

MEETING REPORT

Open Access



CrossMark

# Yale school of public health symposium on lifetime exposures and human health: the exposome; summary and future reflections

Caroline H. Johnson<sup>1\*</sup>, Toby J. Athersuch<sup>2,3</sup>, Gwen W. Collman<sup>4</sup>, Suraj Dhungana<sup>5</sup>, David F. Grant<sup>6</sup>, Dean P. Jones<sup>7</sup>, Chirag J. Patel<sup>8</sup> and Vasilis Vasiliou<sup>1\*</sup>

## Abstract

The exposome is defined as "the totality of environmental exposures encountered from birth to death" and was developed to address the need for comprehensive environmental exposure assessment to better understand disease etiology. Due to the complexity of the exposome, significant efforts have been made to develop technologies for longitudinal, internal and external exposure monitoring, and bioinformatics to integrate and analyze datasets generated. Our objectives were to bring together leaders in the field of exposomics, at a recent Symposium on "Lifetime Exposures and Human Health: The Exposome," held at Yale School of Public Health. Our aim was to highlight the most recent technological advancements for measurement of the exposome, bioinformatics development, current limitations, and future needs in environmental health. In the discussions, an emphasis was placed on moving away from a one-chemical one-health outcome model toward a new paradigm of monitoring the totality of exposures that individuals may experience over their lifetime. This is critical to better understand the underlying biological impact on human health, particularly during windows of susceptibility. Recent advancements in metabolomics and bioinformatics are driving the field forward in biomonitoring and understanding the biological impact, and the technological and logistical challenges involved in the analyses were highlighted. In conclusion, further developments and support are needed for large-scale biomonitoring and management of big data, standardization for exposure and data analyses, bioinformatics tools for co-exposure or mixture analyses, and methods for data sharing.

## Background

Since the completion of the Human Genome Project in 2003, it has become apparent that most chronic diseases are attributable to both genetic and environmental influences; the environment is, however, the major contributor to the global disease burden [1–3]. In 2005, Christopher Wild put forth the concept of the exposome, which highlighted the need for comprehensive environmental exposure assessment tools to better delineate causality and disease [4]. The exposome, originally defined as "the totality of environmental exposures encountered from birth to death" [4], has moved from a concept to a reality and was recently redefined as the "cumulative measure of environmental influences and associated biological responses

throughout the lifespan, including exposures from the environment, behavior, diet, and endogenous processes" [5]. To begin the epic task of attempting to analyze all exposures that an individual may encounter over their lifespan and connect those to biological impact, there have been significant technological advancements to enable the study of the exposome. Herein, the Department of Environmental Health Sciences at Yale School of Public Health (YSPH) brought together leaders in exposome technology advancements to discuss recent developments, limitations, and future needs, also with the intention to bring increased awareness to this growing and important field.

## Findings

The Symposium was initially opened by YSPH Dean Dr. Sten Vermund and National Institute of Environmental Health Sciences (NIEHS) Director Dr. Linda Birnbaum (by

\* Correspondence: caroline.johnson@yale.edu; vasilis.vasiliou@yale.edu

<sup>1</sup>Department of Environmental Health Sciences, Yale School of Public Health, New Haven, CT, USA

Full list of author information is available at the end of the article

prerecorded video), who expressed the importance and value of the exposome in environmental health and epidemiology. The exposome is a highly interdisciplinary holistic approach that intersects environmental exposure monitoring with modern technologies such as genomics and metabolomics. It is a valuable science particularly important for understanding how environmental factors affect children's health and later-life outcomes.

The Symposium discussed some current important considerations, developments, and future perspectives for exposome research which are discussed below.

### **Moving away from a single exposure-disease approach**

Environmental health research has traditionally focused on how a specific chemical, or class of chemicals, influences a specific health outcome. This single exposure-disease approach does not, however, represent the complexity of exposures encountered throughout life. Environmental exposures are complex, widespread, and play a role in human health and the development of disease. Researchers are now beginning to consider the totality of external and internal exposures across the life course, a concept called the exposome. Since the concept was developed, most efforts have focused on defining the exposome and challenges to its application into scientific research. Three studies in Europe—the Health and Environment-wide Associations based on Large population Surveys (HEALS), Human Early-Life Exposome (HELIX), and EXPoSOMICS projects—represent early coordinated efforts to advance exposome research [6, 7]. To further advance and promote implementation of the exposome in environmental health research, NIEHS convened a workshop in 2015 to review the state of the science and develop recommendations in key domains. Takeaways from each group included:

- Biomonitoring: Exploring cumulative exposure history requires a hybrid of traditional (targeted) and exposomic (untargeted) biomonitoring approaches and utilizing advantages of both methods [8].
- Biological response and impact: External and internal exposures interact to alter biological processes and trigger production of new chemical intermediates. Exposomic technologies can link exposures to these downstream effects [9].
- Epidemiology: The exposome is a complement to environmental epidemiology. Untargeted analyses can generate findings that need to be investigated using hypothesis-driven approaches central to epidemiology. Merging data across cohorts with different life stages enables characterization of the exposome across the life course [10].

