Biomarker discovery for asthma phenotyping: From gene expression to the clinic

Wagener, A.H.

Citation for published version (APA):
Wagener, A. H. (2016). Biomarker discovery for asthma phenotyping: From gene expression to the clinic.

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 2
Towards composite molecular signatures in the phenotyping of asthma

Ariane H. Wagener, Ching Yong Yick, Paul Brinkman, Marc P. van der Schee, Niki Fens, Peter J. Sterk

ABSTRACT

The complex biology of respiratory diseases such as asthma is currently feeding the discovery of various disease phenotypes. Although the clinical management of asthma phenotypes by using a single biomarker (e.g. sputum eosinophils) is already very successful, emerging evidence shows the requirement of multi-scale, high-dimensional biological and clinical measurements in order to capture the complexity of various asthma phenotypes. High-throughput ‘omics’ technologies including transcriptomics, proteomics, lipidomics and metabolomics are increasingly standardized for biomarker discovery in asthma. The leading principle is obeying available guidelines on ‘omics’ analysis, thereby strictly limiting false discovery.

In this review we first address the concept of transcriptomics using either microarrays or next-generation RNA sequencing and their applications in asthma, highlighting the strengths and limitations of both techniques. Next we review metabolomics in exhaled air (breathomics) as a non-invasive alternative for sampling the airways directly. These developments will inevitably lead to the integration of molecular signatures in the phenotyping of asthma and other diseases.
BIOMARKER SIGNATURES

Respiratory diseases exhibit far from simple pathophysiology, which has hampered progress in this prominent medical field. The most prevalent respiratory diseases are chronic, exhibiting multiple pathogenetic mechanisms, various pathophysiological pathways in parallel and diverse clinical expressions. This complexity is not even stable and appears to vary during the course of these diseases. Inevitably, this leads to various disease phenotypes. It is increasingly recognised that capturing those phenotypes is a prerequisite for adequate assessment, monitoring and treatment. Nevertheless, most guidelines on chronic respiratory diseases are still based on establishing traditional diagnoses such as asthma and chronic obstructive pulmonary disease (COPD).

Complex phenotypes

A phenotype can be defined as the composite of observable characteristics of an organism, resulting from interaction between its genetic make-up and environmental influences, which is relatively stable but not invariable with time. This definition does not allow a strict separation between adapted (health) and non-adapted (disease) phenotypes. It merely provides the tools for a phenomenological description of living organisms in health or various diseased states. The physicist and Nobelist Erwin Schrödinger recognized about 70 years ago that life must be complex and non-linear, allowing emergent phenomena that might be interpreted healthy and disease states (1). Based on this non-linearity it can be predicted that it takes more than a few clinical and serum markers to capture the true biomedical entity and time fluctuations of complex diseases (2). This is likely to require multi-scale, high-dimensional biological as well as clinical measurements (3). For centuries clinicians have been familiar with integrating clinical information in health and disease (Table 1). When taking asthma as the example, single symptoms (e.g. wheezing) represent the cornerstone of the diagnostic process. Second, multiple disease features in combination (e.g. the asthma control questionnaire ACQ, or a Th2-high profile) are considered to be even more informative. And finally, the most valuable medical information relies on pattern recognition of composite information (e.g. exacerbation or inflammation). The latter pattern recognition is principally empiric and relies on repeated training and validation. All doctors feel they can establish an exacerbation of asthma, whilst after long efforts the definition of an exacerbation has only recently been distilled (4).

Biomarkers

In general, the promise of delivering valuable cellular and molecular biomarkers to the clinic has not been met (5). Asthma seems to be amongst the exceptions to the rule, since sputum eosinophils appear to be predictive in guiding asthma treatment (6).
What is hampering further progress in biomarker development? As indicated above, based on the complex non-linear underlying biology it can be predicted that personal ‘omics’ profiling is required to reveal medical phenotypes (7). Computational pattern recognition of such biomarker signatures will then be analogous to powerful clinical pattern recognition of complex entities (Table 1). Therefore, it can be envisaged that high-throughput ‘omics’ technologies will allow a step change in biomarker discovery and application in respiratory medicine (8). The first evidence that this approach can be successful in differential diagnoses is emerging (9). Subsequently, these technologies need to be applied in the subphenotyping of patients with airways diseases, such as asthma.

