Trends in
Endocrinology & Metabolism

CellPress

Feature Review

An epidemiological introduction to human metabolomic investigations

Amit D. Joshi,[1] Ali Rahnavard,[2] Priyadarshini Kachroo,[3] Kevin M. Mendez,[3] Wayne Lawrence,[4] Sachelly Julián-Serrano,[4,5] Xinwei Hua,[1,6] Harriett Fuller,[7] Nasa Sinnott-Armstrong,[7] Fred K. Tabung,[8] Katherine H. Shutta,[3,9] Laura M. Raffield,[10] Burcu F. Darst[7,*]

Metabolomics holds great promise for uncovering insights around biological processes impacting disease in human epidemiological studies. Metabolites can be measured across biological samples, including plasma, serum, saliva, urine, stool, and whole organs and tissues, offering a means to characterize metabolic processes relevant to disease etiology and traits of interest. Metabolomic epidemiology studies face unique challenges, such as identifying metabolites from targeted and untargeted assays, defining standards for quality control, harmonizing results across platforms that often capture different metabolites, and developing statistical methods for high-dimensional and correlated metabolomic data. In this review, we introduce metabolomic epidemiology to the broader scientific community, discuss opportunities and challenges presented by these studies, and highlight emerging innovations that hold promise to uncover new biological insights.

Highlights

The rapidly emerging field of metabolomic epidemiology presents unique opportunities to gain mechanistic insights into disease risk and identify biomarkers that may inform prevention and screening strategies.

Challenges include dealing with batch effects and drift, the sensitivity of metabolites to environmental exposures and the handling and processing of samples, metabolite identification, harmonizing metabolites across different platforms, analyzing high-dimensional data, the complex correlation structure of metabolites, and integrating metabolomics with other 'omic data types.

We provide a broad introduction to the field for scientists from cross-disciplinary backgrounds, discussing technological, study design, quality control, and statistical considerations, opportunities and challenges in the field, and emerging innovations that hold promise to uncover new biological insights.

Emerging field of metabolomic epidemiology

Metabolites (see Glossary) are small molecules (≤1.5 kDa) involved in the complex set of biochemical reactions that comprise the metabolism of an organism [1]. **Metabolomics**, the study of these molecules, is an emerging field undergoing rapid growth, particularly in its application to epidemiological research to gain unique insights into health-related conditions [2–4], presenting opportunities for improved exposure characterization and biomarker discovery for disease risk and prognosis [5,6]. Metabolites reflect **endogenous** processes, including inherited genetic variation and transcriptional and translational regulation, and the impact of **exogenous** or environmental exposures on these processes; consequently, they are uniquely suited to assess response to dietary, lifestyle, and other environmental factors [6,7] and to serve as preclinical biomarkers of disease outcomes [8]. The potential for metabolomics to provide insights into physiological and pathophysiological processes (Figure 1) has generated remarkable interest among clinical and epidemiological researchers [1] and accelerated the development of methods and statistical tools to address the unique challenges encountered in the acquisition and analysis of metabolomic data.

Inspired by the success of large-scale consortia in genetic epidemiology research, the US National Cancer Institute led an initiative to foster large-scale collaborative research on the human metabolome by creating the COnsortium of METabolomics Studies (COMETS) of 47 worldwide cohorts with blood metabolomic data from over 136 000 individuals [1]. Large-scale consortia, such as COMETS, bring together diverse studies with wide ranges of exposures, considerably improving power to discover and replicate metabolomic associations and advance our understanding of health and disease. COMETS and other initiatives, including the Trans-Omics for Precision Medicine (TOPMed) Program [9], also provide opportunities for multi-omic data

[1]Clinical & Translational Epidemiology Unit, Massachusetts General Hospital, Boston, MA, USA
[2]Computational Biology Institute, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA
[3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[5]Department of Public Health, University of Massachusetts Lowell, Lowell, MA, USA
[6]Department of Cardiology, Peking University Third Hospital, Beijing, China

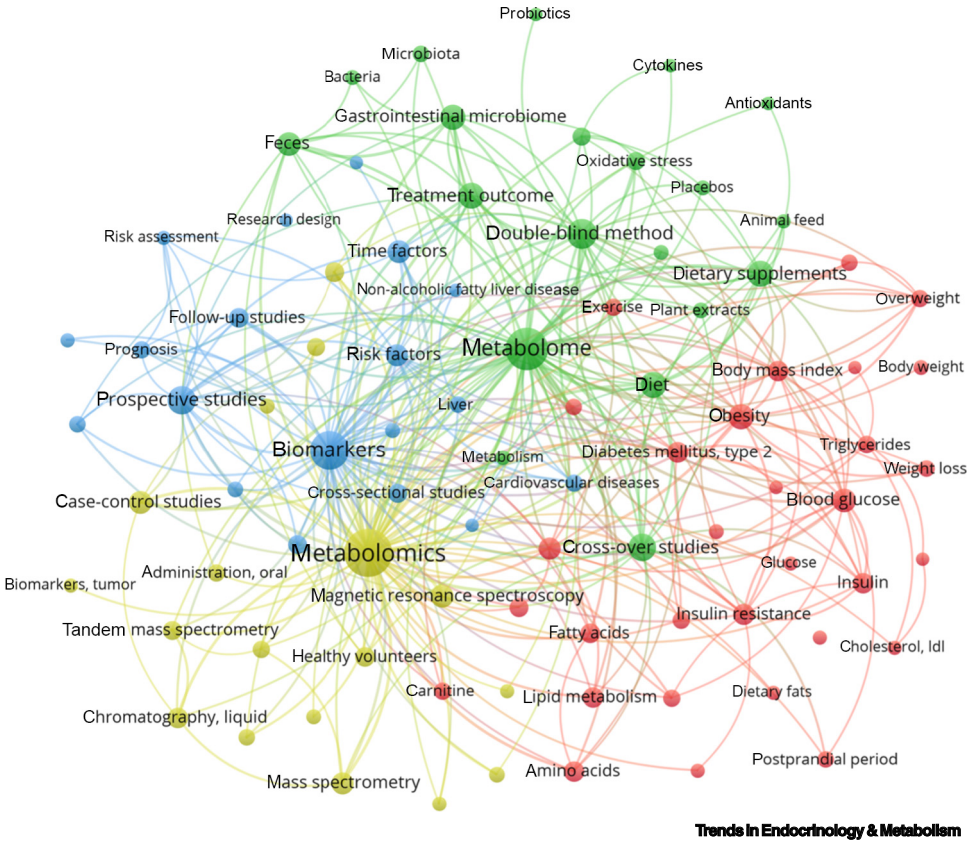

Figure 1. Network analysis of scientific topics that co-occur with 'metabolomics' in PubMed. Data were collected from PubMed using 'metabolomics' as the search keyword and filtering for co-occurring scientific keywords. The number of co-occurrences was used as the edge weight between keywords. A total of 1542 publications, including clinical studies (Phases I–V) and randomized control trials were used in this analysis. From these, 6283 keywords were extracted and 102 keywords with 30 co-occurrences were used to generate the figure using VOSviewer. Each node is a keywork and each edge represents the weight of the co-occurrence between keywords. Colors represent clusters of relevant keywords. Abbreviation: LDL, low-density lipoprotein.

integration (e.g., genetics, transcriptomics, epigenomics, and metabolomics). However, analytical challenges need to be addressed when evaluating metabolomic data generated across different cohorts, time periods, and platforms, and when integrating metabolomics with other 'omic data types. In this review, we provide a broad introduction to the field of **metabolomic epidemiology** for scientists from cross-disciplinary and diverse backgrounds, with a particular focus on analytical strategies, current challenges, and future directions.

Metabolomic technologies: choosing a metabolomics platform

Most metabolomic platforms can be categorized as either **targeted** or **untargeted** [5]. Targeted platforms measure a prespecified set of metabolites, typically selected in a hypothesis-driven fashion based on existing literature. Untargeted platforms profile hundreds to thousands of metabolites in a sample, with the chemical identity of many of these metabolites often unknown. Such metabolites may be referred to as **unidentified** or **unknown metabolites**. Many researchers now take a semitargeted approach, combining multiple technologies to optimize metabolite characterization.

Several techniques can be used to generate metabolomic data; however, proton **nuclear magnetic resonance** (NMR) and **mass spectrometry** (MS) are the most common and versatile

[7]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[8]The Ohio State University College of Medicine and Comprehensive Cancer Center, Columbus, OH, USA
[9]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[10]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

*Correspondence:
bdarst@fredhutch.org (B.F. Darst).

(Figure 2). Both techniques measure metabolite concentrations in a high-throughput fashion, with unique strengths and limitations, as discussed at length in previous publications [10–13]. Briefly, while NMR typically uses a targeted approach and is limited to identifying metabolites at high concentrations, it provides **absolute quantification** and requires minimal sample preparation [2,10]. MS is highly sensitive, enabling the measurement of metabolites at low concentrations and increasing the feasibility of untargeted approaches; however, MS only provides **relative quantification** of metabolites [2,10]. Given the complementary features of NMR and MS, combining NMR with techniques such as gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS) has been used to capture the metabolome more comprehensively [14–16]. In a survey of 47 studies from COMETS, 55% reported acquiring metabolomic data from untargeted platforms (predominantly MS), 18% were from targeted platforms (predominantly NMR), and

