‘OMICS-based’ Biomarkers for Environmental Health Studies

**Almudena Espín-Pérez · Julian Krauskopf ·
Theo M. de Kok · Jos C. Kleinjans**

Published online: 11 September 2014

Abstract The development of biomarkers based on high-throughput techniques, cutting-edge biostatistics and bioinformatics tools is revolutionizing molecular cancer epidemiology. Once validated, such biomarkers will open new and promising perspectives for human health risk assessment and identification of at-risk populations. The application of OMICS to environmental health research is currently resulting in the production of large data sets on gene expression, transcription factors, proteins, metabolites, adducts and epigenetic regulation of the genome in relation to dietary and environmental exposures. The assessment of whole genome transcriptomic and epigenetic profiles (where microRNA analysis has raised special attention in recent years) are regarded as potentially powerful approaches to reduce uncertainties in health risk assessments and to improve strategies for disease prevention, which is a shared aim of the new European Union (EU) research programme, Horizon 2020 (http://ec.europa.eu/programmes/horizon2020/). However, to guarantee appropriate application of OMICS, some challenges need to be addressed in the coming years regarding study design, technicalities in methodology and data analysis.

Keywords Biomarkers · OMICS · Molecular epidemiology · Cancer · Transcriptomics · Epigenetics

A. Espín-Pérez and J. Krauskopf contributed equally to the manuscript

A. Espín-Pérez (✉) · J. Krauskopf · T. M. de Kok · J. C. Kleinjans
Department of Toxicogenomics, Maastricht University,
Universiteitssingel 50, 6200 MD Maastricht, The Netherlands
e-mail: a.espin@maastrichtuniversity.nl

J. Krauskopf
e-mail: j.krauskopf@maastrichtuniversity.nl

T. M. de Kok
e-mail: t.dekok@maastrichtuniversity.nl

J. C. Kleinjans
e-mail: j.kleinjans@maastrichtuniversity.nl

Introduction

Molecular epidemiology in environmental health aims to elucidate the combined role of genetics and environmental risk factors in the development of disease [1].

In human studies, exposure to environmental factors, abnormalities in normal physiological functioning or changes in biological processes can be measured by means of objective indicators referred to as biomarkers. In oncology such biomarkers may represent a molecule or cellular process that can be measured in a wide range of biological samples and can be interpreted as an indicator of increased cancer risk, or as of the presence of cancer, either as a productor response to a malignancy [2]. In order to be useful, biomarkers need to be thoroughly validated. However, so-called “gold standards” concerning the process of validation still need to be defined [3]. Examples of successfully validated biomarkers are measurements of micronuclei as a preclinical marker for carcinogenic risk [4], urinary levels of glucose for diagnosis of diabetes, levels of cardiac troponin in serum for cardiovascular disease, creatinine levels in serum for renal disorders, urinary levels of choriogonadotropin for pregnancy and serum thyrotropin for primary hypothyroidism [5].

Biomarkers for health risk assessment hold promise for elucidating the progression between an environmental exposure and the stage of an associated disease. This knowledge can improve the risk assessment [6] [7] of the substances that people are exposed and in this way determine appropriate policies for prevention [8] [4]. There are different methods (e.g. ExpoCast Programme [9] and Multi-Criteria Decision Analysis (MCDA) [10]) to screen properties of compounds based on human risk.

In 1981 Doll and Peto stated that more than 75 % of cancer deaths from the previous decade in the US were avoidable by means of style of living and other enviromental factors[11]. In 2008, cancer was estimated to be responsible for 12–13 % of the total number of deaths in the US [12]. In view of the expected increases in life expectancy and growth of the population, the proportion of new cancer cases diagnosed in less developed countries is estimated to increase from about 56 % of the world total in 2008 to more than 60 % in 2030 [13]. A relationship between exposure to environmental factors and cancer risk has been established but to which extent specific exposures are causally related to cancer incidence remains unclear. This is where the application of markers of exposure and early biological effect biomarkers in environmental health research are deemed helpful. Measurements of biomarkers of exposure to environmental carcinogens may improve the accuracy of exposure assessment, whereas biological effect markers may be more sensitive as compared actual disease outcomes. This would be the case as they reflect earlier and more subtile molecular and cellular responses which are associated with disease risk. Moreover, such molecular and cellular events are indicative of the modes-of-action, as they identify key processes that explain the development of the toxic effect and thus demonstrate the biological plausibility that exposure to a specific factor is eventually causing the development of cancer. Identification of these markers is suggested to improve risk assessment and prediction of disease development. Integration of biomarkers of exposure and risk with data on inter-individual variability (e.g. genetic polymorphisms), provides the opportunity to describe individual susceptibility in environmental cancer risk.

The term OMICS was given to the high-throughput technology that produces massive and complex data sets from components of biological systems in a relatively short period of time, attempting to understand the system as a whole and aiming at the discovery of novel biomarkers [14]. Lower cost of this technology increased the number of subjects per study, improving reproducibility and data analysis [15] like these illustrative examples show in the association of alterations on gene expression in cord blood from children with arsenic exposure during pregnancy[16] and effect of smoking on bronchial epithelium cells[17]. Furthermore, in view of the high heterogeneity of the disease, it is expected that integration across multiple OMICS platforms may provide better understanding of the various cancer-specific phenotypes [18]. In addition, personalized treatment is considered a priority in cancer research and a key factor that may improve oncological outcomes [19•]. Also a better understanding of the mechanisms underlying disease development based on OMICS research, for instance by gaining insight into gene environment interactions, it is expected to enable personalized cancer prevention strategies [20].