The current requirement is to rigorously standardize the application and analysis of these ‘omics’ technologies in the discovery of biomarker signatures. Fortunately, there are recent recommendations available on the computational and staged validation of ‘omics’ biomarker signatures (10). This also includes strict strategies to limit false discoveries (11). It is envisaged that in this way biomarker signatures will gradually enter clinical medicine, and asthma phenotyping in particular.

**Table 1.** From single to composite biomarkers

<table>
<thead>
<tr>
<th>Molecular</th>
<th>SNP</th>
<th>FE(_{NO})</th>
<th>Periostin</th>
<th>Oxidative stress</th>
<th>Genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular</td>
<td>Eosinophil counts</td>
<td>FACS</td>
<td>Th2-high profile</td>
<td>Differential cell counts</td>
<td></td>
</tr>
<tr>
<td>Histological</td>
<td>Reticular layer thickness</td>
<td>Extracellular matrix composition</td>
<td>Inflammation Remodelling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>FEV(_1)</td>
<td>Variable airways obstruction</td>
<td>Exercise intolerance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Wheezing</td>
<td>ACQ</td>
<td>Exacerbation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From single to composite biomarkers: examples from respiratory research. Diseases and disease states are defined by combining various markers from the clinical level (bottom) towards the molecular level (top). Such disease markers can be derived from singular features (left), panels of combined disease markers (middle) or composite high-dimensional signatures or fingerprints (right). The latter are based on pattern recognition. Such composite patterns (e.g. an ‘exacerbation’) have been, and still are, clinically very effective in disease management. Based on the complex, non-linear biology in health and disease it can be postulated that composite molecular signatures will also provide more useful phenotypic information than singular molecular biomarkers.

SNP, single nucleotide polymorphism; FE\(_{NO}\), fractional concentration of expired nitric oxide; FACS, fluorescence-activated cell sorting; Th2, T-helper type 2 phenotype; FEV\(_1\), forced expiratory volume in one-second as measured by spirometry; ACQ, asthma control questionnaire.

From single to composite biomarkers: examples from respiratory research. Diseases and disease states are defined by combining various markers from the clinical level (bottom) towards the molecular level (top). Such disease markers can be derived from singular features (left), panels of combined disease markers (middle) or composite high-dimensional signatures or fingerprints (right). The latter are based on pattern recognition. Such composite patterns (e.g. an ‘exacerbation’) have been, and still are, clinically very effective in disease management. Based on the complex, non-linear biology in health and disease it can be postulated that composite molecular signatures will also provide more useful phenotypic information than singular molecular biomarkers.

SNP, single nucleotide polymorphism; FE\(_{NO}\), fractional concentration of expired nitric oxide; FACS, fluorescence-activated cell sorting; Th2, T-helper type 2 phenotype; FEV\(_1\), forced expiratory volume in one-second as measured by spirometry; ACQ, asthma control questionnaire.
The current review focuses on the strengths and limitations of transcriptomics by microarrays and next generation RNA-sequencing. In addition, we will shortly address metabolomics in exhaled air as a patient-friendly option for the clinic.

**TRANSCRIPTOMICS ANALYSIS BY MICROARRAYS**

Gene transcription is a determinant of the phenotypic manifestation. Measurement technologies for profiling RNA, using e.g. reverse-transcription PCR and Northern blots, have been available for years. But the ability of high-throughput quantification of the transcriptome (RNA transcripts) was first made possible by microarrays (12), and has been used successfully since with more than 40,000 citations in PubMed (13). Microarrays are based on patches of short oligonucleotide probes which are complimentary to the studied transcripts. In brief, RNA is extracted from the material to be investigated and labeled with fluorescent dyes. The core principal is the hybridization of the studied transcripts to the array. Because the transcripts are fluorescently labeled, the light intensity scanned is a measure of gene expression. In this way, variations in gene expression of a large number of genes can be studied simultaneously.