- Data science: Exposomic approaches generate extensive data to be stored, managed, analyzed, integrated, and shared. Development of community-based data standards and ontologies is critical [3].

While specific challenges exist in each domain, there are cross-cutting themes and needs, central to implementing the exposome into scientific research. Infrastructure support is needed for large-scale untargeted biomonitoring and managing big data. Advancement and standardization of technologies and methods for exposomic analyses as well as data analyses, integration, and sharing are another area requiring considerable support. Training environmental health scientists in big data and use of exposomic approaches in human health research is also critical. Promotion of the concept in research areas that can benefit from the exposome approach, such as epidemiology, can accelerate its implementation.

The NIEHS Children's Health Exposure Analysis Resource (CHEAR) Program addresses many of these challenges. CHEAR laboratory services span the breadth of the exposome and offer targeted and untargeted analyses. The CHEAR Data Center serves as a data repository, providing statistical analyses, data integration, and interpretation services. This data center is also developing community-based data standards and ontologies. NIEHS supports a hybrid approach to investigating the exposome, which depends upon interactive workshops, established resources (e.g., CHEAR), and new research initiatives. In these efforts, NIEHS interacts with other institutes within the National Institutes of Health, as well as with other federal agencies and international consortia as mentioned above. By taking this comprehensive approach, the exposome can be implemented into environmental health research.

### **Inter- and intraindividual variability and analytical advancements**

Many burdensome diseases are complex and result from a combination of hereditary and environmental factors. For example, the heritability, or the phenotypic variation in the population attributable to inherited factors is on average 50%, leaving the other 50% of phenotypic variation to specific influences of the environment [11]. Additional analytic tools and data are needed to discover exposures to explain missing phenotypic variation in the population. In contrast, genomic investigations have accelerated the pace for discovery of inherited factors in disease, and the same advances should be applied to discover the influences of exposures in disease [3]. For exposome studies to deliver the promised benefits and causal understanding we seek, and to inform public health policy and chemical risk assessment, it is important to understand metabolic (and other) networks in a holistic manner to link toxic

exposures, responses, and disease endpoints. However, modeling such networks may be confounded by uncertainty in both the interactions that occur and the variability of network components; some components of urinary and blood metabolomes (pool of low molecular weight compounds in the sample) exhibit high variability, whereas, others appear to be tightly regulated. Contributions of genetic and environmental factors to metabolic phenotypes have been identified, alongside differences in stability and resilience of components of the metabolomes over long periods in individuals [12, 13].

In addition, inter- and intraindividual variability in specific subpopulations or strata that are of importance to exposome studies should be characterized, including those in critical periods of life (including in utero, early life, and old age) where susceptibility to adverse consequences of exposure may be increased. Such subpopulations exist in the European Union FP7-funded HELIX project. This project is focused on studying the early life exposome by combining six mother-child cohorts in Europe: BiB (Born in Bradford, UK), EDEN (Étude des Déterminants pré-et postnataux du développement et de la santé de l'Enfant, France), INMA (INfancia y Medio Ambiente, Spain), KANC (Kaunas Cohort, Lithuania), MoBa (Norwegian Mother and Child Cohort Study, Norway), and the Rhea Mother-Child Cohort Study (Rhea, Greece) [6]. Within these cohorts, a total of 1200 mother-child pairs were selected for exposome characterization using a multitude of analytical approaches, including both internal and external measures. To assess variability in one of the measured parameters, the metabolome, a panel study within the INMA component of the subcohort was conducted to investigate the temporal variability of metabolite concentrations in children's urine over a week, relative to the interindividual variation. Variations attributable to experimental and biological factors (analytics, cohort, etc.) were assessed, giving a better understanding of the sources of variability in these metabolic phenotyping data [14]. The study showed that one of the sources of variation was diurnal and resulted in the recommendation that a pooled sample should be used made from urine samples collected at different times of the day, or a 24-h urine collection for exposome analysis. There was also interindividual variability in metabolites between males and females (citric acid) and in dietary metabolites such as *N*-methylnicotinic acid, which is derived from coffee, soda drinks, and chocolate.

An approach called "exposome-wide association studies" or equivalently "environment-wide association studies" (EWAS) is another methodology to drive discovery of new exposures in disease and the missing phenotypic variation in the population (e.g., [15–18]). EWAS offers a few advantages, including explicitly mitigating false positive findings and assessing the entire database of potential

environmental correlates systematically to avoid a fragmented literature of associations [19–21]. While far from causal (and observational), the associations that emerge are those that can be prioritized to investigate in biological experiments. For example, Patel and colleagues, after using an EWAS-like approach to find exposure factors putatively correlated with telomere length, investigated how the exposure factors potentially influence changes in gene expression using publicly available data from the *Gene Expression Omnibus* [22]. Exposures must influence changes in biological function if causal, and gene expression investigations are among the important approaches to decipher causal routes to disease.

