Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health

Arjun K. Manrai,[1] Yuxia Cui,[2] Pierre R. Bushel,[2] Molly Hall,[3] Spyros Karakitsios,[4] Carolyn J. Mattingly,[5] Marylyn Ritchie,[3,6] Charles Schmitt,[7] Denis A. Sarigiannis,[4] Duncan C. Thomas,[8] David Wishart,[9] David M. Balshaw,[2] and Chirag J. Patel[1,10]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115; email: chirag_patel@hms.harvard.edu

[2]National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709; email: balshaw@niehs.nih.gov

[3]Center for Systems Genomics, The Pennsylvania State University, College Station, Pennsylvania 16802

[4]Environmental Engineering Laboratory, Department of Chemical Engineering, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

[5]Department of Biological Sciences, College of Sciences, North Carolina State University, Raleigh, North Carolina 27695

[6]Geisinger Health System, Danville, Pennsylvania 17821

[7]Renaissance Computing Institute, Chapel Hill, North Carolina 27517

[8]Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089-9011

[9]Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada

[10]Center for Assessment Technology and Continuous Health, Massachusetts General Hospital, Boston, Massachusetts 02114

Annu. Rev. Public Health 2017. 38:279–94

First published online as a Review in Advance on December 23, 2016

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-082516-012737

Keywords

exposures, bioinformatics, genomics, environment-wide association studies, data standards

Abstract

The complexity of the human exposome—the totality of environmental exposures encountered from birth to death—motivates systematic, high-throughput approaches to discover new environmental determinants of disease. In this review, we describe the state of science in analyzing the human exposome and provide recommendations for the public health community to consider in dealing with analytic challenges of exposome-based biomedical

research. We describe extant and novel analytic methods needed to associate the exposome with critical health outcomes and contextualize the data-centered challenges by drawing parallels to other research endeavors such as human genomics research. We discuss efforts for training scientists who can bridge public health, genomics, and biomedicine in informatics and statistics. If an exposome data ecosystem is brought to fruition, it will likely play a role as central as genomic science has had in molding the current and new generations of biomedical researchers, computational scientists, and public health research programs.

INTRODUCTION

The human exposome is defined as the totality of environmental exposures encountered from birth to death and includes a diverse mix of dietary nutrients, pharmaceutical drugs, infectious agents, and pollutants (4, 33, 51, 68, 69). Much as human genetics has benefited from high-throughput profiling in the form of genome-wide association studies (GWASs), exposome science needs a systematic, high-throughput paradigm for systematically and reproducibly discovering the environmental determinants of disease (42). Here we describe emerging analytics and informatics efforts that enable systematic studies to associate exposures with disease. We describe new classes of bioinformatics and biostatistics tools and methods that will be needed for future research, as well as the cross-cutting training that exposome data scientists will need in bioinformatics, statistics, computer science, and public health. We provide recommendations that address the data analytic needs for exposome research (**Table 1**).

THE NEED FOR A NEW DISCOVERY-BASED PARADIGM

Environmental factors have long been implicated as major contributors to the global disease burden. For example, data from several sources indicate that greater than 70% of the nonviolent deaths in the United States can be attributed to cigarette smoking, dietary imbalance, air pollution, adverse drug reactions, and infectious agents, acting alone or in concert with genetic or other host susceptibility factors (7, 34). Yet identification of specific factors, their interactions, and their effects on human health has remained elusive.

Family-based and genomic studies have shown that heritability, the proportion of phenotypic/disease variance that can be ascribed to inherited factors, is often modest (46) and can be overestimated owing to shared exposures (50). For complex diseases such as cancer and type 2 diabetes, heritability ranges from approximately 10% to greater than 50% (56). In a recent meta-analysis across 50 years of twin studies and nearly 18,000 traits, heritability across all traits was 49% (46). Therefore, a significant proportion of phenotypic and disease variance can likely be attributed to nongenetic factors such as environmental exposures. However, aside from a handful of diseases such as lung cancer (of which >80% can be attributed to smoking), we have yet to describe much of the phenotypic variability for most complex diseases.

Humans encounter numerous exposures over the lifespan. For example, recent catalogues contain up to 3,600 toxicants in the Toxic Exposome Database (71) and 13,000 in the Comparative Toxicogenomics Database (9). Through the US Toxic Substances Control Act (TSCA), the US Environmental Protection Agency (EPA) has compiled an inventory of 84,000 chemicals (63). Curated exposures include small-molecule analytes that are by-products of metabolism (e.g., endogenous exposures), nonchemical stressors such as radiation and climate, and complex

Table 1 Data-related recommendations for a human exposome project

Recommendation		Examples
1	Catalog contributions of environmental exposures to disease risk (e.g., susceptibility, variance explained) to strengthen the case for exposome research.	Develop requirements for an exposome-disease association catalog.
2	Identify high-throughput (e.g., 'omics, sensor-based) technologies and gaps to allow agnostic assessment of the exposome.	Develop infrastructure to characterize the variability of the exposome in various populations, akin to the NHANES.
3	Incentivize other parties (e.g., 'omics investigators in other disciplines, funding institutions, industrial entities) to integrate the exposome in their programs and develop high-throughput analytics methods to analyze exposome data.	Develop big data analytics and visualization tools to accelerate exposome-related research (e.g., exposome–phenome association studies). Identify how existing 'omics statistical methods can be extended for exposome research and identify gaps for new method development. Encourage a shift in focus from "one exposure–one phenotype" to multiple exposures, genes, and phenotypes. Develop methods to link the internal and the external exposome. Develop methods to support a variety of study designs (e.g., longitudinal studies) to strengthen inference and causality.
4	Identify data standards for high-throughput exposome research.	Develop data and domain language standards to encourage reuse in exposome-related research in future data collection and retrospective annotation. Formalize the role that ontologies play in integration/analysis.
5	Promote analytics standards and code reuse.	Identify open-source software tools to jump-start exposome analyses. Identify partners to extend existing infrastructure to host repositories
6	Integrate measurement, processing, modeling, and analyses through global initiatives.	Identify requirements to support measurement and raw data analysis workflows to measure individual-level exposomes e.g., connect existing core facilities to measure the exposome, e.g., integrating over NIH Commons initiatives (e.g., metabolomics, microbiome). Determine possibilities for joint funding to assess the robustness between environmental exposures and health status in large populations.
7	Encourage data sharing for reproducible research.	Evaluate strategies for exposome-related data sharing. Work to engage all players involved in the research process, including journal editors and funders.
8	Provide educational and outreach opportunities.	Identify public example data sets for methods development. Sponsor challenges and competitions to promote exposome-driven data analytics development. Develop exposome-related informatics training support akin to NIH Common Fund BD2K K career awards.

