A Section 508–conformant HTML version of this article
is available at https://doi.org/10.1289/EHP9098.

Strengthening Causal Inference in Exposomics Research: Application of Genetic Data and Methods

Christy L. Avery,[1,5] *Annie Green Howard,*[2,5] *Anna F. Ballou,*[1] *Victoria L. Buchanan,*[1] *Jason M. Collins,*[1] *Carolina G. Downie,*[1] *Stephanie M. Engel,*[1] *Mariaelisa Graff,*[1] *Heather M. Highland,*[1] *Moa P. Lee,*[1] *Adam G. Lilly,*[5,6] *Kun Lu,*[4] *Julia E. Rager,*[4] *Brooke S. Staley,*[1] *Kari E. North,*[1] *and Penny Gordon-Larsen*[3,5]

[1]Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[2]Department of Biostatistics, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[3]Department of Nutrition, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[4]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[5]Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[6]Department of Sociology, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Summary: Advances in technologies to measure a broad set of exposures have led to a range of exposome research efforts. Yet, these efforts have insufficiently integrated methods that incorporate genetic data to strengthen causal inference, despite evidence that many exposome-associated phenotypes are heritable.

Objective: We demonstrate how integration of methods and study designs that incorporate genetic data can strengthen causal inference in exposomics research by helping address six challenges: reverse causation and unmeasured confounding, comprehensive examination of phenotypic effects, low efficiency, replication, multilevel data integration, and characterization of tissue-specific effects. Examples are drawn from studies of biomarkers and health behaviors, exposure domains where the causal inference methods we describe are most often applied.

Discussion: Technological, computational, and statistical advances in genotyping, imputation, and analysis, combined with broad data sharing and cross-study collaborations, offer multiple opportunities to strengthen causal inference in exposomics research. Full application of these opportunities will require an expanded understanding of genetic variants that predict exposome phenotypes as well as an appreciation that the utility of genetic variants for causal inference will vary by exposure and may depend on large sample sizes. However, several of these challenges can be addressed through international scientific collaborations that prioritize data sharing. Ultimately, we anticipate that efforts to better integrate methods that incorporate genetic data will extend the reach of exposomics research by helping address the challenges of comprehensively measuring the exposome and its health effects across studies, the life course, and in varied contexts and diverse populations. https://doi.org/10.1289/EHP9098

Introduction

The past decade has witnessed a paradigm shift in the environmental sciences as studies have shifted from examining specific exposures to attempting to comprehensively measure and characterize the broad range of exposures an individual may encounter over the life course. Termed "exposomics," this emerging approach aims to better understand the development and progression of disease by comprehensively measuring exogenous and endogenous environmental exposures (the "exposome") at multiple levels and time periods.[1] The exposome is understandably complex and includes domains that span the chemical environment (e.g., nutrients, medications, and toxicants), health behaviors (e.g., cigarette smoking, sleeping, and physical activity), the social environment (e.g., neighborhood characteristics, social networks, and racism), and the natural and built environment (e.g., air and water pollution, green space)[1–5] (Figure 1). These domains include individual and aggregate exposures (e.g., diet and the built environment) that may be measured directly (e.g., secondhand smoking) or as a biomarker of exposure or effect (e.g., cotinine). The exposome's broad scope and complex correlation structure have elicited comparisons with the Human Genome Project.[3,6,7] However, the dynamic and high-dimensional nature of the exposome makes measurement, characterization, and causal inference—the discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data[8,9]—far more complex.[1]

Emerging statistical methods that integrate genetic data offer several avenues to help address measurement, characterization, and causal inference challenges in exposomics research. These approaches are enabled by the time invariance of germline genetic variants and a growing appreciation that many exposures are heritable, making genetic data a central component of the exposome (Figure 1). For example, although arsenic exposure in itself is not heritable, prior studies demonstrated that biomarkers of arsenic metabolism efficiency, which modulates the absolute and relative amounts of disease associated arsenic metabolites, were highly heritable (h^2 range: 50%–59%).[10] The consequent identification of genetic variants associated with arsenic metabolism efficiency biomarkers has enabled causal inference studies examining the role of arsenic in skin lesion risk[11] and cardiometabolic diseases.[12]

The goal of this commentary is to demonstrate how the application of causal inference methods that integrate genetic data can empower and enrich exposomics research by helping address six challenges:[3,4,7,13] reverse causation and unmeasured confounding, comprehensive examination of phenotypic effects, low efficiency, replication, multilevel data integration, and characterization of tissue-specific effects. To frame this commentary for audiences with varied backgrounds in genetics and exposure science, we provide overviews of genetic variant measurement, core study designs, and consortia, as well as exposomic approaches that leverage advances in high-resolution mass spectrometry (MS).[14] We then describe two core metrics—heritability and genetic risk scores (GRS)—that may be estimated from genetic data and may inform or be employed in several approaches we describe: Mendelian randomization (MR),

Address correspondence to Christy Avery, Gillings School of Global Public Health, 123 W. Franklin St., Suite 410, Chapel Hill, NC 27516 USA. Email: Christy_avery@unc.edu

The authors declare they have no actual or potential competing financial interests.

Received 4 February 2021; Revised 8 April 2022; Accepted 12 April 2022; Published 9 May 2022.

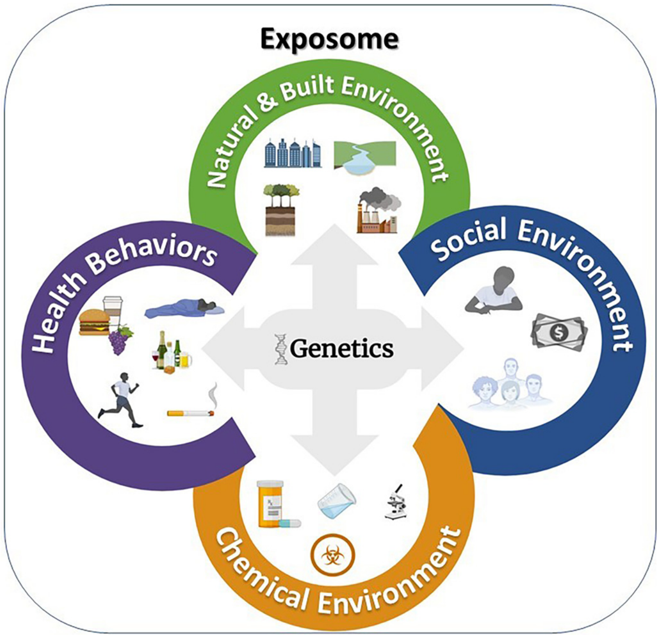

Figure 1. Conceptual diagram of the exposome. By placing genetic data in the middle of four exposome domains (the natural and built environment, the social environment, the chemical environment, and health behaviors), the central role of genetic data is emphasized. Figure adapted from Vermeulen R, Schymanski EL, Barabasi AL, Miller GW. 2020. The exposome and health: Where chemistry meets biology. Science 367:392–396. Reprinted with permission from American Association for the Advancement of Science (AAAS).

phenome-wide association studies (PheWAS), joint models for missing data, cross-study replication, multilevel data integration methods, and tissue-specific biomarker imputation (Table 1). Examples are drawn from studies of biomarkers and health behaviors, exposure domains where the causal inference methods we describe are most often applied. Finally, we acknowledge limitations, including how the utility of methods that integrate genetic data will vary by exposure and may depend on the presence of specific environmental exposures[1] and large sample sizes.[15,16]

Measurement of Genetic Variants, Genome-Wide Association Study (GWAS), and Consortia