### **Recent technological advancements in metabolomics for exposomics**

Exposome studies ultimately promise to help identify causal pathways that link environmental exposures to disease endpoints, by combining a range of state-of-the-art technologies that probe external and internal exposures, and concomitant responses. Central to the task of implementing human exposome studies is to characterize the metabolome, a phenotypic measure that encodes a wealth of information relating to genes, environmental exposures, and their various interactions [23–25]. Accordingly, high-resolution platforms such as ultraperformance liquid chromatography-mass spectrometry (UPLC-MS) and <sup>1</sup>H nuclear magnetic resonance (NMR) spectroscopy for generating metabolic phenotypic data are now commonly included in molecular epidemiological studies of chronic disease [26, 27]. For exposomics, it is essential that these platforms can (1) capture the diversity of the chemical space, (2) assign chemical identity to the associated mass spectral features, and (3) accurately quantify them. Both mass spectrometry (MS) and NMR spectroscopy have been used for analyses of the exposome; however, MS was a predominant focus of this Symposium which highlighted advancements in hardware and software for large-scale exposome analysis.

High-resolution metabolomics (HRM) of plasma has the potential to be a central platform for affordable, high-throughput, biomonitoring of environmental chemical exposures [28, 29]. The platform takes advantage of the ultra-high mass resolution, mass accuracy, sensitivity, and improved scan speed of modern Fourier-transform (FT) mass spectrometers. Improved data extraction methods support routine measurement of more than 20,000 mass spectral features in microliter volumes of plasma and other human samples, and advanced software packages support both knowledge-driven and data-driven approaches for interpretation of these data. These developments, listed below in more detail, create exciting new opportunities for large-scale, systematic exposome research.

### Fourier-transform mass spectrometry

Comisarow and Marshall [30] showed that FT of signals from the free induction decay (FID) of ions responding to an oscillating electric field that is orthogonal to a fixed ion cyclotron resonance (ICR) magnetic field could be used to measure the mass-to-charge ratio ( $m/z$ ) of an ionized metabolite. Application of FT-ICR MS to crude oil showed that 17,000 ions could be detected [31]. These instruments showed linear response characteristics for quantification over at least five orders of magnitude [32], but were limited for high-throughput analyses because of the slow scan speed required, and thus run times needed of 10 minutes or longer. Introduction of a new FT instrument, termed an Orbitrap [33], supported faster scan speeds, and scan speed has been further improved with a high-field version. Importantly, unlike time-of-flight instruments, the mass resolution and mass accuracy are preserved at low ion intensities, and this allows quantification of chemicals over nearly eight orders of magnitude [34]. This extensive dynamic range is an important capability for high-throughput exposome research because environmental chemicals are frequently encountered three-to-five orders of magnitude lower abundance than endogenous metabolites [35]. Early results from gas chromatography (GC)-Orbitrap analyses provide optimism for expansion of high-throughput biomonitoring to include more non-polar, hydrophobic environmental chemicals. Hundreds of environmental chemicals in different pesticide groups have been detected in single analyses of human samples (D.I. Walker, unpublished). Thus, a combination of liquid chromatography (LC) and GC analyses could provide a standardized central platform for longitudinal studies to initiate a human exposome project. Results are already available to show that analyses can be performed on samples stored for decades [36]. Therefore, rapid progress can be made by analysis of existing samples for which health outcome data are available. These high-throughput methods could provide a central resource on which to build a more extensive exposome research capability [28, 29].

### Large-scale, systematic exposome research

Humans are thought to experience more than a million exposures in a lifetime [37], so detailed understanding of life-stage-specific effects represent an enormous challenge. Using the Human Genome Project as a reference, one can focus on several criteria needed for exposome research explained below. These criteria have already been established for HRM, which further supports the use of HRM for exposome research.

First, assays need to be simple enough that they can be easily performed in multiple facilities; samples can now be processed using just a single extraction and centrifugation prior to analysis [38]. Second, assays need to be

high-throughput and automated with extensive coverage; the extraction and injection workflow for LC-MS can be fully automated, and run-times can be as short as 2.5 min (to detect more than 20,000 mass spectral features in batches of samples) [29]. Third, the assays need to be reproducible; analyzing each sample in triplicate, one can verify the reproducibility of an individual signal in an individual sample; however, a second biological sample set is always necessary when available. Fourth, standards are required; reference standardization [34] has been developed to support creation of cumulative data libraries; a pooled reference sample is calibrated against a National Institute of Standards and Technology (NIST) Standard Reference Material (e.g., SRM1950 containing approximately 100 metabolites) and concurrently analyzed with each batch of 20 samples. This analytic structure allows estimation of the absolute concentration of any known compound that is measured in the NIST standard. There are numerous standards available for different environmental chemical classes such as pesticides, organic contaminants from cigarettes, chemicals in drinking water, drugs, and their online spectral library are for approximately 267,376 chemical compounds. However, we know from a recent expansion to the METLIN metabolite database that the number of endogenous and environmental chemical compounds is at least 1 million, so some standards are still needed for complete coverage [39]. Finally, fifth, assays need to be affordable; the use of a dual chromatography [40], and dual ionization approach [41], provides a very broad spectrum of chemicals that can be analyzed in triplicate for about \$100/sample [28], with possibilities to lower this to below \$10/sample [36].

