Functional Genomics and Gene Chips –from Research to Clinical Medicine

Emel Eryüksel, MD[1]; Kirsten Laule-Kilian, MD[2]; Martin H. Brutsche, MD[2]

[1] *Marmara Üniversity Hospital, Department of Pulmonary Medicine, İstanbul, Turkey*
[2] *Division of Respiratory Medicine and Pulmonary Research Laboratory, University Hospital, Basel, Switzerland*

Abstract

In the year 2000, the first complete draft of the human genome was presented by the Human Genome Project Consortium together with Celera Genetics and made the headlines worldwide. Since then, the so-called "post-genome era" has started. The expectations towards novel technology are high, but justified. The last 5 years have seen an exponential number of publications on functional genomics. The use of microarrays for gene expression profiling, genotyping, mutation detection, and gene discovery are leading to remarkable insights into the function of thousands of genes previously known only by their gene sequence. The impact of these technologies on every day clinical practice is not yet clear, although the first clinical applications are ready to be applied in specific clinical situations. This review introduces the principle of gene chip technology and gives an overview of its current and future potential.

Turkish Respiratory Journal, 2003;4:(1):27-31

Key words: *genomics, microarray, respiratory medicine*

Introduction

In the year 2000, the first complete draft of the human genome was presented by the Human Genome Project Consortium together with Celera Genetics (1, 2). The current version of the human genome being built up of $3x10^9$ nucleotides, the bricks of the genome, is not very precise. The full reliable version of the human genome is not expected before 2004 and is expected to contain approximately 50 000 genes. Each gene can be present in different variants, called polymorphisms (or single nucleotide polymorphisms [SNPs]). To date, $3x10^6$ gene polymorphisms are described - a number which increases daily and which is estimated to be greater than $11x10^6$ for the whole human genome. Each individual is composed of a pair of inherited combinations of variants for each of the 50'000 genes explaining the enormous genetic diversity. Only a fraction of these gene polymorphisms are important to human health and disease. However, to find out which ones are important is a major challenge for the next decades.

Not all of these genes are used at a time. Depending on the developmental stage, age of the individual, cell type, organ, environmental factors, etc. a different set of genes is used or transcribed. Genes are transcribed into messenger RNA. The analysis of the expression of genes is called **functional genomics**. It allows to compare the set of used genes in different conditions, e.g. healthy and diseased. Most studies applying gene chips are

Correspondence: *Dr. Martin H. Brutsche*
University Hospital Basel
Petersgraben 4 CH-4031, Basel, Switzerland
Tel: +0041 (61) 265 25 25
Faks +0041 (61) 265 45 87
e-mail: mbrutsche@uhbs.ch

Table 1. Definitions	
Term	**Definition**
Genetics	Analysis of the genome structure and its variations
Functional genomics	Analysis of gene expression of a cell, tissue or organ under given conditions
Proteomics	Analysis of proteins molecules of a cell, tissue or organ under given conditions
Pharmacogenetics	Study of variability in drug responses attributed to hereditary factors in different populations
Pharmacogenomics	Determination and analysis of the genome and its products (RNA and proteins) as they relate to drug response

functional genomic studies.

Not all of the transcribed genes will result in a protein. Also, practically all proteins are modified after the first assembly of amino acids. It is estimated that a protein derived from the same strand of gene can be altered in 10-20 different splits and 3-dimensional forms. Some proteins interact directly with the DNA of genes leading to expression or silencing of genes. Practically all of the proteins interact with other proteins within pathways forming complex multidimensional related networks. The analysis of the proteins is called **proteomics**. Due to the enormous diversity and the difficulty to use robotics for this kind of analysis proteomic applications have only started very recently (3).

If the genotype would automatically lead to a specific condition, so-called phenotype, all identical twins would have exactly the same diseases. Although they considerably resemble each other, they also differ from each other in many ways. Thus, there is not a 100 % match between genotype and phenotype due to environment-gene interactions. Also, many conditions result from different pathogenetic mechanisms, leading to the same phenotype, like asthma or arterial hypertension. In these, so-called complex diseases, a constellation of different susceptibility and disease-modifying genes and not a single one need to be present. Therefore, research failed to identify, e.g., an "asthma gene". To better understand the functional aspects of disease and to bridge the long way between genotype and phenotype, it is necessary to combine genetic, functional genomic and proteomic analyses.