Measurement of genetic variants. Over the past three decades, assays that measured a handful of genetic variants have advanced to today's whole genome sequencing. Whole genome sequencing is considered the gold standard in genome measurement because of its accuracy, scope, and ability to identify new genetic variants.[17] However, cost, storage, and computational feasibility have limited widespread adoption of whole genome sequencing. Instead, the primary source of genetic data remains arrays that genotype 500,000 to 5 million genetic variants. Genotyped genetic variants are then used as a scaffold for high-quality imputation to a wider set (~ 40 million) of genetic variants,[18] helping ensure a common set of genetic variants across studies. Imputation servers [e.g., the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html) and the Trans-Omics Precision Medicine (TOPMed) Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/)] that perform genetic imputation quickly and free of charge have streamlined, simplified, and improved imputation of array data.[19–22] Imputation coverage and accuracy depend crucially on the size of the reference panel used for imputation, the density of genotyped variants, and the genetic distance between reference panel populations and target populations.[23,24] The largest broadly available reference panels are from TOPMed, with $>135,000$ individuals of diverse ancestry.[25,26] These panels provide highly accurate imputation of genetic variants down to a minor allele frequency (MAF) of $\sim 0.1\%$ across multiple ancestries.[26] Although earlier reference panels include individuals from more diverse worldwide populations, these reference panels provide more limited imputation coverage and accuracy because of smaller sample sizes.[27]

The GWAS. Advances in genotyping and imputation have facilitated the rise of the GWAS as a key study design to identify genetic variants associated with complex phenotypes (reviewed by Tam et al.[28]). By design, GWAS are unbiased with respect to mechanistic hypotheses, biological knowledge, and genomic location.[17,29] This design has been remarkably successful in mapping variant–phenotype associations.[30,31] The requirement for large sample sizes and the importance of replication has prompted the formation of numerous large GWAS consortia[32–34] that are well powered to detect common, infrequent and rare variants associated with complex phenotypes. These consortia also routinely share summary-level data (e.g., effect estimates, *p*-values, and allele frequencies) that are publicly available via centralized repositories[30] (https://www.ncbi.nlm.nih.gov/gap/).

Exposomics Consortia

Recognizing the efficiencies enabled by a consortium model, exposomics consortia also have been formed (see Vrijheid et al.[35] and https://emoryhercules.com/). For example, the EXPOsOMICs project is a European exposomics consortium that includes experimental studies, mother–child cohorts, observational studies of adults, and personal exposure monitoring studies.[36] Several studies contributing to EXPOsOMICs have genetic data. By including cohorts across the life course, consortia like EXPOsOMICs enable examination of questions that would be difficult to conduct within a single study; these consortia also provide opportunities for replication or data pooling. Phenotype specific consortia also have been assembled, including the COnsortium of METabolomics (COMETS).[37] COMETS is the world's largest metabolomics

Table 1. Six methods or approaches that leverage genetic data to address challenges facing exposomics research and empower causal inference.

Challenge	Statistical method or approach afforded by genetic data
Reverse causation and unmeasured confounding	Mendelian randomization
Comprehensive phenotype measurement and characterization of phenotypic effects	PheWAS of genetically predicted exposures in large biobanks or populations with EHR.
Decreased efficiency from data missing by design or from detection limits	Joint models that address missing data from exposure measured on subset of participants and detection limits by leveraging the information available from any associations between genetics and covariates with exposomic data
Difficulty replicating findings, particularly if exposure is not measured broadly, not measured with comparable protocols, or unidentified	External replication using genetically predicted exposures.
Multilevel data that may be difficult to integrate	Integrative approaches that use genetic data as a framework to link multi-omic data.
Limited ability to characterize tissue-specific effects	Imputation of tissue-specific biomarkers of exposure and internal dose (e.g., transcripts, methylation, metabolomics, proteins) using publicly available data.

consortium and comprises 47 international cohorts that include >136,000 participants with blood metabolomics data. Genetic data were measured in approximately 68% of COMETS participants.

Core Metrics

Heritability. Heritability (H^2) estimates the proportion of phenotypic variation (range: 0–1) attributable to additive, dominance, and epistatic variance components, i.e., "broad-sense heritability" (reviewed by Zaitlen and Kraft[38]). Traditional methods to estimate heritability use family-based studies and only quantify additive genetic variance ("narrow-sense heritability")—the major contributor to H^2—given longstanding challenges estimating nonadditive genetic variance.[39] Recent innovations have enabled approximation of narrow-sense heritability in population-based studies using genetic data[40]; these approximations typically underestimate narrow-sense heritability due to incomplete linkage disequilibrium (LD) between causal genetic variants and genotyped genetic variants as well as error in genetic variant effect estimates.[15]

Narrow-sense heritability can help gauge the potential utility of genetic data to inform causal inference for a given exposure, because genetic data will have limited value when narrow-sense heritability is low. As an example, an Australian twin study of the metallic elements arsenic, cadmium, copper, mercury, lead, selenium, and zinc measured in erythrocytes estimated narrow-sense heritabilities of moderate size ranging from 0.19 (cadmium) to 0.40 (lead).[41] A follow-up GWAS of copper, selenium, and zinc identified eight highly significant ($p < 5 \times 10^{-8}$) common genetic variants that mapped to loci containing genes with roles in trace element metabolism.[42] These genetic variants accounted for 4%–8% of phenotypic variance in copper, selenium, and zinc. Effects of this magnitude are considerably larger than the majority of genetic variants identified by GWAS to date.[43]

Genetically Predicted Phenotypes

Genetic data can help extend the reach of exposome studies by enabling the estimation of genetically predicted phenotypes. These genetically predicted phenotypes are then substituted for measured phenotypes when conducting association studies or causal inference investigations. As described below, in addition to expanding the number of phenotypes for evaluation, the use of a genetically predicted phenotype can help reduce bias from confounding and reverse causation.[44]

Genetically predicted phenotypes can be constructed using one genetic variant, a limited number of genetic variants, or hundreds to millions of genetic variants. For phenotypes with a monogenic or oligogenic genetic architecture, genetically predicted phenotypes may be constructed using one genetic variant or by aggregating a small number of independent genetic variants.[45,46] For polygenic phenotypes, genetically predicted phenotypes often are constructed by aggregating hundreds to millions of genetic variants (reviewed by Chatterjee et al.[47] and Wray et al.[15]).

Aggregation of genetic variants into a genetically predicted phenotype is accomplished using a GRS. GRS are calculated as a sum of genetic variants that are typically weighted by the magnitude of association between each genetic variant and the phenotype of interest. Numerous approaches are available to estimate GRS, which are distinguished by the method used to select and weigh genetic variants and the method used to account for LD. GRS weights are usually derived from a GWAS[47–49] and then applied in an independent, ancestry-matched target sample for validation.[15,49] In the absence of an independent target sample for validation, methods are emerging[50,51] that estimate GRS using cross-validation to address overfitting. These methods offer efficient alternatives for studies without access to independent data and may be particularly useful when examining a phenotype that is difficult to measure, a phenotype that is uncommonly measured or when conducting research in a unique population.

Exposure Science and the Chemical Exposome

Environmental health studies have undergone a dramatic shift in recent years, with rapid technological advancements enabling broader coverage of the chemical exposome while also expanding the inclusion of nonchemical stressors.[7,52,53] Approaches for chemical exposome characterization include suspect screening and nontargeted analyses, which enable the measurement of many chemicals simultaneously using approaches that rely on high-resolution chemical detection coupled with computational methods to efficiently mine large data sets. Targeted analytical methods also may be employed to evaluate the impacts of exposures to chemical mixtures in the environment.[54] These methods provide more limited coverage of chemicals and thus may not capture exposure information at the "-omic" level. Indeed, an increasing number of global measurement approaches have recently been implemented to characterize exposome signatures within environmental media, including household dust,[55] drinking water,[56] and consumer products.[57] Biological samples, such as blood, saliva, teeth, and urine, also may be analyzed to measure chemicals and their associated metabolites, as well as other exposure biomarkers.[53,58–62]

Suspect screening and nontargeted analyses. Suspect screening and nontargeted analyses leverage MS platforms coupled with compound database matching approaches to identify and potentially confirm chemicals.[63–66] Suspect screening can be implemented using both gas chromatography (GC) and liquid chromatography (LC) separation followed by either low- or high-resolution MS detection. GC-based methods can be aided by the addition of electron ionization, whereas LC-based methods can use softer ionization techniques, such as electrospray ionization or atmospheric pressure chemical ionization, resulting in detailed fragmentation spectra information to better identify tentative chemicals. With suspect screening approaches, resulting spectra are compared against a library of known compounds, and those with matching attributes are identified and prioritized for confirmation. Nontargeted approaches, in contrast, rely on high-resolution MS platforms to acquire accurate mass, isotope profile, and fragmentation spectra. These data are then used to predict chemical structures, and chemicals are tentatively assigned formulas and associated chemical identifier information. Therefore, suspect screening analyses query for known chemicals, whereas nontargeted analyses generate information on chemicals that are potentially completely unknown. Both approaches yield tentatively identified chemicals which require further confirmation, using tandem mass spectrometry (MS/MS) fragmentation information or confirmation via chemical standards.[63,64]

Prioritizing chemicals for confirmation. It is not feasible to confirm all, or even the majority, of chemicals in a given sample. Because of this limitation, it is important to develop and implement methods to prioritize chemicals for final confirmation. Data streams to aid prioritization include chemical exposure estimates and metabolite predictions, which inform the likelihood of a chemical being absent or present in a given sample, as well as toxicity screening and prediction data, which inform the likelihood of a chemical being toxic and therefore of high interest.[55,64] As these methods grow, exposomic measures likely will become increasingly integrated across multiple tiers of data to better address the dynamic nature of the exposome and its overall influence on health and disease.