### Tools for data extraction and analysis

Extraction of ultra-high resolution mass spectral data can be effectively obtained with software language tools such as apLCMS [42] and XCMS [43]. A wrapper function in xMSanalyzer [44] allows re-extraction of the same data with different parameter settings to improve capture of signals with different characteristics, and also provides quality control, batch normalization, and other capabilities. xMSannotator [45] provides an automated computational framework using existing databases (ChemSpider, KEGG, HMDB, T3DB, LipidMaps) and a multistage clustering algorithm to assign confidence scores for chemical identity by evaluating intensity profiles, retention time characteristics, mass defect, isotope patterns, adduct patterns, and correlations with matches to other metabolites in known metabolic pathways. The resulting annotated data are compatible with any of the many pathway analysis tools, e.g., KEGG [46], MetaboAnalyst [47] and MetaCore [Thomson Reuters (<https://portal.genego.com/>)]. To complement these knowledge-based tools, which require knowledge of chemical identities prior to pathway analysis,

*mummichog* is a set of computational algorithms to predict metabolic pathway effects directly from spectral feature tables without prior identification of metabolites [48].

### Chemical identification

One of the major issues for metabolomics in studies of the exposome has been chemical/structural identification. Less than half of the ions detected by mass spectrometry are associated with known chemicals and essentially the rest are the “dark matter of the exposome.” Of critical importance for exposome research, about half of the mass spectral features associated with disease are currently unidentified [37]. It was recognized early on that the structural identification of detected metabolites might be problematic [49]. However, this was considered a temporary problem, easily solved by constructing databases containing all known mammalian metabolites. At the time, the mammalian metabolome was estimated to contain  $\leq 5000$  compounds, so the assertion that a complete metabolite database would be readily available seemed reasonable. Thanks to the diligent work of many individuals, there are now several excellent and freely available metabolite databases (METLIN, HMDB, KEGG, HumanCyc, and others) that contain these 5000 compounds, and many more [50–53]. Unfortunately, despite the availability of databases, high-throughput instrumentation, sophisticated statistical, and cheminformatics tools, the “structure identification” problem has not been solved. What has replaced the initial expectation of identifying the structures of hundreds or thousands of metabolites is the disappointing realization that most discovery metabolomics studies report the chemical structure of fewer than 100 compounds, and very often less than 50 [54]. Thus, the most significant barrier to progress and a seemingly intractable limitation of the field of metabolomics is the inability to identify the structures of most detected compounds.

A novel approach to help overcome this problem is the development of algorithms that predict physical/chemical properties of compounds contained in chemical databases. The properties chosen are those that can be experimentally measured for any unknown compound by LC-MS. These include retention indices, precursor ion survival curves, collision-induced dissociation fragmentation spectra, biological relevance, and collisional cross-sectional area. Compounds in databases (for example PubChem) whose predicted properties most closely match experimentally measured properties are returned as the most likely candidates for the unknown peak. This system is currently being validated in an *in vivo* model of acute trauma and could be of value for exposome studies. In addition, a multilevel framework to include unidentified signals in cumulative databases has been proposed in which ion characteristics

are defined in multivector space, which will aid in overcoming this problem [37].

Orthogonal technologies in MS can also separate and capture the chemical complexity of the samples, aiding structural identification. LC is a proven technology that is orthogonal to MS, and it can be adequately optimized to target the chemical space of interest, for example, hydrophilic interaction liquid chromatography is appropriate for hydrophilic molecules while a reversed-phase method provides optimal separation for lipophilic molecules. The correct chromatography allows separation of isomers and removes isobaric interference in a mass spectrometer and makes identification of unknowns easier. Similar to LC, ion mobility, a gas phase separation of ions based on their mobility, has been shown to be very useful in separating and capturing the components of a complex sample and providing with clean mass spectra needed for identification [55]. It has been recognized that appropriate fractionation of samples at the sample preparation step as well as on-line separations (chromatography and ion mobility) all assist in capturing the chemical complexity of a matrix, such as blood, by decoupling the overlapping interferences. Accurate quantification is the next logical step following detection and identification of exposure-associated molecular species. Molecular species originating from food and drugs fall in the same concentration regime as endogenous metabolites, while the blood concentration levels of environmental pollutants are 1000 times lower. Modern mass spectrometers have a dynamic range of six orders of magnitude, therefore assessment and validation of the dynamic range of the analytical assay used to measure these molecular species are critical [56]. Dynamic range of an analytical method is defined as the concentration range over which a signal response is linear to the concentration. It is important that the dynamic range in exposure analysis be defined in the context of a molecule, especially in MS-based analysis where the ionization of molecular ions can be vastly different. Furthermore, interference in the samples can easily result in ion suppression or enhancement, which can result in inaccurate quantification. Use of hyphenated technologies such as LC-MS is even more important in quantification since chromatography can separate out the interfering ions and alleviate ion suppression or enhancement challenges.