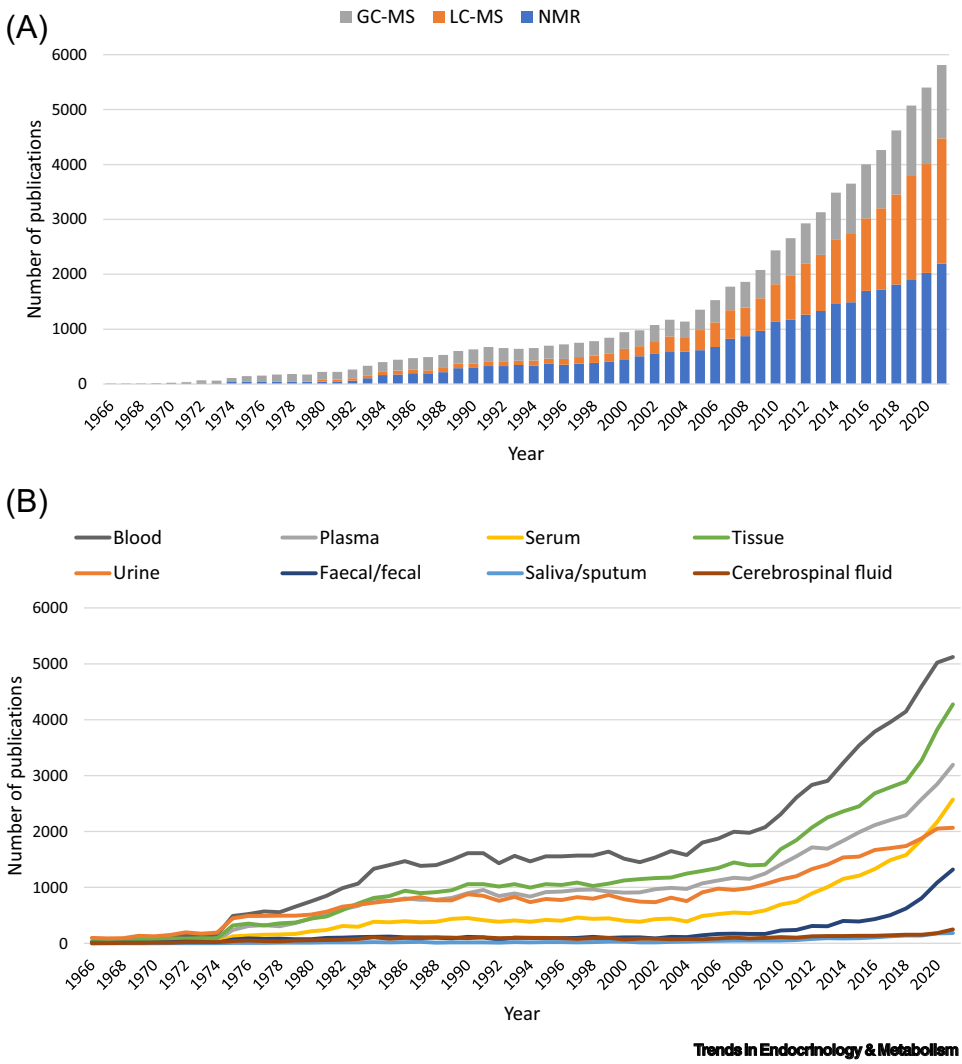

Figure 2. Distribution of metabolomic studies over time by (A) platform and (B) tissue type. PubMed search terms used for (A) included: metabolite*, metabolom*, or metabonom* with the key term NMR, GC-MS, or LC-MS. PubMed search terms used for (B) included: metabolite*, metabolom* or metabonom* with the key term urin*, plasma, serum, saliva/sputum, faecal/fecal, cerebrospinal fluid/CSF, or tissue. Years included are 1966–2021 with searches conducted on November 11, 2022. Abbreviations: GC, gas chromatography; MS, mass spectrometry; NMR, nuclear magnetic resonance.

Glossary

Absolute quantification: measured metabolite concentrations that are determined using the standard curve method. Concentrations can be measured as micromoles (μmol) or millimoles (mmol).
Batch effects: nonbiologically driven variation resulting from processing samples separately in batches.
Confounders: variables that influence both the exposure and outcome and can lead to over- or underestimation of the true association between the exposure and outcome.
Correlation network: undirected graphical model in which nodes represent variables and edges represent correlations between these variables.
Dimensionality reduction: representation of high-dimensional data in fewer dimensions, often applied to highly correlated data, such as metabolomic profiles to mitigate multicollinearity.
Drift: in metabolomics, drift is defined as higher stochastic variation in metabolomic profiles, which can result from temporal sample deterioration or experimental changes.
Endogenous: occurring or originating internally within an organism.
Exogenous: occurring or originating externally outside of an organism.
False discovery rate (FDR): expected proportion of discoveries in a statistical analysis that are incorrect, with the proportion being determined ahead of conducting analyses.
Family-wise error rate (FWER): probability of at least one false positive result when multiple statistical tests are conducted.
Fixed effects meta-analysis: used to estimates a single constant (or 'fixed') effect when associations are assumed to be true for all studies (e.g., assuming no heterogeneity exists between included studies).
Gaussian graphical model (GGM): undirected graphical model in which nodes correspond to variables of interest (such as metabolites) and weighted edges correspond to partial correlations between these variables, where the partial correlation between two variables is conditional on all other variables in the model.
Heteroskedasticity: variance of a variable is not constant across the values of another variable.
Internal standards: compounds that are added to samples to calibrate

27% were from a combination of targeted and untargeted platforms [17]. Furthermore, ionization techniques, such as electrospray ionization (ESI), can be used with MS to operate positive and negative ion run modes, which provides a more comprehensive analysis by enhancing the range of detectable compounds, accommodating variations in the physical and chemical properties of metabolites [18,19].

Quality control and data curation in epidemiological metabolomic studies

Factors affecting metabolite measurements

When designing epidemiological metabolomic studies, one should consider factors that could add noise to metabolite measurements [20], including collection methods (e.g., EDTA versus heparin tubes for plasma samples), lot numbers on sample tubes that indicate groups of tubes manufactured at the same time, run mode (see 'Metabolomic technologies: choosing a metabolomics platform' section), time of sample draw (e.g., season/time of day), participant fasting duration, sample storage duration and conditions, freeze/thaw cycles, and batch (i.e., which samples are processed simultaneously on the same platform), the latter of which is particularly important for MS platforms. Furthermore, metabolite associations can be impacted by **confounders**, including demographics (e.g., age at sample draw, sex, and body anthropometry) and environmental exposure (e.g., diet, alcohol, smoking, physical activity, and air pollution; Table 1) [6,21–29]. Metabolites can also exhibit broad variation in their stability [30,31] (see 'Metabolite stability and reproducibility' section). Gvien that a metabolomic profile is a cross-sectional snapshot of the metabolites present in the sample at a given time, the dynamics of the relationship between the exposure of interest and the observed metabolites could depend on the amount of time between the exposure and the sample collection: for example, a metabolite that has high abundance in an immediate response to an exposure may be so unstable that it appears at low abundance in the subsequently collected sample.

Consideration of these factors at the study design stage and during statistical analyses can help reduce their potential impact on metabolomic measurements and findings and provide an understanding of the limitations of findings.

Batch effects and drift correction

Metabolomic **drift**, defined as higher stochastic variation in metabolomic profiles, can result from temporal sample deterioration or experimental changes (e.g., nonlinear fluctuations in retention times and column aging), whereas **batch effects** are nonbiologically driven variation resulting from processing samples separately [32]. An approach for drift correction includes the use of **internal standards**, which are known metabolites added to samples to quantify drift (Figure 3) [33]. For MS platforms, study samples are ideally processed simultaneously to reduce batch effects. When samples are processed in multiple batches, experimental and computational techniques are required to ensure data quality. For example, pooled reference samples that capture all anticipated metabolites can be included across all batches to correct for measurement differences between batches as well as for drift (Figure 3). Metabolite intensities can be adjusted accordingly to account for drift and batch effects based on comparisons with internal standards and pooled reference samples using techniques such as the nearest neighbor algorithm [34]. It is crucial that samples be randomly distributed across batches in a balanced manner, ensuring that cases and controls (or any potential confounding or noise-increasing features, particularly those highlighted in the 'Factors affecting metabolite measurements' section) are included in each batch to minimize bias.

Once data are generated, known and unknown batch effects and potential noise can be adjusted for using statistical approaches, such as principal component analysis (PCA) and

metabolite measurements and help account for variation that occurs during sample preparation and between batches.
Latent groups: previously unknown subgroups of individuals who share characteristics based on a set of variables.
Limit of detection (LOD): lower limit at which metabolomic platforms are able to quantify metabolite levels.
Mass spectrometry (MS): technique used to measure compounds at high or low concentrations by measuring the mass-to-charge ratio of compounds.
Mega-analysis: statistical method to estimate the effect of an exposure on an outcome by combining individual-level data across studies in a pooled analysis, adjusting for study and other covariates.
Mendelian randomization (MR): statistical method to interrogate the causal effect of a modifiable exposure on an outcome using genetic variants as a proxy for the exposure.
Meta-analysis: statistical method to estimate the effect of an exposure on an outcome by combining summary statistics across studies without the need for individual-level data.
Metabolite harmonization: process of identifying the same metabolite measured on different platforms and/or adjusting metabolite measurements to be comparable between platforms, studies, or batches.
Metabolites: small molecules (≤1.5 kDa) that are involved in the complex set of biochemical reactions comprising the metabolism of an organism.
Metabolomic epidemiology: study of the human metabolome with regards to health-related outcomes or exposures in population-based epidemiologic investigations.
Metabolomics: systematic large-scale study of metabolites.
Nuclear magnetic resonance (NMR): technique used to measure compounds at high concentrations that uses a strong magnetic field and radio waves to identify compounds based on their resonance signal.
One-sample Mendelian randomization: MR conducted using individual-level data from a population in which both the exposure and outcome have been measured.
Polygenic scores (PGS): estimate of an individual's genetic predisposition to a condition or trait estimated by aggregating the effect of many genetic variants.

probabilistic estimation of expression residuals (PEER) factors [35], the latter of which is commonly used in gene expression association studies. PCA and PEER factors are dimensionality reduction techniques that calculate latent factors to account for potential batch effects or noise and can be adjusted for in subsequent analyses. The surrogate variable analysis 'sva' R package offers options to adjust for unknown sources of variation by estimating surrogate variables and known batch effects using ComBat, which is based on an empirical Bayesian framework [36–38]. The Covariates for Multi-phenotype Studies (CMS) approach has been proposed to account for latent batch effects while also utilizing the correlation structure of phenotypic data to increase power to detect associations of phenotypes with genetic factors [39,40]. When batch information is known, batch can be incorporated as covariates in statistical models; however, depending on the study design and magnitude of batch effects, they may still be challenging to overcome or significantly reduce power, warranting modeling of unknown batch effects.