Since microarrays have been widely used by many groups, the increasing experience has revealed the limitations and biases of microarray data, which has led to adaptations in study design and analyses (13). After formulating the research question, the sample needs to be carefully selected and a power calculation must be performed beforehand (14). For conducting a study, new samples can be collected, though existing publicly available data can also be used to answer a specific research question, e.g. NCBI Gene expression Omnibus (GEO). Regardless of how the data is obtained, the original samples used need to correspond to the research question and to specific quality requirements which will be further discussed. First of all, the RNA quality has to be taken into account, because adjustments for variable and low quality RNA have been developed to extend sample suitability for analysis (15). Next, deviations in the sample collection and processing protocols can have large impact on the omics results. Although an accurate clinical design is the most important strategy to minimize varying factors among samples, normalization steps in the data analysis are mandatory to reduce external noise (16).

In gene expression analysis, usually two different groups of tissue samples or conditions are compared, generating a huge amount of data with an expected large number of false positive results. Adjustment for multiple comparisons to control the false discovery rate (FDR) for differentially expressed individual genes is essential, with the Benjamini and Hochberg's method (17) being the most popular form. Traditional statistical tests such as t-test or the Wilcoxon test have been replaced by newly developed tests that are more appropriate for the analysis of microarray data, such as the empirical Bayes
test (18) that borrows information across genes. In addition, the change in gene expression is represented by the log fold change calculated. Although criteria for both p-value and fold change are more and more required to be taken into account for differentially expressed genes, and several approaches to this issue are suggested (19;20), there is no consensus yet as to the exact criteria. A gene that is statistically differentially expressed between two conditions with a larger fold change difference might not necessarily have a larger impact within a molecular pathway.

**Interpretation**

Finally, the interpretation of the results is yet another challenge, because of the usual ‘overdoses’ of data generated. Cluster analysis is one way to reduce the data into subgroups of related genes in an unbiased way, though evaluation of the stability of the obtained clusters is needed (21). Commonly used examples of clustering are hierarchical clustering, k-means clustering and self-organizing maps (22). Different from the analysis on individual gene differences, the analysis of gene signatures combines gene expression of a group of genes and can be used for disease prediction and phenotype discovery. Since many reported molecular signatures have failed reproducibility, specific statistical strategies and validation of a signature is of high importance to reduce false signature discovery (10). Furthermore, network-based approaches are attractive in view of studying function and relationships between genes, instead of studying individual genes. First of all, features can be mapped in known pathways, as used in the Gene Set Enrichment Analysys (GSEA) (23). A list of known pathways is generated with assigned scores, referring to if the pathway is likely to be differentially expressed. Secondly molecular features can be mapped in more unbiased interaction networks, such as protein-protein or protein-DNA networks (Figure 1), showing closely connected genes within a pathway. More methods are available for analyzing gene expression data since this field is rapidly expanding (24).

In pulmonary diseases transcriptomics analysis using microarrays have shown to be promising in biomarker discovery and to improve our understanding of the disease. Concerning biomarker discovery, gene expression profiling of epithelial brushings in asthma identified distinct subtypes of patients with mild to moderate asthma with a Th2-high or Th2-low phenotype (25). Subsequently, patients with severe asthma with high levels of periostin, corresponding to the Th2-high phenotype, appeared to be those who responded to anti-IL13 treatment (26). To examine mechanisms of disease pathogenesis, the effect of cigarette smoke on transcriptome patterns in bronchial epithelial cells of smokers with and without COPD revealed a role for oxidative stress responses in patients with COPD (27). We have recently used microarrays to study the upper and lower airway epithelium transcriptome of patients with asthma and healthy controls in response to a synthetic analog of viral dsRNA, and observed fewer genes differen-
Partially expressed in asthma and a diminished response in the upper airways of patients with asthma as compared to controls (28). Furthermore, gene expression analysis has shown to improve our understanding of the mechanisms of treatment resistance. Gene profiling of bronchoalveolar lavage (BAL) cells of patients with corticosteroid-resistant (CR) asthma and corticosteroid-sensitive (CS) asthma identified genes supporting involvement of endotoxin (LPS) and classical macrophage activation in corticosteroid resistance in asthma (29).

