Human Genomics

CrossMark

Beyond genomics: understanding exposotypes through metabolomics

Nicholas J. W. Rattray[1], Nicole C. Deziel[1], Joshua D. Wallach[2,3], Sajid A. Khan[4,5], Vasilis Vasiliou[1,5], John P. A. Ioannidis[6,7,8,9,10] and Caroline H. Johnson[1,5*] (iD)

Abstract

Background: Over the past 20 years, advances in genomic technology have enabled unparalleled access to the information contained within the human genome. However, the multiple genetic variants associated with various diseases typically account for only a small fraction of the disease risk. This may be due to the multifactorial nature of disease mechanisms, the strong impact of the environment, and the complexity of gene-environment interactions. Metabolomics is the quantification of small molecules produced by metabolic processes within a biological sample. Metabolomics datasets contain a wealth of information that reflect the disease state and are consequent to both genetic variation and environment. Thus, metabolomics is being widely adopted for epidemiologic research to identify disease risk traits. In this review, we discuss the evolution and challenges of metabolomics in epidemiologic research, particularly for assessing environmental exposures and providing insights into gene-environment interactions, and mechanism of biological impact.

Main text: Metabolomics can be used to measure the complex global modulating effect that an exposure event has on an individual phenotype. Combining information derived from all levels of protein synthesis and subsequent enzymatic action on metabolite production can reveal the individual exposotype. We discuss some of the methodological and statistical challenges in dealing with this type of high-dimensional data, such as the impact of study design, analytical biases, and biological variance. We show examples of disease risk inference from metabolic traits using metabolome-wide association studies. We also evaluate how these studies may drive precision medicine approaches, and pharmacogenomics, which have up to now been inefficient. Finally, we discuss how to promote transparency and open science to improve reproducibility and credibility in metabolomics.

Conclusions: Comparison of exposotypes at the human population level may help understanding how environmental exposures affect biology at the systems level to determine cause, effect, and susceptibilities. Juxtaposition and integration of genomics and metabolomics information may offer additional insights. Clinical utility of this information for single individuals and populations has yet to be routinely demonstrated, but hopefully, recent advances to improve the robustness of large-scale metabolomics will facilitate clinical translation.

Keywords: Chemometrics, Exposome, Exposotype, Genomics, Genetic epidemiology, Metabolomics

Background

The main concepts underpinning genetic epidemiology developed rapidly after the delineation of the structure of DNA. Neel and Schull provided the first description of these concepts in 1954 [1, 2]. While the original goal of genetic epidemiology was to understand the nature of population and familial genetic inheritance, it soon became evident that environmental factors and gene-environment interactions were important to consider simultaneously [3].

Currently, the study of the whole genome (genomics) has evolved into a multidisciplinary area of science with highly diverse applications [4, 5]. Improved efficiency of genome technology combined with a sharp decrease in cost has enabled genomic assessments in large study populations [6, 7] using genotyping and next-generation-sequencing (NGS) approaches [8]. Thousands of genome-wide association

* Correspondence: caroline.johnson@yale.edu
[1]Department of Environmental Health Sciences, Yale School of Public Health, Yale University, New Haven, CT, USA
[5]Yale Cancer Center, Yale University School of Medicine, New Haven, CT, USA
Full list of author information is available at the end of the article

Rattray *et al. Human Genomics* (2018) 12:4

Page 2 of 14

studies (GWAS) have tracked relationships between base-pair/gene patterns in genomic loci and hundreds of diseases or exposures [9]. However, the discovered loci from these large-scale studies still explain only the minority of presumed heritability for most phenotypes of interest [10]. Moreover, it has been established that genes alone account for the minority of disease etiology for many important illnesses such as cancer, and environmental and lifestyle influences play a critical role [11]. However, quantifying the myriad of environmental and lifestyle risk factors including diet, smoking, exposure to hazardous chemicals, and pathogenic microorganisms is challenging [12, 13]. An individual can be exposed to a complex mix of chemical and biological contaminants, with multiple sources, for varying durations across their life course. This concept has been termed the "exposome," a framework for the collective analysis, and measurement of an individual's exposures over their lifetime [14]. Moreover, different environmental exposures may be heavily correlated with each other or may act in concert to produce adverse effects, which makes studying them one at a time challenging for assigning causality [15]. Therefore, it is essential to find tools that can measure the cumulative impact of multiple exposures alongside their interactions with the genetic background of individuals. Several multidimensional analytical approaches have been developed, beyond genomics, that try to capture different aspects of this complexity, and their integration into environmental health is discussed in this review.

Application of high-dimensional biology to the environmental health paradigm

Referred to as high-dimensional biology, or a multi-omics/systems-level approach, the combined analysis of data from the genome (genomics), RNA transcription (transcriptomics), proteins/peptides (proteomics), and metabolites (metabolomics) enables researchers to overlay gene information onto complementary datasets towards a more systemic understanding of diseases or other phenotypes of interest [16]. The complexity of high-dimensional datasets becomes even more convoluted when the interaction of environmental exposures is added to the system.