Abbreviations: BD2K, Big Data to Knowledge; NHANES, National Health and Nutrition Examination Survey; NIH, National Institutes of Health.

mixtures such as air pollution. Following these efforts, our first recommendation is to systematically catalog all published associations between environmental exposures and disease (**Table 1**, recommendation 1). A synthesis of the literature in the form of a catalog will enable the scientific community (e.g., investigators and funding agencies) to understand the current state of environmental health research, such as the sizes of associations (e.g., risks) between exposures and disease.

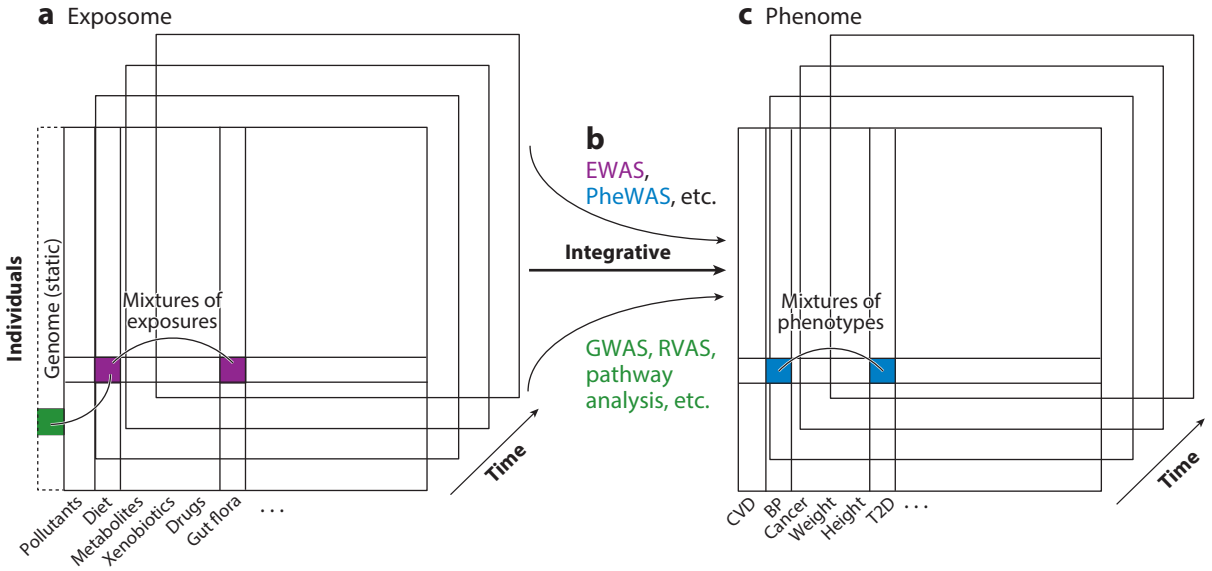

Figure 1

High-throughput data analyses in exposome-related research. (*a*) The exposome is a unified, multimodal, temporally dependent, and comprehensive digital representation of external and internal environmental exposures linked to humans. Individual exposome indicators are depicted in purple. Individual genomes are depicted in green (and are static, in contrast with the exposome). (*b*) Data analytics of the exposome can be used to systematically discover relationships between mixtures of exposures, the genome, and (*c*) the traits and diseases that make up the human phenome. Phenotypes of the phenome are depicted in blue. In *a* and *c*, diet and gut flora are linked with genomic markers to type 2 diabetes (T2D) and blood pressure (BP). Analytic methods to discover exposures (EWAS), genotypes (e.g., GWAS), and phenotypes (e.g., PheWAS) are depicted in purple, green, and blue, respectively.

WHAT IS THE HUMAN EXPOSOME DATA STRUCTURE?

First described in 2005, the exposome has been popularly defined as the "totality of environmental exposures from birth onwards" and a "complement" of the genome (4, 28, 52, 68). Miller & Jones (33) recently refined the definition of the exposome as the "cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment, behavior, diet, and endogenous processes" connecting biological response and exposure.

We use these abstract definitions to inform the characteristics and structure of exposome-related data. Although an individual's inherited genome is a largely static sequence of four nucleic acid bases, exposome data have several notable differences: (*a*) measurement heterogeneity (e.g., biomarkers, external sensors) and type (e.g., continuous, categorical); (*b*) a denser correlation structure; (*c*) time dependence; and (*d*) spatial dependence (23, 41–43) (**Figure 1*a***). These characteristics may influence disease at different life stages, from in utero to adulthood.

