EDITOR'S CHOICE

Application of OMICS technologies in occupational and environmental health research; current status and projections

J Vlaanderen,[1] L E Moore,[2] M T Smith,[3] Q Lan,[2] L Zhang,[3] C F Skibola,[3] N Rothman,[2] R Vermeulen[1]

[1]Utrecht University, Division Environmental Epidemiology, Utrecht, the Netherlands [2]Division of Cancer Epidemiology and Genetics, National Institutes of Health, Bethesda, Maryland, USA [3]Division of Environmental Health Sciences, University of California, Berkeley, California, USA

Correspondence to
Jelle Vlaanderen, Institute for Risk Assessment Sciences, Division of Environmental Epidemiology, University Utrecht, Po Box 80178, 3508 TD, Utrecht, the Netherlands; j.j.vlaanderen@uu.nl

Accepted 12 August 2009

ABSTRACT

OMICS technologies are relatively new biomarker discovery tools that can be applied to study large sets of biological molecules. Their application in human observational studies (HOS) has become feasible in recent years due to a spectacular increase in the sensitivity, resolution and throughput of OMICS-based assays. Although, the number of OMICS techniques is ever expanding, the five most developed OMICS technologies are genotyping, transcriptomics, epigenomics, proteomics and metabolomics. These techniques have been applied in HOS to various extents. However, their application in occupational environmental health (OEH) research has been limited. Here, we will discuss the opportunities these new techniques provide for OEH research. In addition we will address difficulties and limitations to the interpretation of the data that is generated by OMICS technologies. To illustrate the current status of the application of OMICS in OEH research, we will provide examples of studies that used OMICS technologies to investigate human health effects of two well-known toxicants, benzene and arsenic.

In the biological sciences the suffix -omics is used to refer to the study of large sets of biological molecules.[1] The idea that the field of molecular biology needed to move from studying isolated biological molecules towards a broad analysis of large sets of biological molecules was underscored with the completion of the human genome project (HGP) in 2001.[2 3] The HGP demonstrated that a relatively limited number of genes could be identified in the human genome, which substantiated the theory that complex biological processes were regulated on other levels than DNA sequence alone. This realisation triggered the rapid development of several fields in molecular biology that together are described with the term "OMICS". The OMICS field ranges from genomics (focused on the genome) to proteomics (focused on large sets of proteins, the proteome) and metabolomics (focused on large sets of small molecules, the metabolome). We divide the field of genomics into genotyping (focused on the genome sequence), transcriptomics (focused on genomic expression) and epigenomics (focused on epigenetic regulation of genome expression). An overview of the different OMICS fields that will be discussed in this paper is presented in table 1. In this review we define the field of occupational and environmental health (OEH) research as the study of interactions between the following domains: environment (the exposome),[4] individual (genetic) susceptibility (the (epi)genome), and biological outcomes (the responsome)[5] (figure 1). In this context, biological outcomes can be defined as clinical diseases as well as relevant (preclinical) intermediate endpoints. In theory, OMICS technologies have a large potential value for OEH research because the environment is known to influence many of the described processes and therefore OMICS technologies are likely to provide valuable information especially where the three domains overlap. Although the field of OMICS is ever expanding (eg, see http://omics.org), currently five different OMICS fields are well established: genotyping, gene expression profiling, epigenomics, proteomics, and metabolomics. In this paper, we will address the spectacular increase in sensitivity, resolution and throughput of OMICS-based techniques in recent years, and we will discuss the difficulties regarding the interpretation of data generated by these techniques. To illustrate the current status of the application of OMICS in OEH research and the progress that has been made in recent years, we will provide examples of studies that have used OMICS technologies to investigate human health effects of two well-known environmental/occupational toxicants, benzene and arsenic.

OVERVIEW OF OMICS TECHNOLOGIES

Genomics

We divide the field of genomics into genotyping, transcriptomics, and epigenomics.

Genotyping

Genotyping is focused on the identification of the physiological function of genes and the elucidation of the role of specific genes in disease susceptibility.[6] The HGP has provided insight in the number of genes and their location in the human genome.[2 3 7] This knowledge in combination with major technological improvements resulted in the development of assays that are able to assess variability in the DNA sequence of many thousands of genes in a single experiment. This development has opened the possibility to study the combined effect of variability in multiple genes on the development of complex diseases. While several types of genetic variation exist (eg, insertions and deletions of nucleotide base pairs and copy number variations (CNVs)), single nucleotide polymorphisms (SNPs) are the most commonly investigated.[2] At this moment over nine million detected SNPs are available in public databases.[8 9] Because SNPs are highly

Table 1 Overview of the different OMICS technologies

Technology	Molecules of interest	Definition	Temporal variance	Influence by disease status
Genotyping	DNA	Assessment of variability in DNA sequence in the genome	None	No
Epigenomics	Epigenetic modifications of DNA	Assessment of factors that regulate gene expression without changing DNA sequence of the genome	Low/moderate	Probable
Gene expression profiling	RNA	Assessment of variability in composition and abundance of the transcriptome	High	Yes
Proteomics	Proteins	Assessment of variability in composition and abundance of the proteome	High	Yes
Metabolomics	Small molecules	Assessment of variability in composition and abundance of the metabolome	High	Yes

abundant in the human genome, they are commonly used as markers for genetic variation in disease–gene association studies.[10] Due to limited genetic variation and haplotype structure and a high level of linkage disequilibrium within small regions of the genome, a subset of informative SNPs, called tag SNPs, can be genotyped as proxies for haplotype blocks to identify regional associations that influence disease or phenotypes of interest.[11] Fine mapping (eg, sequencing) can further narrow the associated region in the search for the true causal variant(s). However, functional studies are needed to test whether associated SNPs alter the structure or function of DNA, RNA or proteins and influence phenotypes. Among others, functional SNPs might alter peptide sequences, transcription factor binding sites and exonic splicing enhancer/suppressor sites.