Application of OMICS technology to environmental health studies introduced a new generation in molecular epidemiological research, enabling the detection of genetic polymorphisms in genome-wide-association-studies (GWAS) by means of microarrays and high-throughput sequencing. GWAS introduced a change in research approach, moving away from targeted approaches focusing on relatively small sets of candidate genes to the agnostic genome screening with no prior hypothesis [21]. The consequence of this change from hypothesis-driven to hypothesis-free studies is a steep increase in the number of study participants needed to reach sufficient statistical power.

Currently, high-throughput techniques combined with bioinformatics analysis and cutting-edge biostatistics allow quantitative measurement of sets of molecules like gene expression, metabolites, adducts, proteins, transcription factors and epigenetic regulation of the genome. However, it is difficult to establish new links between diseases and exposures based on induced gene expression changes as some exposures only result in relatively small changes. This may result in inaccurate classification of exposures or/and complexity of interactions [22•]. Implementation of these OMICS technologies might thus revolutionize the field but simultaneously, new challenges concerning study design, laboratory protocols, statistics and interpretation need to be addressed. In this review, we will focus mainly on biomarker discovery based on transcriptomics and to some extent, epigenomics in environmental health cancer studies.

Genomics Approaches in Environmental Cancer Risk Assessment

Transcriptomics

The expression pattern of genes will change in response to environmental exposures and can be measured as the abundance of mRNA transcripts in a particular sample. Therefore, gene expression profiling is used to identify genes or transcripts that show different expression as a response to different environmental fluctuations, time points or cell types. In molecular epidemiology studies transcriptome data can be used to compare gene expression profiles between subpopulations, for instance between a group of individuals with similar characteristics, such as a specific exposure or disease, and a reference group. This reference group usually consists of healthy individuals or unexposed individuals that are matched by age and sex. Other relevant characteristics, such as smoking status or exposure to other environmental exposures should also be matched or adjusted for in the analyses. Based on the genes that are differentially expressed in these groups, characteristic gene profiles, often derived from blood samples, are identified as potential biomarkers [23•]. Additionally, a priori knowledge on gene functions or pathwyas in which they operate can be used to interpret these biomarkers in terms of cancer risks, particularly if the gene expression changes lead to deregulation of genetic networks with established or hypothesized roles in cancer development.

At present two different techniques, microarray technology [23•]and high-throughput sequencing [24], are available for studying gene expression. Both techniques have advantages and drawbacks, depending on the experimental conditions. Microarray technology is a powerful tool that enables measurement of expression of thousands of genes simultaneously by hybridization of mRNA to a solid surface [24]. In a classical microarray experiment, mRNA is extracted from cells or tissue, converted to complementary DNA (cDNA) and labeled by incorporation of a fluorescent dye. After the sample has been hybridized on the microarray and the remaining cDNA is washed off, the array is scanned by a laser to obtain the signal intensities for each sequence probe. After background correction and statistical data analysis the signal intensities estimate the expression level of the mRNA. There are several microarray platforms on the market. The most common microarray platforms are from Agilent Technologies and Affymetrix. Microarrays have been used to study transcriptomic responses in environmental health for over 10 years and therefore experimental protocols and data analysis approaches are well established and standardized. Standards like microarray quality control (MAQC), minimum information about a microarray experiment (MIAME) and external RNA control consortium (ERCC) provide standardization to the assay [25•]. Microarrays have become the method of choice for large-scale gene expression studies because the gene representation on chips has increased while the associated costs have decreased. This trend is supported by the increasing number of microarray-based environmental health studies available through public databases such as Gene Expression Omnibus (GEO) [26] and ArrayExpress [27]. Such studies have shown differences in gene expression profiles between diverse populations (e.g., exposed vs. non-exposed; adults vs. children; males vs. females). Based on a microarray study, analyzing gene expression in blood of smokers and non-smokers van Leeuwen et al. identified six significant differentially expressed genes influenced by cigarette smoke that mainly functioned within carcinogen metabolism, oxidative stress response and anti-apoptosis [28]. A study conducted in an area of the Czech Republic with high levels of environmental pollution indicated that children may have higher susceptibility to pollutants based on significant differences in gene expression changes in children as compared to their parents [29]. Smith et al. applied microarrays to study global gene expression in the peripheral blood cells of benzene-exposed workers and identified more than 100 genes that were differentially expressed [30]. Another study investigated genome-wide gene expression in 40 adults exposed to environmental pollutants and identified gender-specific transcriptional profiles significantly modulated in response to environmental exposure to among others PCBs [31]. Finally, Hochstenbach et al. evaluated the global gene expression in cord blood of newborns as a consequence of fetal carcinogenic exposure to dioxin-like compounds and acrylamide, and identified different transcriptomic responses between boys and girls [32]. All these studies relate gene expression to a wide range of exposure and different health outcomes and demonstrate the potential of transcriptomics to environmental health research.

As a result of major improvements in sequencing technology, high-throughput sequencing now represents a more comprehensive, sensitive and increasingly cost-effective approach. Sequencing of mRNA libraries, called RNA-seq, also enables estimations of transcript levels from RNA samples [33]. Before starting a RNA-seq experiment one has to choose between several high-througput sequencing platforms as the output vary and will affect how experiments are interpreted. The key characteristics of these various platforms have been reviewed by Chu et al. [34] . Sequencing-by-synthesis technology (SBS) is used im most RNA-seq studies. It simultaneously sequences millions of short fragments by massively parallel sanger sequencing [35]. Several samples can be distributed over one sequencing lane allowing several samples being sequenced in parallel. Briefly, mRNA is isolated from cell line or tissue and converted to cDNA; after sequencing adapters are ligated and after PCR amplification the samples are sequenced and scanned; finally the data analysis begins with quality control and the subsequent mapping of the sequencing reads to the transcriptome reference sequence; several statistical packages are available to quantify the gene expression based on this mapping [36]. Besides overcoming several of the limitations of microarray analysis, such as background signals, cross hybridization, different hybridization properties or signal saturation, RNA-sequencing provides a complete overview of the expressed genes, including low abundant and novel transcripts as well as splice variants. Furthermore it is not necessary to define genes of interest in an experiment as all transcripts in a sample will be sequenced. RNA sequencing also consists of steps where experimental biases can occur, such as RNA fragmentation, cDNA or PCR amplification (which are prone to transcript length) [37]. However, RNA-seq experiments require a large computational capacity for calculations and storage of the sequencing data. The approach is still in the early stage of development and examples of applying next-generation sequencing to environmental health studies in human populations are not yet available, but RNA-seq is increasingly being applied to in vitro toxicogenomics studies [38] [39].