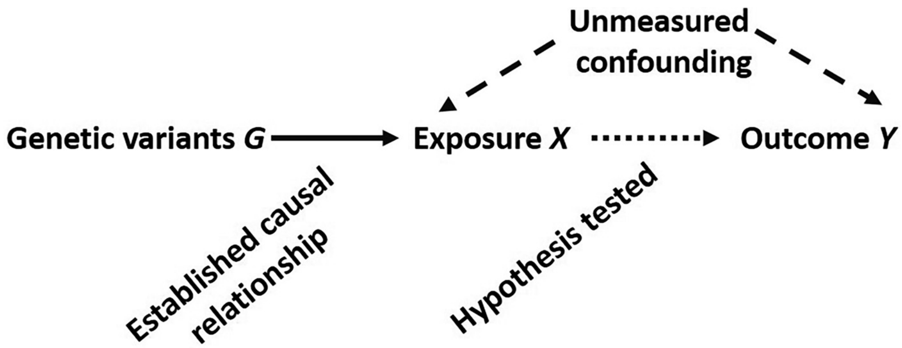

Figure 2. Example causal diagram representing the relationship between genetic variants G, exposure X, and outcome Y. The hypothesis tested by Mendelian randomization is shown by the dotted arrow where G serves as an instrumental variable for X (solid arrow).

How Causal Inference Methods and Study Designs That Use Genetic Data Can Empower Exposomics

In this section we draw on research interrogating a spectrum of exposures to demonstrate how causal inference methods and study designs that integrate genetic data can empower exposomics research. We focus on six challenges that are not necessarily unique to exposomics research, but we consider particularly salient, given the score and dimensionality of the exposome. Although it may not be feasible to comprehensively address all six challenges using causal inference methods and study designs that integrate genetic data, we anticipate that these approaches will help strengthen causal inference for numerous exposome phenotypes.

Method 1: Evaluate reverse causation and unmeasured confounding. Ideally exposomics research would leverage a longitudinal prospective design in which exposures are sampled repeatedly before an outcome occurs.[4] However, many exposome studies use cross-sectional[67] or case–control designs.[68] Reverse causation is a concern with these designs, because disease status may affect the exposure or its measurement.[69–71] Other challenges include unmeasured or poorly measured confounders as well as exposures that are poorly understood, making the identification of confounders difficult. MR, a popular causal inference tool that uses genetic data to investigate associations between potentially modifiable risk factors, including environmental exposures, and outcomes in observational data (reviewed by Davies et al.[44]), has been proposed to help address these challenges.

MR is a form of instrumental variable (IV) analysis[72] based on the concept that if exposure X affects outcome Y, factors affecting X (i.e., inherited genetic variants, G) also must affect Y (Figure 2). G therefore serves as an IV for studies of the X–Y association. Strengths of MR include G–Y associations that are generally robust to confounding from variables other than ancestry, which can be addressed through adjustment.[73] Because G is determined at conception, G precedes Y, also protecting against reverse causation under the assumption that G is associated with X, not Y or an alternative cause of Y.[74] MR also is dependent on the identification of a strong genetic IV (i.e., a genetically predicted phenotype) and assumes an absence of pleiotropy (i.e., the effect of G on Y is not exclusively through X). The development of methods to evaluate these assumptions is an active area of research[75,76] and alternative methods, including mediation analysis, have been proposed when assumptions are violated[77] or when biological mechanisms do not conform to assumptions.[78] If IV assumptions are satisfied, MR can inform on the presence and direction of the association between X and Y. However, numerical MR estimates generally are not informative, because by estimating the G–Y association, MR estimates cannot be interpreted as the predicted real-world influence of changes to X.[79] Although few investigators have successfully used MR to study time-varying exposures,[80] methods are under development.[81]

Despite these challenges, MR has been used to strengthen causal inference in exposomics research across a variety of phenotypes.[82–84] For example, Pierce et al.[11] used MR to confirm the presence and direction of the association between biomarkers of arsenic methylation efficiency and arsenic toxicity.[11] Here, MR helped gauge the degree to which observational findings reflected reverse causation or residual confounding by unmeasured or poorly measured covariates in a process termed "triangulation," i.e., the integration of results from several different approaches, each with different and unrelated key sources of potential bias.[85] Other applications of MR relevant to exposomics research include multivariate MR, which has been used to simultaneously examine causal effects of correlated phenotypes.[86]

Method 2: comprehensively examine phenotypic effects. Few studies have comprehensively studied the health effects of exposome phenotypes.[87] However, advances in large-scale phenotyping through biobanks and linkage to electronic health records (EHR), in combination with genetic data, offer opportunities to help address this research gap via a PheWAS (reviewed by Bush et al.[88]). Benefits of the hypothesis-free PheWAS include broad phenotypic characterization, enabling the identification of potentially novel associations. For example, a recent PheWAS of genetically predicted serum calcium examined associations with 925 disease outcomes constructed from hospital inpatient and mortality data.[89] This PheWAS identified associations with renal, musculoskeletal, and cardiovascular phenotypes, which in part mirror findings from calcium supplementation trials.[90] Unexpected associations with allergy or adverse effects of penicillin also were identified, which may point to an unappreciated role of calcium in immune function.

Limitations of PheWAS include the requirement of a genetically predicted phenotype and large sample sizes with broad phenotypic characterization in the same ancestral population from which the genetically predicted phenotype was constructed.[91,92] In addition, few EHR PheWAS have fully incorporated unstructured exposure, behavioral, or lifestyle data, which are likely highly relevant to exposomics research but are challenging to extract from or missing in clinical free text.[93] Emerging electronic phenotyping approaches[93] and global biobank initiatives[94] offer potential ways forward.

Method 3: increase efficiency. Due to cost or other constraints, biomarkers of exposure and effect may be measured on a subset of study participants. Measuring biomarkers on a subset of participants, thereby introducing missing data due to study design, reduces efficiency in comparison with an analysis of the entire study population, thereby introducing uncertainty and decreasing statistical power. Because genetic data often are available on larger population subsets, statistical methods that use genetic data to infer biomarkers in participants not selected for measurement can help increase efficiency. For example, the Atherosclerosis Risk in Communities (ARIC) study measured serum metabolites on 4,032 (26%) of 15,792 participants at baseline.[95] Although most analyses investigating associations between metabolites and outcomes often restrict to this smaller sample size, imputation methods may increase efficiency. These methods perform well when the sample with measured data is a random or stratified random sample of the larger study population.[96–98]

In addition to missing data due to study design, many biomarkers of exposure and effect are subject to limits of detection and are nondetectable. There are commonly accepted analytic practices for nondetectable data; however, most methods cannot address multiple missing data mechanisms.[99,100] Treating all missing data as originating from one missing data mechanism also can result in grossly inefficient and potentially biased estimates.[101,102]