### Integrated omics methodologies

An important need is to link external exposures to internal body burden, and internal body burden to biologic responses and disease outcomes. A recent example of this analysis continuum was presented in an HRM study of occupational exposure to trichloroethylene (TCE) [57]. This study showed that TCE exposure was associated with known TCE metabolites in the blood and urine

and that other halogenated compounds were correlated both with TCE metabolites and disease biomarkers. Thus, the results showed the use of HRM to connect external exposures, internal body burden, biologic response, and biomarkers of disease outcome. These analyses used a new R script, xMWAS [<https://sourceforge.net/projects/xmwash/>] [58], which advances capabilities to integrate HRM data with any other omics data, as already shown for transcriptome x metabolome [59], microbiome x metabolome [60], cytokine x metabolome [61], and suitable for genome, epigenome, proteome, and other integrated omics analyses that will be essential for mechanistic understanding of the exposome.

## Current limitations in exposomics

### Mixtures of exposure profiles

One major limitation of approaching exposure-disease associations is the number, diversity, and mixture of environmental exposure profiles at the individual level; such an array of exposures is a challenge for both measurement science and in efforts to identify components or combinations that are associated with adverse outcomes. Several publications have shown how perturbations in metabolic networks (and others such as those of gene regulation and protein-protein interactions) can help describe complex phenomena such as multimorbidity; nodes in the underlying biochemical networks connect disease phenotypes [62, 63]. It may, therefore, be useful to consider a set of these network perturbations as endpoint proxies when identifying the effects of exposure. Similarly, the conceptual framework for defining adverse outcome pathways (AOPs), i.e., making causal links from molecular initiating events (MIEs) to subsequent key events (KEs), may help reduce the complexity through identifying common MIEs for multiple environmental stressors [64].

### False-positives

With big data come big analytical challenges. EWASs are observational studies and prone to biases such as confounding and reverse causality (e.g., disease coming before the exposure) [65]. While there are analytic and epidemiological study designs that attempt to mitigate these biases, they have yet to be harnessed in high-throughput exposure investigations. Yet another issue includes multiple significant associations (low  $p$  values) with small effect sizes. A recent “prescription-wide association study” on time-to-cancer in a large and comprehensive Swedish database [66] found that when examining  $> 500$  drugs prescribed, almost a quarter of them were associated with cancer with small effect. Furthermore, these effects changed as a function of analytic method chosen (e.g., case-crossover vs. Cox proportional hazard), leaving an investigator with many potential false positives to sift through. This may be the norm in

exposome-wide analytics, where correlations may be small but seemingly correlated with everything else and it will be a challenge to ascertain causality [67, 68].

## Future perspectives

With respect to the future, the thirst for metabolic phenotype data, to complement other available omics, personal monitoring, and activity data, is only likely to increase, as the concept of the human exposome is developed and deployed around the world. As mentioned previously, methods that can integrate metabolomic data with other system-level data can provide valuable insight for exposomics and provide a way to link these datasets together in a meaningful way. It is therefore timely that several national and international initiatives have been started to provide high-throughput, multiplatform analyses with the capacity to serve the wider scientific community. These initiatives will benefit from a renewed commitment to analytical and data standards harmonization, and opportunities for wider access through open data agreements. If the recent efforts to make large, well-curated, EWAS datasets available alongside tools for their interrogation (e.g., Patel et al. [69]) can be mirrored for MWAS, then progress in exposome studies is likely to move forward at pace.

While recent advances in MS have been enabling, we still need to take advantage of the orthogonally complementing technologies to separate and capture the chemical complexity of the samples. In addition, standardization needs to be addressed, as more exposome studies are developed, reproducibility across sample cohorts will inevitably become an issue. As biobanks are being amassed, there is no current standardization for sample collection, thus data generated will be biased to collection procedure and quality of sample during long-term storage. Currently, data repositories exist for metabolomics and genomics but are missing important additional detailed information such as the type of sample collection tube, time between collection and freezing, number of freeze-thaw cycles, and analytical platform used (vendor/model). Thus, these important points need to be considered for widespread integration of the exposome.