Classical hypothesis-driven research, often analysing a single or a couple of genes or proteins, was and is a very successful and reliable scientific strategy. However, classical research is not able to cope with the number of new discovered genes, proteins and potential interactions. Thus, it is necessary to apply techniques able to allow complex answers. Novel techniques are necessary to more rapidly screen thousands of genes and generate new hypotheses. This exactly is the role of high-throughput technology, like microarrays, also called gene chips. As a hypothesis-generating approach, high-throughput methods can lead to the identification of a set of potentially interesting genes associated with a certain condition, so called candidate genes. Microarray techniques, however, will not replace the classical hypothesis-driven research. Identified candidate genes need to be tested for their function and relevancy by classical approaches.

Technicalities of microarrays

What is a microarray and how does it work?

A microarray - or gene chip - measures the expression level of a gene by determining the amount of messenger RNA (mRNA) that is present. Unlike a conventional Northern blot where one can analyse the abundance of up to 20 mRNAs, a microarray allows the simultaneous analysis of the expression levels of hundreds, thousands, or even tens of thousands of genes in a single experiment (Figure 1). The latest chips carry up to 450 000 spots for the analysis of more than 20 000 genes and gene sequences on a small glass slide (1-2 cm^2).

The gene chip works as follows: Purified RNA from the biological sample, e.g. blood or tissue, is labelled and then hybridised for several hours to the gene chip. RNA consists of a sequence built up of 4 different oligonucleotides (thymidine, guanine, cytosine, uracil), which is specific for each gene. Such a specific RNA sequence is able to bind to a complementary sequence of oligonucleotides only. This specificity of binding is used in gene chips. If the sequence of oligonucleotides matches with the sequence of oligonucleotides on a specific spot of the gene chip, hybridisation occurs. On each of the up to 450 000 spots different intensities of binding occur depending on the concentration of the different genes in the biological sample tested. Thus, the concentration of mRNA (=gene expression) can be measured quantitatively.

Two main methods are used to make microarrays. In the first -cDNA array technology-DNA is spotted onto a glass slide; in the second-oligonucleotide array-oligonucleotides of 15-30 nucleic acid base pairs are synthesised on to a silica slide by a process known as photolithography. Both methods have advantages and disadvantages. It is, however, likely that oligonucleotide arrays will become the gold standard in the foreseeable future, due to advantages in reproducibility, variability and quality.

What can be measured by gene chips?

Gene chip technology can be used for three main applications:

1. ***Gene expression profiling*:** RNA extracted from a biological sample is applied to the microarray. The result reveals the level of expression of tens of thousands of genes, effectively all the genes in the genome, in that sample. This result is known as a gene expression "profile" or "signature".

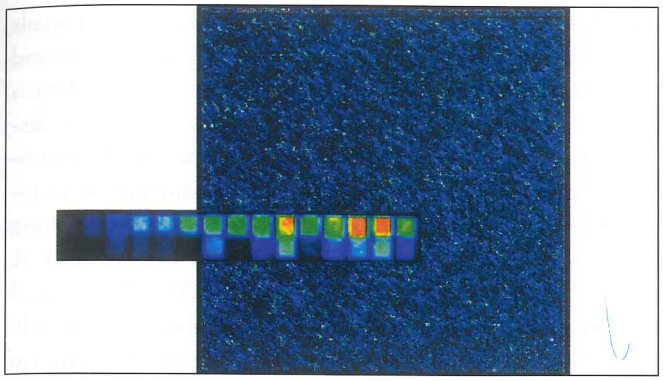

Figure 1. Example of a hybridized GeneChip© by Affimetrix Inc. with enlarged probe set of a gene as an insert bottom right. With this chip 12 626 genes and expressed sequence tags can be assayed in a single experiment. Messenger RNA is copied into labelled cDNA with reverse transcriptase so that the relative abundance of individual mRNAs is reflected in the cDNA product. Thus, the intensity of the hybridisation signal for a given gene product is a result of its relative abundance in the target sample. This method has proven to provide excellent specificity and reproducibility. Messenger RNA species comprising 1:10 000-100 000 of the mass of the target poly(A)+RNA, which corresponds to approximately 1 transcript per 100 000, could readily be detected. The intensity of the hybridisation signal for a given gene is a result of its relative abundance in the RNA-derived DNA probe.

