Exposome, 2024, **4(1)**, osae005
https://doi.org/10.1093/exposome/osae005
Advance Access Publication Date: 24 April 2024
Review

A data-centric perspective on exposomics data analysis

Le Chang [1,†], Jessica Ewald [2,†], Fiona Hui [2], Stéphane Bayen [3], and Jianguo Xia [1,2,*]

[1]Department of Human Genetics, McGill University, Montreal, Canada
[2]Institute of Parasitology, McGill University, Montreal, Canada
[3]Department of Food Science and Agricultural Chemistry, McGill University, Montreal, Canada

*To whom correspondence should be addressed: Email: jeff.xia@mcgill.ca

†These authors contributed equally

Abstract

Exposomics represents a systematic approach to investigate the etiology of diseases by formally integrating individuals' entire environmental exposures and associated biological responses into the traditional genotype-phenotype framework. The field is largely enabled by various omics technologies which offer practical means to comprehensively measure key components in exposomics. The bottleneck in exposomics has gradually shifted from data collection to data analysis. Effective and easy-to-use bioinformatics tools and computational workflows are urgently needed to help obtain robust associations and to derive actionable insights from the observational, heterogenous, and multi-omics datasets collected in exposomics studies. This data-centric perspective starts with an overview of the main components and common analysis workflows in exposomics. We then introduce six computational approaches that have proven effective in addressing some key analytical challenges, including linear modeling with covariate adjustment, dimensionality reduction for covariance detection, neural networks for identification of complex interactions, network visual analytics for organizing and interpreting multi-omics results, Mendelian randomization for causal inference, and cause-effect validation by coupling effect-directed analysis with dose-response assessment. Finally, we present a series of well-designed web-based tools, and briefly discuss how they can be used for exposomics data analysis.

Keywords: multi-omics integration; covariate adjustment; dimensionality reduction; neural networks; Mendelian randomization; dose-response analysis

Overview of exposomics

First envisioned by Christopher Wild in 2005, exposomics is a fast growing field centered around profiling the complete set of exposures individuals encounter across their lifespan.[1] This shift towards evaluating the entire exposure profile and their associated biological responses rather than discrete environmental factors represents a pivotal development, as the etiology of many diseases involves complex interactions between an array of genetic susceptibilities and environmental exposures.[1-6] Traditionally, genetic factors have been the primary focus of many disease studies. While genetics predispose individuals to certain phenotypic development, it's often the exposures from our environment, the lifestyles of individuals, and social factors that lead to their manifestation.[7] It has been estimated that genetics contributes to less than 50% of most complex chronic diseases such as common cancers, type 2 diabetes, heart diseases, *etc.*[7,8] By considering the totality of exposures, we have an improved likelihood at understanding, preventing, and treating various diseases, as well as improving health and quality of life.[9]

Exposomics has become feasible with recent advances in high-throughput omics technologies, remote sensors, wearable devices, etc, which together allow comprehensive collection of exposomic data at scale.[10-12] For example, global or untargeted metabolomics based on liquid chromatography—high-resolution mass spectrometry (LC-HRMS) can detect, from a single blood or urine sample, 1000 s∼10 000 s of features contributed from endogenous metabolites and xenobiotic exposures (ie, food, drugs, gut microbiome, chemical contaminants, *etc*,).[5,13] The recent development of multiplexed, multimodal real-time chemical sensors is poised to have profound impact upon the discovery of biomarkers.[14] When applied to large cohorts, researchers are facing tremendous "big data challenges" due to the volume, variety, and velocity of datasets.

As shown in Figure 1A, a typical exposomic dataset is composed of a series of data matrices that describe the major components of exposomics: environmental exposures (external and internal), individuals (phenotype and genotype), and the interactions between individuals and their environments (various functional 'omics). The environmental exposure (E) data include measurements of external exposures from food, air, water or soil, as well as internal chemical exposures such as xenobiotics and their biotransformation products from microbiome or drug metabolism detected in blood, urine or saliva.[15] The genotype (G) data describe individuals' genetic variants such as single nucleotide polymorphisms (SNPs), copy number variants, or structural rearrangements. The biological response (R) data include one or multiple omics profiles such as transcriptomics, proteomics, and metabolomics in response to environmental exposures. Finally, the phenotype (P) data, often known as sample metadata, contain various observable host traits. This could include variables like sex, age, blood pressure, body mass index (BMI), and other

Received: October 1, 2023. **Revised:** March 14, 2024. **Accepted:** March 23, 2024

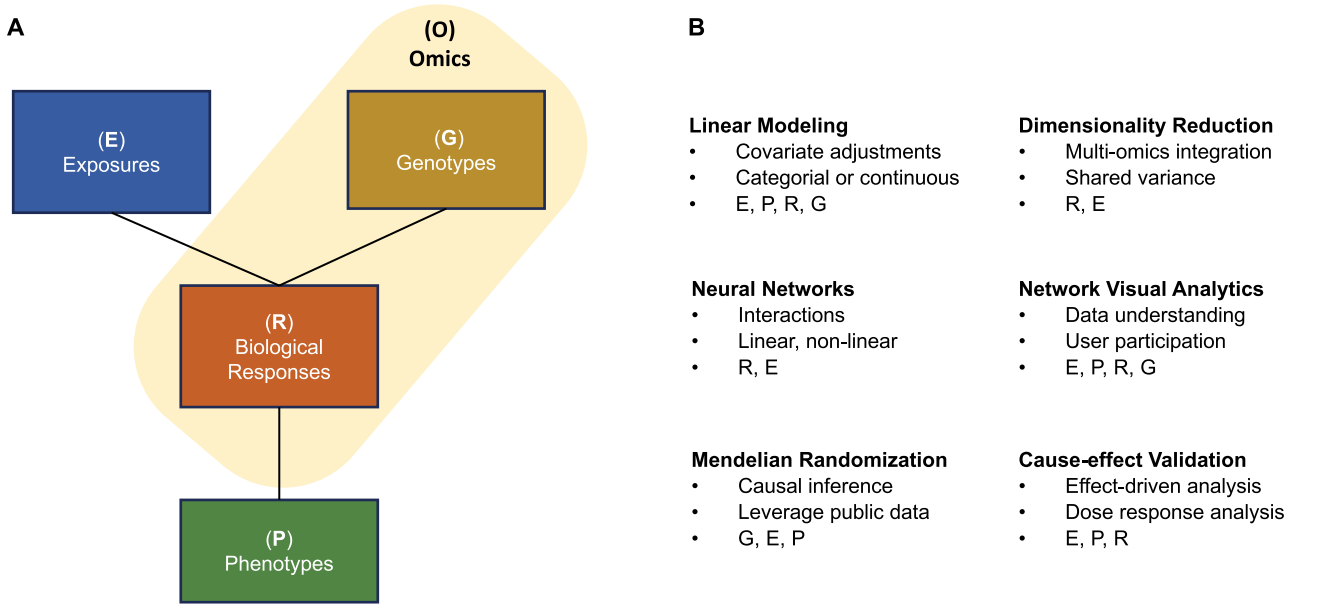

Figure 1. Overview of the main components in exposomics datasets and the selected analysis approaches. (A) Multiple data matrices characterizing environment (E), genotypes (G) and biological responses (R) captured by various omics profiles (O), as well as phenotypes (P). (B) Key features of the six selected approaches for exposomics data analysis.

physiological and clinical measurements, including those directly related to diseases. The main motivation of this perspective is to introduce several well-established as well as promising new methods, developed primarily from the biomedical fields, within the context of exposomics data analysis. To facilitate discussion, we will use omics data (O) to include both individuals' genotypes (G) and biological response (R) measured by genomics and various downstream functional genomics technologies, respectively.