Methods that leverage genetic data can account for biomarker data that are missing due to study design and limits of detection. Using metabolites as an example, these missing-data methods use joint models to model both the association between genetic data and metabolites and the association between the metabolites and the outcome[103]; data from participants with genetic and outcome data are used regardless of whether metabolites were measured. By using data from a larger group of participants (e.g., increasing the sample size from $n = 4{,}032$ to $n = 12{,}773$ in the ARIC Study[104]), these models offer more efficient estimates of exposure–outcome associations. Accounting for two types of missing data also reduces the potential for biased estimates. Simulations have shown scenarios where these methods provided virtually unbiased estimates, whereas methods addressing only one type of missing data can provide estimates that can differ by as much as 20% from the true value.[103]

Method 4: replicate findings. Genetic data also provide a template for replication, defined here as the consistent estimation of effect direction, significance, and potentially magnitude (depending on the phenotype under investigation) in an independent sample from the same source population.[105] Replication is especially important in exposomic studies, because the number of exposures adds an exploratory element and concomitant large potential for false positive findings[106] that may mislead scientists and the public and misdirect the allocation of scientific resources. In GWAS, the potential for a high proportion of false positive results was addressed through imputation to common genotype reference panels, stringent multiple hypothesis testing correction, and replication (reviewed in Weinberg[107] and Chanock et al.[108]). A parallel framework is needed in exposomics research, although many exposures may not be measured broadly. Methods and output also may vary across studies according to study design factors such as, including sample type, instrumentation, analytical conditions, and the domain of chemicals under investigation.

For studies of the exposome where researchers do not have access to an independent replication sample, genetic data may provide a partial solution. To illustrate, a recent study identified associations between manganese, lead, and chromium biomarkers with intelligence quotient (IQ) in adolescents.[67] Although the authors did not attempt replication, partnering with an independent study with measured IQ and genetic data from which to construct measures of genetically predicted manganese, lead, and chromium phenotypes could provide a replication opportunity. Other avenues for replication could include publicly available GWAS summary statistics.[30] Using publicly available GWAS summary statistics, researchers could examine whether genetic variants predictive of the phenotype of interest also were predictive of the outcome. Returning to the Bauer 2020 example, genetic variant rs13107325 was identified in GWAS of manganese[109] as well as intelligence[110] and general cognitive ability,[111] providing independent evidence linking manganese with IQ.

Method 5: multilevel data integration. The exposome includes exposures that are multilevel, complex, and likely affected by genetic, environmental, and gene-by-environment effects. However, when describing these complexities, a majority of exposome studies distinguish between environmental and genetic causes of disease, with few studies considering opportunities to integrate information. Multi-omics studies, which include genomic, epigenomic, transcriptomic, proteomic, microbiomic, and exposomic data, are emerging efforts that attempt to disentangle complex, multilayered exposure effects. Examples of multilevel exposome studies include dimensionality reduction and variable selection approaches that consider the correlation structure between multiple omics.[112] Systems and network analyses also have been used to better assess the complex interplay within and between different omics while accounting for biological functionality.[113,114] Parallel efforts include the modeling of concentration dependency and several tools that accommodate different dose–response trends also have been published.[115,116] Together, these approaches are promising avenues to address cross-omics relationships and their complex dynamics.[117]

Despite emerging interests in multi-omics studies, few studies have integrated genetics data with other omics data, even though genetics is the most mature of the omics fields.[118] One example is provided by research examining atopic dermatitis (AD),[119] which integrated genetic, epigenomic, transcriptomic, and proteomic data to better understand disease heterogeneity. A crucial component of that study approach was the use of GWAS findings to identify priority genes from which candidate disease pathways integrating multilevel data were constructed and tested.

Method 6: characterization of tissue-specific effects. Biomarkers of exposure and effect are a promising tool to evaluate molecular responses to exposures as well as downstream consequences of variation in molecular response. However, direct measurement of biomarkers across relevant tissues is largely infeasible due to expense and tissue accessibility. This evidence gap constrains interpretation of biomarker effects and determination of relevance. The parallel collection of genetic and omics data in varied tissues enables construction of tissue-specific genetically predicted phenotypes; one example is the construction of genetically predicted gene expression.[120,121] Measures of genetically predicted gene expression offer a partial solution to examining downstream effects of variation in tissue-specific gene expression, because models to infer genetically predicted gene expression are publicly available (http://predictdb.org/ and http://gusevlab.org/projects/fusion/). These models also can be used to construct exposures for association testing and to examine evidence of tissue-specific effects. Similar imputation approaches are being developed for other omics, including DNA methylation levels.[122] Although exposomics research examples are currently scarce, emerging research examining genetically predicted omics in inaccessible but highly relevant tissues demonstrates the role of this emerging approach for pathophysiological insight.[123]

Discussion

In this commentary, we described how the increased application of genetic data and methods could strengthen causal inference in exposomics research. These approaches are enabled by the broad availability of genetic data, the active development of causal inference tools and study designs that use genetic data, publicly available data repositories, and a growing appreciation that many exposome phenotypes are heritable.[124] Although the application of genetic data and methods may add analytical complexity, these approaches offer the potential to extend the reach of exposomics research and help address the challenges of comprehensively measuring the exposome and its health effects across studies, the life course, and in varied contexts and in diverse populations.

We acknowledge several limitations to the application of genetic data and methods for causal inference in exposomics research. Few studies have comprehensively cataloged genetic variants that predict diverse exposures. Even when genetic variants have been identified and replicated in independent studies, ascertaining biological impact remains challenging.[125] However, interpretation of a genetic variant's impact is not necessarily required for many of the methods we propose, noting that some of the approaches we describe (e.g., using genetic data to characterize tissue-specific effects) may help illuminate effects in toxicity-relevant tissues and organs. We therefore advocate for expanding the evidence base to examine more comprehensively the genetic architecture of exposure biomarkers[126] and health behaviors,[127,128] the exposome domains that most likely harbor heritable exposures or exposure biomarkers. Another major challenge is the lack of diversity in published GWAS. Although the limited racial/ethnic diversity of GWAS has been the topic of several commentaries,[129] GWAS in populations exposed to specific toxicants or populations capturing crucial life course stages (e.g., infancy and childhood or pregnancy) also remain uncommon. Expanding the diversity of GWAS and the cataloging of genetic variants that predict exposome phenotypes, e.g., by international scientific collaborations that share summary results through established repositories, could help remedy these research gaps. Further, we excluded discussion of gene–environment interaction, instead focusing on genetic data applications that are less well known in exposomics research. Consistent with the other challenges described, methods to enhance gene–environment interaction studies are areas of active research.[130–132] Finally, the sample sizes needed to construct well-powered genetically inferred phenotypes may be infeasible for a single study. Again, the sharing of summary data is a disciplinary norm that can increase statistical power to detect genetic effects and construct predictive genetically inferred phenotypes, particularly when examining phenotypes influenced by many common genetic variants of small effects, phenotypes for which the genetic effects are only observable in the presence of specific exposures that are themselves uncommon, or when studying gene–environment interaction.[1]

Wild proposed the concept of the exposome in 2005,[1] emphasizing the need to balance investments in genetics research with investments in exposomics research. Almost two decades later, distinctions between environmental vs. genetic effects on disease remain common in the exposomics literature,[3,6,133] with few examples of studies that successfully integrate both sources of data. It is noteworthy that many of the perceived hurdles associated with genetic data, including measurement scale, are not new to exposure scientists. Fully leveraging exposomic data also requires embracing biological complexity and systems-level thinking, two core exposure science paradigms.[2] Adding genetic data simply adds one more level of complexity. Ultimately, the success of attempts to integrate genetic data into exposomics research will likely require environmental scientists to expand their large collaborative network to include geneticists and genetic epidemiologists, because the requisite data are largely extant.[36] Through these collaborations, efforts that better integrate genetic and exposomics data to improve human health are achievable.

Acknowledgments

The authors acknowledge funding from R01HL142825 (C.L.A., A.F.B., C.G.D., H.M.H., K.E.N.), R01HL147853 (C.L.A., H.M.H), R01HL143885 (C.L.A., P.G.-L., A.G.H., K.E.N.), R01HG10297 (C.L.A., M.G., H.M.H., K.E.N.), T32HL129982 (V.L.B., B.S.S.), T32HD091058 (A.G.L.), F32HL149256 (M.P.L.), T32HL007055 (J.M.C., M.P.L), and P30ES010126 (S.M.E., K.L., J.E.R.).