The first SNP-based studies focused on ≥1 SNPs per gene in a limited set of candidate genes. However, since the introduction of array-based genotyping techniques, allowing the simultaneous assessment of up to one million SNPs in a single assay, it has become possible to cover, with varying resolution, the entire genome in what are now commonly referred to as genome-wide association studies (GWAS). These GWAS have uncovered, and will continue to uncover, interesting and previously unknown polymorphic variants that are associated with a variety of chronic diseases. The effect sizes of these findings have in general been small (OR 1.2–1.5) fuelling debates on positive interactions between one or more common variants and the environment.[12]

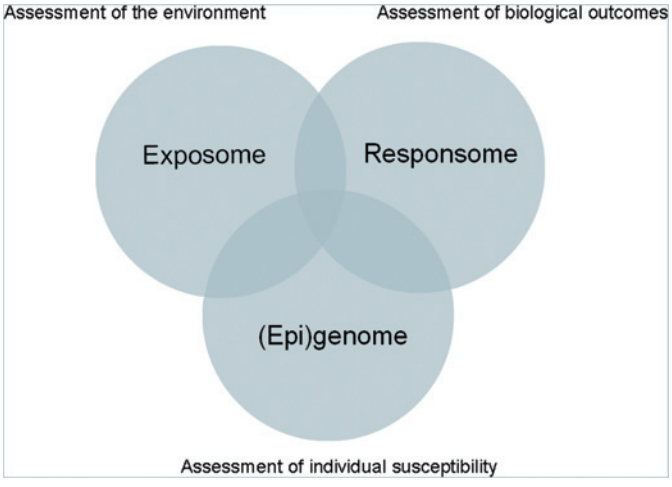

Figure 1 OMICS within the domain of occupational and environmental health. Genotyping operates completely within the domain of genomics. The other OMICS technologies operate in the intersection between the exposome (assessment of the environment), the responsome (assessment of health effects) and the (epi)genome (assessment of individual genetic susceptibility).

Yet, identifying these gene–environment interactions will be difficult in ongoing GWAS given the low prevalence of exposures and/or the poor characterisation of environmental exposures in these large, often multicentre/country studies. As such, OEH research can play an important role in the identification of gene–environment interactions as the exposure is more prevalent and assessed with greater accuracy than in population- or hospital-based case-control studies that have provided most GWAS to date. Of course, sample sizes will likely be much smaller in these studies limiting the statistical power, and therefore the number of SNPs that can be tested simultaneously. Until recently most OEH studies on gene–environment have been focused on candidate genes, where the success depends on previous knowledge and ability for selection of candidate genes.[13] Application of GWAS has been limited except in a study on exposure to environmental tobacco smoke.[14] The application of GWAS to OEH studies will, however, result in some computational challenges as the number of genes that have a possible interaction with the exposure are large. Recently, several papers have proposed new statistical approaches for gene–environment-wide interaction studies which minimise the type 1 error (ie, false positives) while gaining efficiency and power.[15–17]

Although they occur less frequently than SNPs CNVs play an important role in genetic variation.[18] CNVs are caused by genomic structural variations such as insertions, deletions, and duplications and have been defined as "segments of DNA that are 1 kb or larger and present at variable copy number in comparison with a reference genome".[19] CNVs located in gene promoter regions can influence gene expression, and might influence the development of complex disease traits where gene dosage is altered but not abolished.[19] CNVs proximal to genes but not in promoter sequences could perturb the "histone code" and also influence gene expression. Further, CNVs located in exons could result in mis-spliced mRNA with detrimental effects on protein expression. Techniques that have been used to assess CNVs in the genome include comparative genomic hybridisation, a technique that compares labelled DNA from individuals in a study population with differently labelled reference genomic DNA,[20] and SNP-based platforms that use allele intensity ratios to make inferences about CNVs.[19] CNV has been frequently assessed in studies that investigated the effects of the *glutathione* S-*transferase M1* (*GSTM1*) gene on environment–cancer associations.[21 22] To date most studies assessed the effect of having the null genotype (deletion) of *GSTM1* gene versus having at least one copy of the gene. Recent studies were also able to assess gene dosage effects (ie, does having two copies of the *GSTM1* gene result in stronger associations with cancer than having one copy?).[23 24]

Transcriptomics

The abundance of specific mRNA transcripts in a biological sample is a reflection of the expression levels of the

corresponding genes.[25] Gene expression profiling is the identification and characterisation of the mixture of mRNA that is present in a specific sample. An important application of gene expression profiling is to associate differences in mRNA mixtures originating from different groups of individuals to phenotypic differences between the groups.[26] In contrast to genotyping, gene expression profiling allows characterisation of the level of gene expression. Both the presence of specific forms of mRNA and the levels in which these forms occur are parameters that provide information on gene expression.[27] The transcriptome in contrast to the genome is highly variable over time, between cell types and will change in response to environmental changes (table 1). A gene expression profile provides a quantitative overview of the mRNA transcripts that were present in a sample at the time of collection. Therefore, gene expression profiling can be used to determine which genes are differently expressed as a result of changes in environmental conditions. A typical gene expression profiling study includes a group of individuals with similar phenotype (eg, exposure level, disease status) and compares the gene expression profile of this group to the profile of a reference group matched on selected factors such as age and sex to the group of interest. Studies of this type usually report a set of genes that are differently expressed between the groups.