## Conclusions

In a new era of high-throughput environmental exposure assessment, there is an urgent need for new analytic approaches to drive discovery of new exposures associated with disease and phenotype [70]. Herein, the Symposium provided increased awareness of the utility of the exposome in environmental health research, as well as recent updates on technological innovation and challenges that still need to be overcome to enable the study of the exposome, particularly in metabolomics. Despite these challenges, the exposome paradigm may bear fruit in

enhancing discoveries to identify exogenous factors that have up to now been elusive in disease risk. Moving forward, new technologies to measure environmental exposures in large populations in a cost-effective manner are needed, along with harnessing sound epidemiological designs to extract potential signals from noise, with data deposited in the public domain for investigators of all ilk to analyze results.

#### Acknowledgments

We would like to thank the Waters Corporation for their support and sponsorship for the Symposium Drs. Dustin Yaworsky and Ray Chen. We would also like to thank Ms. Donna Cropley and Ms. Damaris Faustine for their invaluable help in organizing the Symposium.

#### Funding

The authors would like to acknowledge the following funding sources: NIH R01 EY017963 (VV), NIH U01 AA021724 (VV), NIH R24 AA022057 (VV), NIH R00 ES023504 (CP), R21 ES025052 (CP), R01 ES023485 (DJ), R21 ES025632 (DJ), P30 ES019776 (DJ), NIH S10 OD018006 (DJ), U2C ES026560 (DJ), Women's Health Research at Yale (CJ), and Yale Comprehensive Cancer Center (CJ),

#### Availability of data and materials

Not applicable.

#### Authors' contributions

All authors were involved in writing and contributing to the manuscript. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Department of Environmental Health Sciences, Yale School of Public Health, New Haven, CT, USA. <sup>2</sup>Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. <sup>3</sup>MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College Norfolk Place, London, UK. <sup>4</sup>Division of Extramural Research and Training, National Institute of Environmental Health Sciences, National Institutes of Health, Morrisville, NC, USA. <sup>5</sup>Waters Corporation, Metabolomics and Translational Research, Milford, MA, USA. <sup>6</sup>Department of Pharmaceutical Sciences, University of Connecticut, Storrs, CT, USA. <sup>7</sup>Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA. <sup>8</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

Received: 1 November 2017 Accepted: 1 December 2017

Published online: 08 December 2017

#### References

1. Willett WC. Balancing life-style and genomics research for disease prevention. *Science*. 2002;296:695–8.
2. Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. *Science*. 2010;330:460–1.
3. Manrai AK, Cui Y, Bushel PR, Hall M, Karakitsios S, Mattingly CJ, Ritchie M, Schmitt C, Sarigiannis DA, Thomas DC, et al. Informatics and data analytics to support exposome-based discovery for public health. *Annu Rev Public Health*. 2017;38:279–94.
4. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev*. 2005;14:1847–50.
5. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. *Toxicol Sci*. 2014;137:1–2.
6. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, Thomsen C, Wright J, Athersuch TJ, Avellana N, et al. The human early-life exposome (HELIx): project rationale and design. *Environ Health Perspect*. 2014;122:535–44.
7. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, Kogevinas M, Kyrtopoulos S, Nieuwenhuijsen M, Phillips DH, et al. The exposome in practice: design of the EXPOsOMICS project. *Int J Hyg Environ Health*. 2017;220:142–51.
8. Dennis KK, Marder E, Balshaw DM, Cui Y, Lynes MA, Patti GJ, Rappaport SM, Shaughnessy DT, Vrijheid M, Barr DB. Biomonitoring in the era of the exposome. *Environ Health Perspect*. 2016;125:502–10.
9. Dennis KK, Auerbach SS, Balshaw DM, Cui Y, Fallin MD, Smith MT, Spira A, Sumner S, Miller GW. The importance of the biological impact of exposure to the concept of the exposome. *Environ Health Perspect*. 2016;124:1504–10.
10. Stingone JA, Buck Louis GM, Nakayama SF, Vermeulen RC, Kwok RK, Cui Y, Balshaw DM, Teitelbaum SL. Toward greater implementation of the exposome research paradigm within environmental epidemiology. *Annu Rev Public Health*. 2017;38:315–27.
11. Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, Posthuma D. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet*. 2015;47:702–9.
12. Nicholson G, Rantalaainen M, Maher AD, Li JV, Malmodin D, Ahmadi KR, Faber JH, Hallgrimsdottir IB, Barrett A, Toft H, et al. Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol Syst Biol*. 2011;7:525.
13. Ghini V, Saccetti E, Tenori L, Assfalg M, Luchinat C. Allostasis and resilience of the human individual metabolic phenotype. *J Proteome Res*. 2015;14: 2951–62.
14. Maitre L, Lau CE, Vizcaino E, Robinson O, Casas M, Siskos AP, Want EJ, Athersuch T, Slama R, Vrijheid M, et al. Assessment of metabolic phenotypic variability in children's urine using 1H NMR spectroscopy. *Sci Rep*. 2017;7: 46082.
15. Patel CJ, Rehkopf DH, Leppert JT, Bortz WM, Cullen MR, Chertow GM, Ioannidis JP. Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States national health and nutrition examination survey. *Int J Epidemiol*. 2013;42:1795–810.
16. Patel CJ, Cullen MR, Ioannidis JP, Butte AJ. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. *Int J Epidemiol*. 2012;41:828–43.
17. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One*. 2010;5:e10746.
18. McGinnis DP, Brownstein JS, Patel CJ. Environment-wide association study of blood pressure in the National Health and Nutrition Examination Survey (1999–2012). *Sci Rep*. 2016;6:30373.
19. Ioannidis JP. Exposure-wide epidemiology: revisiting Bradford Hill. *Stat Med*. 2016;35:1749–62.
20. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*. 2011;22:450–6.
21. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68:1046–58.
22. Patel CJ, Manrai AK, Corona E, Kohane IS. Systematic correlation of environmental exposure and physiological and self-reported behaviour factors with leukocyte telomere length. *Int J Epidemiol*. 2017;46:44–56.
23. Athersuch TJ. The role of metabolomics in characterizing the human exposome. *Bioanalysis*. 2012;4:2207–12.
24. Athersuch T. Metabolome analyses in exposome studies: profiling methods for a vast chemical space. *Arch Biochem Biophys*. 2016;589:177–86.
25. Athersuch TJ, Keun HC. Metabolic profiling in human exposome studies. *Mutagenesis*. 2015;30:755–62.
26. Dona AC, Jimenez B, Schafer H, Humpfer E, Spraul M, Lewis MR, Pearce JT, Holmes E, Lindon JC, Nicholson JK. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem*. 2014;86:9887–94.
27. Lewis MR, Pearce JT, Spagou K, Green M, Dona AC, Yuen AH, David M, Berry DJ, Chappell K, Horneffer-van der Sluis V, et al. Development and