Typical exposomics data analysis workflows

From the study design and data collection perspective, there are three complementary approaches that emphasize different components in exposomics. The environment-first approach starts from measuring external exposures (air, water, soil, personal care products, etc,) or at personal levels (food intake, lifestyle variables, skin and fecal microbiome, etc,) using remote sensors, chemical monitoring and sequencing technologies.[16,17] The phenotype-first approach begins with comprehensive collection of health-related data, particularly electronic health records (EHRs) and precisely quantified clinical phenotypes. Finally, the omics-first approach focuses on profiling various omics data to characterize host genotypes and associated biological responses. The adoptions of the above strategies are based on the perceived main player (E, P or O) to the research questions at hand, the established practices in the respective fields as well as the expertise of individual researchers. For instance, environmental toxicologists tend to choose environment-first approach, clinicians will start from phenotypes, while geneticists tend to adopt omics-first approach. It is important to note that the three approaches can be applied in parallel by different research teams in large-scale collaborative exposomics studies.

The environment-first approach

The environment-first approach starts from collecting and analyzing the extensive data arising from an individual's environmental exposures before examining the complex internal biological responses.[12] According to Zhang et al. the environmental exposures can be organized into three main domains—the general external domain, the specific external domain, and the internal domain.[18] The general external domain includes natural environments, socioeconomical, psychological factors, etc. The specific external domain is concerned with immediate local environment such as air, water, food, lifestyle, and occupation. The internal domain includes processes intrinsic to the body such as metabolic profiles, microbiome, inflammation status, etc, Comprehensively capturing exposure from specific external domain in a convenient and cost-effective manner is a fast-evolving research area, with many innovations and progresses made in recent years. For instance, environmental sensors provide objective, real-time measurements of factors like air and water pollutants, humidity, temperature, and allergens.[19] With improved sensor technologies, granular data on environmental conditions can now be continuously collected. Recent studies integrating both targeted and non-targeted strategies are producing ever larger sets of data on associated chemical profiles in real-life matrices such as air, food, water or personal care products.[20-22] Wearable devices like smartwatches, fitness trackers, and specialized monitors allow constant physiological and contextual data gathering about activity, sleep, ultraviolet exposure, and other metrics, providing invaluable insights on exposures at personal level.[10] Some devices can be combined with geospatial location data and health parameters, providing some direct bridge across datasets.[23] For example, silicon wristbands used for air pollutant monitoring can also be used to measure sweat metabolites.[24,25]

Food and gut microbiome are strong influencers of host development and health trajectories through extensive metabolic interactions with the host.[26] The Earth Microbiome Project[27,28] and Human Microbiome Project[29,30] have significantly improved our knowledge on the microbiome. Many initiatives are producing and organizing data on food constituents such as FooDB, the Periodic Table of Food Initiative, Database on Migrating and Extractable Food Contact Chemicals.[31-33] Comprehensive

chemical fingerprinting of foods is an active field of research. The vast diversity of natural products, macromolecules, and food processing-induced chemicals suggest that significant efforts are needed to map out all the chemical compounds in the foods. Chemical mixture exposome assessment remains an active research area. Many innovative concepts and approaches have been proposed, such as proactive use of the likelihood of co-exposure patterns depending on lifestyle and dietary habits, and effect-directed analysis to prioritize and assess the specific risk of identified mixtures.[34,35]

Overall, the environment-first approach begins with an individual's external or internal environmental exposure data to build the initial framework for unraveling the complex exposome. The current focus is on better characterizing food and microbiome, better capturing real-life chemical mixtures, individual exposure, and temporal variations in the exposome notably during critical windows of susceptibility to health hazards, such as childhood or pregnancy.[36]

The phenotype-first approach

The phenotype-first approach starts by comprehensively profiling health outcomes before linking them to individuals' environmental exposures, omics profiles and their interactions. Key sources for health outcomes data include EHRs, health surveys and various observable pathophysiological measures.[12,37] EHRs contain extensive documentation of an individual's medical history, including diagnoses, medications, medical procedures, and other clinical information compiled over time across healthcare encounters. These records provide a rich context of an individual's health trajectory over time. EHRs are mainly used in exposomics studies conducted in clinical settings. For studies focusing on environmental and public health, common phenotype data are health surveys administered by public health agencies and self-reported population-level data on health conditions, behaviors, lifestyle factors and environmental exposures collected by researchers.[38,39]

Despite the wealth of data, effectively utilizing health outcome information presents significant challenges.[40] Patient privacy is a major concern given the ethical sensitivities around personal medical data. Heterogeneity also hinders synthesis of data from diverse EHR systems and surveys. The variations in terminology, format, structure, and coding make harmonizing disparate data a challenging task. Specialized platforms have been developed to address these challenges by enabling collaborative analysis of heterogeneous health data. For instance, the Observational Health Data Sciences and Informatics (OHDSI) provides an open-source framework to standardize and integrate diverse health datasets for evidence generation.[41]

Overall, the phenotype-first approach relies on accessing and synthesizing comprehensive health data sources through proper privacy protections and data integration techniques. Leveraging the recent developments in ontologies and artificial intelligence (AI) have the potential to transform this area.[42,43] For instance, the latest large language models (LLMs) such as GPT-4 and Gemini can read and interpret both texts and images within the same context. Their potential implementations in healthcare settings, such as to process and interpret EHR, have attracted considerable attention.[44]

The omics-first approach

The omics-first approach starts with applying various omics technologies to obtain comprehensive profiles on individuals' genotypes and biological responses associated with environmental exposures. Genomics encodes an individual's genetic susceptibilities to certain exposures, environmentally induced epigenetic changes (eg, through DNA methylation) can be retained through lifetime or even trans-generationally as a proxy for exposures,[45-47] while miRNAomics, transcriptomics, proteomics, and metabolomics reveals corresponding expression changes at miRNA, mRNA, protein or metabolite levels.[48-50] Genomics are considered static and only need to be measured once. The downstream omics, often known as functional genomics, vary in response to environmental factors and operate at different time scales. They are usually sampled at multiple time points to depict the trajectory of the changes.