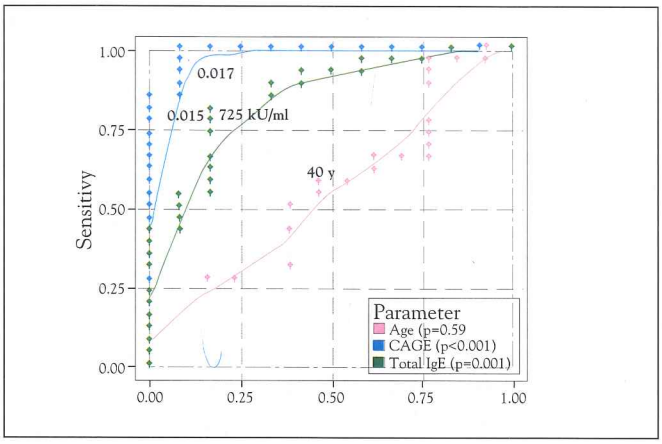

Figure 2. *Receiver-operating characteristic (ROC) analyses to compare the composite atopy gene expression (CAGE) score, total IgE and age for diagnosing atopy. The CAGE score (area=0.96 [0.90-1.00]; p<0.001) was better in differentiating atopic from non-atopic individuals than total IgE (area=0.85 [0.73-0.98]; p=0.001). The ROC curve for age fluctuated around the diagonal and was not significant (area=0.50 [0.30-0.71]; p=0.97). The dots show the relationship between sensitivity and specificity at the respective cut-off values. A smoothing iteration was performed to show continuous ROC curves.*

2. ***Genotyping***: DNA, extracted from a biological sample, is amplified by the polymerase chain reaction and applied to the microarray. The genotype for hundreds or thousands of genetic markers across the genome can be determined in a single experiment. This approach has considerable potential in risk assessment, both in research and clinical practice.
3. ***DNA sequencing***: DNA extracted from a biological sample is amplified and applied to specific "sequencing" microarrays. Thousands of base pairs of DNA can be screened on a single microarray for mutations in specific genes whose normal sequence is already known. This greatly increases the scope for precise molecular diagnosis in single gene and genetically complex diseases.

How to analyse microarray data?

The analysis of microarray data is complex. It involves a stepwise progression of analysis through many levels. These include image segmentation, pixel counting, image analysis, numerical data manipulation and variance statistics, pattern identification, data visualisation, data mining, and data integration. All these processes then hopefully lead to biological interpretation and useful biological insights. The exact path of analysis for each experiment depends on the scale and goals of each project as well as the expectations of the investigator. However, all array methods require the construction of databases for the management of information on the genes represented on the array, the primary results of hybridisation and the construction of algorithms to make it possible to examine the outputs from single and multiple array experiments (4). Software tools have been developed to manage these huge data sets across many experiments; to query, sort, cluster and visualise by time, behaviour, function, chromosomal position, pathway and other phenotypic or experimental parameters; and to link the results to other information databases.

Application of microarrays in human pathology with relevance to respiratory medicine

In the clinical application microarrays have strengths in four areas: 1. It is possible to identify individuals at risk for certain diseases by looking for **disease susceptibility** genes. Such patients can be included in specific disease prevention programs, 2. Microarrays can help to establish the **correct diagnosis efficiently and early** in the disease process, 3. Microarrays can be used to measure **reliable prognostic markers and gene expression scores**, 4. It is possible -in future- to apply **an individualised treatment** according to the patient's gene expression profile. Therefore, the treatment with an optimised effect to side effect potential in each patient will be chosen on an individual basis.