- application of ultra-performance liquid chromatography-TOF MS for precision large scale urinary metabolic phenotyping. *Anal Chem.* 2016;88:9004–13.
28. Jones DP. Sequencing the exposome: a call to action. *Tox Rep.* 2016;3:29–45.
  29. Walker DJ, Go YM, Liu K, Pennell KD, Jones DP. Population screening for biological and environmental properties of the human metabolic phenotype: implications for personalized medicine. In: Holmes E, Nicholson JK, Darzi AW, Lindon JC, editors. *Metabolic Phenotyping in personalized and public healthcare*. London: Elsevier; 2016. p. 167–211.
  30. Comisarow MB, Marshall AG. Fourier-transform ion-cyclotron resonance spectroscopy. *Chem Phys Lett.* 1974;25:282–3.
  31. Rodgers RP, Schaub TM, Marshall AG. Petroleomics: MS returns to its roots. *Anal Chem.* 2005;77:20. A-27 A
  32. Johnson JM, Strobel FH, Reed M, Pohl J, Jones DP. A rapid LC-FTMS method for the analysis of cysteine, cystine and cysteine/cystine steady-state redox potential in human plasma. *Clin Chim Acta.* 2008;396:43–8.
  33. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. *J Mass Spectrom.* 2005;40:430–43.
  34. Go YM, Walker DJ, Liang Y, Uppal K, Soltow QA, Tran V, Strobel F, Quyyumi AA, Ziegler TR, Pennell KD, et al. Reference standardization for mass spectrometry and high-resolution metabolomics applications to exposome research. *Toxicol Sci.* 2015;148:531–43.
  35. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect.* 2014;122:769–74.
  36. Walker DJ, Mallon CT, Hopke PK, Uppal K, Go YM, Rohrbeck P, Pennell KD, Jones DP. Deployment-associated exposure surveillance with high-resolution metabolomics. *J Occup Environ Med.* 2016;58:S12–21.
  37. Uppal K, Walker DJ, Liu K, Li S, Go YM, Jones DP. Computational metabolomics: a framework for the million metabolome. *Chem Res Toxicol.* 2016;29:1956–75.
  38. Johnson JM, Yu T, Strobel FH, Jones DP. A practical approach to detect unique metabolic patterns for personalized medicine. *Analyst.* 2010;135:2864–70.
  39. Warth B, Spangler S, Fang M, Johnson CH, Forsberg EM, Granados A, Martin RL, Domingo-Almenara X, Huan T, Rinehart D, et al. Exposome-scale investigations guided by global metabolomics, pathway analysis, and cognitive computing. *Anal Chem.* 2017;89:11505–13.
  40. Soltow QA, Strobel FH, Mansfield KG, Wachtman L, Park Y, Jones DP. High-performance metabolic profiling with dual chromatography-Fourier-transform mass spectrometry (DC-FTMS) for study of the exposome. *Metabolomics.* 2013;9:S132–43.
  41. Liu KH, Walker DJ, Uppal K, Tran V, Rohrbeck P, Mallon TM, Jones DP. High-resolution metabolomics assessment of military personnel: evaluating analytical strategies for chemical detection. *J Occup Environ Med.* 2016;58:S53–61.
  42. Yu T, Park Y, Johnson JM, Jones DP. apLCMS—adaptive processing of high-resolution LC/MS data. *Bioinformatics.* 2009;25:1930–6.
  43. Libiseller G, Dvorak M, Kleb U, Gander E, Eisenberg T, Madeo F, Neumann S, Trausinger G, Sinner F, Pieber T, Magnes C. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinform.* 2015;16:118.
  44. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, Jones DP. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinform.* 2013;14:15.
  45. Uppal K, Walker DJ, Jones DP. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. *Anal Chem.* 2017;89:1063–7.
  46. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 2014;42:D199–205.
  47. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.* 2012;40:W127–33.
  48. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, Jones DP, Pulendran B. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol.* 2013;9:e1003123.