A key opportunity of the current multi-modality exposomics studies is integrated analysis to enable comprehensive understanding of exposure-response relationships.[51-53] For example, coupling exposomic profiling with epigenomics provides insights into how exposures may modulate epigenetic patterns underlying development and health outcome.[54] Integrating transcriptomic, proteomic and metabolomic data can reveal specific pathways modulated by exposures. At the very downstream of omics cascade, metabolomics occupies a unique place as the interface between individuals and environment exposures. Metabolomics platforms such as LC-HRMS can simultaneously measure both chemical exposures as well as metabolic responses from the same samples. In exposomics, metabolomics is often paired with microbiome profiling and food intake questionnaires to study the interactions between gut microbiome, dietary patterns, and the metabolome.[6,55-58]

In summary, omics technologies have become well established across life sciences over the past two decades. They can efficiently measure host genotypes and capture downstream biological responses, with current focus on multi-omics, single-cell omics and spatial omics. Effective integration of these high-resolution omics profiles with environmental exposure data for translational discovery remains an active research area.[59]

Effective methods for exposomics data analysis

Exposomics data analysis is primarily concerned with detecting robust associations between features in different data matrices, with follow-up work to try to identify mechanistic, causal relationships to inform interventions aimed at improving health outcomes.[8] Over the past decades, many powerful statistical methods and computational algorithms have been developed to help address different aspects in omics data analysis. In this section, we will introduce several promising methods that have proven to be effective in addressing some challenging tasks in exposomics data analysis. Their key features and applications are summarized in Figure 1B.

Using generalized linear models to deal with complex study design and covariates

Exposomics data are primarily observational, therefore the typical dataset contains many covariates that must be accounted for during statistical analysis. The field of transcriptomics has a long history of using generalized linear models in gene expression analysis, which can flexibly handle many different experimental designs.[60-62] The methods have been extended to other omics types such as metabolomics,[63] proteomics,[64] microbiome data,[65] and in theory could be used to analyze any high-dimensional matrices. In this analysis framework, linear models can be configured to include both continuous and categorical variables, and

categorical covariates can be modeled as either fixed or random effects.[63] For cases where there is only a chemical exposure matrix (E) and a sample metadata or phenotype matrix (P), a linear model can be configured to identify chemicals associated with a metadata of interest, while accounting for any number of covariates. For example, if the objective in a biomonitoring study is to identify chemicals associated with occupation type while accounting for relevant covariates, a potential model could be $chemical_i \sim occupation + age + sex + years\ of\ employment$, where the latter three terms are covariates, and p-values and coefficients from the *occupation* term are used to identify the direction and statistical significance of chemical-occupation relationships. This method is considered univariate because each feature is analyzed independently.

In scenarios where omics data (O) are collected in addition to E and P, the objective may be to identify gene-chemical relationships. In this case, either the genes or the chemicals can be treated as sample metadata, and the same modeling approach described above can be used. Here, a potential model could be $gene_j \sim chemical_i + age + sex + BMI$, which will generate results for i*j gene-chemical associations. Since a typical transcriptomics experiment generates measurements for ~20 000 genes, this approach can produce an overwhelming number of results, and therefore it is recommended to apply stringent filters (including both statistical filters such as false discovery rate, and biological filters such as fold change or tissue specificity) to both the O and E matrices and focus only on the relationships of greatest interest. Walker et al. used this approach to identify compounds associated with occupational trichloroethylene (TCE) exposure, while adjusting for age, sex, and BMI.[66] Mass features associated with TCE exposure were carefully examined to separate TCE and TCE-derived compounds from metabolites that represent biological responses. This is a good example of how untargeted metabolomics data can simultaneously survey both the external chemical compounds and internal biological pathways. In the next section, we describe multivariate approaches that may be better suited to analyzing multiple high-dimensional matrices.

Using joint dimensionality reduction to detect shared variance

Dimensionality reduction (DR) methods summarize important sources of variation within high-dimensional datasets into a few component scores that are combinations of the original input feature values. There are many different DR methods, each of which produces sample-level scores and feature-level loadings. The most used method is principal component analysis (PCA), where sets of linear combination coefficients, also called 'loadings', are computed such that the resulting scores capture the maximum variance within the dataset. Different DR methods vary in the types of input features that they accept, the preprocessing steps that they employ, and the statistical criteria that they aim to maximize.[67] Most researchers use DR methods to visualize high-dimensional datasets in low-dimensional space such as 2D or 3D scatter plots. Data points are often annotated with different metadata using colors or text labels, to see if there are patterns correspond to specific experimental factors or covariates. DR can also be used as a feature selection tool. The magnitude of the loading coefficients indicates which features contribute most to each component and highly correlated features will have similar loadings. Therefore, extracting features with the highest magnitude for the top components will retrieve a set of features that are both highly correlated with each other, and that contribute a major source of variability within the dataset. If the scores plots showed that the component is related to a specific metadata variable, then the features in the top loadings can be interpreted as having a relationship with that metadata.

The same concept can be extended to support multi-omics data integration by considering two or more input matrices through joint dimensionality reductions (JDR). In this case, the component scores capture the shared variations across multiple datasets, and the component loadings capture sets of features that have correlated levels both within and across datasets. Multiple co-inertia analysis (MCIA) and multi-omics factor analysis (MOFA) are two commonly used JDR methods without considering sample metadata.[68,69] In comparison, the data integration analysis for biomarker discovery using latent components (DIABLO) is a supervised JDR method.[70] It takes a single metadata variable as input in addition to the two or more feature tables. Components are computed to maximize association with the metadata variable in addition to correlations between features both within and across the feature tables. In general, JDR methods require strictly matching samples and relatively large sample size (at least 20 per group) to ensure the identified patterns are robust.[71] The top components from JDR are meaningful proxies of the original high-dimensional data by capturing their shared variance. They can be included in linear modeling to help identify important features underlay the observed associations with exposure or other metadata of interest. Effectively integrating JDR methods with linear modeling to reveal global patterns and the key underlying features represents a promising direction for developing computational platforms to support exploratory data analysis for exposomics.

Using neural networks to detect complex interactions

Understanding the interplays of the exposome, metabolome and gut microbiome within the context of host genetics has been an active research area.[72,73] It is of great interest to be able to identify potential interactions between exposures and biological responses directly from their corresponding matrices. For example, detecting microbe-metabolite interactions from paired microbiome and metabolomics data have attracted significant research efforts in many recent studies.[74-77] The linear modeling and multivariate DR methods we discussed above can partially address this issue, but they have significant limitations. Linear models consider each feature individually, missing the opportunity to use the shared information across multiple features to boost performance. DR methods are notoriously difficult to interpret and are especially ill suited when the main goal is to obtain individual exposure-biomarker interactions where non-linear and higher-order interactions are expected to be common.[12] Advanced machine learning methods especially neural networks have shown great promise in this direction.

The *mmvec* is probably the first neural network algorithm applied to estimate microbe-metabolite interactions by learning the embeddings of microbial and metabolite features to estimates their co-occurrence probabilities.[74] Another neural network model, MiMeNet (Microbiome-Metabolome Network), uses a multi-layer perceptron (MLP) architecture to model metabolomic profiles based on microbial compositions.[76] The input layer takes the microbial composition profiles and maps them onto the hidden layers, where the neural network learns to extract complex patterns and representations from the input data through a series of mathematical functions, the final output layer produces the prediction of metabolomic profiles. This architecture enables

the modeling of inputs and outputs of different sizes, which is the case for microbial and metabolomic features. A more recent algorithm, mNODE (metabolomic profile predictor using neural ordinary differential equations) has shown great promise in efficiently learning potential interactions from multi-omics data.[77] Similar to MiMeNet, mNODE adopts a MLP architecture, with an additional neural ODE module in between the hidden layers and employs different mathematical operations. The algorithm can incorporate dietary information to further enhance the performance in predicting metabolomic profiles from microbial compositions.