Metabolite missingness and imputation

Missingness in epidemiological metabolomic data is common [1,17] and often due to technical challenges, including the **limit of detection** (LOD)/quantification of the platform, quality control issues, low metabolite abundance, and rare metabolites (i.e., metabolites found only in a subset of individuals). In data obtained from MS platforms, it is often assumed that missing measurements are below the LOD; as such, missing metabolite values are imputed to a fraction (often ½) of the minimum observed value for that metabolite among other participants in the sample. NMR data are often assumed to be missing at random, and one approach for handling missing NMR data is to impute missing values to the mean value for that metabolite [41]. However, if missingness is abundant, one could formally evaluate whether it was indeed at random with respect to the variables of interest or those that could introduce noise or confounding. Metabolite measurements can also be converted into an indicator variable denoting missing or present [17] or such an indicator variable can be adjusted for as a covariate in association tests [42]. Other approaches exploit the high degree of correlations observed in metabolomic data to impute missing values using analytical approaches, such as K-nearest neighbor imputation (KNN), multiple imputation by chained equations (MICE), Markov chain Monte Carlo (MCMC), PCA, or random forest imputation [43–45]. Imputation of missing metabolites may also be feasible from correlations across other assays, such as microbiome and genetic data, both of which contribute substantially to variance in many metabolites [46]. However, metabolites with a large proportion of missing values could reflect measurement concerns, in which case imputation could result in low information content and excluding the metabolite may be more appropriate. High missingness can also occur in exogenous metabolites (e.g., xenobiotics or drugs) and, thus, imputing to zero or evaluating as 'present' or 'absent' may be appropriate. Some studies exclude metabolites with a percent of missingness above a threshold, which can range from 5% to 90% [17], suggesting the need for more consistent imputation methods.

Centering, standardizing, and transforming metabolite levels

Data normalization of metabolomic data is crucial to implement before statistical analyses and includes three key steps: transformation, centering, and scaling [47]. Transformations may be necessary to correct for **heteroskedasticity** or skewed distributions, with log transformation and inverse normal transformation being common variance stabilization approaches. The most common approach to centering and scaling (referred to as standardization) is autoscaling (unit variance scaling), which results in each feature having a mean of zero and standard deviation of one. If batch information is known, standardization can be done within a batch, such that each has a comparable mean and standard deviation.

Postprandial metabolism: period of metabolism after food consumption.
Pleiotropy: when a genetic variant or gene impact more than one unrelated phenotypic traits.
Random effects meta-analysis: used to estimate the average variance of an association across studies when associations are heterogeneous across studies.
Relative quantification: measured metabolite levels that are relative to a reference sample rather than their exact concentrations and lack formal units.
Similarity network: network in which an edge between two nodes represents their pairwise similarity.
Targeted metabolomic platforms: metabolomic platforms that measure a prespecified set of metabolites that are typically selected in a hypothesis-driven fashion based on existing literature.
Two-sample Mendelian randomization: MR conducted using GWAS summary statistics for an exposure and outcome from two independent populations.
Unknown or unidentified metabolites: metabolites the chemical identity of which is not known.
Untargeted metabolomic platforms: metabolomic platforms that measure hundreds to thousands of metabolites, agnostic to metabolite identity.

Table 1. Consistently reported associations between circulating metabolites and disease risk and exposures in prospective investigations

Exposure	Effect direction	Metabolite class	Metabolite	Refs
Disease outcome				
Cardiovascular disease	Positive	Aminoxides	Trimethylamine *N*-oxide (TMAO)	[139]
Cancer	Negative	Carbohydrates and carbohydrate conjugates	Glycerol	[140]
		Cholesterols	Low-density lipoproteins, very-low-density lipoproteins	[140]
	Positive	BCAAs	Valine, leucine, isoleucine, creatinine	[8,140]
		Cholines	Choline	[140,141]
		Hydroxy acids	Lactic acid	[140]
		Lipids	Lysophosphatidylcholine 20:4	[140]
Asthma	Negative	Steroids	DHEA-S	[142,143]
			Cortisone, cortisol	[142,143]
	Positive	Sphingolipids	Ceramide (C18:1)	[142]
		Fatty acids	Palmitoleic acid	[142]
COPD	Negative	Amino acids	3-Methyloxytyrosine, phenylalanine, valine, tyrosine, isoleucine, 3-(4-hydroxyphenyl) lactate, 2-methylbutyrylcarnitine (C5), alpha-hydroxyisovalerate	[144,145]
		Carbohydrates and carbohydrate conjugates	Fructose, lactate, mannose	[144]
		Lipids	7-Hoca, oleoylcarnitine, lathosterol, glycerol	[144]
		Nucleotide s	Pseudouridine, *N*2,*N*2-dimethylguanosine	[144]
		Xenobiotics	Theophylline	[144]
	Positive	Amino acids	5-Oxoproline, hydrocinnamate, glutamine, asparagine, *N*-acetylglycine, glycine	[144]
		Carbohydrates	Glycerate	[144]
		Cofactors and vitamins	Biliverdin	[144]
Type 2 diabetes mellitus	Positive	Amino acids	Leucine, valine, tyrosine, phenylalanine	[146–148]
Alzheimer's disease	Negative	Branched-chain amino acids	Isoleucine, leucine, valine	[149]
	Positive	Amino acids	Glutamine	[149]
Exposure				
Age	Negative	Amino acids	Histidine, creatinine, 4-methyl-2-oxopentanoate, 3-methyl-2-oxovalerate, leucine serine, tryptophan	[150–152]
		Glycerophospholipids	PC ae C42:4, PC ae C42:5, PC ae C44:4	[150]
		Nucleotide	Uridine	[21,151]
	Positive	Amino acids	Glutamine, tyrosine, trans-4-hydroxyproline, kynurenine, ornithine, dimethylarginine, citrulline, *N*-acetyl alanine, *N*-acetyl glycine, *N*-acetyl threonine, urea, 4-acetamidobutanoate	[21,151]
		Carbohydrate	Erythronate, glycerate, glucose, arabinose, mannose	[21,151]
		Cofactors and vitamins	Alpha-tocopherol, pantothenate, biliverdin, pyridoxate, threonate	[21,151]
		Fatty acids	DHA; 22:6n3, EPA; 20:5n3, n3 DPA; 22:5n3, CMPF, stearidonate (18:4n3), 10-heptadecenoate (17:1n7), linolenate [alpha or gamma; (18:3n3 or 6)], 10-nonadecenoate (19:1n9), glycerol, stearate (18:0), 2-hydroxypalmitate,	[21,151]

Table 1. (continued)

Exposure	Effect direction	Metabolite class	Metabolite	Refs
			nonadecanoate (19:0), palmitate (16:0), caprate (10:0), pentadecanoate (15:0), 5-dodecenoate (12:1n7), linoleate (18:2n6), myristate (14:0), myristoleate (14:1n5), dihomo-linoleate (20:2n6), palmitoleate (16:1n7), margarate (17:0), stearate (18:0)	
		Glycerophospholipids	PC aa C28:1	[150,151]
		Nucleotides	Pseudouridine, *N*1-methyladenosine, allantoin, urate	[21,151]
		Sphingolipids	SM C16:1, SM C18:1	[150]
BMI	Negative	Amino acids	Asparagine, glycine	[153–155]
	Positive	Amino acids	Phenylalanine, glutamate, tyrosine	[153–156]
		Branched-chain amino acids	Valine	[154,156]
		Carbohydrates and carbohydrate conjugates	Mannose	[154]
		Organic acids	Lactate	[154]
		Carnitines	Carnitine	[154,155]
Sex	Higher in females	Amino acids	Creatine	[21,153,157]
		Glycerophospholipids	PC aa C32:3, PC aa C28:1, PC aa C40:3, PC aa C30:2	[26]
		Lipids	Glycerol	[21,153,157]
		Sphingolipids	SM(OH) C22:2. SM C18:1, SM C20:2	[26]
	Higher in males	Amino acids	Isoleucine, leucine, creatinine, valine, glutamate, tryptophan	[21,26,157]
		Nucleotides	Urate	[21,157]
Smoking	Positive	Amino acids	Tryptophan	[153,158]
		Prenol lipids	Methanol-glucuronide	[159,160]
		Xenobiotics	Cotinine	[161,162]
Alcohol consumption	Negative	Amino acids	Creatinine	[163]
		Glycerophospholipids	PC ae C30:2, PC ae C36:2, PC ae 36:2, PC ae 38:3	[164]
		Sphingomyelins	SM(OH) C14:1, SM(OH) C16:1, SM(OH) C22:2	[164]
	Positive	Amino acids	Threonine, 2-amino butyrate, 2-hydroxybutyrate	[163]
		Glycerophospholipids	PC aa C32:1, PC aa C34:1	[164]

Statistical approaches for metabolomic analysis

The objective of epidemiological metabolomic studies typically involves elucidating metabolomic patterns that are: (i) signatures of prior exposures (e.g., smoking); (ii) predictive of disease risk/health outcomes (e.g., type 2 diabetes mellitus risk in prediagnostic samples); and (iii) prognostic of disease progression (e.g., colorectal cancer recurrence). The most common metabolomic analytic approaches consider each metabolite separately, while complementary approaches analyze multiple metabolites in conjunction, such as pathway analyses, or estimate summary measurements of overall variation in metabolomic data, such as PCA.

Generalized linear modeling

A standard approach to evaluate the association between a single metabolite and trait of interest in observational epidemiological studies is regression using the generalized linear model (GLM) family with appropriate link functions, such as identity for linear regression, logit for

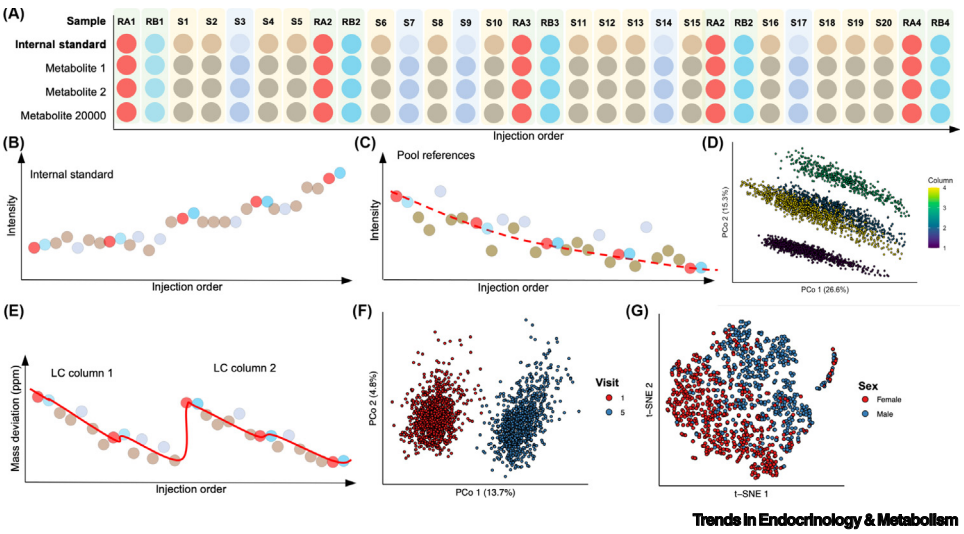

Figure 3. Drift and batch effect correction to increase the biology-to-noise ratio in metabolomic data. Liquid chromatography–mass spectrometry (LC-MS) is shown here as an exemplary technique. (A) We illustrate a common approach for metabolite sample quality control (QC) implemented in a study. Two pooled references for QC (column labels starting with 'RA' shown in red and 'RB' shown in blue) and internal standards (demonstrated by the first row) are used to capture trends introduced during the sequential processing of samples (column labels starting with 'S'). Sample colors represent varying metabolite intensities. Three unique metabolites are represented by rows two, three, and four. The *x*-axis represents the order in which the samples are injected to the ionization source (injection order). (B) Internal standards (also represented by row 1 of A) are specific metabolites that are added to samples as a baseline to adjust for intensity drift introduced by sample injection order. (C) Pooled references are aliquots of all samples used as a baseline for all metabolites we expect to see among samples and are used complementarily with internal standards to adjust for drift introduced by sample injection order. (D) Large studies require more than one column in LC-MS approaches to process samples and introduce a column effect combined with (E) injection order effect that need to be considered during batch effect correction. (F) Samples collected at two different time points can differ significantly and be a source of confounding. (G) After correcting metabolites for batch effects and sample collection time, sex differences can still be observed. Principal coordinate analysis (PCo) reflects the metabolite variation explained by PCo, and *t*-distributed stochastic neighbor embedding (t-SNE) is used to visually demonstrate the metabolite similarity between pairwise samples.