References

1. 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):1847–1850, PMID: 16103423, https://doi.org/10.1158/1055-9965.EPI-05-0456.
2. Wild CP. 2012. The exposome: from concept to utility. Int J Epidemiol 41(1):24–32, PMID: 22296988, https://doi.org/10.1093/ije/dyr236.
3. Dennis KK, Auerbach SS, Balshaw DM, Cui Y, Fallin MD, Smith MT, et al. 2016. The importance of the biological impact of exposure to the concept of the exposome. Environ Health Perspect 124(10):1504–1510, PMID: 27258438, https://doi.org/10.1289/EHP140.
4. DeBord DG, Carreon T, Lentz TJ, Middendorf PJ, Hoover MD, Schulte PA. 2016. Use of the "exposome" in the practice of epidemiology: a primer on -omic technologies. Am J Epidemiol 184(4):302–314, PMID: 27519539, https://doi.org/10.1093/aje/kwv325.
5. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. 2019. The exposome: molecules to populations. Annu Rev Pharmacol Toxicol 59:107–127, PMID: 30095351, https://doi.org/10.1146/annurev-pharmtox-010818-021315.
6. Niedzwiecki MM, Miller GW. 2017. The exposome paradigm in human health: lessons from the Emory Exposome Summer Course. Environ Health Perspect 125(6):064502, PMID: 28669935, https://doi.org/10.1289/EHP1712.
7. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. 2020. The exposome and health: where chemistry meets biology. Science 367(6476):392–396, PMID: 31974245, https://doi.org/10.1126/science.aay3164.
8. Hill J, Stuart EA. 2015. Causal inference: overview. In: *International Encyclopedia of the Social & Behavioral Sciences*. Wright JD, ed. 2nd ed. New York, New York: Elsevier, 255–260.
9. Pearce N, Vandenbroucke JP, Lawlor DA. 2019. Causal inference in environmental epidemiology: old and new approaches. Epidemiology 30(3):311–316, PMID: 30789434, https://doi.org/10.1097/EDE.0000000000000987.
10. Tellez-Plaza M, Gribble MO, Voruganti VS, Francesconi KA, Goessler W, Umans JG, et al. 2013. Heritability and preliminary genome-wide linkage analysis of arsenic metabolites in urine. Environ Health Perspect 121(3):345–351, PMID: 23322787, https://doi.org/10.1289/ehp.1205305.
11. Pierce BL, Tong L, Argos M, Gao J, Farzana J, Roy S, et al. 2013. Arsenic metabolism efficiency has a causal role in arsenic toxicity: mendelian randomization and gene-environment interaction. Int J Epidemiol 42(6):1862–1871, PMID: 24536095, https://doi.org/10.1093/ije/dyt182.
12. Scannell Bryan M, Sofer T, Mossavar-Rahmani Y, Thyagarajan B, Zeng D, Daviglus ML, et al. 2019. Mendelian randomization of inorganic arsenic metabolism as a risk factor for hypertension- and diabetes-related traits among adults in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) cohort. Int J Epidemiol 48(3):876–886, PMID: 30929011, https://doi.org/10.1093/ije/dyz046.
13. VanderWeele TJ. 2017. Outcome-wide epidemiology. Epidemiology 28(3):399–402, PMID: 28166102, https://doi.org/10.1097/EDE.0000000000000641.
14. Andra SS, Austin C, Patel D, Dolios G, Awawda M, Arora M. 2017. Trends in the application of high-resolution mass spectrometry for human biomonitoring: an analytical primer to studying the environmental chemical space of the human exposome. Environ Int 100:32–61, PMID: 28062070, https://doi.org/10.1016/j.envint.2016.11.026.
15. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. 2013. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14(7):507–515, PMID: 23774735, https://doi.org/10.1038/nrg3457.
16. Huffman JE. 2018. Examining the current standards for genetic discovery and replication in the era of mega-biobanks. Nat Commun 9(1):5054, PMID: 30498205, https://doi.org/10.1038/s41467-018-07348-x.
17. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 2017. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101(1):5–22, PMID: 28686856, https://doi.org/10.1016/j.ajhg.2017.06.005.
18. Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406, PMID: 19715440, https://doi.org/10.1146/annurev.genom.9.081307.164242.
19. Das S, Abecasis GR, Browning BL. 2018. Genotype imputation from large reference panels. Annu Rev Genomics Hum Genet 19:73–96, PMID: 29799802, https://doi.org/10.1146/annurev-genom-083117-021602.
20. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. 2016. Next-generation genotype imputation service and methods. Nat Genet 48(10):1284–1287, PMID: 27571263, https://doi.org/10.1038/ng.3656.