The environmental health paradigm (Fig. 1) integrates the knowledge of exposures and environmental health
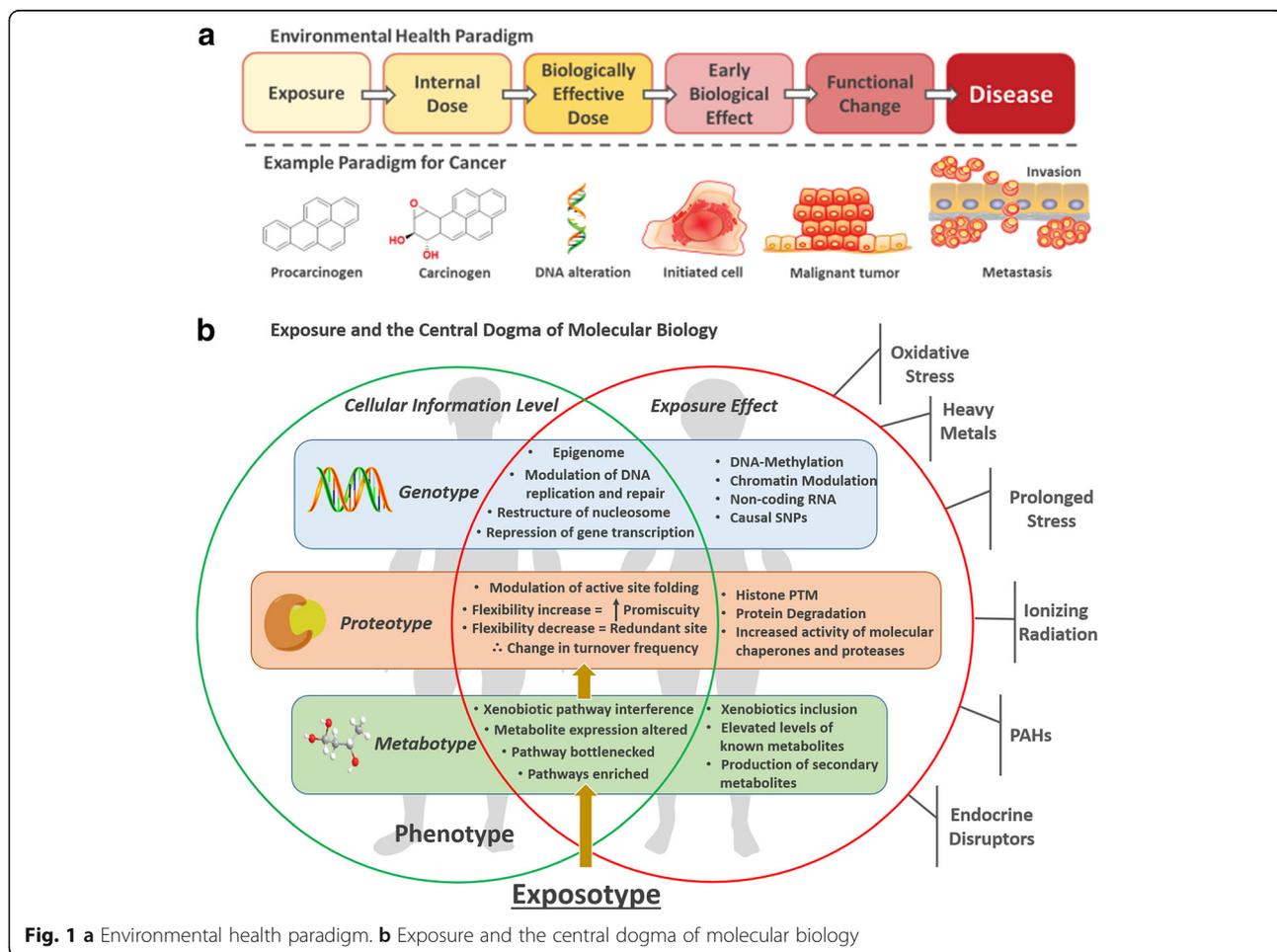
Fig. 1 a Environmental health paradigm. **b** Exposure and the central dogma of molecular biology

Rattray *et al. Human Genomics* (2018) 12:4

Page 3 of 14

sciences to gain a deeper understanding of the consequences of exposure towards expression of a disease phenotype [17]. Exposures can elicit subtle effects at different stages of gene-encoding, protein synthesis, and on circulating metabolites. Multi-omics approaches using combined data from genomics, proteomics, and metabolomics techniques can identify downstream chemical alterations contributing to the development of an exposotype, the exposure phenotype (Fig. 1), that describes the accrued biological changes within a system that has undergone a specific exposure event [18]. Combining information from all levels of protein synthesis and subsequent enzymatic action on metabolite production is an essential step to start comprehending the complex global modulating effect that an exposure event has on an individual phenotype. This may allow for a greater direct understanding of molecular mechanisms that underpin the route of exposure, and the effect of molecular transit on different areas of metabolism, cellular reproduction, and ultimately the resulting exposotype.