Microarray studies in lung disease

Kaminski et al. (5) investigated bleomycin-induced pulmonary fibrosis in mice. While bleomycin induces lung inflammation and fibrosis in wild-type mice, mice deficient in epithelium-restricted integrin beta 6-subunit (beta 6-/-) develop exaggerated inflammation but are protected from pulmonary fibrosis. Comparative analysis of gene expression profiles during bleomycin-induced pulmonary disease in wild-type and beta 6-/- mice distinguishes gene clusters involved in the inflammatory response from gene clusters that mediate specifically fibrotic responses. Luzina et al. (6) analysed gene expression in bronchoalveolar lavage cells from scleroderma patients with and without interstitial lung

disease and found indications for T-cell recruitment and macrophage activation in scleroderma patients at greater risk for lung fibrosis. Own gene chip studies found a phenotype-specific reduction in apoptosis signals in atopy, asthma (7) and sarcoidosis (8) as compared to healthy controls *in vivo*. In another own study in patients with sarcoidosis, the expression of specific growth factor-related genes was associated with progressive disease and therefore predicted outcome (9).

We designed a composite gene expression score (CAGE) to quantify atopy and asthma (10). The CAGE score was better than total IgE in differentiating atopic from non-atopic subjects (sensitivity 96%, specificity 92%; Figure 2). Correlation between the CAGE score and total IgE ($p<0.001$) was found and there was a trend for correlation with asthma severity ($p=0.051$). So far, an individual is considered either being atopic or not. However, as experienced in clinical practice, some patients seem to be more allergic then others. In this situation, the composite atopy gene expression score is helpful to decide upon the best therapy. A further in vivo-study investigated the differences in B-cell isotype control mechanisms in atopy and asthma compared to healthy control subjects (11).

A gene expression analysis was performed on lung tissue to identify the genes expressed in the lung tissue of emphysema patients but not in that of patients with healthy lungs. We detected 142 differentially expressed genes. One of these genes, secreted frizzled-related protein sFRP-1, an inhibitor of Wnt signaling, was further characterized by its role in the pathophysiology of emphysema (12). In order to determine transcriptional programs that are active in human pulmonary fibrosis, we analyzed gene expression patterns using GeneChip microarrays (Affymetrix; Santa Clara, CA) in samples from patients with histologic diagnoses of usual interstitial pneumonia (UIP), samples obtained from lung resections for cancer that were determined normal by histologic examination, and pooled normal-lung RNA obtained commercially. Surprisingly, gene expression patterns were quite homogenous among pulmonary fibrosis and control samples, despite the expected variability in genetic background and condition of subjects. Using the total number of misclassifications score, the information-content score, and Gaussian-error score, we determined that 100 genes were significantly differentially expressed between UIP and control lungs. Groups of genes that were significantly overexpressed included metalloproteases, extracellular matrix-related genes, markers of smooth-muscle differentiation, and antioxidants (5).

The pathological distinction between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung can be cumbersome using established methods. We propose that a simple technique, based on the expression levels of a small number of genes, can be useful in the early and accurate diagnosis of MPM and lung cancer. This method is designed to accurately distinguish between genetically disparate tissues using gene expression ratios and rationally chosen thresholds. Here we have tested the fidelity of ratio-based diagnosis in differentiating between MPM and lung cancer in 181 tissue samples (31 MPM and 150 ADCA). A training set of 32 samples (16 MPM and 16 ADCA) was used to identify pairs of genes with highly significant, inversely correlated expression levels to form a total of 15 diagnostic ratios using expression profiling data. Any single ratio of the 15 examined was at least 90% accurate in predicting diagnosis for the remaining 149 samples (*e.g.*, test set). We then examined (in the test set) the accuracy of multiple ratios combined to form a simple diagnostic tool. Using two and three expression ratios, we found that the differential diagnoses of MPM and lung ADCA were 95% and 99% accurate, respectively. We propose that using gene expression ratios is an accurate and inexpensive technique with direct clinical applicability for distinguishing between MPM and lung cancer. Furthermore, we provide evidence suggesting that this technique can be equally accurate in other clinical scenarios. (13).

