseases and complex traits in human and other animals such as in HIV/AIDs susceptibility (Gonzalez et al. 2005), autoimmune disease (Fanciulli et al. 2007; McKinney et al. 2008), asthma (Brasch-Andersen et al. 2004), Crohn's disease (McCarroll et al. 2008), Osteoporosis (Yang et al. 2008). High-resolution microarray-based CNV analysis is one of the novel and quick way to detect these copy number gains and losses throughout the genome (Shaikh et al. 2007; Miller et al. 2010). In contrast to traditionally techniques like fluorescence *in situ* hybridization (FISH), the high resolution array-based technologies offers robust methods for genome wide search of CNVs with higher resolution and speed (Carson et al. 2006). It opens new insight into microdeletions and microduplications detection as well as uncovering novel CNVs that are undetectable by standard karyotype analysis or fluorescence *in situ* hybridization (Shaikh et al. 2007).

There are two very efficient types of microarrays experiment that typically used for CNV monitoring: either array-based comparative genomic hybridization (aCGH) or SNP-based microarrays (SNP-arrays) (Pinto et al. 2011). Several factors need to be considered in order to apply the most suitable detection system, including resolution desired and ability to customize probe content. One of the most important field the SNP arrays have been used is the mapping of human disease susceptibility loci as published in genome-wide-association studies (GWAS) (Hindorff et al. 2009). In order to facilitate the GWAS, a detailed human haplotype map has been created using over a million SNP (The International HapMap Consortium 2005). High-density SNP arrays contain oligonucleotide probes spotted systematically to detect the two alleles of a specific SNP locus, in which both the homozygous and heterozygous genotypes could be detected.

Array-based comparative genomic hybridization (aCGH) is a rapid method to monitor major DNA copy number changes like deletions or amplifications and provide more accurate information about chromosomal imbalances (Houldsworth et al. 1994).

DNA copy number profiling by microarrays often applied in tumor genomic researches since specific rearrangements usually are characteristic to the individual tumor types and states. aCGH provides a lot of information about genomic balance of tumor cells, mono- or trisomies, amplifications and deletions in a simple experiment. In different cancer types, changes within the chromosome like short deletions, insertions or amplifications are quite frequent (Mitelman et al. 1997). Specific DNA segments containing tumor suppressors (deletions) or oncogenic element (insertions) can be revealed using CGH. Genes mapped to the locations of these rearrangements can play roles in the formation of tumor. Their investigation can contribute to the characterization of the different tumors and tumor stages. To detect these chromosomal abnormalities, genomic DNA (gDNA) from tissue samples should be purified and analyzed. In this case, not the transcriptome but the whole genome is under investigation. (Alkan et al. 2011). aCGH uses similar labeling strategy as gene expression profiling, DNA from the two samples fluorescently labeled with Cy5 or Cy3 dye, denatured and hybridized together onto a DNA chip containing high number of genomic DNA fragments or oligonucleotides complementary to specific regions of already identified genes. After reading with a confocal laser scanner, differences in color can be measured and analyzed with computer software; differences of fluorescent intensities in a spot (mainly red or mainly green spots) refer to altered DNA copy numbers between the two compared samples. The ratio of signal intensity between the test sample and the control is used to determine the copy number changes at specific genomic locations. The results can be easily analyzed with the help of currently available databases that contain significant amount of information about the chromosomal location of the genes identified in the experiments. Great advantage of these experiments that good quality of labeled probe can be obtained even from small amount of paraffin archived material. In contrast to conventional CGH (uses metaphase chromosomes), aCGH applies probes that immobilized onto a solid surface to hybridize to the labeled test and control DNAs. These probes can vary in size from small oligonucleotides (25-85 base pairs) to genomic clones such as bacterial artificial chromsomes (80,000-200,000 base pairs). aCGH technique has been applied to create comprehensive maps of human CNVs (Iafrate et al. 2004; Redon et al. 2006; Wong et al. 2007)

The initial study of aCGH was from Pinkel et al. (1998). They immobilized bacterium artificial chromosomes (BAC) and genomic fragments from human chromosome 20 on glass surface and demonstrated the feasibility of detecting both gains and losses with single copy sensitivity. In spite that aCGH is not suitable for identification of small mutations (SNPs or deletions/amplifications of few nucleotides within genes), it is extremely useful, well applicable high throughput way for the overall analysis of the whole genome. It has the significant advantage of being less sensitive to cell contaminations. A single gene-copy change may be detected from a sample containing up to 60% of normal, healthy cell contamination (Hodgson et al. 2001). In case of adult T-cell leukemia, both deletion and amplification were successfully determined by this method. Testing of 64 patients showed amplification in 14q, 7q, and 3q chromosome regions, while in the regions 6q and 13q, deletions were observed. These chromosome changes were much more frequent in patients with aggressive form of leukemia than in indolent form. An

increased number of chromosome imbalances were detected in patients, where the chance of survival was significantly lower (Tsukasaki et al. 2001).