Although the "black box" characteristic of neural networks may be intimidating to some, their robustness in capturing complex patterns has been widely recognized especially in recent years. Compared to traditional modeling methods, they offer a distinct advantage in automatically learning and extracting relevant features from the data without having to necessarily define the exact number or nature of the features. With the growing acceptance of AI, the neural network-based approaches are expected to find broader applications in exposomics data analysis. In general, neural networks require a reasonably large experimental dataset, with sample size on the scale of hundreds and beyond, to train the model and tune the parameters. This usually will not be a concern for large-scale exposomics studies. When the sample size is small, linear modelling should be preferred due to its robustness and "white box" nature.

Using network visual analytics for data presentation and integration

Networks are natural representations of complex relationships among different entities. Networks can be created based on our prior knowledge (ie, knowledge-based networks), based on the correlations computed from the current datasets (ie, data-driven networks), or using a mixture of both approaches. After creation, they can then be analyzed by applying graph theory to identify important connections and modules to gain insights into the organizing principles of the systems.[78] A key appealing feature for network-based approach is its intuitive visualization, which engages and empowers researchers to identify meaningful patterns. Networks can be viewed in hierarchical structures consisting of multiple layers corresponding to different components of exposomics datasets (Figure 2). Each omics layer can be explored separately and then jointly to gain mechanistic insights. We can intuitively find out the key points of connection between different omics layers as well as their interaction partners within the same layer to help understand pathways, biological processes, as well as to identify potential targets of intervention. This type of visualization can greatly facilitate data understanding and hypothesis generation, making networks a common choice for exposomics data integration and result presentation.[79]

The knowledge-driven networks are created by projecting molecules of interest into a comprehensive knowledge network describing the known relationships (based on experimental evidence or computational prediction) among SNPs, genes, proteins, metabolites, diseases, and chemicals. The data-driven integrative networks can be created *de novo* based on the associations detected from current data using algorithms described in the previous sections. For instance, after covariate adjustments, the significant associations detected based on linear modeling or neural networks can be used to form networks between molecules, exposures and phenotypes. A key feature of the DIABLO-based multi-omics integration is its ability to generate highly interconnected networks with enriched biological themes.[70] In addition to visual exploration, different algorithms such as topology analysis, causal mediation analysis, integer linear programming can be applied to the networks to reveal key patterns, causal links, and mechanistic hypotheses.[80-82]

Using Mendelian randomization for causal inferences

Mendelian randomization (MR) has emerged as a valuable strategy in exposomics research to assess causal relationships between exposures and health outcomes by leveraging genetic data.[83-85] The core premise of MR is using genetic variants like SNPs as instrumental variables to infer the causal effect of a modifiable exposure on an outcome[86] This is possible because the random assortment of alleles during meiosis results in independent allocations of genetic variants. Therefore, associations between genetic variants and outcomes are less susceptible to confounding biases that often challenge conventional observational studies.[87]

Two-sample MR (2SMR) approach has been proven very useful in exposomics by leveraging published data in MR analysis.[88] It utilizes summary statistics from comparable, separate genome-wide association studies (GWAS) for both exposure and outcome. These GWAS are often generated from large cohort studies, with sample sizes ranging from 1000 s to 100 000 s of individuals. For instance, a recent study examined the impact of 4587 environmental exposures on longevity using data from 361,194 individuals in the UK Biobank, and identified factors such as sugar intake, body fat, and various age-related diseases as being causally linked to longevity.[84] The study effectively demonstrated the capability of 2SMR to unravel relationships between exposures and health outcomes. The reliance on publicly available GWAS summary statistics not only enhances the cost-effectiveness of the 2SMR approach but also maximizes the utility of existing datasets, allowing for a broader examination of exposure-outcome relationships in an exposome-wide fashion.[89-91]

The fundamental assumptions underpinning MR is that the genetic variants are strongly associated with the exposure, independent of confounders, and affect the outcome only through the exposure and not via alternative pathways.[90] A critical aspect of its design is the rigorous selection of appropriate instrumental variables.[92,93] Incorrect selection can lead to biased results, such as those arising from pleiotropy. Therefore, conducting sensitivity analysis is crucial to assess the robustness of the MR estimates against potential biases.[94] Moreover, careful interpretation of MR findings is necessary to avoid overstating causal conclusions based solely on observed associations.[95] To ensure the robustness and validity of findings, MR studies must control for Type I and Type II errors. Techniques such as the Bonferroni correction and False Discovery Rate (FDR) control are essential for minimizing the risk of false positives, while careful power analysis is important for reducing the likelihood of false negatives.[96]

Using effect-directed analysis and dose-response assessment to characterize cause and effect

Results from causal inference in an observational setting require further experimental validations to establish causality. For complex mixtures, effect-directed analysis (EDA) can help uncover causal relationships between adverse health effects and chemical contaminants. This approach starts with toxicity screening of complex real-life mixtures. The complexity of the mixture is then reduced through fractionation steps and characterized using non-targeted approaches to narrow down on substances of