One example of a successful effort is the National Health and Nutrition Examination Survey (NHANES), a biannual health survey conducted by the US Centers for Disease Control and Prevention. Measured factors include environmental exposures such as chemicals, nutrients, and infectious agents (measured in human tissue such as blood serum, urine, and hair). For example, quantitative measurements of nutrient (e.g., vitamins, carotenes) and pollutant levels (e.g., heavy metals, polychlorinated biphenyls) in human tissue are obtained using comprehensive analytical

techniques such as liquid chromatography and gas chromatography mass spectrometry. Infectious agents (e.g., bacteria) are measured via immunological assays. The NHANES also includes other indicators of environmental exposures such as self-reported nutrient consumption (based on a food questionnaire), physical activity, and prescribed pharmaceutical drugs. Although these data are an example of the complex array of human exposures, they are neither comprehensive across all exposure domains nor longitudinal. For example, orthogonal external exposure data sources such as air pollution sensors and cell phones generate a wealth of complementary data that present opportunities for new analyses as well as shared challenges, including autocorrelation and missing data. Some challenges and opportunities regarding the use of external sensors to assess the exposome are described in an accompanying review (61). We recommend that the scientific community devote efforts to extend NHANES-like epidemiological biomonitoring surveys to ascertain baseline and reference intervals of exposome-based measurements to capture the extent of variability of the exposome in diverse populations (**Table 1**, recommendation 2).

The human exposome data structure is a high-dimensional collection of highly heterogeneous exposure variables that may change upon repeated sampling during an individual's lifetime (**Figure 1*a***). Time-dependent and high-throughput genome-scaled phenotypic data types such as gene expression, protein expression, or metabolomics data are similar in structure. For example, to measure gene expression, an array of gene probes measures the amount of messenger RNA expressed in a tissue or cell. Similarly, an exposome array would measure the amount of a multitude of time-varying external exposures and biomarkers of exposures. However, we have yet to develop an assay to operationalize an exposome-wide measurement (e.g., 100–1000s of exposures measured with a single measurement platform). To this end, we recommend that the exposome community expand on the integration of 'omics technologies (e.g., metabolomics) that allow for the comprehensive, agnostic, and high-throughput measurement of the personal exposome in humans (**Table 1**, recommendation 2). Notable challenges remain in the metabolomics field, such as determining the exact chemical identity or structure of analytes that are detected in human tissue from a mass spectrum.

Another nontrivial challenge involves the integration of data across the external and internal exposome domains. For example, how do (*a*) sensor-based and physical measures (e.g., air pollution or noise) or (*b*) individual-level sociodemographic factors influence the distribution of internal exposome indicators? Data integration across these domains is often executed by merging across spatiotemporal coordinates, a resource-intensive but usually straightforward database task. There is a basis for data fusion in the environmental epidemiology literature [e.g., in the *Handbook of Spatial Epidemiology* (26)]; however, an outstanding challenge remains in how to fuse multidimensional and longitudinal data streams emerging from external and internal exposome measurements. Methods such as canonical correlation analysis or graphical least angle regression (graphical LASSO) techniques can enable investigators to map one large data set (e.g., external exposures) to another (e.g., internal exposures) (13), but missing are methods to consider longitudinal data. Computational methods enjoying a resurgence in the data science community, such as neural networks, may also be harnessed to assimilate data over different dimensions (11).

SYSTEMATIC AND REPRODUCIBLE EXPOSOME-WIDE ASSOCIATIONS

High-throughput data analytics methods can be used to systematically and reproducibly discover relationships between exposures, the genome, and traits and diseases of interest (**Figure 1*a–c***). We recommend that significant research resources be devoted to the development of high-throughput analytics methods to understand the connections between the exposome and the phenome

(**Table 1**, recommendation 3). These efforts should include methods that enable the discovery of relationships among multiple exposures and multiple phenotypes, allowing researchers to extend the one-exposure-one-phenotype approach in a variety of study designs such as longitudinal and/or case-control studies. Computational methods are also needed to establish associations between the external and internal components of the exposome, providing greater depth to our understanding of the nature of exposures in time and space as well as of their relation to biological dose.

Similarities and Differences Between Exposome and Genome Science

Much can be learned about design and analysis challenges in exposome association studies from parallels with GWASs (23). From a design perspective, GWASs have highlighted the value of data sharing and replication in independent data (25). From an analysis perspective, major concerns include mitigating chances for false-positive findings incurred when executing multiple statistical tests (known as multiple hypothesis testing) to achieve genome-wide significance (e.g., $p < 5 \times 10^{-8}$) and adjustment for the confounding effects of population stratification (47). From a strategic perspective, a third major theme has involved leveraging external information now available in numerous genomic and pathway databases ("ontologies") such as gene set enrichment analysis (57). These analytic lessons are equally applicable to exposome-wide association studies (EWASs).

Identifying associations between the exposome and disease must overcome several analytical challenges not encountered in GWAS. It is highly unlikely that there will ever be a single epidemiological study that monitors a cohort's exposures continuously across an entire lifespan (thus outliving the investigators). Instead, available data will typically comprise a few snapshots of exposure, though often during critical windows of exposure. Repeated longitudinal exposure data provide information about the time course of effects (e.g., modification by intensity, duration, age at, and time since exposure) and about vulnerable ages at exposure and can help avoid the problem of reverse causation. Thus, it is important to think not only in terms of simple dose–response curves that describe disease risk as a function of some summary index of cumulative exposure or average exposure intensity but also in terms of "exposure-time-response relationships" (60).