Fehér et al. (2012) also used aCGH to identify gene copy number alterations predictive of metastatic potential or aggressive transformation in papillary thyroid carcinoma (PTC) which is the most common well-differentiated thyroid cancer. The authors analyzed formalin-fixed and paraffin-embedded samples from primary tumors without metastasis, cases with only regional lymph node metastasis, and cases with distant metastasis, recurrence or extrathyroid extension (a total of 43 PTC cases). Deletion of the EIF4EBP3 and TRAK2 gene loci and amplification of thymosin beta 10 (TB10) and Tre-2 oncogene regions were observed as general markers for PTC. Their study was the first to report TB10 as a specific marker that revealed by genomic amplification. They further revealed that the A-kinase anchor protein 13 (AKAP13) gene region was discriminative markers for metastasis. The article suggests AKAP13 and TB10 regions as potential new genomic markers for PTC and cancer progression (Fehér et al. 2012).

Bonnet et al. (2012) showed a very nice example of CHG application for prediction of metastatic event in breast cancer. They performed a comparative genomic hybridization study on BAC arrays and analyzed 45 patients with metastatic relapse and 95 patients without any recurrence after at least 11 years of follow-up. Using the array-CGH data, the authors established a two-parameter index representative of the global level of aneusomy by chromosomal arm, and of the number of breakpoints throughout the genome. This genomic instability index (G2I) with appropriate thresholds applied in the study allowed to distinguish three classes of tumors highly associated with metastatic relapse.

Methylation pattern analysis

Environmental factors have major contribution to the development of complex diseases like psychological disorders (schizophrenia, major depression) through induction of epigenetic modifications, such as DNA methylation. Epigenetics is one of the most expanding fields in biology, which refers to any process that alters gene activity without changing the actual DNA sequence and leads to regulation of gene expression. Changes in methylation pattern usually generate alterations in gene expression programs. Approximately 60-70% of all human gene promoters overlap with CpG islands (these regions has an elevated GC content and a high frequency of CpG dinucleotide). Gene silencing by DNA methylation of specific gene promoters is a well-known feature of neoplastic cells and plays an important role in normal cell differentiation and development. Aberrant DNA methylation pattern of CpG islands is one of the earliest and most common alterations in human malignancies (Jones et al. 1996). Tumor cells are generally characterized by the hypermethylation of tumor suppressor genes and, in contrast, hypomethylation of the whole DNA molecule. This general hypomethylation can be detected relatively early, before the development of the actual tumor. Correlation between hypomethylation and increased gene expression can be detected in cases of large number of oncogenes (Eads et al. 2000; Esteller et al. 2001). Without changing the primary DNA sequences, DNA methylation occurs mainly at CpG dinucleotide and involves the enzymatic addition of a methyl group to the cytosine residue. Such modifications at regulatory regions (in particular gene promoters), correlate well with the transcriptional state of a gene: DNA methylation represses transcription while DNA unmethylation can lead to increased transcription levels. While DNA methylation is an essential mechanism for normal cellular development, imprinting, X-chromosome inactivation, and maintaining tissue specificity it can also significantly contribute to the progression of various human diseases (Esteller 2007).

Microarray technology is a very effective and high throughput way for genome-wide analysis of methylation, since in one experiment ten or hundred thousands of distinct and identified potential methylation sites can be monitored (Adorján et al. 2002). Analysis of genome wide methylation profile enables to characterize new tumor classes, or to cluster newly diagnosed cases into already existing groups based on methylation pattern. In recent years a new method had emerged for the analysis of methylation pattern extending to the whole genome that is suitable for analyzing large number of genes simultaneously (Toyota et al. 1999; Gitan et al. 2001; Adorján et al. 2002). The starting biological sample in this case is genomic DNA, more precisely the methylated part of the gDNA. There are different ways to analyze methylated DNA fragments: (i) bisulfide treatment, (ii) methylated DNA immunoprecipitation (MeDIP), and (iii) double restriction enzymatic DNA cleavage.