There is also more work to be done to determine the degree of corroboration between microarray and RNA-seq approaches and how historical data sets may be leveraged in light of emerging technologies. A pilot study applying deep-sequencing and microarray technology to study transcriptomic responses of a population exposed to benzene reported significant overlap in differentially expressed transcripts using the two technologies [24]. Comparable to the MAQC initiative for microarray analyses, the Sequence Quality Contol (SeQC) project seeks to establish standards for high-throughput sequencing [23•]. Publicly available sequenced data can be retrieved from several databases such as the GEO or Sequence Read Archive (SRA) [40]. Both techniques, microarray and high-throughput sequencing, are prone to some factors leading to false outcomes and therefore it is recommended to validate the results by another quantitative measurement such as quantitative Polymerase Chain Reaction (qPCR). For both techniques, microarray and high-throughput sequencing, gene expression profiles may be influence by differences in bench time (time between sample collection and processing) or sample storage, particularly when results are combined from different studies [41]. For RNA-seq, insufficient sequencing depth may result in loss of relevant information [42]. For microarray results, the variation factors can as well be due to different labeling or hybridization properties of transcripts which can be statistically adjusted to some extent [43]. The unfolding knowledge produced by high-throughput sequencing technology in the near future will for sure enhance the understanding of the underlying mechanisms of environmental exposure risks, will identify new biomarkers and is clearly expected to replace microarray-based techniques at some point in time.

Epigenomics and microRNAs

To understand environmental interactions with the genome it is essential to also consider the epigenome, which refers to a range of molecular modifications to DNA or histone proteins as well as noncoding RNAs that do not affect the actual sequence of a gene but can significantly alter its expression. It is becoming increasingly well accepted that environmental exposures may alter gene expression by mediating epigenetic modifications. Consequently, these modifications may also become important biomarkers of exposure and environmentally influenced diseases [44••]. Research in environmental cancer epidemiology focuses mainly on the epigenetic mechanisms of DNA methylation [45] and histone modification [46]. Several epidemiological studies have revealed the influence of prenatal and early postnatal environmental factors on cancer risk in later life mediated by epigenetic mechanisms [47]. Sanders et al. examined the relationship between cadmium exposure during pregnancy and DNA-methylation in leukocytes of mother-newborn pairs and identified distinct DNA methylation “footprints” as a result of cadmium exposure [48]. Smeester et al. studied methylomes of arsenic-exposed subjects and identified hypermethylated genes that are linked to diseases like cancer, heart diseases and diabetes [49]. A study investigating air pollution and gene-specific methylation in elderly men showed significant associations between exposure levels and F3, ICAM-1, and TLR-2 hypomethylation, and IFN-γ and IL-6 hypermethylation [50]. Russo et al. studied DNA methylation from bronchial epithelial cells and matching blood of smokers and non-smokers and observed lung cancer-associated methylation changes in both tissues. In this study they proposed the altered DNA methylation patterns as biomarker to predict cancer progression or predisposition [51]. Another publication by Sundar et al. presents usable biomarkers for cigarette-smoke induced chronic lung diseases by posttranslational acetylation and methylation of histone H3 and H4 [52]. Baccarelli et al. studied time-dependent methylation of a cohort exposed to particulate pollutants (black carbon) and found decreased repeated-element methylation after exposure to traffic particles [53]. Another study exploring the association between air pollution, DNA methylation and respiratory outcomes in children identified increased CpG methylation in nitric oxide synthase genes [54]. Herbstmann et al. observed global methylation changes in a prenatal polycyclic aromatic hydrocarbon-exposed population [55]. Recent reviews have summarized the human evidence on the association of environmental exposures with air pollution [56], arsenic [57], and other chemicals.

Recently, microRNAs (miRNAs) have emerged as a potential new type of biomarkers in oncology and are suggested to provide the base of novel clinically accessible molecular monitoring tools for different types of cancer [58]. These small non-coding RNA molecules are involved in the regulation of gene expression and have unique sequences of about 22 nucleotides [59]. They are cell type specific [60] and highly stable in biological fluids such as urine, saliva or blood [61] [62]. These properties make circulating miRNAs ideal biomarkers for both environmental health studies and clinical diagnostics. Both microarray and deep-sequencing technique can be applied to study the expression of miRNAs. Based on a combination of microarray and deep-sequencing data from a cohort possessing breast, lung, ovarian and prostate carcinoma patients, Zadran et al. produced mRNA and miRNA signatures that were able to distinguish with high fidelity cancer patients and noncancerous controls [63]. In a microarray study on human lung tumor and corresponding normal lung samples from highly asbestos-exposed subjects, known and novel asbestos related miRNAs were identified and shown to be inversely correlated with expression of the corresponding target genes. Interestingly, many more miRNAs were differentially expressed between normal lung from either cancer cases and healthy controls than between tumor and corresponding normal samples from the same individual. The authors suggested that miRNAs may be potential biomarkers for early-stage carcinogenesis [64]. Several studies aimed at sequencing blood circulating miRNAs as markers of disease such as lung carcinogenesis [65] or myeloid leukemia (AML) [66]. Zhi et al. sequenced small RNA libraries from serum of AML patients to quantify circulating miRNA levels and identified a 6-miRNA profile to differentiate between AML patients and normal controls and possibly to predict survival [66].