Histopathology is insufficient to predict disease progression and clinical outcome in lung adenocarcinoma. Gene-expression profiles based on microarray analysis can be used to predict patient survival in early-stage lung adenocarcinomas. Genes most related to survival were identified with univariate Cox analysis. Using either two equivalent but independent training and testing sets, or 'leave-one-out' cross-validation analysis with all tumors, a risk index based on the top 50 genes identified low-risk and high-risk stage I lung adenocarcinomas, which differed significantly with respect to survival. This risk index was then validated using an independent sample of lung adenocarcinomas that predicted high- and low-risk groups. This index included genes not previously associated with survival. The identification of a set of genes that predict survival in early-stage lung adenocarcinoma allows delineation of a high-risk group that may benefit from adjuvant therapy (14).

Outlook – Medicine in 2010

It is likely that some diseases will be categorised differently according to phenotypes, which better match with genotypes and with specific gene expression signatures. This will be particularly useful in clinical situations, where different therapeutic strategies apply. Such diseases could be malignant diseases, chronic multi-organ disorders, chronic obstructive lung disease, heart failure, and many others - thus, diseases with different pathogenetic mechanisms but similar phenotypic presentation. Along these lines microarrays could be used to identify individuals at risk for certain

conditions, to establish the exact and early diagnosis, to establish reliable prognosis and to give guidance for therapy. It is possible that the need for some classic diagnostic procedures will be reduced. The increased clinical information provided by microarrays should assure their entry into routine clinical practice within the next three to five years, although the added costs will have to be justified by the clinical benefit.

Due to the complexity of the matter microarray-facilitated medicine will first happen in specialised centres before being introduced broadly. Physicians need to be trained in molecular biology for a successful introduction of gene chips in clinical medicine.

Conclusion

Functional genomics will undoubtedly help to improve screening, early detection/diagnosis, prognostic markers and will enable individualised treatment strategies. Most microarray based tests are still in the development stage, though substantial progress towards commercialisation has occurred in some cases. Like any new diagnostic tool, microarrays will have to be rigorously appraised for sensitivity, specificity, and predictive value. The high costs of microarray based tests will inevitably limit the speed with which they are introduced into clinical practice and initially restrict their use to specialised centres. However, given the huge potential gain in clinically relevant information for individual patients and their diseases, the technology is likely to reach most large hospitals within the next 10 years.

References

1. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304-51.
2. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
3. Peng J, Gygi SP. Proteomics: the move to mixtures. J Mass Spectrom 2001;36:1083-91.
4. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19-23.
5. Kaminski N, Allard JD, Pittet JF, et al. Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. Proc Natl Acad Sci USA 2000;97:1778-83.
6. Luzina IG, Atamas SP, Wise R, et al. Gene expression in bronchoalveolar lavage cells from scleroderma patients. Am J Respir Cell Mol Biol 2002;26:549-57.
7. Brutsche MH, Brutsche IC, Wood P, et al. Apoptosis signals in atopy and asthma measured with cDNA arrays. Clin Exp Immunol 2001;123:181-7.
8. Rutherford RM, Kehren J, Staedtler F, et al. Functional genomics in sarcoidosis-reduced or increased apoptosis? Swiss Med Wkly 2001;131:459-70.
9. Eryüksel E, Rutherford R, Bihl M, et al. Specific pattern of growth factor gene expression in stage I versus stage II/III sarcoidosis. Eur Respir J 2002:A 1414.
10. Brutsche MH, Joos L, Carlen Brutsche IE, et al. Array-based diagnostic gene-expression score for atopy and asthma. J Allergy Clin Immunol 2002; 109:271-3.
11. Brutsche MH, Brutsche IC, Wood P, et al. B-cell isotype control in atopy and asthma assessed with cDNA array technology. Am J Physiol Lung Cell Mol Physiol 2001;280:L627-37.
12. Imai K, D'Armiento J. Differential gene expression of sFRP-1 and apoptosis in pulmonary emphysema. Chest 2002;121:7S.
13. Gordon GJ, Jensen RV, Hsiao LL, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963-7.
14. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816-24.