Sodium bisulfite treatment converts the non-methylated cytosine is to uracil, while the methylated cytosine stays unaffected during the process. The Illumina Methylation Assay is one assay that applies the bisulphite conversion technology on a microarray level to generate genome-wide methylation data. The assay discriminates between the two chemically differentiated loci using two site-specific probes, one designed for the methylated locus (M bead type) and another for the unmethylated locus (U bead type) (http://technology.illumina.com/technology/beadarray-technology/infinium-methylation-assay.html).

Applying immunoprecipitation, sonicated and denatured DNA (fragment size 300-1000bp) first immunoselected by an antibody directed against 5-methyl-cytidine than immunoprecipitated by Protein A Agarose. After elution and precipitation of the pulled-down methylated DNA, whole genome amplification is done to get enough material for further experiments.

The third technique is based on a double enzymatic DNA cleavage followed by QPCR. The DNA samples are digested by both a methylation-sensitive enzyme which will digest unmethylated and partially methylated DNA (the remaining (hyper)-methylated DNA will be detected by real-time PCR) and a methylation dependent enzyme, which will preferentially digest methylated DNA (the remaining unmethylated DNA will be detected by real-time PCR). Using this technique a well-defined selection of DNA segments can be analyzed to reveal the affected DNA fragments.

In a comprehensive study, methylation pattern of twenty samples of advanced ovarian cancer were profiled by Illumina HumanMethylation27 BeadChip technology where 27 578 CpG sites in >14 000 genes were simultaneous analyzed. The goal in this case was to find specific CpG sites that correlated with progression-free interval (PFI) after therapy. They found, that longer survival was associated with both hypomethylated CpG sites (*e.g.*, GREB1, TGIF and TOB1) and hypermethylated ones (*e.g.*, TMCO5, PTPRN and GUCY2C). The affected genes plays role in telomere organization, mesoderm development and immune regulation. The author concluded that this kind of analysis might be of prognostic value (Bauerschlag et al. 2011).

Devaney and co-workers (2013) systematically analyzed the epigenetic defects in prostate cancer (PCa) and tried to find DNA methylation-based biomarkers that may be useful for the early detection and diagnosis of PCa. In their study, methylation status of 485 577 CpG sites from regions with a broad spectrum of CpG densities were examined by Illumina 450K methylation platform. They found numerous candidate novel genes (BNC1, FZD1, RPL39L, SYN2, LMX1B, CXXC5, ZNF783 and CYB5R2) that are frequently methylated and whose methylation was associated with inactivation of gene expression in PCa cell lines (Devaney et al. 2013).

Conclusions

Achievements in the automation technologies, miniaturization, new solid surfaces, fluorescent labeling and detection systems and the expanding sequence data bases made the development of the DNA microarray technology possible. According to many concurrent opinions, DNA microarray technology is as huge step in molecular biology as the development of semiconductor chips was in microelectronics. In both cases, the revolutionary change was the incredible increase of the number of operations per unit of time (in the case of DNA chips the individual hybridizations and specific sequence detection). Although the many advantage of this technology have been discussed in this review, there are some issues to consider and some main pitfalls to carefully think over:

General

A gene-expression experiment reveals the active part of the genome but give no information how it is reflected at the protein level (there does not necessarily have to be a tight correlation between the expression of a gene and the amount of translated protein). Microarrays only present a snapshot of an actual transcriptome in the cell which is continuously changing as the cell responds to cellular and environmental signals. Interplay between genes or groups of genes (*i.e.* mechanisms) cannot be easily decoded. Other methods and experimental tools are needed to illuminate the proteome, understand the varying interactions between genes, and get a more complete picture of cellular behavior. Interpretation of the results is usually hard and the relationships are often difficult to decipher

Experiment design

The goals of the experiment have to be carefully determined. We have to choose the most relevant biological comparisons, taking into account the various sources of variability, choose the most suitable platform, and consider what controls need to be used (internal controls and external controls). Generally, incorporating replicates in experiment design (technical and biological) to generate statistical significance and pooling samples to reduce inherent variation are need to be parts of the design. Technical replicates aim to eliminate procedural variations such as sample preparation and handling, while biological replicates aim to identify variation in the biological system being studied.