logistic regression, and log for Poisson regression. Prospective cohorts and randomized clinical trials generally use Cox proportional hazards modeling. The ability to account for potential confounding or technical variation makes multivariable regression approaches preferable over univariable approaches in observational studies. Some tools perform both data normalization and association testing, such as Microbiome Multivariable Associations with Linear Models (MaAsLin 2) [48].

Investigations of longitudinally measured metabolomic data are advantageous over single timepoint investigations given temporal variation in metabolites (see 'Metabolite stability and reproducibility' section) that occur with changing conditions and exposures, such as diurnal and seasonal variation, fasting versus postprandial variation, medication use, response to aging, and changes in body mass index (BMI) (Table 1). Obtaining measurements over time can account for some of these factors that may be unrelated to the outcome of interest and introduce noise to metabolite measurements. Approaches used to analyze longitudinal measurements require additional considerations due to the non-independence of the samples. While GLMs can be used to evaluate longitudinally measured metabolites that are combined into a single measurement for each participant, such as average metabolite levels or change in metabolite levels, generalized linear mixed models (GLMMs) utilize all observations by adding a random effect to account for the correlation structure of the repeated measurements [49].

Meta-analysis and mega-analysis

Combining metabolomic data across studies, platforms, and batches improves power to detect true associations. Advantages and disadvantages of **meta-analysis** (i.e., combining summary statistics) versus **mega-analysis** (i.e., combining individual level data in a pooled analysis, adjusting for batch and study covariates) should be considered. Meta-analysis allows studies to participate without sharing individual-level data, which are often subjected to extensive regulations regarding participant privacy. Meta-analysis may be more appropriate in the presence of platform heterogeneity, because it allows for joint analysis of the semi-quantitative data often produced by **untargeted metabolomic platforms** (where a given metabolite value may correspond to different concentrations across batches and studies). Mega-analysis may be preferred when data are generated using the same platform and methods at multiple timepoints (e.g., the UK Biobank metabolomics project; see 'Biobank metabolomics' section) or across different studies, because this approach enables the consistent consideration of covariates and facilitates investigations of interactions. In mega-analyses, accounting for batch effects and study-specific confounders is required for robust analysis.

Meta-analysis of metabolomic data is typically performed using **fixed effects** or **random effects** models. Fixed-effects models are recommended when associations are assumed to be consistent across studies, while random-effects models are recommended when associations are expected to be heterogeneous across studies [50]. The assumption of heterogeneity can be tested using statistics such as I^2, which describes the proportion of variation across study estimates attributable to heterogeneity [51]. Evaluating heterogeneity can be highly informative to understanding the relationship between an exposure and outcome, because it could point to sources of differences in associations, such as populations, genetic ancestries, environmental factors, and sampling strategies, among other potential factors. Mega-analysis uses typical regression approaches, but mixed models may be more appropriate if notable interstudy differences exist.

An important consideration when meta- or mega-analyzing metabolomic data between studies or platforms is the requirement to 'harmonize' (match) metabolites across studies when metabolites are annotated inconsistently (see 'Metabolite harmonization' section).

Dimensionality reduction methods

Dimensionality reduction methods are a broad class of statistical approaches designed to represent multivariate data in fewer dimensions, compressing data based on similarity across the originally measured variables. Such approaches are particularly useful for metabolomic data that are high-dimensional and multicollinear due to the high correlation between many metabolites. Dimensionality reduction is one way to potentially mitigate challenges such as overfitting (i.e., learning features specific to the training data that may not generalize to other data), a common risk of applying the popular approach of multivariate linear modeling to multicollinear metabolomic data. They are also useful when testing highly correlated metabolites independently, either as predictors or outcomes, because such approaches can reduce the number of tests conducted to help control for multiple testing (an example of this using PCA is provided in the 'Correction for multiple hypothesis testing' section) and they can lead to variables that may be more informative compared with a given individual metabolite. Common dimensionality reduction methods include: (i) unsupervised methods, such as PCA or factor analysis, where the algorithm is not informed of the outcome; (ii) supervised methods, such as partial least squares-discriminant analysis (PLS-DA) and the orthogonal projections to latent structures-discriminant analysis (OPLS-DA) [52]; and (iii) approaches developed within the GLM framework for dimensionality reduction via feature selection, the most commonly used being the least absolute shrinkage and

selection operator (LASSO) [53] and elastic net regression [54]. For instance, LASSO was applied to 100 NMR metabolites to identify 13 metabolites predictive of coronary heart disease (CHD) [55]. A metabolite score developed from these 13 metabolites was highly predictive of incident CHD in an independent cohort [55].

Other supervised learning algorithms may also help in dimensionality reduction, but few investigations have adopted them for metabolomic data analysis (e.g., k-nearest neighbors, naive Bayes classifiers, and decision trees, such as random forests and neural networks, reviewed with example R code in the freely available book *An Introduction to Statistical Learning* [56]).

Network and pathway enrichment analyses

Network and pathway enrichment analyses of metabolites are biologically informative and powerful tools. Network analyses refer to investigations based on graphical models constructed with metabolites as nodes and edges representing some type of pairwise relationship, such as correlation of abundance or a known metabolic reaction. Pathway enrichment analyses refer to explorations that integrate sets of metabolites, such as known metabolic pathways or groups of metabolites participating in the same cellular function, with observed data to highlight areas of enriched function based on data from multiple metabolites. Metabolites weakly associated with an outcome of interest may be part of important systems-level biological processes that can only be detected with a network or pathway lens.

Two popular tools for network analyses of metabolites are **correlation networks** and **Gaussian graphical models** (GGMs). These network models represent metabolites as nodes connected by edges indicating pairwise associations learned from the distribution of the data. Systems-level characteristics of a particular biological condition can be explored in these networks using node centrality measures, community detection, and other network science approaches [57,58]. Advantageously, both correlation networks and GGMs can be applied in cross-sectional data sets where temporality is unknown.

Correlation networks model metabolites as nodes connected by edges representing their correlations. A popular extension of correlation network analysis is Weighted Gene Correlation Network Analysis (WGCNA) [59]. WGCNA applies a soft thresholding approach to estimate weighted edges based on correlation between nodes. While WGCNA was initially proposed in the context of gene co-expression, it has been applied to metabolomic and proteomic studies with a modified protocol [60]. WGCNA can be applied to untargeted metabolomic analyses to identify clusters of intercorrelated metabolites and interrogate these clusters for pathway or metabolite class enrichment. A follow-up analysis involves calculating the 'eigenmetabolite' for each module (a score based on the first principal component of the module; called 'eigengene' in the original WGCNA paper) and investigating the association of the eigenmetabolite values with traits of interest in the study population [59,60]. For example, a study of aging and the healthy lifespan (healthspan) applied WGCNA to 2957 metabolites measured in 14 younger adults and 29 older adults to identify 20 metabolite modules, 18 of which had eigenmetabolites associated with healthspan and six of which had eigenmetabolites associated with aging [61]. Follow-up analyses of the metabolite classes enriched in these modules found that amino acid and lipid metabolites are of particular importance in the plasma metabolomic signature of aging and healthspan [61].

GGMs, also known as partial correlation networks, are undirected graphical models in which nodes correspond to variables of interest (such as metabolites) and weighted edges correspond to partial correlations between these variables. The partial correlation between two metabolites is a measure of a conditional Pearson correlation, conditioning on all other metabolites in the model

[62,63]. In the Gaussian setting, zero partial correlation corresponds to conditional independence; therefore, GGMs distinguish between direct and indirect dependence patterns among a set of metabolites [64]. A GGM is typically sparser than a correlation network, making GGM attractive for metabolomic data, where it is common to observe dense correlation networks due to high pairwise correlations between metabolites involved in the same biological processes [65]. GGMs have been shown to successfully characterize the roles of certain metabolites and construct biologically meaningful metabolite networks [66]. Tools implementing GGM estimation are the *huge* and *bootnet* R packages, both of which use a variety of GGM methods [62,67].

Pathway enrichment analyses can be classified into three categories: over-representation analysis, functional class scoring, and pathway topology-based approaches [68]. Several methodological advances have been developed to facilitate pathway enrichment analysis. For example, WGCNA [59] and omeClust [69] apply clustering algorithms to identify biologically meaningful groups of metabolites, which can be used to evaluate enrichment of metabolite pathways. Mummichog predicts functional activities by applying network analysis to untargeted metabolomics data to predict metabolic modules and pathways to which metabolites belong [70]. MetaboAnalyst and Chemical Similarity Enrichment Analysis (ChemRich) provide web-based interfaces to perform pathway enrichment and other statistical analyses [71,72]. *deepath* performs pathway enrichment analysis using previously calculated effect estimates [73], enabling adjustment of covariates via multivariable regression models. The potential for such approaches to elucidate biological function was demonstrated by an investigation that performed a pathway enrichment analysis with MetaboAnalyst, which implicated the glycerophospholipids pathway in age-related macular degeneration [74], a pathway that had been previously linked to Alzheimer's disease, another neurodegenerative condition. This finding led to a possible mechanism for glycerophospholipids in the pathology of age-related macular degeneration based on the role of these molecules in the eye.