**Figure 1. Microarray data analysis**
Example of a direct interaction network (Genespring GX12) of genes differentially expressed in the upper airway epithelium of patients with asthma in response to ex-vivo dsRNA as obtained by microarray analysis (Affymetrix Human Genome U133+ PM Genechip Array). The figure shows connected genes (yellow dots) involved in the extracellular matrix, the cellular membrane, the intracellular compartment, and the nucleus (from top to bottom). The network analysis revealed a number of genes that were highly connected to other genes in the network. These highly connected genes, the key components in the network, are encircled and enlarged.

MMP1, matrix metallopeptidase-1; MMP9, matrix metallopeptidase-9; IL6, interleukin-6; IL8, interleukin-8; STAT5A, signal transducer and activator of transcription-5A; CDK1, cyclin-dependent kinase-1; IRF1, interferon regulatory factor-1; IL15, interleukin-15; STAT1, signal transducer and activator of transcription-1; TLR3, toll-like receptor-3; CCL5, chemokine ligand-5.
Taken together, microarray analysis of biological samples is providing highly useful data for hypothesis generating purposes in asthma. This technology may eventually also serve clinical subphenotyping, since already 3 subtypes related to both clinical asthma status and airway inflammation were identified by induced sputum gene expression profiles, including the inflammatory pathways involved (30). Table 2 summarizes the strengths and limitations of transcriptomics analysis by microarrays.

TRANSCRIPTOMICS ANALYSIS BY GENE SEQUENCING

With gene sequencing the sequence of the nucleotide bases that form the DNA is determined. Already in 1977 gene sequencing was introduced by Frederick Sanger (31). This first-generation sequencing method is also called the dideoxy or chain termination sequencing. In short, the template of the DNA molecule is copied repeatedly. Modified nucleotides, also called chain terminators, are added to the reaction, which will terminate the copy process when incorporated in the copied DNA. Due to the fact that copying begins at a fixed location of the template DNA, but terminates randomly through the incorporation of chain terminators by chance, copy DNA with various lengths will be obtained. By comparing these DNA fragments also known as reads, the original DNA nucleotide sequence can be determined.

In the meantime gene sequencing has undergone many technological advances resulting in the development of next-generation high-throughput gene sequencing techniques, which embroider on the Sanger method. One of the most powerful recent next-generation gene sequencing techniques is transcriptome sequencing (RNA-Seq), which allows a detailed characterization of gene expression profiles at the tissue level (32). Particularly in complex disorders such as asthma and COPD this technology has the potential to discover gene expression profiles that are characteristic of the disease and thus can improve our understanding of the cellular and molecular pathways involved in disease (33;34). In contrast to microarray chips, RNA-Seq allows an unbiased analysis of the transcriptome as it is not dependent on predefined probe sets and is therefore not limited to a selection of known genes or nucleotide sequences (13). Consequently, RNA-Seq facilitates the discovery and characterization of novel, disease-related genes.