Epigenomics

The focus of epigenomics is to study epigenetic processes on a large (ultimately genome-wide) scale.[28 29] Epigenetic processes are mechanisms other than changes in DNA sequence that are involved in local activity states such as gene transcription and gene silencing.[30–32] Although the range of epigenetic mechanisms that are discovered is expanding, epigenomics is mainly based on two most comprehensively studied mechanisms, DNA methylation and histone modification.[28 33–39] However, in recent years RNA interference of gene expression by non-coding RNAs such as microRNA and siRNA has acquired considerable attention.[31 40 41] Changes in DNA methylation, histone modification and RNA interference are often associated and it is believed that interaction exists between these epigenetic processes.[31] Here, the focus will be on DNA methylation and histone modification. DNA methylation is the addition of a methyl group to cytosine in a CpG dinucleotide. A distinction is made between global methylation and CpG island-specific methylation. About 70% of the CpG dinucleotides in the human genome are methylated. However, CpG dinucleotides in CpG islands are predominantly unmethylated.[38] Hypermethylation of CpG islands located in promoter regions of genes is related to gene silencing. Under normal conditions gene silencing is related to phenomena such as genomic imprinting, x-chromosome inactivation and tissue-specific gene expression.[28 36] Altered gene silencing plays a causal role in human disease.[31 34 37 38 42] The effect of hypomethylation of the genome outside CpG islands is less well understood but may be involved in chromosomal instability.[32 38] Histone proteins are involved in the structural packaging of DNA in the chromatin complex. Post-translational histone modifications such as acetylation and methylation are believed to regulate chromatin structure and therefore gene expression.[34 37]

Proteomics

In general the function of cells can be described by the proteins that are present in the intra- and intercellular space and the abundance of these proteins.[43] Although all proteins are based on mRNA precursors, post-translational modifications (PTMs) and environmental interactions make it impossible to predict abundance of specific proteins based on gene expression analysis alone. The proteome consists of all proteins present in specific cell types or tissue. In contrast to the genome, the proteome is highly variable over time, between cell types and will change in response to changes in its environment.[44] Proteomics provides insights into the role proteins have in biological systems. A major challenge is the high variability in proteins and protein abundance in certain types of biological samples (eg, the concentration of proteins in plasma ranges up to nine orders of magnitude).[45] This requires the development of technologies that can detect a wide range of proteins in samples from different origins.[46] Many proteomic technologies are currently available but broadly a distinction can be made between approaches that are based on detection by mass spectrometry and protein microarrays using capturing agents such as antibodies. An important focus is the identification of proteins including the presence of PTMs of proteins and identification of proteins interacting in protein complexes.[43 44] Another focus of proteomics is quantification of the protein abundance. Protein expression levels represent the balance between translation and degradation of proteins in cells. It is therefore assumed that the abundance of a specific protein is related to its role in cell function. However, the high dynamic range (ie, the ratio between the smallest and largest concentration and/or mass value) of proteins complicates this type of proteomic analysis.[43 44]

Metabolomics

Metabolic phenotypes are the by-products that result from the interaction between genetic, environmental, lifestyle and other factors.[47] The metabolome consists of small molecules (eg, lipids or vitamins) that are also known as metabolites.[48] Metabolites are involved in the energy transmission in cells (metabolism) by interacting with other biological molecules following metabolic pathways. Metabolomics is defined as the study of metabolic profiles in easily collected biological samples such as urine, saliva or plasma.[48] The metabolome is highly variable and time dependent, and it consists of a wide range of chemical structures (table 1). An important challenge of metabolomics is to acquire qualitative and quantitative information concerning the metabolites that occur under normal circumstances in order to be able to detect perturbations in the complement of metabolites as a result of changes in environmental factors.

CHALLENGES FOR THE APPLICATION OF OMICS IN OEH

The development of new OMICS technologies is an important first step towards implementation of OMICS markers in OEH. However, similar to other (bio)markers of exposure, susceptibility and effect, the successful implementation of OMICS markers in OEH requires appropriate study designs, thorough validation of markers, and careful interpretation of study results.[49–51]

Study design

As indicated in table 1 the transcriptome, proteome and metabolome are highly variable over time and are likely to be influenced by the disease process. This indicates that great care should be given to the timing of biological sample collection and adequate processing (eg, field stabilisation of mRNA) of the sample to minimise measurement error and to avoid potential differential misclassification biases. In table 2 the advantages and disadvantages of the different human observational study (HOS) designs with regard to the collection and use of biological markers are given. In general, it can be stated that hospital-based case-control studies are the least suitable for the application of

Table 2 Comparison of advantages and limitations relevant to the collection of biological specimens and data interpretation in molecular epidemiology study designs (adapted from Garcia-Closas *et al*, 2006[49])

Study design	Advantages	Limitations
Cross-sectional	► Facilitates intense collection and timely processing of specimens (eg, freshly frozen samples, cryopreserved lymphocytes) ► Allows detailed collection of exposure and confounder information	► Relevance of intermediate endpoints altered by current exposures in healthy individuals not always clear
Hospital-based case control	► Facilitates intense collection and timely processing of specimens (eg, freshly frozen samples, cryopreserved lymphocytes) ► Participation rates for biological collections might be enhanced ► Facilitates follow-up of cases for treatment response and survival	► More prone to selection and differential biases than other designs ► Some biomarkers might be affected by disease process or hospital stay
Population-based case control	► Less subject to biases (eg, selection, exposure misclassification) than hospital-based studies	► Some biomarkers might be affected by disease process ► May be more difficult to obtain high participation rates for biological collection than hospital-based designs ► Implementation of intense, specialised blood and tumour collection and processing protocols can be challenging
Prospective cohort	► Allows study of multiple disease endpoints ► Allows study of transient biomarkers and biomarkers affected by disease status ► Selection bias and differential misclassification are avoided: non-differential misclassification might be reduced for some exposures ► Nested case-control or case-cohort studies can be used to improve efficiency of the design	► Implementation of intense, specialised collection and processing protocols for the entire cohort can be challenging ► Obtaining tissue samples and following cases for treatment response and survival can be challenging in many cohorts

these technologies in HOS research, as they are more prone to selection and differential bias, while prospective studies or cross-sectional studies seem most suitable for such approaches. Moreover, hospital case-control studies are problematic as it is impossible to determine if changes in biomarkers are the cause or consequence of a disease. Semi-longitudinal studies might be extremely powerful for some OMICS technologies such as transcriptomics, proteomics and metabolomics where biological measures are taken before and after exposure or change in disease status. In these study designs each individual serves as their own control eliminating the influence of population variance.