The potential value of circulating miRNAs as biomarkers of cancer progression may also be very promising for environmental cancer studies. To date, sample sizes in these studies have been small so further evaluation of feasibility is needed. Possibly a combination of conventional biomarkers and circulating miRNAs will increase the specificity in disease detection.

Other OMICS and Integration

In addition to transcriptomics and epigenetics, additional OMICS technologies may help to elucidate environmental risks for and biomarkers of cancer. Proteins are encoded by mRNA but their presence and enzymatic activity cannot be accurately predicted by transcriptome analysis. Proteomics analysis in oncology provides not only information about functional proteins that are involved in the transformation to malignancy but also biomarkers for prediction and therapeutic efficacy. The proteomics platforms have evolved considerably in the last few years in terms of the development of separation and identification techniques by multi-dimensional sample fractionation methods, mass spectrometry and microarrays for proteins. [67]. Metabolomics is another expanding area that enables analysis of small-molecule metabolites that result from cellular processes. Metabolomics is growing into a powerful, fast and accurate tool to detect spectrum of metabolites, revealing novel biomarkers that are essential to understand tumor mechanisms [68].

In environmental health studies, the integration of data from proteomics and metabolomics platforms with emerging novel approaches (such as adductomics [69] and lipidomics[70]) may eventually explain the full spectrum of a cellular response to an exposure and provide better insights into the associated biological outcome. Nevertheless, the introduction of technical artefacts from different platforms as well as variations in data produced in different laboratories complicate the cross-OMICS analyses.

The methodology for integrating OMICS data sets into a systems biology approach comprises identification of a network scaffold at the first place (the recognition of interactions between cellular components), followed by the decomposition of the scaffold (splitting the network into modules in order to identify active components) and finally modelling of the cellular network (analyze and simulate the data into a system model) (Fig 1) [71]. This requires advanced biostatistics such as Bayesian network analysis. Multi-level Ontology Analysis (MONA) is an example of such an analysis tool where thealgorithm for the analysis of integrated data from different OMICS levelsprovides a flexible framework that allows different ontologies and high yields concerning results even for complex models for the fine-tuning of mRNA by microRNAs [72]. Integrative Clustering of Multiple Genomic Data Types (iCluster) is another tool that allows integration of independent sets of clusters or groups of genes in which expression changes are observed under a certain condition [73].

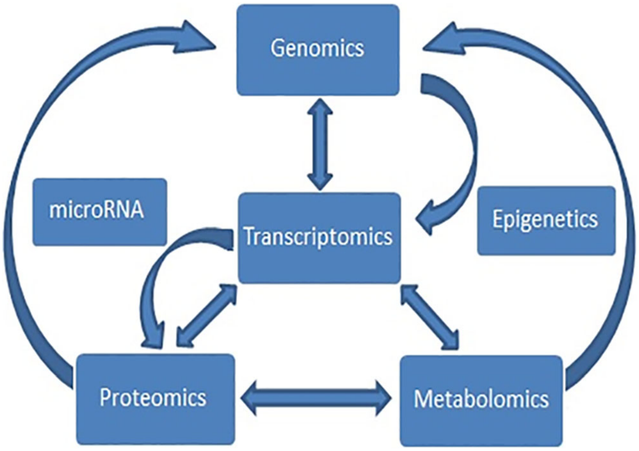

Fig. 1 Integration of heterogeneous data sets is based on the recognition of interactions between cellular components merging information across OMICS. Epigenome and microRNAs influences gene expression and protein expression, leading to changes in metabolism.

Application of Genomics in Molecular Epidemiology

Over the last 10 years the number of publications in environmental cancer research, applying high-throughput OMICS technologies, has increased gradually (Fig. 2).

Earlier environmental health risk assessments have focused on biomarkers of risks, neglecting the fact that humans are exposed to a wide range of adverse environmental factors. These environmental factors, called the external exposome[74], include diverse agents like air pollutants, chemical contaminants, diet or life style factors (tobacco, alcohol). These external factors interact with the internal exposome that consists of processes inside the cell like metabolism, gene expression and epigenetic as consequence of external exposure[44••]. These processes can be measured by the variety of OMICS technologies and hence have led to the novel paradigm of exposomics studies introducing methods for internal and external exposure assessment [75]. An ongoing European Union (EU) project, entitled EXPOsOMICS (http://www.exposomicsproject.eu/), aims to predict health risk related to environmental exposures such as air pollution or water contaminants, during all critical periods of life. A major goal within this project is the development of novel biomarkers and measures of environmental exposures or of risk of health effects. This goal will be achieved by combining: a large amount of available health data from longitudinal cohorts; new data from individual external exposures; new technological devices such as sensors, smartphones and satellites; and cross-omics profiling including transcriptomics and epigenomics. The analysis and integration of data from different omics platforms and the measurements of the external exposome still remain a challenge. However, in the future these efforts will offer new insights into the underlying molecular mechanisms of clinical outcomes, and eventually the transcriptomic profiles and pathways will add to biomarker discovery for health risks.