Samples requirements

Growing, collecting and preparing biological samples, isolating and purifying RNA are crucial. The tissue samples usually have to be processed rapidly to maintain RNA integrity (use of RNA stabilizing solution like RNAlater® (Life Technologies) can resolve this problem). Quality and the amount of RNA is a major challenge, the quality of purified RNA should be carefully verified (determining the RNA Integrity Number (RIN) using Agilent Bioanalyzer system is a good choice to check RNA quality). False microarray data can be generated from degraded mRNA. Working with small sample size (laser dissection or biopsy) can be a difficulty (special amplification steps or labeling protocols should be applied). The amplification techniques should be selected carefully because it can distort the results particularly when multiple amplification steps needed. Heterogeneity of most of the biological samples, typically tumor samples is also a pitfall need to overcome (applying single cell protocols).

Experimental procedure

When cDNA microarrays applied, cross-hybridization can cause false positive or false negative results. Using oligonucleotide DNA chips, products of low activity genes often enable to be detected, or only with high variance (sensitivity issue). The number of false results is bigger in the case of hybridization techniques that use RNA probes because the exact circumstances and energetics of DNA-RNA hybridization is not precisely known. It is especially true when one-nucleotide changes are detected. In case of *in situ* prepared DNA chips, oligonucleotides in one spot are usually mixed because of the inefficiency of the coupling. The applied fluorescent dyes (Cy3 and Cy5) have different spectral and chemical features (quenching, absorbance, acid and ozone sensitivity). This often results false data. Color flip or dye flip experiments (labeling both the control and the sample nucleic acids with both of the dyes) are usually applied to avoid this distortion.

Reproducibility and standardization

There are multiple sources of variability in a microarray experiment: arrays, dye labeling, efficiency in reverse transcription, and hybridization. Since there are numerous error-prone steps in a microarray experiments, the procedures need to be replicated in order to eliminate sources of error. Standardization of protocols and validation of current profiles will have to ensure that gene-expression profiles are reliable and reproducible. It is essential, therefore, that experiments are tightly regulated and quality controlled. All results should be validated by independent methods (quantitative PCR, digital PCR or Western blot). There is a standard for reporting microarray experiments called *Minimum Information About a Microarray Experiment* (MIAME) created by Functional Genomics Data Society which specifies a series of standards on collecting and analyzing microarray data (raw data, normalized data, sample annotation, expreimental design, microarray annotation, data processing protocols). This document is designed to allow data generated by microarray experiments to be interpreted and reproduced with certainty. There are two big repositories such as the Gene Expression Omnibus (GEO) created by the National Center for Biotechnology Information (NCBI) and ArrayExpress created by the EBI to store and share gene expression data compiled according MIAME standards.

Bioinformatics

The application of the right biostatistical methods is a key element of the procedure. Usually analysis of such a large quantities of data generated from microarray tests predispose the results to misinterpretation. Without careful selection of statistical and normalization methods, hundreds of false data can be produced. It is important to consider, that although the threshold of the expression change is usually 2-fold, genes with lower expression ratio can have serious biological effects. Multiple professional data analysis software is necessary to extract the results.

Although DNA microarray technology was only developed a little bit more than two decades ago, nowadays it is wildly spread and applied in almost all the fields of biology and medicine including oncology, microbiology, neurology, developmental biology, psychiatry, diagnostic, forensic medicine and it is the basis of new comprehensive disciplines: functional genomics, toxicogenomics, and pharmacogenomics. The information value of microarray experiment data is unquestionable: discovery new biochemical pathways, identification of genes coding for drug resistance, responsible or predispose for diseases, identification of disease specific molecular genetic markers, analysis of the effects of abiotic or biotic stress, display genetic elements of embryogenesis, revealing molecular genetic background of genetically modified organisms.

Acknowledgments

This work was supported by the TÁMOP-4.1.1.C-13/1/KONV-2014-0001 program entitled „Practice-oriented, student-friendly modernization of the biomedical education for strengthening the international competitiveness of the rural Hungarian universities".

References

Adorján P, Distler J, Lipscher E, Model F, Müller J, Pelet C, Braun A, Florl AR, Gütig D, Grabs G, Howe A, Kursar M, Lesche R, Leu E, Lewin A, Maier S, Müller V, Otto T, Scholz C, Schulz WA, Seifert HH, Schwope I, Ziebarth H, Berlin K, Piepenbrock C, Olek A (2002) Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res 30:e21.