Validation of biomarkers

The value of an OMICS-based biomarker in OEH depends on the reliability of an assay to qualitatively and quantitatively assess the biomarker and on the association between the biomarker and the biological endpoint of interest (exposure, susceptibility or health effect). The reliability of an assay can be tested by investigating the variability of an assay within and between laboratories and comparing results to the variability of existing assays (standards). A necessary step towards an increase in the reliability of OMICS assays is standardisation. Several initiatives have developed standards for new OMICS assays with regards to comparison to existing techniques (microarray quality control (MAQC)), data formats to describe experimental details (minimum information about a microarray experiment (MIAME)) and assessment of sample quality (external RNA controls consortium (ERCC)).[52 53] Once the reliability of assays has been established in the laboratory transitional studies that assess the association between biomarkers and biological endpoints in humans are needed.[49] To achieve an accurate estimate of the association between a biomarker and a biological endpoint reliable and valid measurements of exposure and covariates are needed as well.

A true association between a biomarker and a biological endpoint can be obscured by measurement error. To acquire insight in impact of measurement error on the observed association between a biomarker and a biological endpoint a repeated sampling design, at least on part of the population, is necessary. Repeated sampling on individuals will allow researchers to compare biomarker variability within individuals to biomarker variability between individuals. One measure that can be used to assess the variability of biomarkers within and between individuals is the intraclass correlation coefficient, which represents the proportion of the total variance that can be attributed to the between-individual variance.[49] The level of measurement error that is acceptable for a biomarker depends on the magnitude of the true association between the biomarker and the biological endpoint of interest. For biomarkers with a dichotomous outcome (eg, genotyping) the accuracy of the biomarker is based on the sensitivity (eg, probability of correctly identifying an SNP) and the specificity (eg, probability of incorrectly identifying an SNP) of the biomarker.

Interpretation of study results

In recent years technological developments have had a major impact on the development of new types of study designs of OMICS-based studies. One trend that has been seen consistent within the different OMICS fields is the enormous increase in resolution of the assays (the number of "endpoints" that can be assessed in a single assay) and throughput of the assays (the number of samples that can be analysed per time period). Many of the improvements are based on the introduction of chip-based assays such as DNA microarrays. A major implication of the possibility to investigate multiple endpoints (eg, up to 1 000 000 SNPs in a single assay) in large populations is the possibility for researchers to move away from hypothesis-based studies (focused on a limited set of endpoints) towards hypothesis-free (agnostic) types of study designs (including much larger sets of endpoints). Although the hypothesis-free studies might contribute considerably to the elucidation of the complex biological processes that underlie clinically manifested health effects, it is important to realise that the interpretation of data generated by these types of studies requires a different approach than the interpretation of data generated by more traditional hypothesis-based studies. In hypothesis-based study designs "frequentist" measures such as 95% confidence intervals or p values provide a reasonably good measure to assess the statistical significance of the study's finding. However, the interpretation of such measures is based on the inclusion of a limited number of hypotheses for which the researchers assume that there is a good possibility that the null hypothesis might be rejected (ie, there is a high prior probability of a true positive finding). In a hypothesis-free analytic approach, a study is initiated without a well-defined hypothesis for each included endpoint investigated (ie, a flat prior probability for each finding). However, as a result of chance, the increased number of possible endpoints in a study is accompanied by higher probability of the possibility of detecting

statistically significant false-positive results.[54] Therefore, the traditional statistical approaches that are commonly used in epidemiology are of less value in hypothesis-free studies. A current challenge for the OMICS field is the development of (statistical) approaches that can be used for the interpretation of the high-dimensional data generated by these high-throughput techniques. Several statistical strategies (and also approaches in study designs) have been developed to reduce the probability of false-positive results. Examples are the Bonferroni adjustment for multiple significance testing or more sophisticated Bayesian approaches which include estimation of the false-positive report probability.[15–17 54 55] However, replication of the initial findings in follow-up studies remains the strongest safeguard against false-positive results. Studies that incorporate thousands of biological endpoints should therefore primarily be seen as discovery studies that can aid the generation of new hypotheses. Therefore, new OMICS studies should incorporate strategies for built-in replication of the study findings. Application of a different analytical technique to test the hypothesis a priori in a second/validation set of samples will reduce the possibility that the initial finding was an artefact of the technology used. A potential strategy for built-in replication is to perform the initial analysis on a subset of well-characterised samples matched on potential confounders and effect modifiers and confirm the findings by using alternative analysis methods on the remaining often larger sample set. A potential problem in OEH research is, however, that replication is often complicated as there are often only a limited number of relatively small studies on a single exposure. Even if another large study can be found on a single exposure replication might still be complicated by the fact that the populations are exposed to different levels.

In addition to aspects that contribute to random error, systematic error (bias) is also a potential threat to the validity of HOS utilising OMICS technologies.[56–58] The types of bias that might occur will be largely similar to types of bias that might occur in all HOS. However, issues such as sample collection, handling and storage of samples and analysis technique-specific biases might be especially relevant for studies applying OMICS technologies.[57 59 60] Very recently guidelines for the reporting of genetic association studies (STREGA) have been published.[61] These guidelines underline the necessity of detailed reporting in publications on genetic association studies to allow scientists to assess the potential of bias in study outcomes. Development of similar guidelines for the other OMICS fields will contribute to the identification of relevant types of bias.