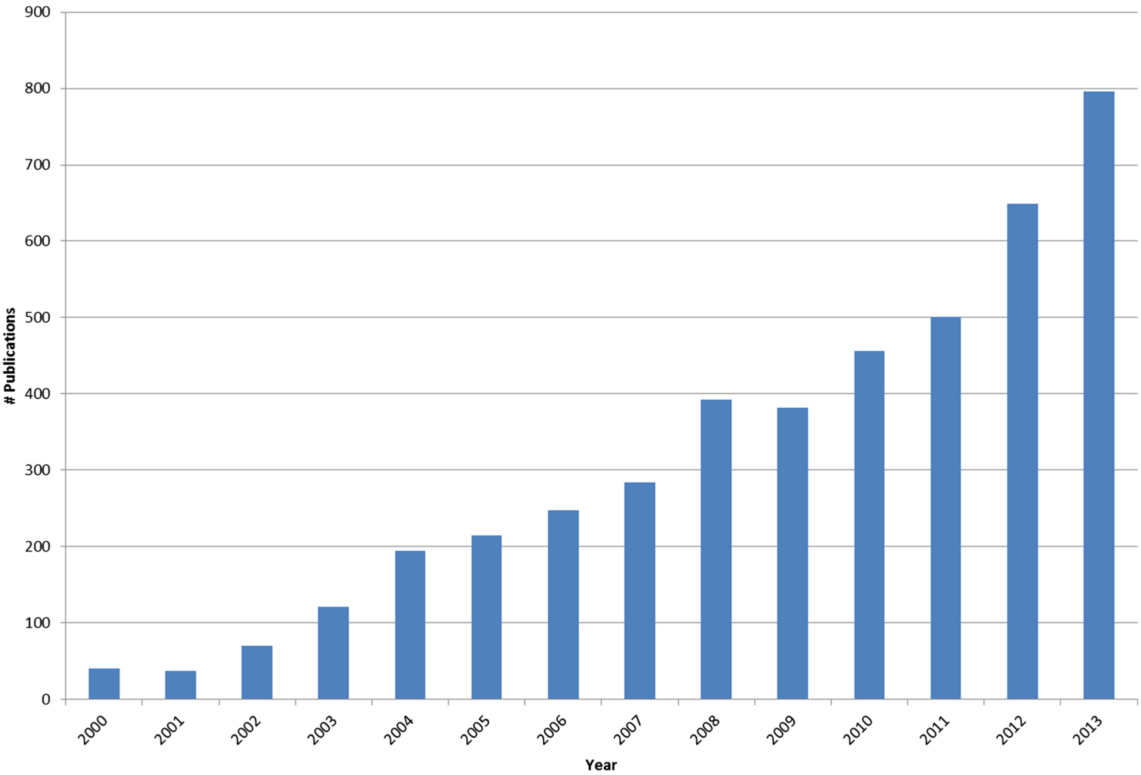

Fig. 2 In the last 10 years the number of publications related to epidemiological cancer research has been increased drastically. The numbers are based on a pubmed search including the following search term: ((((((high-throughput sequencing) OR next-generation sequencing) OR RNA-sequencing) OR microarray) OR transcriptome) OR transcriptomics’) AND ((((((epidemiology) OR epidemiologic) OR population-based) OR case–control study) OR cohort study) OR cross-sectional study) AND cancer AND human

There are other large ongoing projects that also aim to assess the impact of environmental and lifestyle risk factors on human health such as Health and Environment Alliance (HEAL) (http://www.env-health.org/) that identifies environmental threats in order to establish protective policies; Human Early-Life Exposome (HELIX) (http://www.projecthelix.eu/en), that investigates the role of the exposome on mothers’ and childrens’ health; and Envirogenomarkers (EGM) (http://www.envirogenomarkers.net/), which investigates the influence of environmental exposure in human health by developing new cross-omics biomarkers.

Challenges in Environmental Cancer Risk Assessment

For an appropriate application of the multi-OMICS approach there are several aspects that will need improvement in the following years; these mainly concern study design, methodology and data analysis.

Study Design

Use of OMICS-based biomarkers demands suitable study designs and they can be usable depending on the research question. The standard epidemiological study designs, such as cohort or case–control designs, and also hybrid designs, like case-cohort or nested case–control designs, are relevant to environmental cancer risk assessment. A recent review stressed the advantages of nested case–control design and gave recommendations for best practices in the use of biomarker discovery [76]. Case–control design is the approach followed in the EGM project, where the differences in transcriptomic profiles between cases and controls are associated to intermediate biomarkers (changes that represent signs of early effect) and in the same way as to exposures. This so-called “Meet-in-the-middle” approach starts with studying the relationship between exposure and disease, followed by establishing associations between exposure and intermediate OMICS biomarkers and ultimately, investigations of the link between disease and the mentioned intermediate biomarkers [77••].

Additionally, the repeated sampling design improves understanding of the impact that measurement error has on such relationships between biomarkers and endpoints. Repeated information from the same individual enables comparison of the variability within and between individuals, allowing an estimation of the inter-individual variance over the total variance [25•].

Life stage can be included in the assessment of exposures introducing time as a variable in causal inferences since either early or late exposures may have an effect on the development of the disease.

Technical Aspects

When OMICS analyses are applied to biobank or stored samples from epidemiological studies several practical concerns about sample suitability need to be considered. Each OMICS approach has its own set of variabilities that need to be controlled. As OMICS profiles in human blood samples can change over time at room temperature or during storage on ice, the bench time (time from collection of samples until storage in a freezer), needs to be strictly controlled, certainly for the case of transcriptomics if no RNA stabilizer has been added [41]. In case of multi-center studies, laboratory protocols should be comparable and meet the same standards for sample handling (e.g. stabilization of mRNA and of microRNA, use of anticoagulants, long-term storage conditions). Having standardized approaches across fieldwork centers and laboratories reduces technical variation in the dataset and thus improves the assessment of more robust biomarkers [25•] [78].

Interpretation

The OMICS technologies available for environmental health studies at the moment allow for analyzing a substantial number of samples per run with great resolution ("endpoints" that can be evaluated per assay). Data produced by OMICS requires a different interpretation from the traditional hypothesis-based perspective. Since a hypothesis is no longer pre-defined it is more likely that false positives will emerge. Consequently, appropriate statistical tools are required [25•, 79].