The workflow of a typical RNA-Seq process and the method used for the subsequent analysis of the sequence data are highly dependent on the research question. In general, the workflow for RNA-Seq can be divided into 3 steps. In the 1st step, RNA is isolated from which cDNA-libraries are constructed. Various methods can be used including the Ovation RNA-Seq System (NuGEN, San Carlos, CA, US), which has been successfully applied in our gene expression studies (34;35). This is followed by the actual sequencing of the
### Table 2. Strengths and limitations per technique

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Transcriptomics analysis by microarrays</th>
<th>Transcriptomics analysis by gene sequencing</th>
<th>Breathomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maturity of experimental design and analyses</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-throughput, quantitative gene analysis</td>
<td>✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low running costs</td>
<td>✓ (in comparison to RNA-seq)</td>
<td></td>
<td>✓ (applies for eNose)</td>
</tr>
<tr>
<td>Unbiased, not bound to known gene transcripts</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High dynamic range for detection of very low or high expressed genes</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requires only small amount of RNA</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-invasive method allowing measurements in infants and elderly</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signals in asthma have demonstrated adequate repeatability</td>
<td>✓ (49;54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design tailor made devices for specific diseases based on previous results</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Limitations</th>
<th>Transcriptomics analysis by microarrays</th>
<th>Transcriptomics analysis by gene sequencing</th>
<th>Breathomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited to known gene transcripts</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enormous amount of clinically significant and insignificant data generated</td>
<td>✓ ✓ (in case of GC-MS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitive processing with high technical background noise</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extensive demand on bioinformatics and – statistics for storage, processing analysis and interpretation</td>
<td>✓ (55)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insufficient sequencing depth/coverage may hamper the interpretation</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High costs, although decreasing</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time-consuming</td>
<td>✓ ✓ (in case of GC-MS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No international standard for sampling method</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eNose sensor signals are not identical between devices, complicating mapping of breathprints between centres</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table summarizes strengths and limitations per reviewed technique. RNA-seq, transcriptome sequencing; eNose, electronic nose; GC-MS, gas chromatography mass spectrometry.
1. Sequencing

2. Mapping

Figure 2. Workflow RNA-Seq

This figure shows the typical workflow of a RNA-Seq study. The first step comprises the preparation of amplified and purified cDNA from the isolated RNA from the study samples. Next, libraries of cDNA are constructed, followed by the actual sequencing of the samples yielding sequence reads. The figure depicted in step 1 shows a summary of the pyrosequencing chemical reaction, which forms the basis of the gene sequencing method by the GS FLX System (454/Roche). Millions of copies of a single clonal fragment are contained on each DNA capture bead (left). In the second step, the sequence reads are mapped against the reference DNA, e.g. the human genome, enabling the identification of the specific genes present in the sequenced samples and the quantification of those genes. APS, adenosine 5’ phosphosulfate; PP, pyrophosphate.

cDNA-libraries, which will yield sequence reads of various base pair (bp) lengths (Figure 2). In step 2, these reads are mapped to the reference DNA, e.g. the human genome. This enables identification of the specific genes that were present in the sequenced samples in step 1. Furthermore, it allows a quantification of those identified genes, because the more reads of a specific gene are present, the more that gene is expressed. One of the methods used for mapping is the consensus approach, which is in its concept similar to the serial analysis of gene expression (SAGE) method (33). In short, each sequence read is aligned with the reference genome. However, a sequence read may align to multiple locations of the reference genome due to their short bp length compared to the reference. To identify the actual site of origin in the reference genome, the sequence read is put together with other sequence reads that contain overlapping DNA sequences to undergo multiple alignments. This will result in the formation of a contig, which is a contiguous sequence constructed from many clone sequences (36). The eventual bp length of a contig will thus be determined by the length of the sequence tags with overlapping DNA sequences. With an overall longer bp length than the individual sequence tags with which it was constructed, the contig will point towards the actual site of origin.
when aligned to the reference genome. After it has been ascertained which genes were expressed in the sequenced samples, statistical analyses are performed during step 3. One way to analyse the data is to determine the function of an individual gene and associate it with the disease under investigation, e.g. asthma. Additionally, as already discussed above in the Interpretation section concerning microarray data, the set of differentially expressed genes can be used to identify gene networks by pathway analyses. These gene networks may clarify what biochemical processes are differently regulated leading to the manifestation of asthma.