One challenge specific to pathway enrichment analysis in the context of metabolomics is the lack of comprehensively annotated metabolites and metabolite pathways. To this end, data-driven pathways, such as the sets of metabolites identified from the network structure of a GGM [65] or the modules identified by WGCNA, can serve two purposes. First, they may provide additional confidence in expert-curated pathways or provide evidence that such pathways could be refuted. Second, they may identify novel pathways that have not yet been explored in the literature.

Multi-omic integration

Technological advancements enabling the generation of large-scale multi-omic data (e.g., genomics, epigenomics, transcriptomics, proteomics, and metabolomics) have provided opportunities to gain more comprehensive understandings of disease risk. Different 'omic data types provide complementary viewpoints of complex biological processes, with genomics representing upstream biological processes and metabolites representing downstream products of these biological systems as well as environmental influences [75]. Although many challenges exist to integrate multi-omic data [76–79], statistical methods are emerging, as described below, and offer a means to more comprehensively understand the factors that contribute to disease risk and trait variation.

Network, factor, and cluster analyses for multi-omic data

Network analyses can be extended beyond single-omic analyses (see 'Pathway and network analyses' section) to identify groups of biologically meaningful features across multi-omic data. This can provide insights into shared and distinct biological pathways across 'omic data types that may collectively impact disease risk. Examples of extensions of such network analyses include MiBiOmics, which builds on the WGCNA approach to link groups of variables across

multi-omic data sets to a trait of interest [80] and OmicsNet, which offers a web-based platform for creation and 3D visualization of biological networks of genes/proteins, transcription factors, miRNAs, and metabolites and is linked to publicly available molecular interaction databases [81].

Factor analyses decompose biological and technical variation in a data set by identifying latent factors explaining a large proportion of this variation. Each latent factor has 'loading' values for each of the multi-omic measurements, or features, in the data set that correlate with how much the feature contributes to the given factor. For instance, if a factor analysis is conducted with 100 features from a metabolomic and proteomic data set, one factor may be heavily loaded by ten metabolites and proteins while another factor is heavily loaded by five other metabolites and proteins. Conducting subsequent analyses with these factors could reveal the importance of the metabolites and proteins loading to each factor. An example of such a method is Multi-Omics Factor Analysis (MOFA), an unsupervised approach that identifies latent factors capturing major sources of variation across multi-omic data sets [82]. Resulting factors can be sparse, facilitating the identification of specific 'omic features contributing to a given factor. Furthermore, MOFA can impute missing values either within a particular 'omics assay or between 'omic assays for samples that are completely missing an 'omic data type.

Clustering of multi-omic data can identify groups of individuals with profiles associated with disease risk and distinct combinations of factors impacting disease risk. Latent Unknown Clustering with Integrated Data (LUCID) identifies **latent groups** of individuals, where each group has a unique profile differentially associated with the outcome of interest [83]. Similarity network fusion (SNF) creates a **similarity network** for each 'omic data type, where nodes correspond to individual samples and edges between two nodes represent a pairwise measure of how similar the two samples are (e.g., correlation) [84]. An iterative message-passing approach is then applied to integrate these networks into a single fused network. This resulting network can be used to identify clusters of individuals with similar 'omic profiles. An attractive feature of SNF is that the construction of similarity measures within each 'omic data type allows the user to effectively circumvent potential issues of differing scales between data types.

Polygenic scores and metabolomics

Genetic predisposition to diseases and traits are often measured using **polygenic scores** (PGS). PGS are calculated as a sum of genetic variants associated with a trait of interest, typically weighted by variant-specific effect estimates that indicate the magnitude of the association between the variant and the trait. In recent years, PGS have proven highly predictive of many conditions and traits, as documented in the publicly available PGS Catalog resource [85]. Integrating PGS with metabolomic and other 'omic data offers an opportunity to investigate biological mechanisms impacted by genetic predisposition to diseases and traits. For example, a recent proteomic study found that PGS of coronary artery disease, type 2 diabetes mellitus, ischemic stroke, and chronic kidney disease were associated with 49 proteins, many of which mediated the relationship between genetic risk and disease [86]. PGS have also been used to understand how metabolomic networks are dysregulated during disease progression by identifying presymptomatic metabolic alterations in disease-free individuals with high genetic risk and confirming the role of such alterations in individuals with disease [87]. Another study found that a BMI PGS was associated with 24 metabolites, including branched-chain amino acids (BCAAs), lipoprotein lipids, and inflammation-related glycoprotein acetyls [88].

Mendelian randomization

Another way in which metabolomics can be integrated with genomics is through **Mendelian randomization** (MR), which estimates causal associations between an exposure and outcome

using genetic variants as a proxy for an exposure, referred to as an 'instrumental variable'. A 'valid' instrumental variable is: (i) robustly associated with the exposure of interest, referred to as the 'relevance assumption'; (ii) not a confounder of the exposure–outcome association, referred to as the 'independence assumption'; and (iii) only impacts the outcome via the exposure and not alternative mechanisms, referred to as the 'exclusion restriction assumption' or the 'absence of **pleiotropy**' [89]. When an instrumental variable impacts the outcome variable beyond the exposure–outcome association, horizontal pleiotropy is present and MR estimates are biased. Instrumental variables can also be associated with traits downstream of the exposure along the causal pathway to the outcome, known as vertical (or 'spurious') pleiotropy, which is what MR seeks to identify [90]. In its simplest form, MR is performed by evaluating the association between a single instrumental variable and an outcome. However, the availability of high-dimensional 'omic data has led to the development of more complex statistical techniques that enable MR to be performed with multiple genetic variants and intermediates. MR can be performed using genome-wide association study (GWAS) summary statistics from both exposures (e.g., GWAS of metabolite levels) and outcomes (e.g., GWAS of BMI), referred to as '**two-sample MR**' (as opposed to **one-sample MR**, which uses individual-level data), using approaches such as the inverse-variance weighted (IVW) method [91]. In the presence of horizontal pleiotropy, approaches such as Egger regression (MR Egger) [92] and MR Pleiotropy Residual Sum and Outlier Detection (MR-PRESSO) [93] are appropriate choices to account for this pleiotropy. Given the increasing availability of GWAS summary statistics, MR has become a popular method to evaluate casual associations between exposures and outcomes. For example, a recent MR study evaluated putative causal effects of metabolites on 45 common diseases and found evidence for 30 metabolites having a causal effect, predominantly on risk of CHD and primary sclerosing cholangitis [94].

Correction for multiple hypothesis testing

Given the large number of metabolites that can be measured in metabolomic investigations, correction for multiple testing needs to be considered to control type 1 error rates [95–98]. The number of tests performed and commensurate with the elevated probability of reporting false positive results can be accounted for by: (i) adjusting the significance level to lower the probability of falsely rejecting the null hypothesis; or (ii) adjusting the *P*-value distribution itself. Two popular approaches involve controlling the **family-wise error rate** (FWER), with methods such as the Bonferroni correction [99], and controlling the **false discovery rate** (FDR), with methods such as the Benjamini–Hochberg procedure [98].

Bonferroni correction adjusts the nominal significance level by dividing it by the number of tests performed. It may be considered too conservative for metabolomic studies, because it assumes that all tests are independent (generally an incorrect assumption for highly correlated metabolomic data), potentially resulting in a high false negative rate. Adjustment could instead be based on the number of independent metabolites tested (i.e., accounting for correlation between metabolites), which could be determined by using the number of principal components accounting for >95% of the total variation in the metabolomic data [100] or by matrix spectral decomposition [101]. Another approach to correct for multiple testing in the setting of highly correlated metabolites is a permutation-based approach implemented by Westfall and Young [102], which obtains a distribution of minimum *P*-values given the data and sets the *P*-value level to an *a priori* determined cutoff (usually 5%).

The FDR is the expected proportion of discoveries that are incorrect. Controlling the FDR is typically a less conservative, more powerful approach compared with controlling the FWER, making FDR control an attractive option for metabolomic and other high-dimensional data sets

[96,103,104]. The proportion of false positives allowed is determined *a priori*, with 5% being commonly used. The p.adjust function in R implements the approaches described here and other multiple testing procedures.

Researchers often incorporate different forms of prior knowledge into multiple testing procedures to improve precision and understanding of results. Such approaches may include: (i) use of penalty weights; (ii) the use of prior weights; (iii) partitioning hypotheses into groups; and (iv) incorporating knowledge of the dependence structure of the data [105]. While most studies are able to apply only one or two strategies simultaneously, p-filter is an algorithm that provides a unified framework to integrate these four strategies while controlling for desired group and individual hypothesis FDR [105].

Challenges and future directions

Metabolite identification

A key challenge in the metabolomics field is identifying metabolites, particularly in untargeted MS experiments. This process includes matching MS spectrum acquired from biological samples to authentic compounds from established spectral libraries and is directly related to metabolite quantification, which is impacted by sample preparation, chromatin separation, and MS data acquisition [106]. Measurement error and limited spectral libraries pose challenges to identifying metabolites, and it is common for untargeted MS experiments to result in 'unknown' metabolites. One successful strategy uses GWAS to gain clues about unknown metabolites, because many top GWAS hits for metabolites tag genes encoding transporters or enzymes with known links to a given metabolite, allowing mapping of an unknown metabolite to a specific enzymatic pathway [107]. For example, an unknown metabolite from an untargeted LC/MS analysis associated with hepatic fat was identified as dimethylguanidino valerate based on a strong GWAS signal near the gene encoding the enzyme that produces it [108].

New methods are being developed to identify unknown metabolites, such as Metabolite annotation and Dysregulated Network Analysis (MetDNA), a metabolic reaction network-based algorithm that identifies metabolites based on their reaction-paired neighbor metabolites, which tend to have similar spectra [109]. Network-based approaches have been developed that directly utilize untargeted metabolite signatures to predict functional activity without the need for metabolite identification, such as the popular Mummichog [70] (see 'Network and pathway enrichment analyses' section). The MetaMap R package uses the Kyoto Encyclopedia of Genes and Genomes (KEGG) and PubChem databases to assess metabolite associations when the exact metabolite annotation is unknown [110].