There are several methods frequently used to analyze high-dimensional OMICS data. One of them takes independently each variable in a predictor matrix so that the relationship between outcome and individual predictor is tested using the same model (univariate approach). ANOVA, X^2, GAMs and Mixed Models are examples of this approach; the latter is an improvement over the others since it comprises random effect as part of the variability. Family-Wise Error Rate (FWER) and False Discovery Rate (FDR) are strategies to correct for multiple testing. Another method to analyze OMICS data is based on the search for general patterns of association between predictors (multivariate approach) [80]. Discriminant Analysis of Principal Components (DAPC) is a multivariate analysis that combines the methods Principal Components Analysis (PCA), widely used to summarize into a more visual way the information enclosed in large OMICS data sets and Discriminant Analysis (DA), applied to identify different classes into groups from data sets [81].

It is important to take into consideration that many diseases are interconnected. Consequently, it will be increasingly important to take an inclusive approach that integrates data from multiple OMICS levels to ensure simultaneous and unbiased assessment of diverse biological processes [82, 83]. To have a better understanding of the outcome it is required to establish a complete interrelated diagram of cellular components that are influenced by genes and their products, rather than simply studying genes that are known to cause a disease [71]. There could be underlying molecular networks that are not included in the current classifications of diseases, so we should not be limited by known causes of disease [84]. Knowledge and statistical evidence can be comprised in the structure of a causal diagram (representation of variables linked by lines and arrows), allowing the visualization of combined data sets [85].

Pathway analysis is widely used to understand the underlying biology of OMICS measurements. Still, a number of challenges related to approach need to be addressed including incomplete annotations and/or poor resolution in databases due to lack of exact transcrips and SPNs (according to the high resolution data from genomics and proteomics) [86] and the description of dynamics of genomics responses measured over time.

Biomarkers in Preventive Cancer Research

"Discovery" phase is the first step for biomarker development. It involves the identification of potential biomarkers by comparison between cancerous and normal tissues. In the second phase called "Validation" the biomarkers from the previous stage are measured in clinical assays in order to discriminate between a cancerous or healthy status.

Especially for early diagnosis, biomarkers do not provide sufficient sensitivity for detection at low levels and/or specificity to reveal preclinical disease [2]. A biomarker should be a molecule that is present in a well detectable amount and be specific to a tissue of interest in order to be explicit. More data needs to be available since biomarkers require to be regularly monitored over the duration of an individual's disease by suitable tests with quality control and overcome at least one independent validation study [5]. Differences between methodologies implemented in different laboratories is hampering the validation of useful biomarkers and thus consensus on the use of specific techniques is needed [87].

Also OMICS biomarkers need to be validated along similar lines in dedicated studies like those from Bonassi et al. investigating micronuclei as cancer risk biomarker through prospective cohort studies [88] and from Peluso et al. studying bulky DNA adducts in a nested case–control study with lung cancer patients [89].

Current technologies allow us to perform simultaneous evaluations of environmental health risks in association with of a wide range of exposures and lifestyle factors. It remains a challenge to determine whether and how the huge OMICS data sets can be applied to assess their validity and utility in disease prevention, including assessment of at-risk groups, impact of age and gender, cost-effectiveness of OMICS biomarkers with their ethical and social implications.

Conclusions

OMICS has enabled the expansion of molecular epidemiology and it is contributing to advance personalized cancer medicine. Given the critical role of statistics and bioinformatics in all of these studies, implementation of these new high-throughput techniques require scientists trained to use the emerging tools [90] and proper integration of the different knowledge from discovery to implementation of the outcome and public health decision making [91].

Molecular epidemiology has played a key role in the identification of carcinogenic compounds like cadmium, lead, polychlorinated biphenyls, p,p'-dichlorodiphenyldichloroethylene and hexachlorobenzene. However, biomarkers of exposure to many other carcinogens and their mechanisms of action are still unknown. Once novel biomarkers are identified to be linked to carcinogens they still should pass an appropriate process of validation before being widely used in research or in clinical and prevention activities[92].

The issues and challenges discussed above and related to study design, technical aspects like bench times and standardized approaches, as well as the need for improved and integrated data analysis and interpretation still need to be addressed. In the next decade a massive expansion of information is expected, which needs to be translated to improvement of the state of health and management of environment [93]. Despite the significant improvement in technology and computation achieved in the last years, continuous efforts are needed in order to identify and validate biomarkers for cancer risk assessment.