Associating Exposures with Disease

"Environment-wide association studies" or equivalently "exposome-wide association studies," which are analogous methodologically to genome-wide association studies (GWASs), are a recently proposed analytic approach to systematically associate exposures with disease (**Figure 1*b***). In EWASs, multiple exposures are assessed simultaneously for their association with a phenotype or disease of interest. The false discovery rate (1) is controlled to adjust for multiple testing, and significant associations are validated in independent data (38–40, 62). The main advantage of this approach is that it systematically investigates an array of exposures and adjusts for multiple testing, thus avoiding selective reporting while enabling discovery. Just as the literature for genetic associations in disease has become more reproducible owing to standardized and extensively validated analytical procedures (22, 64), we propose that an analogous process to associate the exposome with disease and health outcomes will result in more robust environmental associations. EWASs may be the tip of the analytic iceberg: Humans are a mixture of phenotypic traits (e.g., **Figure 1*b–c***) and perhaps emerging PheWAS [phenome-wide association study (e.g., 10, 18)] approaches (**Figure 1*b–c***) may be developed to understand how exposures are associated with multiple phenotypes and diseases simultaneously.

While an EWAS produces a set of prioritized and possibly replicated individual correlates, this study design does not typically yield causal factors. After an EWAS, experiments must still be designed to infer causality.

Modeling Challenges

Analyzing multiple environmental exposures of the exposome in concert may enable investigators to consider mixtures versus individual components of the environment in relation to health. Relating complex mixtures in the external environment to health outcomes has been recognized as a major challenge in environmental health and public health policy for some time (12), and a variety of statistical methods have been developed to address this problem (2, 58). The essential difficulty is that with highly correlated variables, standard multiple regression models produce highly unstable parameter estimates, and simple variable selection techniques such as stepwise regression can select the wrong variables. Analyses that fail to correct for these issues could lead to predictions that would support potentially counterproductive policy recommendations (e.g., regulating the wrong pollutant or source). Dimension reduction techniques such as regression on principal components analysis, partial least squares, clustering, or kernel machine regression offer some approaches to address these problems, but their interpretation is not straightforward. However, variable selection methods that extend regression models while accounting for correlation of variables, such as elastic net regression (73), may enable the selection of entire clusters of correlated exposures with the interpretability of a simple regression model.

For some agents, the physiological processes involved are understood well enough to enable sophisticated mathematical modeling of the toxicokinetic processes that determine their concentration in human tissues using information about external exposures and the toxicodynamic processes that determine their health effects. These approaches are generally known as physiologically based pharmacokinetic (PBPK) or pharmacodynamic (PBPD) models. These approaches have been extended to allow for interindividual variability in the underlying rate parameters (66) and, more recently, to incorporate specific genetic determinants of these rates (6). A long-term hurdle to the widespread use of PBPK models for exposure/risk assessment is the lack of a standardized modeling framework. Several research groups are developing generic PBPK models, either as standalone tools such as PK-Sim (70) and Indus-Chem (24) or incorporated within integrated computational platforms for exposure assessment such as MENTOR (Modeling ENvironment for TOtal Risk Studies) (15). The development of generic and validated PBPK models for many individual chemical exposures is supported by recent advances in quantitative structure–property relationships (QSPRs) (45, 48, 53). However, the integration, validation, and evaluation of these models into a multiple agent exposome profile are outstanding challenges. For example, applying such approaches to the full spectrum of exposures of the exposome may be metabolized by many different pathways, a challenging proposition for the state of the art (36, 54).

Finally, the problem of exposure measurement error has been widely discussed in the statistical and epidemiological literature (5). In the classical error model, in which the measurement errors are uncorrelated with the true value of the exposure, the general effect in single exposure models is to dilute a true association (biasing its magnitude toward the null and reducing power). A variety of statistical methods are available to correct for this bias if the distribution of measurement errors is known, although the loss of power generally cannot be recovered by purely statistical methods. Further errors in measurement of a causal exposure can spill over into noncausal exposures, producing spurious associations. This phenomenon is very likely to happen in exposome studies where either the true exposures or their measurement errors are highly correlated (72). Study designs

that include multiple measurements or reference substudies with gold standard measurement tools may be essential to approach these problems. Validation through independent replication will be critical in data-driven studies.

EXPOSOME MEETS GENOME: UNCOVERING INTERACTIONS BETWEEN GENES AND EXPOSURE IN DISEASE

A portion of complex disease risk is likely due to the interaction of inherited genetic and non-inherited environmental factors (known as G × E) (30, 59). To date, success has been limited in finding specific G × E interactions in population-based studies. Large sample sizes and accurate measurements of the exposome, genome, and disease are essential, even for studying a priori hypotheses about specific exposures and specific genes. These problems are magnified in the context of agnostic scans across the entire genome for interactions with a specific exposure [genome-wide interaction studies (GWISs)] and even more so for agnostic scans across the entire exposome [gene–environment wide interaction studies (GEWISs)] (59). A systematic search of environmental exposures and genetic loci would consist of G × E tests with a corresponding multiple testing burden and a reduction in statistical power.

Nevertheless, interaction effects have been an active area of statistical research, combining advances in both design and analysis. From a design perspective, investigators have proposed various two-step approaches that combine a preliminary scan for promising interactions (e.g., assuming gene–environment independence) followed by formal testing of interactions on a limited subset of candidates using standard case-control methods (14, 21). From an analysis perspective, investigators have developed empirical Bayes compromises between case-only and case-control estimators that offer the power advantage of the former and the robustness of the latter (27, 35). Still other approaches include selecting individual exposures and genes that have strong main effects (e.g., from GWAS and EWAS, respectively) (39). The key advantage of this strategy, like the two-step approaches above, is the stepwise paring down of the number of interactions to test the increasing power of detection at the cost of missing interactions between factors that do not have strong main effects (i.e., exposures and genetic factors that are not significant in EWAS or GWAS). Furthermore, power assumptions and methods for these tests must be extended to consider the breadth of possible environmental factors investigated in an interaction study.