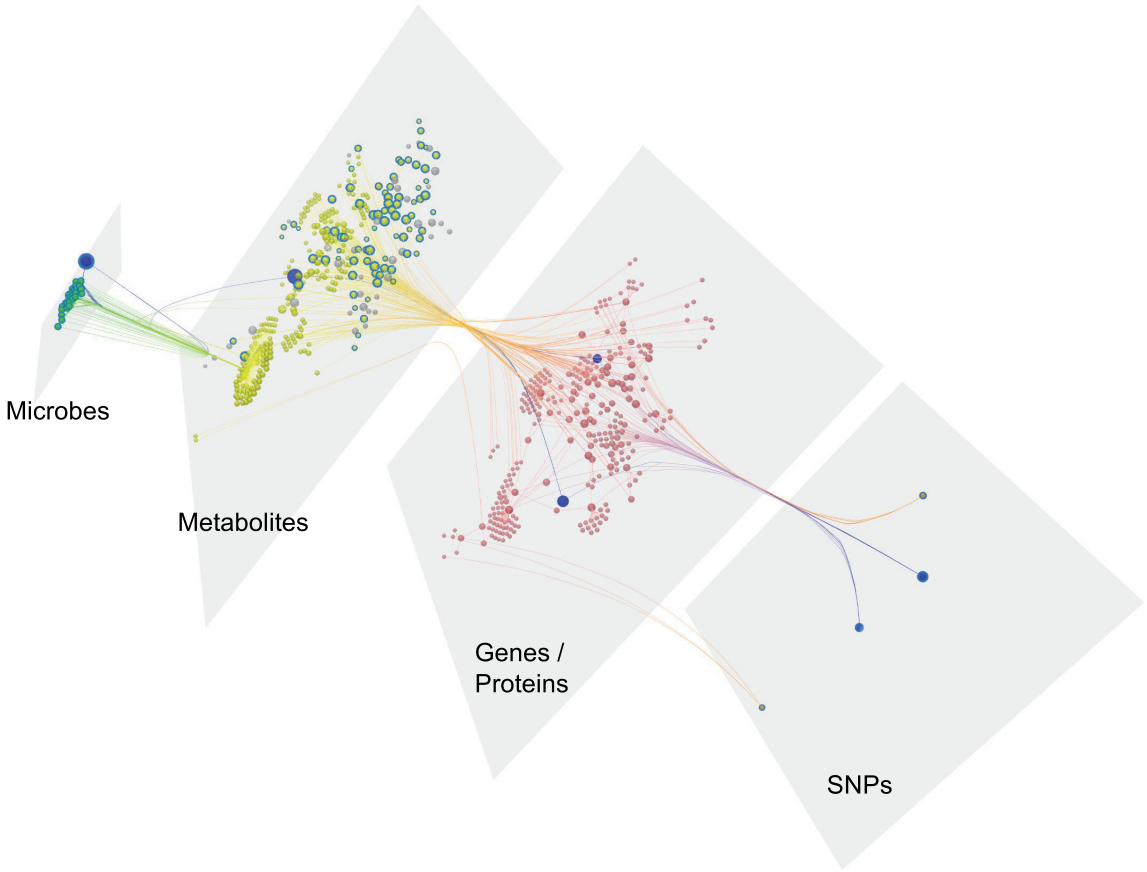

Figure 2. An illustration of network visual analytics for multi-omics integration, understanding and hypothesis generation. This 3D network was generated by integrating results from host genetics, transcriptomics, metabolomics, and microbiome using the OmicsNet platform. Each layer was created using the significant features identified from a particular omics type and their known interaction partners. Cross-omics links are established based on known relationships between SNPs, genes, proteins, metabolites, and microbes.

concern.[35] Finally, dose-response studies are applied to the shortlist of candidate compounds to pinpoint the compounds that are cause bioactivity upon exposure.[97,98]

Dose-response relationship is a foundational concept in toxicology and chemical risk assessment. Recent years have seen the growing applications of dose-response analysis to omics data, particularly transcriptomics data.[99] In this approach, exposures are conducted at multiple concentrations of a chemical, typically include a control group and at least three different dose groups with the same number of replicates in each group. The sample sizes vary from a few dozens to hundreds depending on the design, cost, and effect sizes. After data collection, a suite of linear and non-linear curve models are fitted to the levels of each omics feature. Each curve is analyzed to compute a feature-level benchmark dose (BMD), which is the minimum concentration of a substance that produces a clear, low level health risk relative to the control group.[100] The collection of benchmark doses can be summarized at the pathway level or across the entire experiment (commonly called the transcriptomic point-of-departure, or tPOD). When *in vitro* assays are coupled with high-throughput omics platforms, it becomes feasible for a typical research lab to use transcriptomics dose-response analysis to assess 10∼100 s of chemicals.[97,101,102] The omics-based dose-response study not only allows researchers to establish causal links between specific chemical exposures and general bioactivity (ie, which chemicals in the exposome 'matter'), but also to disentangle specific biological processes that are perturbed by different chemicals.

Other approaches

The previous sections have introduced a handful of methods chosen based on their perceived utilities in exposomics data analysis. Beyond these, many other advanced statistical methods are also used for similar purposes, especially in the field of epidemiology.[12] General machine learning approaches such as random forest, support vector machines, and gradient boosting can be generally applied for biomarker selection and predictive modeling. Bayesian additive regression trees is a more recent method that has the added benefit of quantifying the uncertainty of predictions.[103] Other methods have been developed for assessing causal effects of complex mixture exposures on clinical outcomes and phenotypes, a blend of the objectives that we presented for 2SMR and EDA. Some examples include G-computation associated with super learners,[104] grouped weighted quantile sum (GWQS) regression,[105] Bayesian kernel machine regression,[106] etc,

Bioinformatics tools for exposomics data analysis

A major bottleneck of the current exposomics data analysis is the requirement of advanced knowledge in statistics and programming. There have been tremendous progresses over the past two decades in the development of bioinformatics platforms for processing and analyzing various types of omics data. Despite the inherent differences in the technological platforms that

generate such data, different types of omics data share some core properties that can be dealt with through a coherent conceptual workflow and user interface. Omics data analysis workflow can be organized into four general stages. The 1st stage is raw data processing based on algorithms that are usually specific to the underlying technology (ie, next-generation sequencing versus mass spectrometry). This stage produces high-dimensional data matrices containing the abundance values of different omics features. In the 2nd stage, the data matrices are subject to statistics and machine learning algorithms for comparative, clustering or classification analysis. In the 3rd stage, different functional analysis methods are employed to shed light on the potential pathways or biological processes associated with the identified molecular patterns. The 4th stage involves analyses that are unique to specific domains, such as biomarker analysis for clinical studies[107] or dose-response analysis for toxicology studies.[108]

Analyzing diverse omics data requires specialized platforms tailored to each data type and the unique needs of target users. Based on the general workflow described above, we have developed a series of web-based platforms. For instance, MetaboAnalyst is dedicated for streamlined metabolomic data processing, statistics, visualization and functional analyses[109]; MicrobiomeAnalyst provides a platform for microbiome data processing, community profiling, and functional inference[110]; while ExpressAnalyst enables comprehensive transcriptomics data analysis for both model and non-model species.[111-113] To support integrative analysis of multi-omics data, we have developed multiple tools through knowledge-based networks, including miRNet for miRNA-centric data integration,[114-116] NetworkAnalyst for gene/protein-centric integration,[117-119] and OmicsNet as a general purpose platform for multi-omics integration and network visualization.[120-122] For data-driven multi-omics integrations, we have implemented OmicsAnalyst that offers several well-established statistical and machine learning methods integrated with advanced visual analytics capacities.[123] Collectively, these platforms offer a coherent, user-friendly web interface to help transform complex, heterogeneous omics data into meaningful patterns and biological insights.

Special efforts have been made to support exposomics data analysis in recent releases. For instance, to accommodate complex study designs, these tools now allow users to upload the omics data together with their associated metadata tables and apply different statistics or machine learning methods that can take into account of multiple factors or covariates. Users can also perform transcriptomics and metabolomics dose-response analysis using ExpressAnalyst or MetaboAnalyst, respectively. MetaboAnalyst also supports causal analysis based on 2SMR by leveraging the large collection of mGWAS data.[91,124] Finally, users can upload multiple data matrices to OmicsAnalyst and perform data-driven integration using JDR methods, correlation analysis or multi-view clustering algorithms.[71]

Conclusion and future perspectives

In this perspective, we first provided an overview of the main components of the datasets commonly seen in exposomics studies. We then discussed three workflows in dealing with these datasets, which are primarily driven by the research questions and the established practices. After readers become familiar with exposomics data structures and analysis objectives, we set out to introduce six approaches, including linear models for dealing with covariates, joint dimensionality reductions for detecting shared variance, neural networks for identification of robust interactions, networks visual analytics for understanding complex relationships, Mendelian randomization for causal inferences, as well as integrating effect-directed analysis and dose-response study to identify and characterize key toxicants within chemical mixtures. These methods are originated from statistics, genetics, computing science, epidemiology, environmental toxicology, respectively. They are routinely used by individual research groups within their respective areas but are little known to researchers outside.