Pathway analysis and systems biology

OMICS technologies will enable researchers to look at the complete complement, expression, and regulation of genes, proteins and metabolites. However, at the present time, most statistical analyses are often based on a (simplistic) one-by-one comparison of markers between exposure and/or disease groups. Recently, analytical tools/databases have become available to perform more integrated analyses of biological functions and changes in biological functions as a result of environmental factors. Examples of such approaches are gene ontology (GO), pathway analysis and structural equation modelling (SEM).[62–65] GO is based on a library that consists of gene profiles that are associated with biological processes.[66] Gene sets that are identified in microarray experiments as differently expressed are tested for their association with a profile in the GO library.[63] In pathway analysis, not only the profile of genes associated with a specific biological process is tested, but also the functional interactions between genes in a profile.[62] While still large gaps in the knowledge of biological pathways exist, each new study will contribute to build a base of knowledge necessary for these types of analyses. SEM is a statistical approach that can be used to simultaneously model multiple genes and multiple SNPs within a gene in a hierarchical manner that reflects their underlying role in a biological system.[65]

The increasing knowledge of biological pathways will facilitate the integration of the separate OMICS fields into systems biology approaches. System biology has been described as a global quantitative analysis of the interaction of all components in a biological system to determine its phenotype.[67–69] This integration is facilitated by a continuous increase in computing power and possibilities for data sharing.

EXAMPLES OF THE USE OF OMICS IN OCCUPATIONAL AND ENVIRONMENTAL HEALTH RESEARCH

In table 3 a number of studies are listed to illustrate the current application of OMICS technologies in OEH research. Benzene and arsenic were chosen as examples because of the large populations with potential exposure to these agents in both the occupational and environmental setting and the relatively large

Table 3 Examples of the use of OMICS technologies in occupational and environmental studies that investigate health effects in human populations exposed to benzene or arsenic

OMICS field	Exposure	Topic	References
Genotyping	Benzene	Interaction between SNPs and benzene-induced toxicity	71–73 77–80
Genotyping	Arsenic	Interaction between SNPs and arsenic-induced skin lesions	81 82
Genotyping	Arsenic	Interaction between SNPs and arsenic metabolism	83 84
Genotyping	Arsenic	Interaction between SNPs and exposure to arsenic in relation to non-melanoma skin cancer	85
CNV	Arsenic	Interaction between DNA CNV and exposure to arsenic in relation to transitional cell carcinoma	86
CNV	Arsenic	Interaction between DNA CNV and exposure to arsenic in relation to bladder tumours	87
Epigenomics	Benzene	Relation between gene-specific hypermethylation and exposure to benzene	88
Epigenomics	Arsenic	Relation between epigenetic silencing of tumour suppressor genes and exposure to both tobacco and arsenic	89
Epigenomics	Arsenic	Relation between genomic methylation and exposure to arsenic	90 91
Transcriptomics	Benzene	Relation between gene expression and exposure to benzene	92
Transcriptomics	Arsenic	Interaction between exposure to arsenic and arsenical skin lesions in relation to genome-wide gene expression	74
Transcriptomics	Arsenic	Relation between gene expression and exposure to arsenic	93 94
Proteomics	Benzene	Impact of exposure to benzene on the composition of the proteome	95 96
Proteomics	Arsenic	Impact of exposure to arsenic on the composition of the proteome	97 98

CNV, copy number variation; SNP, single nucleotide polymorphism.

number of studies on these agents that have applied OMICS technologies. It should be noted that inclusion of the example studies was not intended as a systematic overview of studies applying OMICS in OEH research in these specific areas but merely to provide a resource of studies that are indicative of the potential of these new technologies. We highlight three studies from table 3 in some more detail to illustrate the progress in the OMICS field that has been made in recent years. A nice illustration of the progress of the use of genotyping methods in OEH research is a study on haematological effect among a cohort of 250 workers exposed to benzene and 140 controls.[70–72] Initial gene–environment analyses in this study were based on candidate gene approaches focusing on genes involved in the metabolism of benzene (four genes, four SNPs),[72] DNA double strand break repair (seven genes, 24 SNPs),[71] and cytokine and cellular adhesion molecule pathways (20 genes, 40 SNPs).[70] In a more recent analysis of the same study population, Lan *et al* used a chip-based assay (GoldenGate assay) for genotyping which allowed for a larger number of SNPs to be assessed (414 genes, 1433 SNPs).[73] These SNPs were selected from the SNP500Cancer database, and were, therefore, hypothesised to be involved in the development of cancer. However, the influence of these SNPs on benzene-induced haematotoxicity was largely unknown for most SNPs. This study should therefore primarily be seen as hypothesis generating and indeed has provided information on several putative genes involved in benzene haematotoxicity that went well beyond the more classical focus in OEH research on metabolic genes. Although the authors addressed issues of multiple comparisons to reduce the chance of false-positive findings due to the large number of SNPs included in the analysis, it is still critical that the results are replicated in subsequent independent studies.

An example of a hypothesis-free approach towards the assessment of the transcriptome comes from a study by Argos *et al*.[74] In this micro-array-based study ~22 000 genome-wide gene transcripts were measured in 25 subjects with arsenic-induced skin lesions and 15 controls. A false discovery rate of 1% was defined a priori to reduce the risk of chance findings. A set of 486 genes that were differentially expressed between cases and controls was reported. The gene transcripts were also analysed with the use of gene ontology and pathway analysis approaches to elucidate the biological pathways that are involved in arsenic-induced skin lesions. Similar to the genotyping results of the studies discussed above, results from the genome-wide assessment of the transcriptome should be interpreted with great care and require replication in independent studies before they can be used as valid exposure or effect markers.[75 76]

Way forward

It is clear that there have been great technological advances in the different OMICS fields. Some of these technologies have and are starting to be applied in OEH research and will undoubtedly lead to numerous new insights in the near future. With the development of validated technologies, appropriate study designs, better sample handling and advanced statistical methods for data interpretation, OMICS techniques will eventually contribute significantly to OEH and will help the field progress towards an integrated view of the interaction between environment and human health. To achieve this integrated view it will be important to not only focus on genetic variants but also on more functional measures of the phenotype and accurate assessment of exposure. The challenge in this effort will be that the closer one gets to a functional measure of the phenotype (ie, proteomics, metabolomics) the more complex it will be to capture physiologically relevant variability and the more crucial the development of advanced study designs, sampling collection procedures, measurement techniques, and methods for statistical analysis will be to allow interpretation of these parameters.