Metabolite harmonization

The overlap and comparability of metabolites across platforms and studies is often limited, particularly due to differences in platform technologies and coverage of metabolites as well as different strategies used to quantify and identify metabolites. Identifying metabolites that are identical across platforms and adjusting metabolite levels so that they are comparable across platforms, studies, and batches (collectively referred to as '**metabolite harmonization**') can facilitate meta- and mega-analysis, both of which can improve power to detect true associations (see 'Meta-analysis and mega-analysis' section). However, metabolite harmonization remains a central challenge for large-scale metabolomic investigations and reproducibility. A COMETS investigation of >47 cohorts evaluated the overlap of five common platforms (Metabolon Inc., the Broad Institute's Metabolomics Platform, Biocrates, the West Coast Metabolomics Center, and Nightingale Health) and found modest overlap; for example, only ~10% of Metabolon-measured metabolites were matched to metabolites measured on the four other platforms based on unique

identifiers from the Human Metabolome Database (HMDB), PubChem, and other online databases [1]. COMETS also compared metabolite values on 40 duplicate samples assessed by the two most widely used metabolomic MS platforms at the time (Broad Institute and Metabolon); for the overlapping 111 metabolites (a small subset of those measured on each platform), the median Spearman correlation was 0.79 (interquartile range: 0.56–0.89), suggesting very good, but not perfect, concordance [1].

Replication of metabolic pathways that are associated with an outcome as opposed to exact metabolites may increase the utility of metabolomic data across platforms. Identification of unknown metabolites from untargeted MS approaches will also likely continue to improve with better and larger reference data sets and higher performance instrumentation, aiding harmonization across studies and improving identification rates [111]. Use of common reference/control samples and increasing consistency of metabolite naming and identification (see 'Metabolite identification' section) systems across platforms could also help improve cross-platform harmonization [112]. Efforts funded by the Common Fund of the US National Institutes of Health, including the Metabolomics Workbench [113], have helped standardize approaches for metabolomic quantification, identification, analysis, and visualization. This effort has also funded five Compound Identification Development Cores across the USA to increase rates of compound identification in a coordinated manner.

Beyond metabolite harmonization, other factors can impact the successful replication of metabolomic findings. Even across studies using the same platform, differences can result due to pre- and postanalytical processing, relative quantification in MS studies depending on metabolite distributions within studies, and population differences across studies [114] (see 'Metabolite-specific confounders and mediators' section). While replication is a key challenge in metabolomics, careful planning during the study design phase, including sample handling and processing in discovery and validation studies, using the same platform, and including common samples between studies to allow for measurement calibration (see 'Batch effects and drift correction' section), could maximize the discovery of replicable true findings.

Metabolite stability and reproducibility

With the increasing availability of longitudinal metabolomics data, it has become more feasible to assess the stability and reproducibility of commonly assayed metabolites. These measures can add confidence to exposure–outcome associations that are observed in cross-sectional data. For example, data from the Nurses' Health Study and the Health Professionals Follow-Up Study have been used to investigate interassay reproducibility, the stability of metabolites to a processing delay of 24 or 48 h, and within-person reproducibility of metabolite levels over 1 or 2 years, identifying a deleterious effect of processing delays on the measurements of carbohydrates and purine/pyrimidine derivatives [115]. The within-person stability of metabolites over a 10-year period has also recently been investigated in the Nurses' Health Study, identifying lipid, lipid-related, and polar metabolites as reasonably stable over the timescale of a decade [116]. Future efforts such as these will have an important role in assessing confidence in results from longitudinal metabolomics studies.

Metabolomics and diet

Given the influence of dietary factors on the metabolome, metabolomics has the potential to inform nutritional epidemiology. While most studies aim to investigate the fasting metabolome, where metabolites are not as strongly influenced by immediate dietary factors, others investigate regulation of **postprandial metabolism** (i.e., the period after food consumption) [117]. This could be informative to determine how long certain dietary components take to metabolize and

whether variation in this duration is associated with certain conditions, such as diabetes. However, such studies require sequential sampling at defined time intervals, which is costly and labor intensive to collect at scale. Furthermore, nutritional epidemiology studies often rely on self-reported food frequency questionnaires, which can be impacted by recall bias and measurement error [118,119]. The metabolome offers a means to objectively measure dietary factors, but is limited by several challenges, including the short half-lives of some metabolites (see 'Metabolite stability and reproducibility' section), interindividual variation at multiple steps in biochemical pathways, and our limited knowledge of dietary factors impacting individual metabolites [120,121]. Thus, despite recent methods to address some of these issues, it is difficult to accurately assess diet through metabolomic analysis.

Tissue-specific metabolomics

Future work should also explore the utility of generating metabolomic data from disease-relevant tissues, which may pose novel analytical concerns [122]. For example, urine metabolomics, while popular due to ease of access and disease relevance for many conditions (including kidney disease), has much higher sample-to-sample variability in the same individual compared with blood metabolomics [122], and failure to account for underlying differences in albuminuria/proteinuria due to potential damage in the glomerular filtration barrier could lead to non-informative results [123]. Increased availability of tissues directly influenced by disease processes (such as kidney tissue for chronic kidney disease [123] or brain tissue or cerebrospinal fluid for Alzheimer's disease [124]) will likely provide novel biological insights, but large sample sizes will remain difficult to obtain outside of accessible tissues, such as blood, urine, and saliva (Figure 2). In some cases, integrative analyses have found shared signatures across tissue-specific and blood-based metabolomic analyses. For example, sphingolipid and glycerophospholipid associations discovered in a large blood-based metabolomics study were replicated in a subset of postmortem brain samples [125]. Analyses such as these may provide a way to leverage both larger sample sizes in accessible tissues and smaller sample sizes in less accessible tissues.

Single cell metabolomics

Single cell metabolomics enables metabolites to be measured from individual cells and is particularly valuable in studying diverse cell populations, such as tumor cells or cells at different stages of development [126]. Spatial metabolomics maps the distribution of metabolites in tissues or cells, providing a spatial context to the metabolic information [127]. This provides an understanding of how metabolic processes vary across different regions of tissue or within cellular compartments, which can be crucial to studying diseases such as cancer, where spatial organization and tissue architecture have a significant role. Coupling single cell and spatial metabolomics provides insights into the metabolic activity of individual cells, revealing heterogeneity and dynamic changes within cell populations that can be obscured in traditional bulk analyses. While an in-depth discussion on these topics is beyond the scope of this review, advances and challenges of this field have been reviewed in previous publications [126–128].

Biobank metabolomics

Increasing the availability of metabolomic data in biobanks will allow metabolomic investigations to be performed at an unprecedented scale. For example, the UK Biobank will soon release NMR-measured metabolomic data for over 200 metabolites measured on the Nightingale platform in baseline blood samples (N ~500 000) and at the first repeat assessment visit (N ~20 000). The first 120 000 samples were recently released to UK Biobank researchers. Early investigations of these data illustrate the potential for biobank-scale metabolomic analyses, such as the recent identification of correlated multi-metabolite scores for severe pneumonia and Coronavirus 2019 (COVID-19) [129]. While widespread availability of biobank metabolomic data

presents an exciting research frontier, it is likely to create challenges in terms of disparate quality control and analysis across studies, which will be important for researchers to address.

Metabolomic studies in diverse populations

While the field of metabolomics offers promising potential for biomarker discovery, guiding clinical interventions, and improving diagnostics, the stark lack of participant diversity in metabolomic investigations could exacerbate existing health disparities. For example, the population distribution of the ~82 000 participants in the COMETS consortium as of 2019 was 70% European, 18% Asian, 6% African, and 2% Hispanic [1]. Although some evidence has demonstrated broad agreement in metabolomic signatures of BMI and glycemia across pregnant women from diverse populations [130] and across dietary metabolites independent of self-reported race or ethnicity [131], other diverse studies have found that population background modifies associations between metabolomic profiles, insulin resistance, and metabolism [132–134]. Insufficient statistical power across populations is currently a limitation for most diverse metabolomic studies, and further research in larger, more diverse populations is needed to understand the potential metabolic differences between populations and the influence of sociocultural factors on these differences. A recent analysis of 1251 serum metabolites in a deeply phenotyped cohort found that diet and gut microbiome factors were most predictive of metabolite levels [46], both of which widely differ based on geography and sociocultural factors. Thus, in addition to the need for increased representation of diverse populations within the USA, representation of individuals across the globe is needed to comprehensively understand metabolomic variation across populations.

Multipopulation studies may help reveal the complexities that contribute to health disparities in disease risk. To this end, diverse metabolomic studies will need to implement a comprehensive framework that includes social determinants of health, such as structural racism, discrimination, stress, employment status, insurance coverage, access to care, immigration status, income, and language, among other factors, as potential contributors to metabolite variation [135]. To increase participation of under-represented individuals, it will be essential for investigators to engage community members early on at the study design stage to ensure that research questions addressed are relevant to the community and to establish strategies for biobanking and data governance to ensure that the community benefits from research findings [136–138].

Concluding remarks

In this review, we provide an introduction to the field of metabolomic epidemiology, discussing technological, study design, quality control, and statistical considerations, opportunities and challenges in the field, and emerging innovations that hold promise to uncover new biological insights. The metabolome has the potential to improve our mechanistic understanding of disease while also allowing for the identification of biomarkers that may inform prevention and screening strategies. While significant progress has been made to address the unique challenges of the field of metabolomic epidemiology, many remain that will need to be addressed in coming years (see Outstanding questions). Furthermore, efforts need to be made now to ensure that this rapidly emerging field will lead to equitable improvements in healthcare across diverse populations.

Outstanding questions

What additional insights can be gained from metabolites measured from less accessible tissues that may be more relevant to a given disease and how well do those levels correlate with more accessible circulating metabolites?

How can the process of harmonizing metabolites across platforms be eased to increase study power?

How can the large number of 'unknown' or 'unidentified' metabolites contribute to biological interpretations of metabolomic epidemiology investigations?

How can the complex correlation structure of metabolites be optimally leveraged in statistical analyses?

Are metabolomic findings in one ancestral population transferable to other ancestral populations?

Will meta-analyses of metabolomic epidemiology studies prove to be an effective way to increase study power, as has been found with GWAS?

Acknowledgments

This review was supported by the National Institutes of Health (NIH: grants R00 CA246063 awarded to B.F.D., R03DK127148 and K01-DK110267 awarded to A.D.J., KL2TR002490 to L.M.R., and grant P01HL132825; K.H.S. was funded, in part, by R01LM013444 and by NIH NHLBI 2T32HL007427), the Andy Hill Cancer Research Endowment Distinguished Researcher Grant (to B.F.D.), the National Science Foundation (grants DEB-2028280 and DEB-2109688 awarded to A.R.), and the Bill and Melinda Gates Foundation (grant INV-016930 awarded to A.R.). F.K.T. was supported by a

Research Scholar Grant from the American Cancer Society (# RSG-20-124-01-CCE). The opinions expressed by the authors are their own and this material should not be interpreted as representing the official viewpoint of the US Department of Health and Human Services, the NIH, or the National Cancer Institute.