There are several limitations in this perspective. For instance, power and sample size calculations are important considerations before conducting exposomics studies. Missing value estimation, batch effect adjustment, and data normalization are practical issues facing researchers during data analysis. These topics are not discussed here. Instead, we refer readers to the best practices established in their respective omics fields. MetaboAnalyst, MicrobiomeAnalyst and ExpressAnalyst are equipped with comprehensive collections of distinct, yet partially overlapping methods in their respective omics data analysis workflows to deal with these issues.

Comprehensive exposomics data collections are limited to a few large-scale studies, and most current exposomics studies collect a single omics data and measure a few phenotypes. It is of great interest to leverage the datasets from large cohorts to facilitate discovery of robust patterns and enable mechanistic understanding. The power of such approach has already been illustrated by 2SMR, which leverages public genetic data to establish causal links between exposures and outcomes.[90] Given that environmental data and health outcomes might be collected in different cohorts or at different times, one can use GWAS results that link genetic variants to exposure levels such as smoking, air pollution exposure markers, or dietary habits, from one study and then correlate to health outcomes like heart diseases, respiratory conditions, or metabolic disorders using results from another study.

The recent breakthroughs in generative AI technologies have provided exciting opportunities. With the huge sample sizes and extensive data pools now available in the modern biobanks, such as the UK Biobank containing the imaging, genotypes, lifestyles, and EHR,[125] it has become an active research field to build biomedical foundation models tailored to different application domains.[126] These models will have tremendous potential to transform exposomics data analysis and interpretation to inform decision making in the years to come.

Funding

Canadian Foundation for Innovation (CFI), Genome Canada, and Natural Sciences and Engineering Research Council of Canada (NSERC).

Author contributions

Le Chang (Conceptualization [equal], Investigation [equal], Methodology [equal], Writing—original draft [lead]), Jessica Ewald (Conceptualization [equal], Investigation [equal], Methodology [equal], Writing—original draft [equal]), Fiona Hui (Methodology [supporting], Writing—original draft [supporting]), Stephane Bayen (Conceptualization [supporting], Investigation [equal], Methodology [equal], Writing—original draft [supporting]), and Jianguo Xia (Conceptualization [lead], Funding

acquisition [lead], Project administration [lead], Supervision [lead], Writing—original draft [equal], Writing—review & editing [lead])

Data availability

No new data were generated or analyzed in support of this research.

Conflict of interest statement

J.X. is the founder of XiaLab Analytics.

References

1. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005;14(8):1847-1850.
2. Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. Thorax 2014;69(9):876-878.
3. Miller GW. (2013) The Exposome: A Primer. Elsevier.
4. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. Toxicol Sci. 2014;137(1):1-2.
5. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: Where chemistry meets biology. Science. 2020;367(6476):392-396.
6. Manrai AK, Cui Y, Bushel PR, *et al.* Informatics and data analytics to support exposome-based discovery for public health. Annu Rev Public Health. 2017;38:279-294.
7. Rappaport SM. Genetic factors are not the major causes of chronic diseases. PLoS One. 2016;11(4):e0154387.
8. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The exposome: molecules to populations. Annu Rev Pharmacol Toxicol. 2019;59:107-127.
9. Buck Louis GM, Sundaram R. Exposome: time for transformative research. Stat Med. 2012;31(22):2569-2575.
10. Jiang C, Wang X, Li X, *et al.* Dynamic human environmental exposome revealed by longitudinal personal monitoring. Cell 2018;175(1):277-291.e231.
11. Gao P. The exposome in the era of one health. Environ Sci Technol. 2021;55(5):2790-2799.
12. Maitre L, Guimbaud JB, Warembourg C; Exposome Data Challenge Participant Consortium, *et al.* State-of-the-art methods for exposure-health studies: results from the exposome data challenge event. Environ Int. 2022;168:107422.
13. Flasch M, Fitz V, Rampler E, Ezekiel CN, Koellensperger G, Warth B. Integrated exposomics/metabolomics for rapid exposure and effect analyses. JACS Au. 2022;2(11):2548-2560.
14. Sempionatto JR, Lasalde-Ramirez JA, Mahato K, Wang J, Gao W. Wearable chemical sensors for biomarker discovery in the omics era. Nat Rev Chem. 2022;6(12):899-915.
15. 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(8):769-774.
16. Ayeni KI, Berry D, Wisgrill L, Warth B, Ezekiel CN. Early-life chemical exposome and gut microbiome development: African research perspectives within a global environmental health context. Trends Microbiol. 2022;30(11):1084-1100.
17. Hyytiäinen H, Kirjavainen PV, Täubel M, *et al.* Microbial diversity in homes and the risk of allergic rhinitis and inhalant atopy in two European birth cohorts. Environ Res. 2021;196:110835.
18. Zhang X, Gao P, Snyder MP. The exposome in the era of the quantified self. Annu Rev Biomed Data Sci. 2021;4:255-277.
19. Doherty BT, Koelmel JP, Lin EZ, Romano ME, Godri Pollitt KJ. Use of exposomic methods incorporating sensors in environmental epidemiology. Curr Environ Health Rep. 2021;8(1):34-41.
20. Huhn S, Escher BI, Krauss M, Scholz S, Hackermüller J, Altenburger R. Unravelling the chemical exposome in cohort studies: routes explored and steps to become comprehensive. Environ Sci Eur. 2021;33(1):17.
21. Simonnet-Laprade C, Bayen S, McGoldrick D, *et al.* Evidence of complementarity between targeted and non-targeted analysis based on liquid and gas-phase chromatography coupled to mass spectrometry for screening halogenated persistent organic pollutants in environmental matrices. Chemosphere 2022;293:133615.
22. Simonnet-Laprade C, Bayen S, Le Bizec B, Dervilly G. Data analysis strategies for the characterization of chemical contaminant mixtures. Fish as a case study. Environ Int. 2021;155:106610.
23. Cui Y, Eccles KM, Kwok RK, Joubert BR, Messier KP, Balshaw DM. Integrating multiscale geospatial environmental data into large population health studies: challenges and opportunities. Toxics 2022;10(7):403.
24. Mofidfar M, Song X, Kelly JT, Rubenstein MH, Zare RN. Silicone wristband spray ionization mass spectrometry for combined exposome and metabolome profiling. Isr J Chem. 2023;63(7-8):e202200116.
25. Rohlman D, Dixon HM, Kincl L, *et al.* Development of an environmental health tool linking chemical exposures, physical location and lung function. BMC Public Health. 2019;19(1):854.
26. Perler BK, Friedman ES, Wu GD. The role of the gut microbiota in the relationship between diet and human health. Annu Rev Physiol. 2023;85:449-468.
27. Gilbert JA, Jansson JK, Knight R. The Earth microbiome project: successes and aspirations. BMC Biol. 2014;12:69.
28. Thompson LR, Sanders JG, McDonald D; Earth Microbiome Project Consortium, *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. Nature 2017;551(7681):457-463.
29. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207-214.
30. Proctor LM, Network IHiR. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe 2014;16(3):276-289.
31. Mohammed Taha H, Aalizadeh R, Alygizakis N, *et al.* The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. Environ Sci Eur. 2022;34(1):104.
32. Tolosa J, Serrano Candelas E, Valles Pardo JL, *et al.* MicotoXilico: An Interactive Database to Predict Mutagenicity, Genotoxicity, and Carcinogenicity of Mycotoxins. Toxins (Basel) 2023;15(6)
33. Geueke B, Groh KJ, Maffini MV, *et al.* Systematic evidence on migrating and extractable food contact chemicals: Most chemicals detected in food contact materials are not listed for use. Crit Rev Food Sci Nutr. 2023;63(28):9425-9435.