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-376.

Alwine JC, Kemp DJ, Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci USA 74:5350-5354.

Bauerschlag DO, Ammerpohl O, Bräutigam K, Schem C, Lin Q, Weigel MT, Hilpert F, Arnold N, Maass N, Meinhold-Heerlein I, Wagner W (2011) Progression-free survival in ovarian cancer is reflected in epigenetic DNA methylation profiles. Oncology 80:12-20.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple

testing. J Royal Stat Soc B 57:289-300.
Bilban M, Buehler LK, Head S, Desoye G, Quaranta V (2002) Normalizing DNA microarray data. Curr Issues Mol Biol 4:57-64.
Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536-540.
Blohm DH, Guiseppi-Elie A (2001) New developments in microarray technology. Curr Opin Biotechnol 12:41-47.
Bonnet F, Guedj M, Jones N, Sfar S, Brouste V, Elarouci N, Banneau G, Orsetti B, Primois C, de Lara CT, Debled M, de Mascarel I, Theillet C, Sévenet N, de Reynies A, MacGrogan G, Longy M (2012) An array CGH based genomic instability index (G2I) is predictive of clinical outcome in breast cancer and reveals a subset of tumors without lymph node involvement but with poor prognosis. BMC Med Genomics 5:54.
Boone PM, Bacino CA, Shaw CA, Eng PA, Hixson PM, Pursley AN, Kang S-HL, Yang Y, Wiszniewska J, Nowakowska BA, del Gaudio D, Xia Z, Simpson-Patel G, Immken LL, Gibson JB, Tsai AC-H, Bowers JA, Reimschisel TE, Schaaf CP, Potocki L, Scaglia F, Gambin T, Sykulski M, Bartnik M, Derwinska K, Wisniowiecka-Kowalnik B, Lalani SR, Probst FJ, Bi W, Beaudet AL, Patel A, Lupski JR, Cheung SW, Stankiewicz P (2010) Detection of clinically relevant exonic copy-number changes by array CGH. Hum Mutat 31:1326-1342.
Brasch-Andersen C, Christiansen L, Tan Q, Haagerup A, Vestbo J, Kruse TA (2004) Possible gene dosage effect of glutathione-S-transferases on atopic asthma: Using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. Hum Mutat 24:208-214.
Carson AR, Feuk L, Mohammed M, Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. Hum Genomics 2:403-414.
Csont T, Bereczki E, Bencsik P, Fodor G, Görbe A, Zvara A, Csonka C, Puskás LG, Sántha M, Ferdinandy P (2007) Hypercholesterolemia increases myocardial oxidative and nitrosative stress thereby leading to cardiac dysfunction in apoB-100 transgenic mice. Cardiovasc Res 76:100-109.
DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genet 14:457-460.
Devaney JM, Wang S, Funda S, Long J, Taghipour DJ, Tbaishat R, Furbert-Harris P, Ittmann M, Kwabi-Addo B (2013) Identification of novel DNA-methylated genes that correlate with human prostate cancer and high-grade prostatic intraepithelial neoplasia. Prostate Cancer Prostatic Dis 16:292-300.
Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW (2000) MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res 28:e32.
Eisen MB, Brown PO (1999) DNA arrays for analysis of gene expression. Meth Enzymol 303:179-205.
Erdi B, Nagy P, Zvara A, Varga A, Pircs K, Ménesi D, Puskás LG, Juhász G. (2012) Loss of the starvation-induced gene Rack1 leads to glycogen deficiency and impaired autophagic responses in *Drosophila*. Autophagy 8:1124-1135.
Ernst T, Hergenhahn M, Kenzelmann M, Cohen CD, Bonrouhi M, Weninger A, Klären R, Gröne EF, Wiesel M, Güdemann C, Küster J, Schott W, Staehler G, Kretzler M, Hollstein M, Gröne HJ (2002) Decrease and gain of gene expression are equally discriminatory markers for prostate carcinoma: a gene expression analysis on total and microdissected prostate tissue. Am J Pathol 160:2169-2180.
Esteller M, Corn PG, Baylin SB, Herman JG (2001) A gene hypermethylation profile of human cancer. Cancer Res 61:3225-3229.
Esteller M (2007) Epigenetic gene silencing in cancer: the DNA hypermethylome. Hum Mol Genet 6:R50-59.
Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI, Froguel P, Owen CJ, Pearce SH, Teixeira L, Guillevin L, Graham DS, Pusey CD, Cook HT, Vyse TJ, Aitman TJ (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721-723.
Fehér LZ, Pocsay G, Krenács L, Zvara A, Bagdi E, Pocsay R, Lukács G, Győry F, Gazdag A, Tarkó E, Puskás LG (2012) Amplification of thymosin beta 10 and AKAP13 genes in metastatic and aggressive papillary thyroid carcinomas. Pathol Oncol Res 18:449-458.
Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85-97.
Fischer RL, Bennett AB (1991) Role of cell wall hydrolases in fruit ripening. Annu Rev Plant Physiol Plant Mol Biol 42:675-703.
Fonseca S, Hackler L, Zvara A, Ferreira S, Balde A, Dudits D, Pals MS, Puskas LG (2004) Monitoring gene expression along pear fruit development, ripening and senescence using cDNA microarrays. Plant Sci 167:457-469.
Gillespie D, Spiegelman S (1965) A quantitative assay for DNA-RAN hybrids with DNA immobilised on membrane. J Mol Biol 12:829-842.
Gitan RS, Shi H, Chen CM, Yan PS, Huang TH (2001) Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. Ge-