Acknowledgements This work was performed as part of the work package "integrated risk assessment" of the ECNIS Network of Excellence (Environmental Cancer Risk, Nutrition and Individual Susceptibility), operating within the European Union 6th Framework Program, Priority 5: "Food Quality and Safety" (FOOD-CT-2005-513943).

Funding European Union 6th Framework Program "ECNIS" (FOOD-CT-2005-513943). MTS, LZ and CFS were supported by NIH grants P42ES004705, R01 ES006721, R01 CA122663, and U54 ES016115.

Competing interests MTS has received consulting and expert testimony fees from law firms representing both plaintiffs and defendants in cases involving exposure to benzene.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Smith MT,** Vermeulen R, Li G, *et al*. Use of 'Omic' technologies to study humans exposed to benzene. *Chem Biol Interact* 2005;**153–154**:123–7.
2. **Sachidanandam R,** Weissman D, Schmidt SC, *et al*. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;**409**:928–33.
3. **Venter JC,** Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science* 2001;**291**:1304–51.
4. **Wild CP.** Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 2005;**14**:1847–50.
5. **Smith MT.** Future perspectives of molecular cancer epidemiology. *EJC Supplements* 2008;**6**:188.
6. **Syvänen A-C.** Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2001;**2**:930–42.
7. **Lander ES,** Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
8. **Eichler EE,** Nickerson DA, Altshuler D, *et al*. Completing the map of human genetic variation. *Nature* 2007;**447**:161–5.
9. **Rocha D,** Gut I, Jeffreys AJ, *et al*. Seventh international meeting on single nucleotide polymorphism and complex genome analysis: 'ever bigger scans and an increasingly variable genome'. *Hum Genet* 2006;**119**:451–6.
10. **Engle LJ,** Simpson CL, Landers JE. Using high-throughput SNP technologies to study cancer. *Oncogene* 2006;**25**:1594–601.
11. **Patil N,** Berno AJ, Hinds DA, *et al*. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;**294**:1719–23.
12. **Bodmer W,** Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;**40**:695–701.
13. **Castro-Giner F,** Kauffmann F, de Cid R, *et al*. Gene-environment interactions in asthma. *Occup Environ Med* 2006;**63**:776–86, 761.
14. **Colilla S,** Kantoff PW, Neuhausen SL, *et al*. The joint effect of smoking and AIB1 on breast cancer risk in BRCA1 mutation carriers. *Carcinogenesis* 2006;**27**:599–605.
15. **Mukherjee B,** Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* 2008;**64**:685–94.
16. **Murcray CE,** Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 2009;**169**:219–26.
17. **Wakefield J,** De Vocht F, Hung RJ. Bayesian mixture modeling of gene-environment and gene-gene interactions. *Genet Epidemiol* In press.
18. **Shlien A,** Malkin D. Copy number variations and cancer. *Genome Med* 2009;**1**:62.
19. **Redon R,** Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. *Nature* 2006;**444**:444–54.
20. **Costa JL,** Meijer G, Ylstra B, *et al*. Array comparative genomic hybridization copy number profiling: a new tool for translational research in solid malignancies. *Semin Radiat Oncol* 2008;**18**:98–104.
21. **Benhamou S,** Lee WJ, Alexandrie AK, *et al*. Meta- and pooled analyses of the effects of glutathione S-transferase M1 polymorphisms and smoking on lung cancer risk. *Carcinogenesis* 2002;**23**:1343–50.
22. **Carlsten C,** Sagoo GS, Frodsham AJ, *et al*. Glutathione S-transferase M1 (GSTM1) polymorphisms and lung cancer: a literature-based systematic HuGE review and meta-analysis. *Am J Epidemiol* 2008;**167**:759–74.
23. **Crosbie PA,** Barber PV, Harrison KL, *et al*. GSTM1 copy number and lung cancer risk. *Mutat Res* 2009;**664**:1–5.
24. **Sorensen M,** Raaschou-Nielsen O, Brasch-Andersen C, *et al*. Interactions between GSTM1, GSTT1 and GSTP1 polymorphisms and smoking and intake of fruit and vegetables in relation to lung cancer. *Lung Cancer* 2007;**55**:137–44.