Declaration of interests

None declared by authors.

References

1. Yu, B. *et al.* (2019) The Consortium of Metabolomics Studies (COMETS): metabolomics in 47 prospective cohort studies. *Am. J. Epidemiol.* 188, 991–1012
2. Marshall, D.D. and Powers, R. (2017) Beyond the paradigm: combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Prog. Nucl. Magn. Reson. Spectrosc.* 100, 1–16
3. Tolstikov, V. *et al.* (2020) Current status of metabolomic biomarker discovery: impact of study design and demographic characteristics. *Metabolites* 10, 224
4. Lasky-Su, J. *et al.* (2021) A strategy for advancing for population-based scientific discovery using the metabolome: the establishment of the Metabolomics Society Metabolomic Epidemiology Task Group. *Metabolomics* 17, 45
5. van Roekel, E.H. *et al.* (2019) Metabolomics in epidemiologic research: challenges and opportunities for early-career epidemiologists. *Metabolomics* 15, 9
6. Guertin, K.A. *et al.* (2014) Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *Am. J. Clin. Nutr.* 100, 208–217
7. Jones, D.P. *et al.* (2012) Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu. Rev. Nutr.* 32, 183–202
8. Mayers, J.R. *et al.* (2014) Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nat. Med.* 20, 1193–1198
9. Taliun, D. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299
10. Emwas, A.H. (2015) The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol. Biol.* 1277, 161–193
11. Gromski, P.S. *et al.* (2015) A tutorial review: metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* 879, 10–23
12. Wishart, D.S. (2016) Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* 15, 473–484
13. Tebani, A. *et al.* (2016) Clinical metabolomics: the new metabolic window for inborn errors of metabolism investigations in the post-genomic era. *Int. J. Mol. Sci.* 17, 1167
14. Bouatra, S. *et al.* (2013) The human urine metabolome. *PLoS ONE* 8, e73076
15. Perez de Souza, L. *et al.* (2021) Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research. *Nat. Methods* 18, 733–746
16. Gathungu, R.M. *et al.* (2020) The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices. *Mass Spectrom. Rev.* 39, 35–54
17. Playdon, M.C. *et al.* (2019) Metabolomics Analytics Workflow for Epidemiological Research: Perspectives from the Consortium of Metabolomics Studies (COMETS). *Metabolites* 9, 145
18. Farag, M.A. *et al.* (2007) Metabolic profiling and systematic identification of flavonoids and isoflavonoids in roots and cell suspension cultures of Medicago truncatula using HPLC-UV-ESI-MS and GC-MS. *Phytochemistry* 68, 342–354
19. Nordstrom, A. *et al.* (2008) Multiple ionization mass spectrometry strategy used to reveal the complexity of metabolomics. *Anal. Chem.* 80, 421–429
20. Stevens, V.L. *et al.* (2019) Pre-analytical factors that affect metabolite stability in human urine, plasma, and serum: a review. *Metabolites* 9, 156
21. Darst, B.F. *et al.* (2019) Longitudinal plasma metabolomics of aging and sex. *Aging (Albany NY)* 11, 1262–1282
22. Chaleckis, R. *et al.* (2016) Individual variability in human blood metabolites identifies age-related differences. *Proc. Natl. Acad. Sci. U. S. A.* 113, 4252–4259
23. Trabado, S. *et al.* (2017) The human plasma-metabolome: reference values in 800 French healthy volunteers; impact of cholesterol, gender and age. *PLoS ONE* 12, e0173615
24. Lau, C.E. *et al.* (2018) Determinants of the urinary and serum metabolome in children from six European populations. *BMC Med.* 16, 202
25. Alzharani, M.A. *et al.* (2020) Metabolomics profiling of plasma, urine and saliva after short term training in young professional football players in Saudi Arabia. *Sci. Rep.* 10, 19759
26. Mittelstrass, K. *et al.* (2011) Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet.* 7, e1002215
27. Holmes, E. *et al.* (2008) Metabolic phenotyping in health and disease. *Cell* 134, 714–717
28. Menni, C. *et al.* (2013) Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. *Metabolomics* 9, 506–514
29. Jin, L. *et al.* (2021) Use of untargeted metabolomics to explore the air pollution-related disease continuum. *Curr. Environ. Health Rep.* 8, 7–22
30. Vuckovic, D. (2018) Improving metabolome coverage and data quality: advancing metabolomics and lipidomics for biomarker discovery. *Chem. Commun. (Camb.)* 54, 6728–6749
31. Gil, A. *et al.* (2015) Stability of energy metabolites-an often overlooked issue in metabolomics studies: a review. *Electrophoresis* 36, 2156–2169
32. Watrous, J.D. *et al.* (2017) Visualization, quantification, and alignment of spectral drift in population scale untargeted metabolomics data. *Anal. Chem.* 89, 1399–1404
33. Roberts, L.D. *et al.* (2012) Targeted metabolomics. *Curr. Protoc. Mol. Biol.* 30 1–30.2.24
34. Han, S. *et al.* (2022) TIGER: technical variation elimination for metabolomics data using ensemble learning architecture. *Brief. Bioinform.* 23, bbab535
35. Stegle, O. *et al.* (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* 6, e1000770
36. Leek, J.T. *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883
37. Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–1735
38. Johnson, W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127
39. Aschard, H. *et al.* (2017) Covariate selection for association screening in multiphenotype genetic studies. *Nat. Genet.* 49, 1789–1795
40. Gallois, A. *et al.* (2019) A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat. Commun.* 10, 4788
41. Ritchie, S.C. *et al.* (2023) Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Sci. Data* 10, 64
42. Chiou, S.H. *et al.* (2019) The missing indicator approach for censored covariates subject to limit of detection in logistic regression models. *Ann. Epidemiol.* 38, 57–64
43. Kokla, M. *et al.* (2019) Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics* 20, 492

44. Wei, R. *et al.* (2018) Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.* 8, 663
45. Do, K.T. *et al.* (2018) Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* 14, 128
46. Bar, N. *et al.* (2020) A reference map of potential determinants for the human serum metabolome. *Nature* 588, 135–140
47. van den Berg, R.A. *et al.* (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7, 142
48. Mallick, H. *et al.* (2021) Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* 17, e1009442
49. Fitzmaurice, G.M. *et al.* (2012) *Applied Longitudinal Analysis*, John Wiley & Sons
50. Borenstein, M. *et al.* (2010) A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* 1, 97–111
51. Higgins, J.P. and Thompson, S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21, 1539–1558
52. Worley, B. and Powers, R. (2013) Multivariate analysis in metabolomics. *Curr. Metabolomics* 1, 92–107
53. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288
54. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat.Methodol.)* 67, 301–320
55. Vaarhorst, A.A. *et al.* (2014) A metabolomic profile is associated with the risk of incident coronary heart disease. *Am. Heart J.* 168, e47
56. James, G. *et al.* (2021) *An Introduction to Statistical Learning with Applications in R* (2nd edn), Springer
57. Girvan, M. and Newman, M.E. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7821–7826
58. Koschutzki, D. and Schreiber, F. (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene. Regu. Syst. Bio.* 2, 193–201
59. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559
60. Pei, G. *et al.* (2017) WGCNA application to proteomic and metabolomic data analysis. *Methods Enzymol.* 585, 135–158
61. Johnson, L.C. *et al.* (2018) Amino acid and lipid associated plasma metabolomic patterns are related to healthspan indicators with ageing. *Clin. Sci. (Lond.)* 132, 1765–1777
62. Epskamp, S. *et al.* (2018) Estimating psychological networks and their accuracy: a tutorial paper. *Behav. Res. Methods* 50, 195–212
63. Shutta, K.H. *et al.* (2022) Gaussian graphical models with applications to omics analyses. *Stat. Med.* 41, 5150–5187
64. Uhler, C. (2017) Gaussian graphical models: an algebraic and geometric perspective. *arXiv* 2017. arXiv:1707.04345
65. Krumsiek, J. *et al.* (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* 5, 21
66. Albrecht, E. *et al.* (2014) Metabolite profiling reveals new insights into the regulation of serum urate in humans. *Metabolomics* 10, 141–151
67. Zhao, T. *et al.* (2012) The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* 13, 1059–1062
68. Khatri, P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8, e1002375
69. Rahnavard, A. *et al.* (2021) Omics community detection using multi-resolution clustering. *Bioinformatics* 37, 3588–3594
70. Li, S. *et al.* (2013) Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* 9, e1003123
71. Pang, Z. *et al.* (2021) MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res.* 49, W388–W396
72. Barupal, D.K. and Fiehn, O. (2017) Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Sci. Rep.* 7, 14567
73. Stearrett, N. *et al.* (2021) Expression of human endogenous retroviruses in systemic lupus erythematosus: multiomic integration with gene expression. *Front. Immunol.* 12, 661437
74. Lains, I. *et al.* (2019) Human plasma metabolomics in age-related macular degeneration: meta-analysis of two cohorts. *Metabolites* 9, 127
75. Horgan, R.P. and Kenny, L.C. (2011) 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet. Gynaecol.* 13, 189–195
76. Wu, C. *et al.* (2019) A selective review of multi-level omics data integration using variable selection. *High Throughput* 8, 4
77. Chu, S.H. *et al.* (2019) Integration of metabolomic and other omics data in population-based study designs: an epidemiological perspective. *Metabolites* 9, 117
78. Worheide, M.A. *et al.* (2021) Multi-omics integration in biomedical research - a metabolomics-centric review. *Anal. Chim. Acta* 1141, 144–162
79. Cavill, R. *et al.* (2016) Transcriptomic and metabolomic data integration. *Brief. Bioinform.* 17, 891–901
80. Zoppi, J. *et al.* (2021) MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC Bioinformatics* 22, 6
81. Zhou, G. and Xia, J. (2018) OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res.* 46, W514–W522
82. Argelaguet, R. *et al.* (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* 14, e8124
83. Peng, C. *et al.* (2020) A latent unknown clustering integrating multi-omics data (LUCID) with phenotypic traits. *Bioinformatics* 36, 842–850
84. Wang, B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337
85. Lambert, S.A. *et al.* (2021) The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 53, 420–425
86. Ritchie, S.C. *et al.* (2021) Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *Nat Metab.* 3, 1476–1483
87. Wainberg, M. *et al.* (2020) Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. *Proc. Natl. Acad. Sci. U. S. A.* 117, 21813–21820
88. Wurtz, P. *et al.* (2014) Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med.* 11, e1001765
89. Burgess, S. *et al.* (2017) A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* 26, 2333–2355
90. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23, R89–R98
91. Burgess, S. *et al.* (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37, 658–665
92. Bowden, J. *et al.* (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525
93. Verbanck, M. *et al.* (2018) Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50, 693–698
94. Qin, Y. *et al.* (2020) Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases. *medRxiv* Published online August 4, 2020. https://doi.org/10.1101/2020.08.01.20166413
95. Korthauer, K. *et al.* (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* 20, 118
96. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300
97. Rouam, S. (2013) False discovery rate (FDR). In *Encyclopedia of Systems Biology* (Dubitzky, W. *et al.*, eds), pp. 731–732, Springer