An emerging area of relevance to human health that integrates exposome and genome science is microbiomics. For example, David et al. (8) recently showed that the human gut microbiome responds rapidly to dietary intake. Coupling these insights with external biological, environmental, and clinical data (e.g., epigenetic profiling or gene expression) will give us a more complete picture of a disease (e.g., type 2 diabetes) and modifiable intermediary (sub)clinical phenotypes (e.g., obesity). Computational methods to integrate these disparate data types will be required to extract signal from noise.

DATA STANDARDS AND INFRASTRUCTURE REQUIREMENTS FOR HUMAN EXPOSOMIC STUDIES

The exposome presents new challenges to data integration and harmonization. With these challenges come new opportunities to leverage and apply legacy data with relevance to the exposome and/or to align or build standards so that relevant studies performed in the future are accessible and interoperable. We recommend implementing and enforcing data standards as a critical component of exposome-related research (**Table 1**, recommendation 4). In developing the data infrastructure, standards, and catalogs required for exposome research, one needs to consider the characteristics

of the data being generated; the computing infrastructure needed to curate, anonymize, share, and analyze the data; and the capabilities of existing technological infrastructure. Such infrastructure needs to anticipate multiple types of investigators and data generators. For example, data generators, including epidemiologists or toxicologists, represent an inherently interdisciplinary group of researchers who have historically worked independently or collaborated minimally and who may not have anticipated the future application of their work to the exposome concept.

Lessons from Genomics

The critical role of standards is perhaps best demonstrated by the field of genomics, which largely drove the establishment of official nomenclatures for biological entities as curated by the National Center for Biotechnology Information (NCBI), such as gene names standardized by the HUGO Gene Nomenclature Committee (HGNC). Data standards have fueled discovery via GWAS that now have sample sizes that are >100,000 individuals, leading to reproducible discovery. Large sample sizes are possible in GWASs owing to well-established genetics data standards. Investigators can pool their data with one another to assemble large data sets powered for discovery. Another example is MIAME (the Minimum Information About a Microarray Experiment), which has become the de facto standard for microarray-based gene expression studies (3). Similar data standards are needed to enable the integration of exposome measurements across cohort data sets for large-scale analyses (42).

Genome science has also illustrated the value of open-source standardized analytic frameworks such as PLINK (49). In addition, standardization in data representation and file formats [e.g., variant call format (VCF) files for sequencing data] has enabled widespread reuse of genomic data. The development of similar analytical tools and resources for exposome science would standardize and accelerate research efforts. To this end, we recommend the establishment of data analytics standards and analytics code reuse (**Table 1**, recommendation 5). Specifically, we recommend the creation of public software code repositories and the development of open-access software libraries to conduct standardized exposome-wide association studies.

As some earlier efforts have shown, standards should not become a hurdle to exposome-related researchers. For example, the >$300 million Cancer Bioinformatics Grid (caBIG), billed as an open-access data network, imposed data standards that ultimately were not adopted by the cancer genomics community (31). The caBIG one-size-fits-all policy was too onerous for its clients; it imposed inflexible standards on complex data sets to ensure compatibility with their software tools. What works best? According to Masys and colleagues (31), the best standards and resources are those that are easy to adopt and that produce a benefit, especially when they represent a complex system. Most importantly, standards must have buy-in from external clients/investigators who will use the resource. Data analytics methods and standards should be simple and should clearly benefit the clients of the data system. They must also change when uses of the data change. Therefore, at present, we recommend a bottom-up approach to implementing data standards for the exposome. In a bottom-up approach, standards are formulated iteratively using available data (as opposed to data that could exist). The advantage of this approach is that it includes quick adoption of data standards and immediate applications for addressing timely biomedical questions. However, to accommodate new data modalities and ensure relevance, these standards need to be updated by all players within the exposome research community, including data generators and consumers.

Creating and Sharing Resources to Benefit the Entire Exposome Community

Several human exposome projects have been established, including the EU-funded Human Early-Life Exposome (HELIX) (65), the EU-funded Health and Environment-wide Associations

based on Large population surveys (HEALS; **http://www.heals-eu.eu**), the EU-funded EXPOsO-MICS (**http://www.exposomicsproject.eu**), the National Institutes of Health (NIH)-funded Exposome Research Center (**http://emoryhercules.com**), and now the Children's Health Exposure Analysis Resource (CHEAR; **http://www.niehs.nih.gov/research/supported/exposure/chear/**).

We recommend that significant investment be made, leveraging these ongoing projects, to support the integration of existing, disparate platforms to build a unified human exposome measurement platform across institutions and funding bodies (**Table 1**, recommendation 6). Such an effort should include a network of core facilities that each exhibit expertise in measuring a component of the exposome, such as metabolomics. For example, such an effort would support exposome-related research that integrates data from platforms developed in the NIH Common Fund initiatives such as the Core Metabolomics. We recommend that funding bodies, scientific journals, investigators, and study participants consider methods and protocols for universal data sharing to promote reproducible research (**Table 1**, recommendation 7). Specifically, many details will need to be understood and demonstrated to reproduce findings, including the data sets used for analysis, the analysis method/software tools, and the analysis parameters. Computational exposome research promises to be a complex exercise, and standardized measurement platforms and data analytic provenance will be instrumental for the development of computational methods and for reproducible research.

We recommend that studies akin to the NHANES establish background variability, geotemporal dependence, and prevalence of exposome factors of different populations around the world (e.g., race/ethnicity groups, sex, age). Such data sets can also provide a substrate to develop data standards and promote the development of analytical methods. For example, through efforts such as the HapMap project (16), human geneticists have measured and cataloged the genetic diversity of human populations. This characterization has enabled large-scale investigations such as GWASs to understand the relationship between genes and disease. To understand the relationship between the exposome and health, we must understand how the exposome varies through space and time and in different segments of the human population. Such a resource should also document how external environment indicators, such as sensor-based measures of air pollution or individual socioeconomic differences, map to or associate with internal environmental exposure dosage (40).