nome Res 12:158-164.
Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell R J, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434-1440.
Gotea V, Ovcharenko I (2008) DiRE: identifying distant regulatory elements of co-expressed genes. Nucleic Acids Res 36:W133-139.
Hautaniemi S, Järvinen AK, Edgren H, Mitra SK, Astola J Berger JA (2004) Optimized LOWESS normalization parameter selection for DNA microarray data. BMC Bioinformatics 5:194.
Hedrick SM, Cohen DI, Nielsen EA, Davis MM (1984) Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature 308:149-153.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362-9367.
Hodgson G, Hager JH, Volik S, Hariono S, Wernick M, Moore D, Nowak N, Albertson DG, Pinkel D, Collins C, Hanahan D, Gray JW (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nat Genet 29:459-64.
Holm S (1979) A simple sequentially rejective bonferroni test procedure. Scand J Statist 6:65-70.
Houldsworth J, Chaganti RS (1994) Comparative genomic hybridization: an overview. Am J Pathol 145:1253-60.
Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 19:342-347.
Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951.
Jones PA (1996) DNA methylation errors and cancer. Cancer Res 65:2463-2467.
Kafatos FC, Jones CW, Efstratiadis A (1979) Determination of nucleic acid sequence homologies and relative concentrations by a dot blot hybridization procedure. Nucleic Acids Res 24:1541-1552.
Kitajka K, Puskás LG, Zvara A, Hackler L Jr, Barceló-Coblijn G, Yeo YK, Farkas T (2002) The role of n-3 polyunsaturated fatty acids in brain: Modulation of rat brain gene expression by dietary n-3 fatty acids. Proc Natl Acad Sci USA 99:2619-2624.
Kitajka K, Sinclair AJ, Weisinger RS, Weisinger HS, Mathai M, Jayasooriya AP, Halver JE, Puskás LG (2004) Effects of dietary omega-3 polyunsaturated fatty acids on brain gene expression. Proc Natl Acad Sci USA 101:10931-10936.
Kopper L, Timar J (2002) Gene expression profiles in the diagnosis and prognosis of cancer. Magy Onkol 46:3-9.
Kozian DH, Kirschbaum BJ (1999) Comparative gene-expression analysis. Trends Biotechnol 17:73-78.
Liang P, Pardee A (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967-971.
Lin YM, Furukawa Y, Tsunoda T, Yue CT, Yang KC, Nakamura Y (2002) Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. Oncogene 21:4120-4128.
Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675-1680.
McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, Duerr RH, Silverberg MS, Taylor KD, Rioux JD, Altshuler D, Daly MJ, Xavier RJ (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 40:1107-1112.
McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, Jones PB, McLean L, O'Donnell JL, Pokorny V, Spellerberg M, Stamp LK, Willis J, Steer S, Merriman TR (2008) Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis 67:409-413.
Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86:749-764.
Mitelman F, Mertens F, Johansson B (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. Nat Genet 15:417-474.
Murphy D (2002) Gene expression studies using microarrays: Principles, problems, and prospects. Adv Physiol Educ 26:256-270.
Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I

(2007) Predicting tissue-specific enhancers in the human genome. Genome Res 2:201-211.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207-211.