Compliance with Ethics Guidelines

Conflict of Interest Almudena Espín-Pérez, Julian Krauskopf, Dr. Theo M. de Kok, and Dr. Jos C. Kleinjans are all employed by the University of Maastricht.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Semenza JC, Weasel LH. Molecular epidemiology in environmental health: the potential of tumor suppressor gene p53 as a biomarker. Environ Health Perspect. 1997;105 Suppl 1:155–63.
2. Wagner PD, Verma M, Srivastava S. Challenges for biomarkers in cancer detection. Ann N Y Acad Sci. 2004;1022:9–16.
3. Firestein GS. A biomarker by any other name. Nat Clin Pract Rheumatol. 2006;2(12):635.
4. Bonassi S, Au WW. Biomarkers in molecular epidemiology studies for health risk prediction. Mutat Res. 2002;511(1):73–86.
5. Diamandis EP. Cancer biomarkers: can we turn recent failures into success? J Natl Cancer Inst. 2010;102(19):1462–7.
6. Owen R et al. Biomarkers and environmental risk assessment: guiding principles from the human health field. Mar Pollut Bull. 2008;56(4):613–9.
7. Vainio H. Use of biomarkers in risk assessment. Int J Hyg Environ Health. 2001;204(2–3):91–102.
8. Bonassi S, Neri M, Puntoni R. Validation of biomarkers as early predictors of disease. Mutat Res. 2001;480–481:349–58.
9. Cohen Hubal EA et al. Advancing exposure characterization for chemical evaluation and risk assessment. J Toxicol Environ Health B Crit Rev. 2010;13(2–4):299–313.
10. Giubilato E et al. A risk-based methodology for ranking environmental chemical stressors at the regional scale. Environ Int. 2014;65C:41–53.
11. Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst. 1981;66(6):1191–308.
12. Schottenfeld D et al. Current perspective on the global and United States cancer burden attributable to lifestyle and environmental risk factors. Annu Rev Public Health. 2013;34:97–117.
13. Jemal A et al. Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol Biomarkers Prev. 2010;19(8):1893–907.
14. Kyrtopoulos SA. Making sense of OMICS data in population-based environmental health studies. Environ Mol Mutagen. 2013;54(7):468–79.
15. Wild CP. Environmental exposure measurement in cancer epidemiology. Mutagenesis. 2009;24(2):117–25.
16. Fry RC et al. Activation of inflammation/NF-kappaB signaling in infants born to arsenic-exposed mothers. PLoS Genet. 2007;3(11):e207.
17. Spira A et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A. 2004;101(27):10143–8.
18. Le Cao KA, Gonzalez I, Dejean S. Integromics: an R package to unravel relationships between two omics datasets. Bioinformatics. 2009;25(21):2855–6.
19.• Roukos DH. Integrated clinical genomics: new horizon for diagnostic and biomarker discoveries in cancer. Expert Rev Mol Diagn. 2013;13(1):1–4. *Tools to guide personalized cancer and limitations in cancer research.*
20. Chen R et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293–307.
21. Kitsios GD, Zintzaras E. Genome-wide association studies: hypothesis-"free" or "engaged"? Transl Res. 2009;154(4):161–4.
22.• Bonassi S, Taioli E, Vermeulen R. Omics in population studies: a molecular epidemiology perspective. Environ Mol Mutagen. 2013;54(7):455–60. *Overview about the evolution of exposure biomarkers and perspective of omics biomarkers in epidemiological studies.*
23.• McHale CM et al. Analysis of the transcriptome in molecular epidemiology studies. Environ Mol Mutagen. 2013;54(7):500–17. *Review of transcriptome analysis in molecular epidemiology studies.*
24. Thomas R et al. Global gene expression response of a population exposed to benzene: a pilot study exploring the use of RNA-sequencing technology. Environ Mol Mutagen. 2013;54(7):566–73.
25.• Vlaanderen J et al. Application of OMICS technologies in occupational and environmental health research; current status and projections. Occup Environ Med. 2010;67(2):136–43. *Study design,*

validation of biomarkers and interpretation of results as challenges of omics in enviromental studies.

26. Barrett T et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41(Database issue):D991–5.
27. Rustici G et al. ArrayExpress update–trends in database growth and links to data analysis tools. Nucleic Acids Res. 2013;41(Database issue):D987–90.
28. van Leeuwen DM et al. Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. Carcinogenesis. 2007;28(3):691–7.
29. van Leeuwen DM et al. Genomic analysis suggests higher susceptibility of children to air pollution. Carcinogenesis. 2008;29(5): 977–83.
30. Smith MT et al. Use of 'Omic' technologies to study humans exposed to benzene. Chem Biol Interact. 2005;153:123–7.
31. De Coster S et al. Gender-specific transcriptomic response to environmental exposure in Flemish adults. Environ Mol Mutagen. 2013;54(7):574–88.
32. Hochstenbach K et al. Global gene expression analysis in cord blood reveals gender-specific differences in response to carcinogenic exposure in utero. Cancer Epidemiol Biomarkers Prev. 2012;21(10):1756–67.
33. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
34. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 2012;22(4):271–4.
35. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92(5): 255–64.
36. Marioni JC et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
37. Bullard JH et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinforma. 2010;11:94.
38. van Delft J et al. RNA-Seq provides new insights in the transcriptome responses induced by the carcinogen benzo[a]pyrene. Toxicol Sci. 2012;130(2):427–39.
39. Su Z et al. Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. Chem Res Toxicol. 2011;24(9):1486–93.
40. Kodama Y et al. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40(Database issue):D54–6.
41. Hebels DG et al. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environ Health Perspect. 2013;121(4):480–7.
42. Robles JA et al. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics. 2012;13:484.
43. McHale CM et al. Global gene expression profiling of a population exposed to a range of benzene levels. Environ Health Perspect. 2011;119(5):628–34.
44.•• Wild CP, Scalbert A, Herceg Z. Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. Environ Mol Mutagen. 2013;54(7):480–99. *Concept of exposome, transcriptomics/epigenetics changes and challenges in cancer epidemiology.*
45. Lopez-Serra L, Esteller M. Proteins that bind methylated DNA and human cancer: reading the wrong words. Br J Cancer. 2008;98(12): 1881–5.
46. Fraga MF et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat Genet. 2005;37(4):391–400.
47. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007;8(4):253–62.
48. Sanders, A.P., et al., Cadmium exposure and the epigenome: exposure-associated patterns of DNA methylation in leukocytes from mother-baby pairs. Epigenetics, 2013. 9(2).
49. Smeester L et al. Epigenetic changes in individuals with arsenicosis. Chem Res Toxicol. 2011;24(2):165–7.
50. Bind MA, et al. Air pollution and gene-specific methylation in the Normative Aging Study: Association, effect modification, and mediation analysis. Epigenetics, 2014. 9(3).
51. Russo AL et al. Differential DNA hypermethylation of critical genes mediates the stage-specific tobacco smoke-induced neoplastic progression of lung cancer. Clin Cancer Res. 2005;11(7):2466–70.
52. Sundar IK, et al. Cigarette Smoke Induces Distinct Histone Modifications in Lung Cells: Implications for the Pathogenesis of COPD and Lung Cancer. J Proteome Res, 2013.
53. Baccarelli A et al. Rapid DNA methylation changes after exposure to traffic particles. Am J Respir Crit Care Med. 2009;179(7):572–8.
54. Breton CV et al. Particulate matter, DNA methylation in nitric oxide synthase, and childhood respiratory disease. Environ Health Perspect. 2012;120(9):1320–6.
55. Herbstman JB et al. Prenatal exposure to polycyclic aromatic hydrocarbons, benzo[a]pyrene-DNA adducts, and genomic DNA methylation in cord blood. Environ Health Perspect. 2012;120(5): 733–8.
56. Breton C, Marutani A. Air pollution and epigenetics: recent findings. Curr Environ Health Reports. 2014;1(1):35–45.
57. Bailey KA, Fry RC. Arsenic-associated changes to the epigenome: what Are the functional consequences? Curr Environ Health Rep. 2014;1:22–34.
58. Berger F, Reiser MF. Micro-RNAs as potential New molecular biomarkers in oncology: have they reached relevance for the clinical imaging sciences? Theranostics. 2013;3(12):932–41.
59. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.
60. Liang Y et al. Characterization of microRNA expression profiles in normal human tissues. BMC Genomics. 2007;8:166.
61. Turchinovich A, Weiz L, Burwinkel B. Extracellular miRNAs: the mystery of their origin and function. Trends Biochem Sci. 2012;37(11):460–5.
62. Arroyo JD et al. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci U S A. 2011;108(12):5003–8.
63. Zadran S, Remacle F, Levine RD. miRNA and mRNA cancer signatures determined by analysis of expression levels in large cohorts of patients. Proc Natl Acad Sci U S A. 2013;110(47): 19160–5.
64. Nymark P et al. Integrative analysis of microRNA, mRNA and aCGH data reveals asbestos- and histology-related changes in lung cancer. Genes Chromosom Cancer. 2011;50(8):585–97.
65. Wu JJ et al. Alteration of serum miR-206 and miR-133b is associated with lung carcinogenesis induced by 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. Toxicol Appl Pharmacol. 2013;267(3):238–46.
66. Zhi F, et al. Identification of Circulating MicroRNAs as Potential Biomarkers for Detecting Acute Myeloid Leukemia. Plos One, 2013. 8(2).
67. Lopez E et al. Clinical proteomics and OMICS clues useful in translational medicine research. Proteome Sci. 2012;10(1):35.
68. Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: current trends and future perspectives. J Pharm Biomed Anal. 2014;87:1–11.
69. Balbo S, Turesky RJ, Villalta PW. DNA adductomics. Chem Res Toxicol. 2014;27(3):356–66.