21. Fuchsberger C, Abecasis GR, Hinds DA. 2015. minimac2: faster genotype imputation. Bioinformatics 31(5):782–784, PMID: 25338720, https://doi.org/10.1093/bioinformatics/btu704.
22. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44(8):955–959, PMID: 22820512, https://doi.org/10.1038/ng.2354.
23. Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. 2017. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet 25(7):869–876, PMID: 28401899, https://doi.org/10.1038/ejhg.2017.51.
24. Roshyara NR, Scholz M. 2015. Impact of genetic similarity on imputation accuracy. BMC Genet 16(1):90, PMID: 26193934, https://doi.org/10.1186/s12863-015-0248-2.
25. Quick C, Anugu P, Musani S, Weiss ST, Burchard EG, White MJ, et al. 2020. Sequencing and imputation in GWAS: cost-effective strategies to increase power and genomic coverage across diverse populations. Genet Epidemiol 44(6):537–549, PMID: 32519380, https://doi.org/10.1002/gepi.22326.
26. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, et al. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. 2019. Use of >100,000 NHLBI Trans-Omics for precision medicine (TOPMed) consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet 15(12): e1008500, Dec, PMID: 31869403, https://doi.org/10.1371/journal.pgen.1008500.
27. Belsare S, Levy-Sakin M, Mostovoy Y, Durinck S, Chaudhuri S, Xiao M, et al. 2019. Evaluating the quality of the 1000 genomes project data. BMC Genomics 20(1):620, PMID: 31416423, https://doi.org/10.1186/s12864-019-5957-x.
28. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. 2019. Benefits and limitations of genome-wide association studies. Nat Rev Genet 20(8):467–484, PMID: 31068683, https://doi.org/10.1038/s41576-019-0127-1.
29. Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. Am J Hum Genet 90(1):7–24, PMID: 22243964, https://doi.org/10.1016/j.ajhg.2011.11.029.
30. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. 2019. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47(D1):D1005–D1012, PMID: 30445434, https://doi.org/10.1093/nar/gky1120.
31. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. 2017. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res 45(D1):D896–D901, PMID: 27899670, https://doi.org/10.1093/nar/gkw1133.
32. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, et al. CHARGE Consortium. 2009. Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium: design of prospective Meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2(1):73–80, PMID: 20031568, https://doi.org/10.1161/CIRCGENETICS.108.829747.
33. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al. 2017. The OncoArray consortium: a network for understanding the genetic architecture of common cancers. Cancer Epidemiol Biomarkers Prev 26(1):126–135, PMID: 27697780, https://doi.org/10.1158/1055-9965.EPI-16-0106.
34. Sullivan PF. 2010. The psychiatric GWAS consortium: big science comes to psychiatry. Neuron 68(2):182–186, PMID: 20955924, https://doi.org/10.1016/j.neuron.2010.10.003.
35. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. 2014. The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect 122(6):535–544, PMID: 24610234, https://doi.org/10.1289/ehp.1307204.
36. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, et al. EXPOsOMICS Consortium. 2017. The exposome in practice: design of the EXPOsOMICS project. Int J Hyg Environ Health 220(2 pt A):142–151, PMID: 27576363, https://doi.org/10.1016/j.ijheh.2016.08.001.
37. Yu B, Zanetti KA, Temprosa M, Albanes D, Appel N, Barrera CB, et al. 2019. The consortium of metabolomics studies (COMETS): metabolomics in 47 prospective cohort studies. Am J Epidemiol 188(6):991–1012, PMID: 31155658, https://doi.org/10.1093/aje/kwz028.
38. Zaitlen N, Kraft P. 2012. Heritability in the genome-wide association era. Hum Genet 131(10):1655–1664, PMID: 22821350, https://doi.org/10.1007/s00439-012-1199-6.
39. Hivert V, Sidorenko J, Rohart F, Goddard ME, Yang J, Wray NR, et al. 2021. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. Am J Hum Genet 108(5):786–798, PMID: 33811805, https://doi.org/10.1016/j.ajhg.2021.02.014.
40. Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76–82, PMID: 21167468, https://doi.org/10.1016/j.ajhg.2010.11.011.
41. Whitfield JB, Dy V, McQuilty R, Zhu G, Heath AC, Montgomery GW, et al. 2010. Genetic effects on toxic and essential elements in humans: arsenic, cadmium, copper, lead, mercury, selenium, and zinc in erythrocytes. Environ Health Perspect 118(6):776–782, PMID: 20053595, https://doi.org/10.1289/ehp.0901541.
42. Evans DM, Zhu G, Dy V, Heath AC, Madden PAF, Kemp JP, et al. 2013. Genome-wide association study identifies loci affecting blood copper, selenium and zinc. Hum Mol Genet 22(19):3998–4006. Oct 1, PMID: 23720494, https://doi.org/10.1093/hmg/ddt239.
43. Zhang Y, Qi G, Park JH, Chatterjee N. 2018. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet 50(9):1318–1326, PMID: 30104760, https://doi.org/10.1038/s41588-018-0193-x.
44. Davies NM, Holmes MV, Davey Smith G. 2018. Reading mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601, PMID: 30002074, https://doi.org/10.1136/bmj.k601.
45. Trinder M, Uddin MM, Finneran P, Aragam KG, Natarajan P. 2021. Clinical utility of lipoprotein(a) and LPA genetic risk score in risk prediction of incident atherosclerotic cardiovascular disease. JAMA Cardiol 6(3):287–295, https://doi.org/10.1001/jamacardio.2020.5398.
46. Bonifacio E, Beyerlein A, Hippich M, Winkler C, Vehik K, Weedon MN, et al. 2018. Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: a prospective study in children. PLoS Med 15(4): e1002548, PMID: 29614081, https://doi.org/10.1371/journal.pmed.1002548.
47. Chatterjee N, Shi J, García-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet 17(7):392–406, PMID: 27140283, https://doi.org/10.1038/nrg.2016.27.
48. Wand H, Lambert SA, Tamburro C, et al. 2007. Improving reporting standards for polygenic scores in risk prediction studies. medrxiv. Preprint posted online March 23, 2007. https://doi.org/10.1101/2020.04.23.20077099.
49. Martin AR, Daly MJ, Robinson EB, Hyman SE, Neale BM. 2019. Predicting polygenic risk of psychiatric disorders. Biol Psychiatry 86(2):97–109, PMID: 30737014, https://doi.org/10.1016/j.biopsych.2018.12.015.
50. Mefford J, Park D, Zheng Z, Ko A, Ala-Korpela M, Laakso M, et al. 2020. Efficient estimation and applications of Cross-Validated genetic predictions to polygenic risk scores and linear mixed models. J Comput Biol 27(4):599–612, PMID: 32077750, https://doi.org/10.1089/cmb.2019.0325.
51. Mak TSH, Porsch RM, Choi SW, Sham PC. 2018. Polygenic scores for UK Biobank scale data. bioRxiv 252270. Preprint posted online October 5, 2018. https://doi.org/10.1101/252270.
52. Clougherty JE, Rider CV. 2020. Integration of psychosocial and chemical stressors in risk assessment. Curr Opin Toxicol 22:25–29, https://doi.org/10.1016/j.cotox.2020.07.005.
53. Nakamura J, Mutlu E, Sharma V, Collins L, Bodnar W, Yu R, et al. 2014. The endogenous exposome. DNA Repair (Amst) 19:3–13, PMID: 24767943, https://doi.org/10.1016/j.dnarep.2014.03.031.
54. Rager JE, Clark J, Eaves LA, Avula V, Niehoff NM, Kim YH, et al. 2021. Mixtures modeling identifies chemical inducers versus repressors of toxicity associated with wildfire smoke. Sci Total Environ 775:145759, PMID: 33611182, https://doi.org/10.1016/j.scitotenv.2021.145759.
55. Rager JE, Strynar MJ, Liang S, McMahen RL, Richard AM, Grulke CM, et al. 2016. Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. Environ Int 88:269–280, PMID: 26812473, https://doi.org/10.1016/j.envint.2015.12.008.
56. Newton SR, McMahen RL, Sobus JR, Mansouri K, Williams AJ, McEachran AD, et al. 2018. Suspect screening and non-targeted analysis of drinking water using point-of-use filters. Environ Pollut 234:297–306, PMID: 29182974, https://doi.org/10.1016/j.envpol.2017.11.033.
57. Phillips KA, Yau A, Favela KA, Isaacs KK, McEachran A, Grulke C, et al. 2018. Suspect screening analysis of chemicals in consumer products. Environ Sci Technol 52(5):3125–3135, PMID: 29405058, https://doi.org/10.1021/acs.est.7b04781.
58. Bloch R, Schütze S-E, Müller E, Röder S, Lehmann I, Brack W, et al. 2019. Non-targeted mercapturic acid screening in urine using LC-MS/MS with matrix effect compensation by postcolumn infusion of internal standard (PCI-IS). Anal Bioanal Chem 411(29):7771–7781, PMID: 31667563, https://doi.org/10.1007/s00216-019-02166-6.
59. Andra SS, Austin C, Arora M. 2016. The tooth exposome in children's health research. Curr Opin Pediatr 28(2):221–227, PMID: 26859286, https://doi.org/10.1097/MOP.0000000000000327.
60. 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):769–774, PMID: 24659601, https://doi.org/10.1289/ehp.1308015.
61. Bessonneau V, Pawliszyn J, Rappaport SM. 2017. The saliva exposome for monitoring of individuals' health trajectories. Environ Health Perspect 125(7):077014, PMID: 28743678, https://doi.org/10.1289/EHP1011.
62. Manuck TA, Lai Y, Ru H, Glover AV, Rager JE, Fry RC, et al. 2021. Metabolites from midtrimester plasma of pregnant patients at high risk for preterm birth. Am J Obstet Gynecol MFM 3(4):100393, PMID: 33991707, https://doi.org/10.1016/j.ajogmf.2021.100393.