25. **Manning AT,** Garvin JT, Shahbazi RI, *et al*. Molecular profiling techniques and bioinformatics in cancer research. *Eur J Surg Oncol* 2007;**33**:255—65.
26. **Nachtomy O,** Shavit A, Yakhini Z. Gene expression and the concept of the phenotype. *Stud Hist Philos Biol Biomed Sci* 2007;**38**:238—54.
27. **Celis JE,** Kruhoffer M, Gromova I, *et al*. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett* 2000;**480**:2—16.
28. **Callinan PA,** Feinberg AP. The emerging science of epigenomics. *Hum Mol Genet* 2006;**15**:R95—101.
29. **Feinberg AP.** Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;**447**:433—40.
30. **Bird A.** Perceptions of epigenetics. *Nature* 2007;**447**:396—8.
31. **Egger G,** Liang G, Aparicio A, *et al*. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 2004;**429**:457—63.
32. **Suzuki MM,** Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 2008;**9**:465—76.
33. **Bibikova M,** Lin Z, Zhou L, *et al*. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* 2006;**16**:383—93.
34. **Esteller M.** Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* 2007;**8**:286—98.
35. **Higgs DR,** Vernimmen D, Hughes J, *et al*. Using genomics to study how chromatin influences gene expression. *Ann Rev Genomics Hum Genet* 2007;**8**:299—325.
36. **Jirtle RL,** Skinner MK. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* 2007;**8**:253—62.
37. **Jones PA,** Baylin SB. The epigenomics of cancer. *Cell* 2007;**128**:683—92.
38. **Liu L,** Wylie RC, Andrews LG, *et al*. Aging, cancer and nutrition: the DNA methylation connection. *Mech Ageing Dev* 2003;**124**:989—98.
39. **Moore LE,** Huang WY, Chung J, *et al*. Epidemiologic considerations to assess altered DNA methylation from environmental exposures in cancer. *Ann N Y Acad Sci* 2003;**983**:181—96.
40. **Costa FF.** Non-coding RNAs: lost in translation? *Gene* 2007;**386**:1—10.
41. **de Fougerolles A,** Vornlocher HP, Maraganore J, *et al*. Interfering with disease: a progress report on siRNA-based therapeutics. *Nat Rev Drug Discov* 2007;**6**:443—53.
42. **Moss TJ,** Wallrath LL. Connections between epigenetic gene silencing and human disease. *Mutat Res* 2007;**618**:163—74.
43. **Sellers TA,** Yates JR. Review of proteomics with applications to genetic epidemiology. *Genet Epidemiol* 2003;**24**:83—98.
44. **Fliser D,** Novak J, Thongboonkerd V, *et al*. Advances in urinary proteome analysis and biomarker discovery. *J Am Soc Nephrol* 2007;**18**:1057—71.
45. **Hanash SM,** Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. *Nature* 2008;**452**:571—9.
46. **Wingren C,** Borrebaeck CA. Antibody microarrays: current status and key technological advances. *Omics* 2006;**10**:411—27.
47. **Holmes E,** Loo RL, Stamler J, *et al*. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 2008;**453**:396—400.
48. **Claudino WM,** Quattrone A, Biganzoli L, *et al*. Metabolomics: available results, current research projects in breast cancer, and future applications. *J Clin Oncol* 2007;**25**:2840—6.
49. **Garcia-Closas M,** Vermeulen R, Sherman M, *et al*. Application of biomarkers in cancer epidemiology. In: Schottenfeld D**,** Fraumeni JF Jr, eds. *Cancer epidemiology and prevention*. New York: Oxford University Press, 2006:70—88.
50. **Vineis P,** Gallo V. The epidemiological theory: principles of biomarker validation. In: Vineis P**,** Gallo V, eds. Epidemiological concepts of validation of biomarkers for the identification/quantification of environmental carcinogenic exposures. Lodz: Nofer Institute of Occupational Medicine, 2007.
51. **Duncan MW.** Omics and its 15 minutes. *Exp Biol Med (Maywood)* 2007;**232**:471—2.
52. **Morrison N,** Wood AJ, Hancock D, *et al*. Annotation of environmental OMICS data: application to the transcriptomics domain. *Omics* 2006;**10**:172—8.
53. **Wilkes T,** Laux H, Foy CA. Microarray data quality - review of current developments. *Omics* 2007;**11**:1—13.
54. **Wacholder S,** Chanock S, Garcia-Closas M, *et al*. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;**96**:434—42.
55. **Bland JM,** Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;**310**:170.
56. **Semmes OJ.** The "omics" haystack: defining sources of sample bias in expression profiling. *Clin Chem* 2005;**51**:1571—2.
57. **Ransohoff DF.** Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005;**5**:142—9.
58. **Dumeaux V,** Borresen-Dale AL, Frantzen JO, *et al*. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Res* 2008;**10**:R13.
59. **Attia J,** Ioannidis JP, Thakkinstian A, *et al*. How to use an article about genetic association: B: Are the results of the study valid? *JAMA* 2009;**301**:191—7.
60. **Yesupriya A,** Evangelou E, Kavvoura FK, *et al*. Reporting of human genome epidemiology (HuGE) association studies: an empirical assessment. *BMC Med Res Methodol* 2008;**8**:31.
61. **Little J,** Higgins JP, Ioannidis JP, *et al*. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. *Eur J Epidemiol* 2009;**24**:37—55.
62. **Draghici S,** Khatri P, Tarca AL, *et al*. A systems biology approach for pathway level analysis. *Genome Res* 2007;**17**:1537—45.
63. **Khatri P,** Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005;**21**:3587—95.
64. **Weniger M,** Engelmann JC, Schultz J. Genome expression pathway analysis tool—analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics* 2007;**8**:179.
65. **Nock NL,** Larkin EK, Morris NJ, *et al*. Modeling the complex gene x environment interplay in the simulated rheumatoid arthritis GAW15 data using latent variable structural equation modeling. *BMC Proc* 2007;**1**(Suppl 1):S118.
66. **The Gene Ontology Consortium.** Creating the gene ontology resource: design and implementation. *Genome Res* 2001;**11**:1425—33.
67. **Lemberger T.** Systems biology in human health and disease. *Mol Syst Biol* 2007;**3**:136.
68. **Studer SM,** Kaminski N. Towards systems biology of human pulmonary fibrosis. *Proc Am Thorac Soc* 2007;**4**:85—91.
69. **Lund E,** Dumeaux V. Systems epidemiology in cancer. *Cancer Epidemiol Biomarkers Prev* 2008;**17**:2954—7.
70. **Lan Q,** Zhang L, Shen M, *et al*. Polymorphisms in cytokine and cellular adhesion molecule genes and susceptibility to hematotoxicity among workers exposed to benzene. *Cancer Res* 2005;**65**:9574—81.
71. **Shen M,** Lan Q, Zhang L, *et al*. Polymorphisms in genes involved in DNA double-strand break repair pathway and susceptibility to benzene-induced hematotoxicity. *Carcinogenesis* 2006;**27**:2083—9.
72. **Lan Q,** Zhang L, Li G, *et al*. Hematotoxicity in workers exposed to low levels of benzene. *Science* 2004;**306**:1774—6.
73. **Lan Q,** Zhang L, Shen M, *et al*. Large-scale evaluation of candidate genes identifies associations between DNA repair and genomic maintenance and development of benzene hematotoxicity. *Carcinogenesis* 2009;**30**:50—8.
74. **Argos M,** Kibriya MG, Parvez F, *et al*. Gene expression profiles in peripheral lymphocytes by arsenic exposure and skin lesion status in a Bangladeshi population. *Cancer Epidemiol Biomarkers Prev* 2006;**15**:1367—75.
75. **Gillis B,** Gavin IM, Arbieva Z, *et al*. Identification of human cell responses to benzene and benzene metabolites. *Genomics* 2007;**90**:324—33.
76. **Smith MT.** Misuse of genomics in assigning causation in relation to benzene exposure. *Int J Occup Environ Health* 2008;**14**:144—6.
77. **Chen Y,** Li G, Yin S, *et al*. Genetic polymorphisms involved in toxicant-metabolizing enzymes and the risk of chronic benzene poisoning in Chinese occupationally exposed populations. *Xenobiotica* 2007;**37**:103—12.
78. **Gu SY,** Zhang ZB, Wan JX, *et al*. Genetic polymorphisms in CYP1A1, CYP2D6, UGT1A6, UGT1A7, and SULT1A1 genes and correlation with benzene exposure in a Chinese occupational population. *J Toxicol Environ Health A* 2007;**70**:916—24.
79. **Kim YJ,** Paek D, *et al*. Association of the NQO1, MPO, and XRCC1 polymorphisms and chromosome damage among workers at a petroleum refinery. *J Toxicol Environ Health A* 2008;**71**:333—41.
80. **Zhang Z,** Wan J, Jin X, *et al*. Genetic polymorphisms in XRCC1, APE1, ADPRT, XRCC2, and XRCC3 and risk of chronic benzene poisoning in a Chinese occupational population. *Cancer Epidemiol Biomarkers Prev* 2005;**14**:2614—9.
81. **McCarty KM,** Chen YC, Quamruzzaman Q, *et al*. Arsenic methylation, GSTT1, GSTM1, GSTP1 polymorphisms, and skin lesions. *Environ Health Perspect* 2007;**115**:341—5.
82. **Breton CV,** Zhou W, Kile ML, *et al*. Susceptibility to arsenic-induced skin lesions from polymorphisms in base excision repair genes. *Carcinogenesis* 2007;**28**:1520—5.
83. **Schlawicke Engstrom K,** Broberg K, Concha G, *et al*. Genetic polymorphisms influencing arsenic metabolism: evidence from Argentina. *Environ Health Perspect* 2007;**115**:599—605.
84. **Steinmaus C,** Moore LE, Shipp M, *et al*. Genetic polymorphisms in MTHFR 677 and 1298, GSTM1 and T1, and metabolism of arsenic. *J Toxicol Environ Health A* 2007;**70**:159—70.
85. **Applebaum KM,** Karagas MR, Hunter DJ, *et al*. Polymorphisms in nucleotide excision repair genes, arsenic exposure, and non-melanoma skin cancer in New Hampshire. *Environ Health Perspect* 2007;**115**:1231—6.
86. **Hsu LI,** Chiu AW, Pu YS, *et al*. Comparative genomic hybridization study of arsenic-exposed and non-arsenic-exposed urinary transitional cell carcinoma. *Toxicol Appl Pharmacol* 2008;**227**:229—38.
87. **Moore LE,** Smith AH, Eng C, *et al*. Arsenic-related chromosomal alterations in bladder cancer. *J Natl Cancer Inst* 2002;**94**:1688—96.
88. **Bollati V,** Baccarelli A, Hou L, *et al*. Changes in DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer Res* 2007;**67**:876—80.
89. **Marsit CJ,** Karagas MR, Schned A, *et al*. Carcinogen exposure and epigenetic silencing in bladder cancer. *Ann N Y Acad Sci* 2006;**1076**:810—21.
90. **Chanda S,** Dasgupta UB, Guhamazumder D, *et al*. DNA hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy. *Toxicol Sci* 2006;**89**:431—7.
91. **Pilsner JR,** Liu X, Ahsan H, *et al*. Genomic methylation of peripheral blood leukocyte DNA: influences of arsenic and folate in Bangladeshi adults. *Am J Clin Nutr* 2007;**86**:1179—86.
92. **Forrest MS,** Lan Q, Hubbard AE, *et al*. Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. *Environ Health Perspect* 2005;**113**:801—7.