As discussed earlier, a simple first step toward sharing data includes sharing summarized findings that emerge from exposome research such as EWASs and GEWISs, analogous to the National Human Genome Research Institute (NHGRI)-hosted GWAS Catalog (67) (**Table 1**, recommendation 1). By archiving experimental findings in one place, follow-up investigations can be quickly planned, new meta-analyses can be performed, and a global view of exposome-related investigations can be attained. Because only summary-level information is being shared, individual-level privacy concerns are avoided.

However, individual-level exposome data sets must also be archived and shared for their continued use and integration for data-driven discovery (e.g., 44). Progress in archiving and sharing these types of data is dependent on the continued development, adoption, and dissemination of standardized data dictionaries and ontologies that enable exposure nomenclature (analogous to gene names) and that catalog how exposures are measured (e.g., **Figure 2**). One example of such a catalog is the PhenX Toolkit (19). Such a tool kit provides standard measures for association studies, including the name of the exposure (e.g., cigarette smoke), related exposures that may be of interest, and protocols for an exposure's measurement in a study. For example, a recent EWAS of type 2 diabetes (17) successfully ascertained self-reported indicators of exposure by implementing the measurements cataloged in PhenX. Although the current focus of PhenX is to disseminate

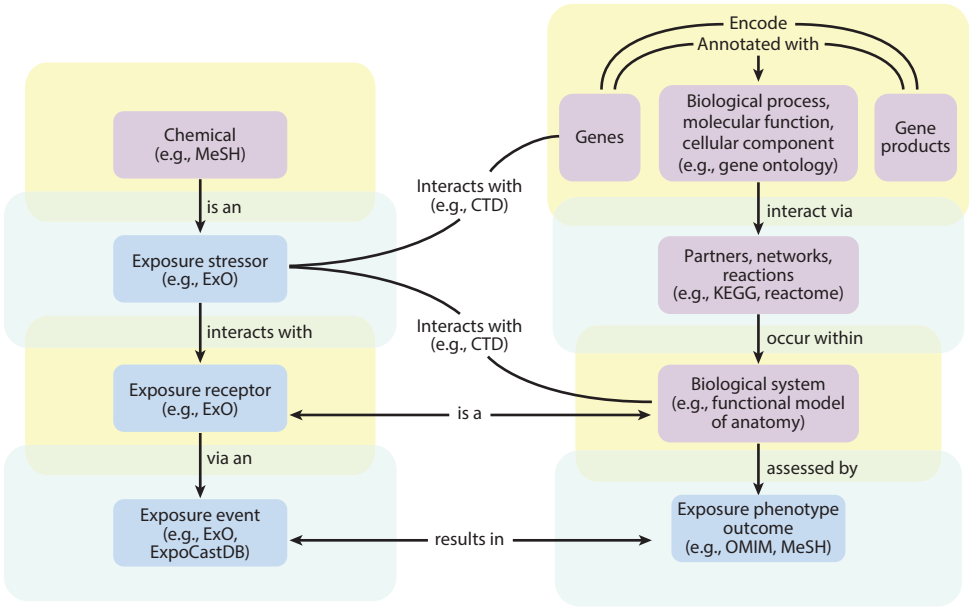

Figure 2

High-level schematic of ExO integration within a broader biological context (adapted from Reference 32). CTD, Comparative Toxicogenomics Database; ExO, exposure ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MeSH, Medical Subject Headings; OMIM, Online Mendelian Inheritance in Man.

self-reported standard instruments to assess exposure, this resource can be expanded to include quantitative markers of the exposome.

Considerations for a Universal Data Repository to Support Exposome-Related Investigation

In conceptualizing an approach and determining the feasibility and usefulness of a data repository to share individual-level exposome data, all parties, be it funding bodies, researchers, and study participants, need to consider issues common to sharing individual-level genomic information, such as ownership and sustainability (e.g., who will own and maintain repositories), granularity (e.g., which study data are stored and how to account for heterogeneity of multiple exposome measurement platforms), and compliance (e.g., how will repositories guarantee confidentiality and handle data use agreements). The NCBI hosts repositories [e.g., databases of genotypes and phenotypes (dbGaP) (29)] which have taken some of these considerations into account, such as data use compliance. However, it remains to be seen how both funding agencies and PIs will support the use and dissemination of exposome data resources.

KNOWLEDGE RESOURCES TO SUPPORT THE INTERPRETATION OF EXPOSOME-BASED FINDINGS

Information to support the interpretation of primary findings will require the continued development of ontologies and databases that are relevant to the exposome. Two established databases that will be instrumental for developing standards for nomenclature and toxicological knowledge

are the Toxic Exposome Database (T3DB; **http://www.t3db.ca**) (71) and the Comparative Toxicogenomics Database (CTD; **http://ctdbase.org**) (9). T3DB currently provides information for more than 3,600 compounds and 2,000 targets, expression data sets for more than 15,000 genes, and extensive information on chemical concentrations in biofluids and referential chemical spectra data. CTD aims to help elucidate the molecular mechanisms by which environmental exposures contribute to disease etiologies by providing manually curated information about chemical–gene–disease interactions. It is currently expanding its curated data to include comprehensive information (>50 annotation fields) from exposure publications, including demographic details, routes of exposure, exposure levels (where measured), and statistical metrics. Relevant ontologies exist in areas that include genomics (e.g., gene ontology), pathways (e.g., Kyoto Encyclopedia of Genes and Genomes), chemical (including T3DB classes), and even cigarette smoke (the Cigarette Smoke Exposure Ontology) and the exposure ontology (ExO) (32). However, the current mechanisms to access and evaluate ontologies present several challenges for the exposure community, including (*a*) assessing the degree to which existing ontologies adequately cover the full semantics of the exposome (i.e., the exposome semantic space); (*b*) determining canonical ontologies and sets of ontologies to cover the exposome semantic space; (*c*) harmonizing across ontologies; and (*d*) determining best practices for the community to maintain and update relevant ontologies, including those not exclusive to the environmental health community. One significant challenge for exposome-based research includes continuing to evolve, harmonize, and raise awareness of the different standards for describing exposome data through community-driven processes that promote consistent use and ongoing maintenance and development. Software repositories such as GitHub (**http://github.com**), development tools such as WebProtege (20), and portals such as BioPortal (37) and oboFoundry (55) are existing tools that can enable the communal and open-source development of ontologies.