63. Rager JE, Bangma J, Carberry C, Chao A, Grossman J, Lu K, et al. 2020. Review of the environmental prenatal exposome and its relationship to maternal and fetal health. Reprod Toxicol 98:1–12, PMID: 32061676, https://doi.org/10.1016/j.reprotox.2020.02.004.
64. Sobus JR, Wambaugh JF, Isaacs KK, Williams AJ, McEachran AD, Richard AM, et al. 2018. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J Expo Sci Environ Epidemiol 28(5):411–426, PMID: 29288256, https://doi.org/10.1038/s41370-017-0012-y.
65. Xue J, Lai Y, Liu CW, Ru H. 2019. Towards mass spectrometry-based chemical exposome: current approaches, challenges, and future directions. Toxics 7(3):41, PMID: 31426576, https://doi.org/10.3390/toxics7030041.
66. Jones DP. 2016. Sequencing the exposome: a call to action. Toxicol Rep 3:29–45, PMID: 26722641, https://doi.org/10.1016/j.toxrep.2015.11.009.
67. Bauer JA, Devick KL, Bobb JF, Coull BA, Bellinger D, Benedetti C, et al. 2020. Associations of a metal mixture measured in multiple biomarkers with IQ: evidence from Italian adolescents living near ferroalloy industry. Environ Health Perspect 128(9):97002, PMID: 32897104, https://doi.org/10.1289/EHP6803.
68. Eaton JE, Juran BD, Atkinson EJ, Schlicht EM, Xie X, de Andrade M, et al. 2015. A comprehensive assessment of environmental exposures among 1000 North American patients with primary sclerosing cholangitis, with and without inflammatory bowel disease. Aliment Pharmacol Ther 41(10):980–990, PMID: 25783671, https://doi.org/10.1111/apt.13154.
69. Allen JG, Gale S, Zoeller RT, Spengler JD, Birnbaum L, McNeely E. 2016. PBDE flame retardants, thyroid disease, and menopausal status in U.S. women. Environ Health 15(1):60, PMID: 27215290, https://doi.org/10.1186/s12940-016-0141-0.
70. Leijs MM, Koppe JG, Olie K, van Aalderen WM, de Voogt P, ten Tusscher GW. 2009. Effects of dioxins, PCBs, and PBDEs on immunology and hematology in adolescents. Environ Sci Technol 43(20):7946–7951, PMID: 19921918, https://doi.org/10.1021/es901480f.
71. Lang IA, Galloway TS, Scarlett A, et al. 2008. Association of urinary bisphenol a concentration with medical disorders and laboratory abnormalities in adults. JAMA 300(11):1303–1310, PMID: 18799442, https://doi.org/10.1001/jama.300.11.1303.
72. Didelez V, Sheehan N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res 16(4):309–330, PMID: 17715159, https://doi.org/10.1177/0962280206077743.
73. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. 2019. Genetic analyses of diverse populations improves discovery for complex traits. Nature 570(7762):514–518, PMID: 31217584, https://doi.org/10.1038/s41586-019-1310-4.
74. Burgess S, Butterworth AS, Thompson JR. 2016. Beyond mendelian randomization: how to interpret evidence of shared genetic predictors. J Clin Epidemiol 69:208–216, PMID: 26291580, https://doi.org/10.1016/j.jclinepi.2015.08.001.
75. Hemani G, Bowden J, Davey Smith G. 2018. Evaluating the potential role of pleiotropy in mendelian randomization studies. Hum Mol Genet 27(R2):R195–R208, PMID: 29771313, https://doi.org/10.1093/hmg/ddy163.
76. Cinelli C, LaPierre N, Hill BL, Sankararaman S, Eskin E. 2020. Robust Mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy. bioRxiv. Preprint posted online October 21, 2020. https://doi.org/10.1101/2020.10.21.347773.
77. VanderWeele TJ, Tchetgen EJ, Cornelis M, Kraft P. 2014. Methodological challenges in Mendelian randomization. Epidemiology 25(3):427–435, PMID: 24681576, https://doi.org/10.1097/EDE.0000000000000081.
78. Tobi EW, van Zwet EW, Lumey LH, Heijmans BT. 2018. Why mediation analysis trumps mendelian randomization in population epigenomics studies of the Dutch famine. bioRxiv 362392. Preprint posted online July 5, 2018. https://doi.org/10.1101/362392.
79. Burgess S, O'Donnell CJ, Gill D. 2020. Expressing results from a mendelian randomization analysis: separating results from inferences. JAMA Cardiol 6(1):7–8, PMID: 32965465, https://doi.org/10.1001/jamacardio.2020.4317.
80. Labrecque JA, Swanson SA. 2019. Interpretation and potential biases of mendelian randomization estimates with time-varying exposures. Am J Epidemiol 188(1):231–238, PMID: 30239571, https://doi.org/10.1093/aje/kwy204.
81. Cao Y, Rajan SS, Wei P. 2016. Mendelian randomization analysis of a time-varying exposure for binary disease outcomes using functional data analysis methods. Genet Epidemiol 40(8):744–755, PMID: 27813215, https://doi.org/10.1002/gepi.22013.
82. Liu G, Shi M, Mosley JD, Weng C, Zhang Y, Lee MTM, et al. 2021. A mendelian randomization approach using 3-HMG-Coenzyme-a reductase gene variation to evaluate the association of statin-induced low-density lipoprotein cholesterol lowering with noncardiovascular disease phenotypes. JAMA Netw Open 4(6):e2112820, PMID: 34097045, https://doi.org/10.1001/jamanetworkopen.2021.12820.
83. Zhang X, Theodoratou E, Li X, Farrington SM, Law PJ, Broderick P, et al. 2021. Genetically predicted physical activity levels are associated with lower colorectal cancer risk: a mendelian randomisation study. Br J Cancer 124(7):1330–1338, PMID: 33510439, https://doi.org/10.1038/s41416-020-01236-2.
84. Schooling CM, Johnson GD, Grassman J. 2019. Effects of blood lead on coronary artery disease and its risk factors: a Mendelian randomization study. Sci Rep 9(1):15995, PMID: 31690775, https://doi.org/10.1038/s41598-019-52482-1.
85. Lawlor DA, Tilling K, Davey Smith G. 2016. Triangulation in aetiological epidemiology. Int J Epidemiol 45(6):1866–1886, PMID: 28108528, https://doi.org/10.1093/ije/dyw314.
86. Burgess S, Thompson SG. 2015. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am J Epidemiol 181(4):251–260, PMID: 25632051, https://doi.org/10.1093/aje/kwu283.
87. Braun JM, Kalloo G, Kingsley SL, Li N. 2019. Using phenome-wide association studies to examine the effect of environmental exposures on human health. Environ Int 130:104877, PMID: 31200158, https://doi.org/10.1016/j.envint.2019.05.071.
88. Bush WS, Oetjens MT, Crawford DC. 2016. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat Rev Genet 17(3):129–145, PMID: 26875678, https://doi.org/10.1038/nrg.2015.36.
89. Zhou A, Morris HA, Hyppönen E. 2019. Health effects associated with serum calcium concentrations: evidence from MR-PheWAS analysis in UK Biobank. Osteoporos Int 30(11):2343–2348, PMID: 31392400, https://doi.org/10.1007/s00198-019-05118-z.
90. Jackson RD, LaCroix AZ, Gass M, Wallace RB, Robbins J, Lewis CE, et al. 2006. Calcium plus vitamin D supplementation and the risk of fractures. N Engl J Med 354(7):669–683, PMID: 16481635, https://doi.org/10.1056/NEJMoa055218.
91. Pendergrass SA, Buyske S, Jeff JM, Frase A, Dudek S, Bradford Y, et al. 2019. A phenome-wide association study (PheWAS) in the Population Architecture using Genomics and Epidemiology (PAGE) study reveals potential pleiotropy in African Americans. PLoS One 14(12):e0226771, PMID: 31891604, https://doi.org/10.1371/journal.pone.0226771.
92. Hebbring SJ. 2014. The challenges, advantages and future of phenome-wide association studies. Immunology 141(2):157–165, PMID: 24147732, https://doi.org/10.1111/imm.12195.
93. Pendergrass SA, Crawford DC. 2019. Using electronic health records to generate phenotypes for research. Curr Protoc Hum Genet 100(1):e80, PMID: 30516347, https://doi.org/10.1002/cphg.80.
94. Zhou W, Kanai M, Wu K-HH, et al. 2021. Global biobank meta-analysis initiative: powering genetic discovery across human diseases. medrxiv. Preprint posted online November 19, 2021. https://doi.org/10.1101/2021.11.19.21266436.
95. Sun D, Tiedt S, Yu B, Jian X, Gottesman RF, Mosley TH, et al. 2019. A prospective study of serum metabolites and risk of ischemic stroke. Neurology 92(16):e1890–e1898, PMID: 30867269, https://doi.org/10.1212/WNL.0000000000007279.
96. Cai T, Cai TT, Zhang A. 2016. Structured matrix completion with applications to genomic data integration. J Am Stat Assoc 111(514):621–633, PMID: 28042188, https://doi.org/10.1080/01621459.2015.1021005.
97. Voillet V, Besse P, Liaubet L, San Cristobal M, González I. 2016. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. BMC Bioinformatics 17(1):402, PMID: 27716030, https://doi.org/10.1186/s12859-016-1273-5.
98. Lin D, Zhang J, Li J, Xu C, Deng HW, Wang YP. 2016. An integrative imputation method based on multi-omics datasets. BMC Bioinformatics 17:247, PMID: 27329642, https://doi.org/10.1186/s12859-016-1122-6.
99. Diao G, Lin DY. 2006. Semiparametric variance-component models for linkage and association analyses of censored trait data. Genet Epidemiol 30(7):570–581, PMID: 16858699, https://doi.org/10.1002/gepi.20168.
100. Epstein MP, Lin X, Boehnke M. 2003. A tobit variance-component method for linkage analysis of censored trait data. Am J Hum Genet 72(3):611–620, PMID: 12587095, https://doi.org/10.1086/367924.
101. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. 2018. Missing value imputation approach for mass spectrometry-based metabolomics data. Sci Rep 8(1):663, PMID: 29330539, https://doi.org/10.1038/s41598-017-19120-0.
102. Schisterman EF, Vexler A, Whitcomb BW, Liu A. 2006. The limitations due to exposure detection limits for regression models. Am J Epidemiol 163(4):374–383, PMID: 16394206, https://doi.org/10.1093/aje/kwj039.
103. Lin DY, Zeng D, Couper D. 2020. A general framework for integrative analysis of incomplete multiomics data. Genet Epidemiol 44(7):646–664, PMID: 32691502, https://doi.org/10.1002/gepi.22328.
104. The ARIC Investigators. 1989. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. The ARIC investigators. Am J Epidemiol 129(4):687–702, PMID: 2646917.
105. Feofanova EV, Chen H, Dai Y, Jia P, Grove ML, Morrison AC, et al. 2020. A genome-wide association study discovers 46 loci of the human metabolome in the Hispanic community health study/study of Latinos. Am J Hum Genet 107(5):849–863, PMID: 33031748, https://doi.org/10.1016/j.ajhg.2020.09.003.
106. Lash TL. 2017. The harm done to reproducibility by the culture of null hypothesis significance testing. Am J Epidemiol 186(6):627–635, PMID: 28938715, https://doi.org/10.1093/aje/kwx261.