93. **Fry RC,** Navasumrit P, Valiathan C, *et al*. Activation of inflammation/NF-kappaB signaling in infants born to arsenic-exposed mothers. *PLoS Genet* 2007;**3**:e207.
94. **Wu MM,** Chiou HY, Ho IC, *et al*. Gene expression of inflammatory molecules in circulating lymphocytes from arsenic-exposed human subjects. *Environ Health Perspect* 2003;**111**:1429—38.
95. **Joo WA,** Sul D, Lee DY, *et al*. Proteomic analysis of plasma proteins of workers exposed to benzene. *Mutat Res* 2004;**558**:35—44.
96. **Vermeulen R,** Lan Q, Zhang L, *et al*. Decreased levels of CXC-chemokines in serum of benzene-exposed workers identified by array-based proteomics. *Proc Natl Acad Sci U S A* 2005;**102**:17041—6.
97. **Zhai R,** Su S, Lu X, *et al*. Proteomic profiling in the sera of workers occupationally exposed to arsenic and lead: identification of potential biomarkers. *Biometals* 2005;**18**:603—13.
98. **Hegedus CM,** Skibola CF, Warner M, *et al*. Decreased urinary beta-defensin-1 expression as a biomarker of response to arsenic. *Toxicol Sci* 2008;**106**:74—82.