98. Peluso, A. *et al.* (2021) Multiple-testing correction in metabolome-wide association studies. *BMC Bioinformatics* 22, 67
99. Armstrong, R.A. (2014) When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* 34, 502–508
100. Kettunen, J. *et al.* (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* 7, 11122
101. Li, J. and Ji, L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb.)* 95, 221–227
102. Westfall, P.H. and Young, S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley
103. Glickman, M.E. *et al.* (2014) False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* 67, 850–857
104. Yoav, B. and Daniel, Y. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188
105. Aaditya, K.R. *et al.* (2019) A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Stat.* 47, 2790–2821
106. Xiao, J.F. *et al.* (2012) Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Anal. Chem.* 32, 1–14
107. Rueedi, R. *et al.* (2017) Metabomatching: using genetic association to identify metabolites in proton NMR spectroscopy. *PLoS Comput. Biol.* 13, e1005839
108. O'Sullivan, J.F. *et al.* (2017) Dimethylguanidino valeric acid is a marker of liver fat and predicts diabetes. *J. Clin. Invest.* 127, 4394–4402
109. Shen, X. *et al.* (2019) Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.* 10, 1516
110. Grapov, D. *et al.* (2015) MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. *Bioinformatics* 31, 2757–2760
111. Tarazona, S. *et al.* (2020) Harmonization of quality metrics and power calculation in multi-omic studies. *Nat. Commun.* 11, 3092
112. Pinu, F.R. *et al.* (2019) Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites* 9, 76
113. Sud, M. *et al.* (2016) Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44, D463–D470
114. Gertsman, I. and Barshop, B.A. (2018) Promises and pitfalls of untargeted metabolomics. *J. Inherit. Metab. Dis.* 41, 355–366
115. Townsend, M.K. *et al.* (2013) Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin. Chem.* 59, 1657–1667
116. Zeleznik, O.A. *et al.* (2022) Intrapersonal stability of plasma metabolomic profiles over 10 years among women. *Metabolites* 12, 372
117. Sandler, V. *et al.* (2017) Associations of maternal BMI and insulin resistance with the maternal metabolome and newborn outcomes. *Diabetologia* 60, 518–530
118. Kirkpatrick, S.I. *et al.* (2022) Measurement error affecting web- and paper-based dietary assessment instruments: insights from the multi-cohort eating and activity study for understanding reporting error. *Am. J. Epidemiol.* 191, 1125–1139
119. Satija, A. *et al.* (2015) Understanding nutritional epidemiology and its role in policy. *Adv. Nutr.* 6, 5–18
120. Guasch-Ferre, M. *et al.* (2018) Use of metabolomics in improving assessment of dietary intake. *Clin. Chem.* 64, 82–98
121. Rafiq, T. *et al.* (2021) Nutritional metabolomics and the classification of dietary biomarker candidates: a critical review. *Adv. Nutr.* 12, 2333–2357
122. Smith, L. *et al.* (2020) Important considerations for sample collection in metabolomics studies with a special focus on applications to liver functions. *Metabolites* 10, 104
123. Dubin, R.F. and Rhee, E.P. (2020) Proteomics and metabolomics in kidney disease, including insights into etiology, treatment, and prevention. *Clin. J. Am. Soc. Nephrol.* 15, 404–411
124. Desai, R.J. *et al.* (2020) Targeting abnormal metabolism in Alzheimer's disease: The Drug Repurposing for Effective Alzheimer's Medicines (DREAM) study. *Alzheimers Dement (N Y)* 6, e12095
125. Varma, V.R. *et al.* (2018) Brain and blood metabolite signatures of pathology and progression in Alzheimer disease: a targeted metabolomics study. *PLoS Med.* 15, e1002482
126. Duncan, K.D. *et al.* (2019) Advances in mass spectrometry based single-cell metabolomics. *Analyst* 144, 782–793
127. Taylor, M.J. *et al.* (2021) Spatially resolved mass spectrometry at the single cell: recent innovations in proteomics and metabolomics. *J. Am. Soc. Mass Spectrom.* 32, 872–894
128. Lanekoff, I. *et al.* (2022) Single-cell metabolomics: where are we and where are we going? *Curr. Opin. Biotechnol.* 75, 102693
129. Julkunen, H. *et al.* (2021) Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *Elife* 10, e63033
130. Jacob, S. *et al.* (2017) Targeted metabolomics demonstrates distinct and overlapping maternal metabolites associated with BMI, glucose, and insulin sensitivity during pregnancy across four ancestry groups. *Diabetes Care* 40, 911–919
131. de Souza, R.J. *et al.* (2020) Maternal diet and the serum metabolome in pregnancy: robust dietary biomarkers generalizable to a multiethnic birth cohort. *Curr. Dev. Nutr.* 4, nzaa144
132. Palmer, N.D. *et al.* (2015) Metabolomic profile associated with insulin resistance and conversion to diabetes in the Insulin Resistance Atherosclerosis Study. *J. Clin. Endocrinol. Metab.* 100, E463–E468
133. Lee, C.C. *et al.* (2016) Branched-chain amino acids and insulin metabolism: The Insulin Resistance Atherosclerosis Study (IRAS). *Diabetes Care* 39, 582–588
134. Kadakia, R. *et al.* (2019) Maternal metabolites during pregnancy are associated with newborn outcomes and hyperinsulinaemia across ancestries. *Diabetologia* 62, 473–484
135. Duggan, C.P. *et al.* (2020) Race, ethnicity, and racism in the nutrition literature: an update for 2020. *Am. J. Clin. Nutr.* 112, 1409–1414
136. Hudson, M. *et al.* (2020) Rights, interests and expectations: indigenous perspectives on unrestricted access to genomic data. *Nat. Rev. Genet.* 21, 377–384
137. Erves, J.C. *et al.* (2017) Needs, priorities, and recommendations for engaging underrepresented populations in clinical research: a community perspective. *J. Community Health* 42, 472–480
138. Fox, K. (2020) The illusion of inclusion - The 'All of Us' Research Program and Indigenous Peoples' DNA. *N. Engl. J. Med.* 383, 411–413
139. Qi, J. *et al.* (2018) Circulating trimethylamine N-oxide and the risk of cardiovascular diseases: a systematic review and meta-analysis of 11 prospective cohort studies. *J. Cell. Mol. Med.* 22, 185–194
140. Bamji-Stocke, S. *et al.* (2018) A review of metabolism-associated biomarkers in lung cancer diagnosis and treatment. *Metabolomics* 14, 81
141. Huang, J. *et al.* (2022) Metabolomic profile of prostate cancer-specific survival among 1812 Finnish men. *BMC Med.* 20, 362
142. Reinke, S.N. *et al.* (2017) Metabolomics analysis identifies different metabotypes of asthma severity. *Eur. Respir. J.* 49, 1601740
143. Huang, M. *et al.* (2021) Maternal metabolome in pregnancy and childhood asthma or recurrent wheeze in the Vitamin D Antenatal Asthma Reduction Trial. *Metabolites* 11, 1234–1241
144. Yu, B. *et al.* (2019) Metabolomics identifies novel blood biomarkers of pulmonary function and COPD in the general population. *Metabolites* 9, 61
145. Cruickshank-Quinn, C.I. *et al.* (2018) Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. *Sci. Rep.* 8, 17132
146. Wang, T.J. *et al.* (2011) Metabolite profiles and the risk of developing diabetes. *Nat. Med.* 17, 448–453
147. Walford, G.A. *et al.* (2014) Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. *Diabetes Care* 37, 2508–2514
148. Floegel, A. *et al.* (2013) Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* 62, 639–648

149. Jiang, Y. *et al.* (2019) Metabolomics in the development and progression of dementia: a systematic review. *Front. Neurosci.* 13, 343
150. Yu, Z. *et al.* (2012) Human serum metabolic profiles are age dependent. *Aging Cell* 11, 960–967
151. Menni, C. *et al.* (2013) Metabolomic markers reveal novel pathways of ageing and early development in human populations. *Int. J. Epidemiol.* 42, 1111–1119
152. Rist, M.J. *et al.* (2017) Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLoS ONE* 12, e0183228
153. Dunn, W.B. *et al.* (2015) Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics* 11, 9–26
154. Moore, S.C. *et al.* (2014) Human metabolic correlates of body mass index. *Metabolomics* 10, 259–269
155. Ho, J.E. *et al.* (2016) Metabolomic profiles of body mass index in the Framingham Heart Study reveal distinct cardiometabolic phenotypes. *PLoS ONE* 11, e0148361
156. Ottosson, F. *et al.* (2018) Connection between BMI-related plasma metabolite profile and gut microbiota. *J. Clin. Endocrinol. Metab.* 103, 1491–1501
157. Krumsiek, J. *et al.* (2015) Gender-specific pathway differences in the human serum metabolome. *Metabolomics* 11, 1815–1833
158. Wang, J. and Li, M.D. (2010) Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation. *Neuropsychopharmacology* 35, 702–719
159. Benowitz, N.L. *et al.* (2010) Urine menthol as a biomarker of mentholated cigarette smoking. *Cancer Epidemiol. Biomark. Prev.* 19, 3013–3019
160. Hsu, P.C. *et al.* (2017) Metabolomic profiles of current cigarette smokers. *Mol. Carcinog.* 56, 594–606
161. Benowitz, N.L. (1996) Cotinine as a biomarker of environmental tobacco smoke exposure. *Epidemiol. Rev.* 18, 188–204
162. Gu, F. *et al.* (2016) Cigarette smoking behaviour and blood metabolomics. *Int. J. Epidemiol.* 45, 1421–1432
163. Harada, S. *et al.* (2016) Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. *Environ. Health Prev. Med.* 21, 18–26
164. van Roekel, E.H. *et al.* (2018) Circulating metabolites associated with alcohol intake in the European Prospective Investigation into Cancer and Nutrition Cohort. *Nutrients* 10, 654