BIG DATA ANALYTICS AND THE EXPOSOME: IMPLICATIONS FOR CROSS-CUTTING TRAINING

Discovery research with the human exposome is a big data analytics integration challenge that cuts across statistics, computer science, biomedicine, and public health. Thus, data-driven exposome research requires the training of a new breed of researchers who can bridge multiple fields of investigation and work in consortium/team science capacities. We recommend promoting exposome research among scientists in these different fields through educational outreach programs administered by both research funders and institutions (**Table 1**, recommendation 8). Primary tasks for outreach include identifying example and publicly available data sets (e.g., US NHANES) to develop new methods and to support research in a classroom setting. If an exposome data ecosystem is fully realized, it will likely be as impactful as genomic science has been in molding the current and new generations of biomedical researchers, computational scientists, and research programs.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was supported, in part, by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS). C.J.P. was

supported by R00 ES023504, R21 ES025052, NIH/BD2K U54 HG007963, and a PhRMA Foundation fellowship. D.C.T. was supported by R01 ES019876, P30 ES 07048, and P30 CA014089. A.K.M. was supported by NIH/BD2K U54 HG007963. C.J.M. was supported by R01 ES014065 and R01 ES019604. M.H. was supported by F31 HG008588.

LITERATURE CITED

1. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57(1):289–300
2. Billionnet C, Sherrill D, Annesi-Maesano I. 2012. Estimating the health effects of exposure to multipollutant mixture. *Ann. Epidemiol.* 22(2):126–41
3. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29(4):365–71
4. Buck Louis GM, Sundaram R. 2012. Exposome: time for transformative research. *Stat. Med.* 31(22):2569–75
5. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. *Measurement Error in Nonlinear Models*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.
6. Cortessis V, Thomas DC. 2004. Toxicokinetic genetics: an approach to gene-environment and gene-gene interactions in complex metabolic pathways. *IARC Sci. Publ.* (157):127–50
7. Danaei G, Ding EL, Mozaffarian D, Taylor B, Rehm J, et al. 2009. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLOS Med.* 6(4):e1000058
8. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, et al. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505:559–63
9. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, et al. 2015. The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic Acids Res.* 43(Database issue):D914–20
10. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26(9):1205–10
11. Di Q, Rowland S, Koutrakis P, Schwartz J. 2017. A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. *J. Air Waste Manag. Assoc.* 67:39–52
12. Dominici F, Peng RD, Barr CD, Bell ML. 2010. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology* 21(2):187–94
13. Friedman J, Hastie T, Tibshirani R. 2007. Sparse inverse covariance estimation with the lasso. *Biostatistics* 9:432–41
14. Gauderman WJ, Zhang P, Morrison JL, Lewinger JP. 2013. Finding novel genes by testing G × E interactions in a genome-wide association study. *Genet. Epidemiol.* 37(6):603–13
15. Georgopoulos PG, Lioy PJ. 2006. From a theoretical framework of human exposure and dose assessment to computational system implementation: the Modeling ENvironment for TOtal Risk Studies (MENTOR). *J. Toxicol. Environ. Health. B Crit. Rev.* 9(6):457–83
16. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, et al. 2003. The International HapMap Project. *Nature* 426:789–96
17. Hall MA, Dudek SM, Goodloe R, Crawford DC, Pendergrass SA, et al. 2014. Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank. *Pac. Symp. Biocomput.* 2014:200–11
18. Hall MA, Verma A, Brown-Gentry KD, Goodloe R, Boston J, et al. 2014. Detection of pleiotropy through a phenome-wide association study (pheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLOS Genet.* 10(12):e1004678
19. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, et al. 2011. The PhenX toolkit: Get the most from your measures. *Am. J. Epidemiol.* 174(3):253–60
20. Horridge M, Tudorache T, Nuylas C, Vendetti J, Noy NF, Musen MA. 2014. WebProtégé: a collaborative Web-based platform for editing biomedical ontologies. *Bioinformatics* 30(16):2384–85

21. Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. 2012. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet. Epidemiol.* 36(3):183–94
22. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. *Nat. Genet.* 29(3):306–9
23. Ioannidis JPA, Loy EY, Poulton R, Chia KS. 2009. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci. Transl. Med.* 1(7):7ps8
24. Jongeneelen FJ, Ten Berge WF. 2011. A generic, cross-chemical predictive PBTK model with multiple entry routes running as application in MS Excel; design of the model and comparison of predictions with experimental results. *Ann. Occup. Hyg.* 55(8):841–64
25. Kraft P, Zeggini E, Ioannidis JPA. 2009. Replication in genome-wide association studies. *Stat. Sci.* 24(4):561–73
26. Lawson AB, Banerjee S, Haining RP, Ugarte MD. 2016. *Handbook of Spatial Epidemiology*. Boca Raton, FL: Chapman & Hall/CRC
27. Li D, Conti DV. 2009. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am. J. Epidemiol.* 169(4):497–504
28. Lioy PJ, Rappaport SM. 2011. Exposure science and the exposome: an opportunity for coherence in the environmental health sciences. *Environ. Health Perspect.* 119(11):a466–67
29. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39(10):1181–86
30. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–53
31. Masys DR, Harris PA, Fearn PA, Kohane IS. 2012. Designing a public square for research computing. *Sci. Transl. Med.* 4(149):149fs32
32. Mattingly CJ, McKone TE, Callahan MA, Blake JA, Cohen Hubal EA. 2012. Providing the missing link: the exposure science ontology ExO. *Environ. Sci. Technol.* 46(6):3046–53
33. Miller GW, Jones DP. 2014. The nature of nurture: refining the definition of the exposome. *Toxicol. Sci.* 137(1):1–2
34. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. 2004. Actual causes of death in the United States, 2000. *JAMA* 291(10):1238–45
35. Mukherjee B, Ahn J, Gruber SB, Chatterjee N. 2012. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.* 175(3):177–90
36. Nagashima K, Sato Y, Noma H, Hamada C. 2013. An efficient and robust method for analyzing population pharmacokinetic data in genome-wide pharmacogenomic studies: a generalized estimating equation approach. *Stat. Med.* 32(27):4838–58
37. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, et al. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 37(Suppl. 2):W170–73
38. Patel CJ, Bhattacharya J, Butte AJ. 2010. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLOS ONE* 5(5):e10746
39. Patel CJ, Chen R, Kodama K, Ioannidis JPA, Butte AJ. 2013. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum. Genet.* 132(5):495–508
40. Patel CJ, Cullen MR, Ioannidis JPA, Butte AJ. 2012. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. *Int. J. Epidemiol.* 41(3):828–43
41. Patel CJ, Ioannidis JPA. 2014. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J. Epidemiol. Community Health* 68(11):1096–100
42. Patel CJ, Ioannidis JPA. 2014. Studying the elusive environment in large scale. *JAMA* 311(21):2173–74
43. Patel CJ, Manrai AK. 2015. Development of exposome correlation globes to map out environment-wide associations. *Pac. Symp. Biocomput.* 2015:231–42
44. Patel CJ, Pho N, McDuffie M, Easton-Marks J, Kothari C, et al. 2016. A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Sci. Data* 3:160096
45. Peyret T, Krishnan K. 2011. QSARs for PBPK modelling of environmental contaminants. *SAR QSAR Environ. Res.* 22(1–2):129–69

46. Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, et al. 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47(7):702–9
47. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8):904–9
48. Price K, Krishnan K. 2011. An integrated QSAR-PBPK modelling approach for predicting the inhalation toxicokinetics of mixtures of volatile organic chemicals in the rat. *SAR QSAR Environ. Res.* 22(1–2):107–28
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81(3):559–75
50. Rappaport SM. 2016. Genetic factors are not the major causes of chronic diseases. *PLOS ONE* 11(4):e0154387
51. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. 2014. The blood exposome and its role in discovering causes of disease. *Environ. Health Perspect.* 122(8):769–74
52. Rappaport SM, Smith MT. 2010. Environment and disease risks. *Science* 330:460–61
53. Sarigiannis D, Gotti A, Karakitsios S. 2011. A computational framework for aggregate and cumulative exposure assessment. *Epidemiology* 22:S96–97
54. Sarigiannis D, Gotti A, Reale GC, Marafante E. 2009. Reflections on new directions for risk assessment of environmental chemical mixtures. *Int. J. Risk Assess. Manag.* 13:216
55. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. 2007. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11):1251–55
56. SNPedia. 2014. Heritability. *SNPedia*, updated July 12. **http://www.snpedia.com/index.php/Heritability**
57. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102(43):15545–50
58. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, et al. 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ. Health* 12(1):85
59. Thomas D. 2010. Gene–environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* 11(4):259–72
60. Thomas DC. 1988. Models for exposure-time-response relationships with applications to cancer epidemiology. *Annu. Rev. Public Health* 9:451–82
61. Turner M, Nieuwenhuijsen M, Anderson K, Balshaw D, Cui Y, et al. 2017. Assessing the exposome with external measures: commentary on the state of the science and research recommendations. *Annu. Rev. Public Health* 38:215–39
62. Tzoulaki I, Patel CJ, Okamura T, Chan Q, Brown IJ, et al. 2012. A nutrient-wide association study on blood pressure. *Circulation* 126(21):2456–64
63. US EPA (US Environ. Prot. Agency). 2014. *Toxic Substances Control Act (TSCA) search*. Updated Sept. 22, US EPA, Washington, DC. **http://www.epa.gov/enviro/facts/tsca/tsca_search.html**
64. Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am. J. Hum. Genet.* 90:7–24
65. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, et al. 2014. The human early-life exposome (HELIX): project rationale and design. *Environ. Health Perspect.* 122(6):535–44
66. Wakefield J. 1996. The Bayesian analysis of population pharmacokinetic models. *J. Am. Stat. Assoc.* 91:62–75
67. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42(D1):D1001–6
68. Wild CP. 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev.* 14(8):1847–50
69. Wild CP. 2012. The exposome: from concept to utility. *Int. J. Epidemiol.* 41(1):24–32
70. Willmann S, Lippert J, Sevestre M, Solodenko J, Fois F, Schmitt W. 2003. PK-Sim®: a physiologically based pharmacokinetic "whole-body" model. *BIOSILICO* 1(4):121–24

71. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, et al. 2015. T3DB: the toxic exposome database. *Nucleic Acids Res*. 43(D1):D928–34
72. Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, et al. 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ. Health Perspect.* 108(5):419–26
73. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67(2):301–20