Next-generation high-throughput gene sequencing has become increasingly accessible for research and even clinical purposes mainly due to the significant increase in sample numbers that can be simultaneously sequenced, and decrease in costs and time needed to perform a sequencing analysis (37). Therefore, gene sequencing presents a multitude of possibilities to increase our understanding of many more complex diseases besides asthma and COPD (3;38). When applying gene sequencing, we have recently found 46 genes to be differentially expressed between endobronchial biopsies from asthmatics and controls. This included periostin, pendrin and BLC2 with 10 gene networks (34). The 3 networks with the highest network scores comprised: a) BLC2, MAPK1, NF-κB, p38MAPK, TGF-β, b) STAT3, periostin, STAU2, and c) EGFR and SLC26A4. In a second study we examined whether systemic glucocorticoids affect gene expression of bronchial smooth muscle in asthma, demonstrating 15 significantly changes genes (35). The changes in two of those appeared to be associated with accompanying changes in bronchial hyperresponsiveness, strongly suggesting that glucocorticoids also exert their beneficial effects through activity on bronchial smooth muscle. Hence, by implementing such a sophisticated state-of-the-art biochemical technology in characterizing a disease, breakthroughs in the treatment of diseases may be reached through the development of targeted therapies.

Although RNA-Seq holds many advantages, there are also some drawbacks to this novel technique compared to e.g. qPCR methods. Table 2 points out both limitations and strengths of this technique.

**BREATHOMICS**

Respiratory medicine is in a privileged position when it concerns non-invasive access to composite molecular samples. Exhaled breath contains a complex gas mixture of volatile organic compounds (VOCs) (39) as well as a composite of non-volatile compounds derived from exhaled breath condensate (EBC) (40). These metabolites are derived from both systemic and local metabolic, inflammatory and oxidative processes. The term ‘breathomics’ has recently been coined to cover metabolomics approaches in exhaled
air. Even though multiple labs have had trouble in the validation of specific biomarkers in EBC (40), recent metabolomics strategies in EBC have been very successful. This includes metabolomics by NMR spectroscopy (41) as well as metabolomics by liquid chromatography and mass spectrometry (LC-MS) (42). Hence, ‘omics’ technologies have opened new avenues for EBC, which require stringent validation.

The advantage for breathomics is likely to reside in the assessment of exhaled VOCs. The standard for detecting individual molecular compounds in a VOC mixture is gas chromatography mass spectrometry (GC-MS) (39). GC-MS has been employed in the discovery of biomarkers for inflammatory airway diseases, such as asthma (43) and COPD (44). The VOCs that have been associated with the presence of asthma, COPD or lung cancer are visualized in Figure 3. This shows that these diseases are characterized with partially overlapping combinations of multiple VOCs, thereby highlighting the need for establishing molecular signatures. Interestingly, in asthma and COPD the exhaled VOC profiles appear to be associated with the inflammatory phenotype (eosinophilic

Figure 3. Volatile organic compounds in pulmonary diseases
Exhaled volatile organic compound associated with asthma, COPD and/or lung cancer. Compilation of data published in the literature. The lines between the compounds and the diseases represent associations published in the literature. The type of compound (hydrocarbons, aldehydes, ketones, alcohols, cyclic) is mentioned next to the compounds.
or neutrophilic) (44;45). This suggests that breathomics is suitable for non-invasive sub-phenotyping of inflammatory airway diseases.