Pinto D, Darvishi K, Shi XH, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, MacDonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29:U512-U576.

Prashar Y, Weissman S (1996) Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. Proc Nat Acad Sci USA 93:659-663.

Puskás LG, Zvara A, Hackler L Jr, Van Hummelen P (2002a) RNA amplification results in reproducible microarray data with slight ratio biases. Biotechniques 32:1330-1342.

Puskás LG, Zvara A, Hackler L Jr, Micsik T, van Hummelen P (2002b) Production of bulk amounts of universal RNA for DNA-microarrays. Biotechniques 33:898-900, 902, 904.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human genome. Nature 444:444-454.

Saiki RK, Walsh PS, Levenson CH, Elrich HA (1989) Genetic analysis of amplified DNA with immobilized sequence-specific olygonucleotide probes. Proc Natl Acad Sci USA 86:6230-62334.

Shaikh TH (2007) Oligonucleotide arrays for high-resolution analysis of copy number alteration in mental retardation/multiple congenital anomalies. Genet Med 9:617-625.

Shaikh TH, O'Connor RJ, Pierpont ME, McGrath J, Hacker AM, Nimmakayalu M, Geiger E, Emanuel BS, Saitta SC (2007) Low copy repeats mediate distal chromosome 22q11.2 deletions: sequence analysis predicts breakpoint mechanisms. Genome Res 17:482-491.

Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) DAVID knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8:426.

Southern E (1975) Detection of specific sequences among DNS fragments separated by gel electrophoresis. J Mol Biol 98:503-517.

Stremmel C, Wein A, Hohenberger W, Reingruber B (2002) DNA microarrays: a new diagnostic tool and its implications in colorectal cancer. Int J Colorectal Dis 17:131-136.

Szatmári Á, Zvara Á, Móricz ÁM, Besenyei E, Szabó E, Ott PG, Puskás LG, Bozsó Z (2014) Pattern triggered immunity (PTI) in tobacco: isolation of activated genes suggests role of the phenylpropanoid pathway in inhibition of bacterial pathogens PLoS One 9:e102869.

Szűcs A, Jäger K, Jurca ME, Fábián A, Bottka S, Zvara A, Barnabás B, Fehér A (2010) Histological and microarray analysis of the direct effect of water shortage alone or combined with heat on early grain development in wheat (*Triticum aestivum*). Physiol Plant 2:174-88.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-1320.

Toyota M, Ho C, Ahuja N, Jair KW, Li Q, Ohe-Toyota M, Baylin SB, Issa JP (1999) Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. Cancer Res 59:2307-2312.

Tsukasaki K, Krebs J, Nagai K, Tomonaga M, Koeffler HP, Bartram CR, Jauch A (2001) Comparative genomic hybridization analysis in adult T-cell leukemia/lymphoma: correlation with clinical course. Blood 97:3875-3881.

van Hal NL, Vorst O, van Houwelingen AM, Kok EJ, Peijnenburg A, Aharoni A, van Tunen AJ, Keijer J (2000) The application of DNA microarrays in gene expression analysis. J Biotechnol 78:271-280.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:464-467.

Virok DP, Kis Z, Szegedi V, Juhász G, Zvara A Jr, Müller G, Lévay G, Hársing LG, Rajkó R, Penke B, Janka Z, Janáky T, Puskás LG (2011) Functional changes in transcriptomes of the prefrontal cortex and hippocampus in a mouse model of anxiety. Pharmacol Rep 63:348-361.

Vögeli-Lange R, Bürckert N, Boller T, Wiemken A (1996) Rapid selectionand classification of positive clones generated by mRNA differential display. Nucleic Acids Res 24:1385-1386.

Westfall PH, Young SS (1993) Resampling-based Multiple Testing. Wiley, NewYork.

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL (2007) A comprehensive analysis of com-

mon copy-number variations in the human genome. Am J Hum Genet 80:91-104.

Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G (2004) A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. J Biomol Tech 15:276-284

Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ, Papasian CJ, Recker RR, Song XP, Cheng J, Deng HW (2008) Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet 83:663-674.

Zvara A, Szekeres G, Janka Z, Kelemen JZ, Cimmer C, Sántha M, Puskás LG (2005) Over-expression of dopamine D2 receptor and inwardly rectifying potassium channel genes in drug-naive schizophrenic peripheral blood lymphocytes as potential diagnostic markers. Dis Markers 21:61-69.