34. Tralau T, Oelgeschläger M, Kugler J, *et al.* A prospective whole-mixture approach to assess risk of the food and chemical exposome. Nat Food. 2021;2(7):463-468.
35. Tian Z, McMinn MH, Fang M. Effect-directed analysis and beyond: how to find causal environmental toxicants. Exposome. 2023;3(1):osad002.
36. Buck Louis GM, Smarr MM, Patel CJ. The exposome research paradigm: an opportunity to understand the environmental basis for human health and disease. Curr Environ Health Rep. 2017;4(1):89-98.
37. Radezova Trifunovska M, Jolevski I, Ristevski B, Savoska S. (2021) Environmental data as exposome and opportunity of combining with cloud-based personal health records. In: *The 14-th conference on Information Systems and Grid Technologies, May 28-29, Sofia, Bulgaria*.
38. Ferrante G, Fasola S, Cilluffo G, Piacentini G, Viegi G, La Grutta S. Addressing exposome: an innovative approach to environmental determinants in pediatric respiratory health. Front Public Health. 2022;10:871140.
39. Andrianou XD, Pronk A, Galea KS, *et al.* Exposome-based public health interventions for infectious diseases in urban settings. Environ Int. 2021;146:106246.
40. Honeyford K, Expert P, Mendelsohn EE, *et al.* Challenges and recommendations for high quality research using electronic health records. Front Digit Health. 2022;4:940330.
41. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015; 216:574-578.
42. Fries JA, Steinberg E, Khattar S, *et al.* Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun. 2021;12(1):2017.
43. Xu J, Mazwi M, Johnson AEW. AnnoDash, a clinical terminology annotation dashboard. JAMIA Open. 2023;6(3):ooad046.
44. Yang X, Chen A, PourNejatian N, *et al.* A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.
45. Colwell ML, Townsel C, Petroff RL, Goodrich JM, Dolinoy DC. Epigenetics and the Exposome: DNA methylation as a proxy for health impacts of prenatal environmental exposures. Exposome 2023;3(1):osad001.
46. Cadiou S, Bustamante M, Agier L, *et al.* Using methylome data to inform exposome-health association studies: an application to the identification of environmental drivers of child body mass index. Environ Int. 2020;138:105622.
47. Siklenka K, Erkek S, Godmann M, *et al.* Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. Science. 2015;350(6261):aab2006.
48. Sarigiannis D. Transcriptomics within the exposome paradigm. In: Dagnino, S., Macherone, A. (eds) Unraveling the Exposome: A Practical View 2019:183-214.
49. Vrijens K, Bollati V, Nawrot TS. MicroRNAs as potential signatures of environmental exposure or effect: a systematic review. Environ Health Perspect. 2015;123(5):399-411.
50. Walker DI, Valvi D, Rothman N, Lan Q, Miller GW, Jones DP. The metabolome: a key measure for exposome research in epidemiology. Curr Epidemiol Rep. 2019;6(2):93-103.
51. Price EJ, Vitale CM, Miller GW, *et al.* Merging the exposome into an integrated framework for "omics" sciences. iScience 2022; 25(3):103976.
52. Miller GW. Integrating the exposome into a multi-omic research framework. Exposome 2021;1(1):osab002.
53. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017;8:84.
54. Carreras-Gallo N, Cáceres A, Balagué-Dobón L, *et al.* The early-life exposome modulates the effect of polymorphic inversions on DNA methylation. Commun Biol. 2022;5(1):455.
55. Bagheri M, Shah RD, Mosley JD, Ferguson JF. A metabolome and microbiome wide association study of healthy eating index points to the mechanisms linking dietary pattern and metabolic status. Eur J Nutr. 2021;60(8):4413-4427.
56. Tang ZZ, Chen GH, Hong QL, *et al.* Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. Front Genet. 2019;10:454.
57. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
58. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85-97.
59. Gao P, Shen X, Zhang X, *et al.* Precision environmental health monitoring by longitudinal exposome and multi-omics profiling. Genome Res. 2022;32(6):1199-1214.
60. Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
61. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140.
62. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
63. Pang Z, Zhou G, Ewald J, *et al.* Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. Nat Protoc. 2022; 17(8):1735-1761.
64. van Ooijen MP, Jong VL, Eijkemans MJC, *et al.* Identification of differentially expressed peptides in high-throughput proteomics data. Brief Bioinform. 2018;19(5):971-981.
65. Mallick H, Rahnavard A, McIver LJ, *et al.* Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol. 2021;17(11):e1009442.
66. Walker DI, Uppal K, Zhang L, *et al.* High-resolution metabolomics of occupational exposure to trichloroethylene. Int J Epidemiol. 2016;45(5):1517-1527.
67. Cantini L, Zakeri P, Hernandez C, *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021;12(1):124.
68. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014;15:162.
69. Argelaguet R, Velten B, Arnol D, *et al.* Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124.
70. Singh A, Shannon CP, Gautier B, *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055-3062.
71. Ewald JD, Zhou G, Lu Y, *et al.* Web-based multi-omics integration using the analyst software suite. Nat Protoc. 2024.
72. Neveu V, Nicolas G, Amara A, Salek RM, Scalbert A. The human microbial exposome: expanding the exposome-explorer database with gut microbial metabolites. Sci Rep. 2023;13 (1):1946.