Pattern recognition of exhaled VOCs can also be accomplished by electronic noses. Electronic noses (eNoses) are based on arrays of nano-sensors that capture various combinations of VOCs, which allows exhaled air fingerprinting (breathprints) rather than identification of individual chemical constituents (46;47). The nano-sensors are sensi-

---

**Figure 4. Breathprint**

Spider chart of the breathprints of exhaled air collected from 1 patient with severe asthma and 1 healthy control. The exhaled air is analysed by a composite eNose platform that integrates different types of eNoses to measure breathprints in parallel, developed for the U-BIOPRED group. The platform array consists of 190 sensors from 4 different types of eNoses using: 1) carbon-polymer sensors, 2) quartz microbalance metalloporphyrins sensors, 3) metal oxide semiconductor sensors, and 4) field asymmetric ion mobility spectrometry. Every marker in the chart is the signal of one sensor. The signals of the 190 sensors are displayed, normalized towards an arbitrary unit at a scale between 0 (centre) and 100 (outer circle). The different breathprints of the two subjects can be distinguished.
tive to partly overlapping fractions of the VOC mixtures (breathprint) and are based on conducting polymers, metal oxide, metal oxide field effect transistors, surface or bulk acoustic waves, optical sensors, colorimetric sensors, ion mobility spectrometry, infrared spectroscopy, gold nanoparticles, and GC-MS (Figure 4) (46). The pattern recognition algorithms require training and validation and provide probabilistic evidence in favour of (positive predictive value) or against (negative predictive value) particular medical conditions. Such probabilistic evidence is well suited for clinical diagnostics, phenotyping and monitoring, making eNoses a potentially cheap and real-time metabolomics tool in the clinic.

The research application of breathomics in (respiratory) medicine is rapidly expanding, in particular in infectious diseases, lung cancer and airway diseases (47;48). When focusing on asthma, studies conducted by GC-MS or various eNose sensor systems indicate that asthmatics can be discriminated from healthy controls with accuracies between 80-100% (49). Interestingly, asthma patients can also be discriminated from those with COPD. When selecting the non-overlapping extremes amongst patients with a gold-standard diagnosis of asthma or COPD the accuracy by eNose reached 95% (50). Subsequently, when allowing overlap between patients with asthma and COPD (both featuring fixed airflow limitation) the accuracy of separating the two diseases by eNose remained 88% (51). Such external validation in newly recruited (overlapping) patients from different hospitals is essential for limiting the false discovery rate and establishing diagnostic accuracy (10;11). With regard to the capabilities of eNoses to phenotype patients with asthma, preliminary data indicate that eosinophilic and non-eosinophilic asthma can be discriminated when using a composite eNose platform that integrates different types of eNoses in parallel (52). And notably, in a recent study focusing on the prediction of oral glucocorticoid responsiveness amongst patients with asthma, eNose measurements performed even better than sputum eosinophils (53). This suggests that composite molecular signatures can perform better than single biomarkers in the phenotyping of patients with asthma. As indicated above, this may not be surprising. Table 2 summarizes the strengths and limitations of breathomics.

**CONCLUSIONS**

When considering the complex biology of health and disease it can be envisaged that composite molecular fingerprints have the best prospect as biomarkers in the phenotyping of patients. Indeed, recent application of ‘omics’ technologies (e.g. transcriptomics, breathomics) has not only provided signatures of asthma but also of relevant subphenotypes of the disease. For signature discovery, the current state-of-the-art method is transcriptomics analysis using microarrays, especially because of its maturity.
of experimental design and analysis, and available data for comparisons and validation. Though, because RNA sequencing enables higher dynamic detection ranges and a more unbiased analysis, this method will eventually be preferable when costs have decreased and experience has increased. Breathomics is a none-invasive method with diagnostic potential. However, sampling methods and devices need to be standardized to enable comparisons with other centers and validate results. Each of the applied platforms has its strengths and limitations that need to be taken into account in each and every study. The leading principle for this is obeying available guidelines on the process of molecular signature discovery in medicine, thereby strictly limiting false-positive results (10;11). This will inevitably lead to the integration of molecular signatures in the phenotyping of asthma and other diseases (3;7;8).
REFERENCES

(1) Schrödinger E. What is life? Cambridge University Press; 1944.
(13) Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 2011;9:34.
(22) D’haeseleer P. How does gene expression