70. Arafah K, et al. Lipidomics for Clinical Diagnosis: Dye-Assisted Laser Desorption/Ionization (DALDI) Method for Lipids Detection in MALDI Mass Spectrometry Imaging. OMICS, 2014.

71. Mitra K et al. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14(10):719–32.

72. Sass S et al. A modular framework for gene set analysis integrating multilevel omics data. Nucleic Acids Res. 2013;41(21):9622–33.

73. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.

74. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005;14(8):1847–50.

75. Vineis P et al. The impact of new research technologies on our understanding of environmental causes of disease: the concept of clinical vulnerability. Environ Health. 2009;8:54.

76. Rundle A, Ahsan H, Vineis P. Better cancer biomarker discovery through better study design. Eur J Clin Invest. 2012;42(12):1350–9.

77.•• Vineis P et al. Advancing the application of omics-based biomarkers in environmental epidemiology. Environ Mol Mutagen. 2013;54(7):461–7. *The "meet-in-the-middle" concept and challenges to be addressed in omics in the coming years*.

78. Abu-Asab MS et al. Biomarkers in the age of omics: time for a systems biology approach. OMICS. 2011;15(3):105–12.

79. Manning AT et al. Molecular profiling techniques and bioinformatics in cancer research. Eur J Surg Oncol. 2007;33(3):255–65.

80. Chadeau-Hyam M et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutagen. 2013;54(7):542–57.

81. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11:94.

82. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol. 2006;7(3):198–210.

83. Gibbs DL et al. Multi-omic network signatures of disease. Front Genet. 2014;4:309.

84. Barabasi AL. Network medicine–from obesity to the "diseasome". N Engl J Med. 2007;357(4):404–7.

85. Joffe M et al. Causal diagrams in systems epidemiology. Emerg Themes Epidemiol. 2012;9(1):1.

86. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

87. Vydelingum NA et al. Standards in molecular diagnostics for the discovery and validation of clinically useful cancer biomarkers. Expert Rev Mol Diagn. 2013;13(5):421–3.

88. Bonassi S et al. An increased micronucleus frequency in peripheral blood lymphocytes predicts the risk of cancer in humans. Carcinogenesis. 2007;28(3):625–31.

89. Peluso M et al. DNA adducts and lung cancer risk: a prospective study. Cancer Res. 2005;65(17):8042–8.

90. Arts IC, Weijenberg MP. New training tools for new epidemiologists. Environ Mol Mutagen. 2013;54(7):611–5.

91. Spitz MR, Caporaso NE, Sellers TA. Integrative cancer epidemiology–the next generation. Cancer Discov. 2012;2(12): 1087–90.

92. Wagner PD, Srivastava S. New paradigms in translational science research in cancer biomarkers. Transl Res. 2012;159(4):343–53.

93. Peitsch MC, de Graaf D. A decade of Systems Biology: where are we and where are we going to? Drug Discov Today, 2013.