107. Weinberg CR. 2017. Invited commentary: can issues with reproducibility in science be blamed on hypothesis testing? Am J Epidemiol 186(6):636–638, PMID: 28938713, https://doi.org/10.1093/aje/kwx258.
108. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. 2007. Replicating genotype-phenotype associations. Nature 447(7145):655–660, PMID: 17554299, https://doi.org/10.1038/447655a.
109. Ng E, Lind PM, Lindgren C, Ingelsson E, Mahajan A, Morris A, et al. 2015. Genome-wide association study of toxic metals and trace elements reveals novel associations. Hum Mol Genet 24(16):4739–4745, PMID: 26025379, https://doi.org/10.1093/hmg/ddv190.
110. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, de Leeuw CA, et al. 2018. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet 50(7):912–919, PMID: 29942086, https://doi.org/10.1038/s41588-018-0152-6.
111. Davies G, Lam M, Harris SE, et al. 2018. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. Nat Commun 9(1):2098, PMID: 29844566, https://doi.org/10.1038/s41467-018-04362-x.
112. Alfano R, Chadeau-Hyam M, Ghantous A, Keski-Rahkonen P, Chatzi L, Perez AE, et al. 2020. A multi-omic analysis of birthweight in newborn cord blood reveals new underlying mechanisms related to cholesterol metabolism. Metabolism 110:154292, PMID: 32553738, https://doi.org/10.1016/j.metabol.2020.154292.
113. Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H. 2019. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. NPJ Syst Biol Appl 5:22, PMID: 31312515, https://doi.org/10.1038/s41540-019-0099-y.
114. Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. 2019. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites 9(4):76, PMID: 31003499, https://doi.org/10.3390/metabo9040076.
115. Larras F, Billoir E, Baillard V, Siberchicot A, Scholz S, Wubet T, et al. 2018. DRomics: a turnkey tool to support the use of the dose-response framework for omics data in ecological risk assessment. Environ Sci Technol 52(24):14461–14468, PMID: 30444611, https://doi.org/10.1021/acs.est.8b04752.
116. Smetanová S, Riedl J, Zitzkat D, Altenburger R, Busch W. 2015. High-throughput concentration-response analysis for omics datasets. Environ Toxicol Chem 34(9):2167–2180, PMID: 25900799, https://doi.org/10.1002/etc.3025.
117. Laine JE, Bodinier B, Robinson O, Plusquin M, Scalbert A, Keski-Rahkonen P, et al. 2020. Prenatal exposure to multiple air pollutants, mediating molecular mechanisms, and shifts in birthweight. Environ Sci Technol 54(22):14502–14513, PMID: 33124810, https://doi.org/10.1021/acs.est.0c02657.
118. Hasin Y, Seldin M, Lusis A. 2017. Multi-omics approaches to disease. Genome Biol 18(1):83, PMID: 28476144, https://doi.org/10.1186/s13059-017-1215-1.
119. Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. 2018. Leveraging multilayered "omics" data for atopic dermatitis: a road map to precision medicine. Front Immunol 9:2727, PMID: 30631320, https://doi.org/10.3389/fimmu.2018.02727.
120. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48(3):245–252, PMID: 26854917, https://doi.org/10.1038/ng.3506.
121. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47(9):1091–1098, PMID: 26258848, https://doi.org/10.1038/ng.3367.
122. Rawlik K, Rowlatt A, Tenesa A. 2016. Imputation of DNA methylation levels in the brain implicates a risk factor for Parkinson's disease. Genetics 204(2):771–781, PMID: 27466229, https://doi.org/10.1534/genetics.115.185967.
123. Gamazon ER, Zwinderman AH, Cox NJ, Denys D, Derks EM. 2019. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. Nat Genet 51(6):933–940, PMID: 31086352, https://doi.org/10.1038/s41588-019-0409-8.
124. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. 2017. Phenome-wide heritability analysis of the UK Biobank. PLoS Genet 13(4):e1006711, PMID: 28388634, https://doi.org/10.1371/journal.pgen.1006711.
125. Warrington NM, Zhu G, Dy V, Heath AC, Madden PAF, Hemani G, et al. 2015. Genome-wide association study of blood lead shows multiple associations near ALAD. Hum Mol Genet 24(13):3871–3879, PMID: 25820613, https://doi.org/10.1093/hmg/ddv112.
126. Hagenbeek FA, Pool R, van Dongen J, Draisma HHM, Jan Hottenga J, Willemsen G, et al. 2020. Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. Nat Commun 11(1):39, PMID: 31911595, https://doi.org/10.1038/s41467-019-13770-6.
127. Xu K, Li B, McGinnis KA, Vickers-Smith R, Dao C, Sun N, et al. 2020. Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. Nat Commun 11(1):5302, PMID: 33082346, https://doi.org/10.1038/s41467-020-18489-3.
128. Cole JB, Florez JC, Hirschhorn JN. 2020. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. Nat Commun 11(1):1467, PMID: 32193382, https://doi.org/10.1038/s41467-020-15193-0.
129. Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. Nature 538(7624):161–164, PMID: 27734877, https://doi.org/10.1038/538161a.
130. Kim J, Ziyatdinov A, Laville V, Hu FB, Rimm E, Kraft P, et al. 2019. Joint analysis of multiple interaction parameters in genetic association studies. Genetics 211(2):483–494, PMID: 30578273, https://doi.org/10.1534/genetics.118.301394.
131. Moore R, Casale FP, Jan Bonder M, Horta D, Franke L, Barroso I, et al. 2019. A linear mixed-model approach to study multivariate gene-environment interactions. Nat Genet 51(1):180–186, PMID: 30478441, https://doi.org/10.1038/s41588-018-0271-0.
132. Thomas D. 2010. Gene–environment-wide association studies: emerging approaches. Nat Rev Genet 11(4):259–272, PMID: 20212493, https://doi.org/10.1038/nrg2764.
133. Cui Y, Balshaw DM, Kwok RK, Thompson CL, Collman GW, Birnbaum LS. 2016. The exposome: embracing the complexity for discovery in environmental health. Environ Health Perspect 124(8):A137–A140, PMID: 27479988, https://doi.org/10.1289/EHP412.