  49. Walker DJ, Uppal K, Zhang L, Vermeulen R, Smith M, Hu W, Purdue MP, Tang X, Reiss B, Kim S, et al. High-resolution metabolomics of occupational exposure to trichloroethylene. *Int J Epidemiol.* 2016;45:1517–27.
  50. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. METLIN—a metabolite mass spectral database. *Ther Drug Monit.* 2005;27:747–51.
  51. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu YF, Djoumbou Y, Mandal R, Aziz F, Dong E, et al. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.* 2013;41:D801–7.
  52. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012;40:10.
  53. Trupp M, Altman T, Fulcher CA, Caspi R, Krummenacker M, Paley S, Karp PD. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. *Genome Biol.* 2010;11:O12.
  54. Zamboni N, Saghatelian A, Patti GJ. Defining the metabolome: size, flux, and regulation. *Mol Cell.* 2015;58:699–706.
  55. Paglia G, Astarita G. Metabolomics and lipidomics using traveling-wave ion mobility mass spectrometry. *Nat Protoc.* 2017;12:797–813.
  56. Metz TO, Baker ES, Schymanski EL, Renslow RS, Thomas DG, Causon TJ, Webb IK, Hann S, Smith RD, Teagueguarden JG. Integrating ion mobility spectrometry into mass spectrometry-based exposome measurements: what can it add and how far can it go? *Bioanalysis.* 2017;9:81–98.
  57. Castle AL, Fiehn O, Kaddurah-Daouk R, Lindon JC. Metabolomics standards workshop and the development of international standards for reporting metabolomics experimental results. *Brief Bioinform.* 2006;7:159–65.
  58. Uppal K, Ma C, Go YM, Jones DP. xMWAS: a data-driven integration and differential network analysis tool. *Bioinformatics.* 2017;btw656.
  59. Roede JR, Uppal K, Park Y, Tran V, Jones DP. Transcriptome-metabolome wide association study (TMWAS) of maneb and paraquat neurotoxicity reveals network level interactions in toxicologic mechanism. *Toxicol Rep.* 2014;1:435–44.
  60. Cribbs SK, Uppal K, Li S, Jones DP, Huang L, Tipton L, Fitch A, Greenblatt RM, Kingsley L, Guidot DM, et al. Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome.* 2016;4:3.
  61. Chandler JD, Hu X, Ko EJ, Park S, Lee YT, Orr M, Fernandes J, Uppal K, Kang SM, Jones DP, Go YM. Metabolic pathways of lung inflammation revealed by high-resolution metabolomics (HRM) of H1N1 influenza virus infection in mice. *Am J Physiol Regul Integr Comp Physiol.* 2016;311:R906–16.
  62. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A.* 2008;105:9880–5.
  63. Sturemberg JP, Bennett JM, Martin CM, Picard M. ‘Multimorbidity’ as the manifestation of network disturbances. *J Eval Clin Pract.* 2017;23:199–208.
  64. Sturla SJ, Boobis AR, FitzGerald RE, Hoeng J, Kavlock RJ, Schirmer K, Whelan M, Wilks MF, Peitsch MC. Systems toxicology: from basic research to risk assessment. *Chem Res Toxicol.* 2014;27:314–29.
  65. Ioannidis JP, Loy EY, Poulton R, Chia KS. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci Transl Med.* 2009;1:7ps8.
  66. Patel CJ, Ji J, Sundquist J, Ioannidis JP, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. *Sci Rep.* 2016;6:31308.
  67. Patel CJ, Ioannidis JP. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Community Health.* 2014;68:1096–100.
  68. Patel CJ, Manrai AK. Development of exposome correlation globes to map out environment-wide associations. In: Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, Klein TE, editors. *Biocomputing 2015 Proceedings of the Pacific Symposium: Kohala Coast, Hawaii, USA: 4–8 January 2015*. New Jersey: World Scientific Pub Co Inc; 2015. p. 231–42.
  69. Patel CJ, Pho N, McDuffie M, Easton-Marks J, Kothan C, Kohane IS, Avillach P. A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Sci Data.* 2017;3.
  70. Patel CJ, Ioannidis JP. Studying the elusive environment in large scale. *JAMA.* 2014;311:2173–4.