73. Tu P, Chi L, Bodnar W, *et al.* Gut microbiome toxicity: connecting the environment and gut microbiome-associated diseases. Toxics 2020;8(1):19.
74. Morton JT, Aksenov AA, Nothias LF, *et al.* Learning representations of microbe-metabolite interactions. Nat Methods. 2019; 16(12):1306-1314.
75. Mallick H, Franzosa EA, McLver LJ, *et al.* Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun. 2019;10(1):3136.
76. Reiman D, Layden BT, Dai Y. MiMeNet: Exploring microbiome-metabolome relationships using neural networks. PLoS Comput Biol. 2021;17(5):e1009021.
77. Wang T, Wang X-W, Lee-Sarwar KA, *et al.* Predicting metabolomic profiles from microbial composition through neural ordinary differential equations. Nat Mach Intell. 2023;5(3):284-293.
78. Zhou G, Li S, Xia J. Network-based approaches for multi-omics integration. Methods Mol Biol 2020;2104:469-487.
79. Maitre L, Bustamante M, Hernández-Ferrer C, *et al.* Multi-omics signatures of the human early life exposome. Nat Commun. 2022;13(1):7024.
80. Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. NPJ Syst Biol Appl. 2019;5:15.
81. Dugourd A, Kuppe C, Sciacovelli M, *et al.* Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol Syst Biol 2021;17:e9730.
82. Huang YT, Yang HI. Causal mediation analysis of survival outcome with multiple mediators. Epidemiology. 2017;28 (3):370-378.
83. Vineis P, Robinson O, Chadeau-Hyam M, Dehghan A, Mudway I, Dagnino S. What is new in the exposome? Environ Int. 2020; 143:105887.
84. Huang S-Y, Yang Y-X, Chen S-D, *et al.* Investigating causal relationships between exposome and human longevity: a Mendelian randomization analysis. BMC Med. 2021;19(1):150.
85. Li H-Q, Feng Y-W, Yang Y-X, *et al.* Causal relations between exposome and stroke: a mendelian randomization study. J Stroke. 2022;24(2):236-244.
86. Smith GD, Ebrahim S. Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22.
87. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89-98.
88. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG; Consortium, E.-I. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol. 2015;30(7):543-552.
89. Elsworth B, Lyon M, Alexander T, *et al.* The MRC IEU OpenGWAS data infrastructure. Preprint. 2020. bioRxiv, 2020.2008.2010.244293.
90. Hemani G, Zheng J, Elsworth B, *et al.* The MR-base platform supports systematic causal inference across the human phenome. eLife 2018;7:e34408.
91. Chang L, Zhou G, Xia J. mGWAS-Explorer 2.0: Causal analysis and interpretation of metabolite-phenotype associations. Metabolites 2023;13(7):826.
92. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res. 2017;26(5):2333-2355.
93. Swerdlow DI, Kuchenbaecker KB, Shah S, *et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. Int J Epidemiol. 2016;45(5):1600-1616.
94. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol. 2017;32(5):377-389.
95. Lor GCY, Risch HA, Fung WT, *et al.* Reporting and guidelines for mendelian randomization analysis: A systematic review of oncological studies. Cancer Epidemiol. 2019;62:101577.
96. Brion MJ, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. Int J Epidemiol. 2013;42(5):1497-1501.
97. Rowan-Carroll A, Reardon A, Leingartner K, *et al.* High-throughput transcriptomic analysis of human primary hepatocyte spheroids exposed to per- and polyfluoroalkyl substances as a platform for relative potency characterization. Toxicol Sci. 2021;181(2):199-214.
98. Nyffeler J, Willis C, Lougee R, Richard A, Paul-Friedman K, Harrill JA. Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling. Toxicol Appl Pharm 2020;389:114876.
99. Phillips JR, Svoboda DL, Tandon A, *et al.* BMDExpress 2: enhanced transcriptomic dose-response analysis workflow. Bioinformatics. 2019;35(10):1780-1782.
100. Farmahin R, Williams A, Kuo B, *et al.* Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. Arch Toxicol. 2017;91(5):2045-2065.
101. Harrill J, Shah I, Setzer RW, *et al.* Considerations for Strategic Use of High-Throughput Transcriptomics Chemical Screening Data in Regulatory Decisions. Curr Opin Toxicol. 2019; 15:64-75.
102. Basu N, Crump D, Head J, *et al.* EcoToxChip: a next-generation toxicogenomics tool for chemical prioritization and environmental management. Environ Toxicol Chem. 2019;38 (2):279-288.
103. Zhang T, Geng G, Liu Y, Chang HH. Application of bayesian additive regression trees for estimating daily concentrations of PM(2.5) components. Atmosphere (Basel) 2020;11(11):1233.
104. Le Borgne F, Chatton A, Leger M, Lenain R, Foucher Y. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. Sci Rep. 2021;11(1):1435.
105. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Assessment of grouped weighted quantile sum regression for modeling chemical mixtures and cancer risk. Int J Environ Res Public Health 2021;18(2):504.
106. Devick KL, Bobb JF, Mazumdar M, *et al.* Bayesian kernel machine regression-causal mediation analysis. Stat Med. 2022;41 (5):860-876.
107. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013;9(2):280-299.
108. Ewald J, Soufan O, Xia J, Basu N. FastBMD: an online tool for rapid benchmark dose-response analysis of transcriptomics data. Bioinformatics. 2021;37(7):1035-1036.
109. Pang Z, Chong J, Zhou G, *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Res. 2021;49(W1):W388-w396.
110. Lu Y, Zhou G, Ewald J, Pang Z, Shiri T, Xia J. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. Nucleic Acids Res. 2023;51 (W1):W310-w318.
111. Liu P, Ewald J, Pang Z, *et al.* ExpressAnalyst: a unified platform for RNA-sequencing analysis in non-model species. Nat Commun. 2023;14(1):2995.

112. Ewald J, Zhou G, Lu Y, Xia J. Using expressanalyst for comprehensive gene expression analysis in model and non-model organisms. Curr Protoc 2023;3(11):e922.
113. Liu P, Ewald J, Galvez JH, *et al.* Ultrafast functional profiling of RNA-seq data for nonmodel organisms. Genome Res. 2021;31 (4):713-720.
114. Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res. 2020;48(W1):W244-w251.
115. Chang L, Xia J. MicroRNA Regulatory Network Analysis Using miRNet 2.0. Methods Mol Biol. 2023;2594:185-204.
116. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia J. miRNet—dissecting miRNA-target interactions and functional associations through network-based visual analysis. Nucleic Acids Res. 2016;44(W1):W135-141.
117. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 2019;47(W1):W234-W241.
118. Xia J, Benner MJ, Hancock RE. NetworkAnalyst—integrative approaches for protein-protein interaction network analysis and visual exploration. Nucleic Acids Res. 2014;42(Web Server issue):W167-174.
119. Xia J, Gill EE, Hancock RE. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc. 2015;10(6):823-844.
120. Zhou G, Pang Z, Lu Y, Ewald J, Xia J. OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics. Nucleic Acids Res. 2022;50(W1):W527-W533.
121. Zhou G, Xia J. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Res. 2018;46(W1):W514-W522.
122. Zhou G, Xia J. Using OmicsNet for network integration and 3D visualization. Curr Protoc Bioinformatics 2019;65(1):e69.
123. Zhou G, Ewald J, Xia J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Res. 2021;49(W1):W476-w482.
124. Chang L, Zhou G, Ou H, Xia J. mGWAS-explorer: linking SNPs, genes, metabolites, and diseases for functional insights. Metabolites 2022;12(6):526.
125. Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. Nature 2018; 562(7726):203-209.
126. Moor M, Banerjee O, Abad ZSH, *et al.* Foundation models for generalist medical artificial intelligence. Nature 2023;616 (7956):259-265.

Exposome, 2024, 4, 1–11
https://doi.org/10.1093/exposome/osae005
Review