Metabolites are the substrates and products of metabolism that drive essential cellular processes such as energy production, and signal transduction [19]. Of all the molecular entities (genes, transcripts, proteins, metabolites), metabolites have the closest relationship to expressed phenotype as they are the final end-points of upstream biochemical processing. Quantitative readouts of metabolite abundance reflect both this cellular processing and xenobiotics (foreign substances such as environmental chemicals, pollutants, drugs, food additives, dyes) that are physico-chemically distinct from molecular entities that originate in the host. Xenobiotics can be processed by enzymatic machinery, and metabolomics also allows quantification of these metabolites. Therefore, metabolomics can simultaneously analyze both exogenous chemicals and their metabolites, and changes to the endogenous metabolome, to allow assessment of broadly defined exposures and their biological impact [20–23]. One such example was a recent study of occupational exposure to trichloroethylene (TCE) [24]. TCE metabolites were identified in human plasma and associated with changes to endogenous metabolites that were known to be involved in immunosuppression, hepatotoxicity, and nephrotoxicity. This allowed the investigation into how the toxic effects of TCE exposure were manifested [24]. Another study, from the EXPOsOMICS project (http://www.exposomicsproject.eu/), examined human biofluids and exhaled breath for exposure to swimming pool disinfection by-products (DBPs) and for concomitant changes to endogenous metabolites. The study revealed a possible association between DBPs and perturbations to metabolites in the tryptophan pathway [25]. However, these studies and others which have measured exposures in relation to the metabolome highlight the challenge of attempting to unravel the effect of one circumscribed exposure versus combinations of different environmental exposures on the metabolome [26, 27].

One of the major bottlenecks of metabolomics is metabolite identification. However, the expansion and development of metabolite databases have eased this issue. Tens of thousands of metabolites have been identified and uploaded onto metabolite databases such as The Human Metabolome Database (HMDB) (http://www.hmdb.ca/metabolites), which to date houses 114,113 metabolites with associated chemical, clinical, and biochemical information. HMDB also hosts four additional databases including the Toxic Exposome Database (T3DB) (http://www.t3db.ca/) which contains information on 3763 toxins [28, 29]. METLIN (https://metlin.scripps.edu), another large database containing 961,829 metabolites, recently expanded due to the integration of xenobiotics from the United States Environmental Protection Agency's "Distributed Structure-Searchable Toxicity (DSSTox)" database [30, 31]. The Exposome-Explorer database was recently designed to contain information on biomarkers of exposure to environmental risk factors for diseases. This database has information on 692 dietary and pollutant biomarkers, and importantly concentration values measured in biospecimens, with correlation values to assess quality of the biomarkers [32]. These databases, and others that house both xenobiotics and endogenous metabolites, appear in Table 1 [33–38]. With the recent expansion of these databases to include xenobiotics, metabolomics can facilitate both biomonitoring of exposures, assessment of biological impact, and identification of exposotypes [39]. However, one potential gap in these databases still exists, the prediction of phase I and phase II biotransformed metabolites of xenobiotics which can be used as proxy biomarkers for the chemical exposure. Metabolomics has revealed numerous novel metabolites of previously well-characterized pharmaceutical drugs such as acetaminophen [40], dietary supplements [41], and the genotoxic heterocyclic amine 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) [42], present in meats cooked at high temperatures. Metabolomics provides a window to identifying these new metabolites, as the biotransformed metabolite will only be present in a sample from an exposed individual. Secondly, there is typically more than one biotransformation metabolite present for each xenobiotic, which will have a similar covariance and correlation within the biological sample examined, thus making it possible to easily map out the related metabolites. One way to overcome this gap in the metabolite databases would be to have a tool housed on these databases that could automatically predict any potential biotransformations, and display the resultant

Rattray *et al. Human Genomics* (2018) 12:4

Page 4 of 14

Table 1 Mass spectrometry metabolite databases for identification of environmental exposures

Database name	Description	URL
Human metabolome database (HMDB)	114,113 xenobiotic and endogenous metabolites with chemical, biochemical, and clinical information.	http://www.hmdb.ca/ [33]
Toxic exposome database (T3DB)	3767 toxic compounds, targets and gene expression data, part of the HMDB suite.	http://www.t3db.ca/ [28]
METLIN	961,829 xenobiotic and endogenous metabolites with chemical information. Contains information from DSSTox.	https://metlin.scripps.edu/ [34]
Exposome-Explorer	692 dietary and pollutant biomarkers, with concentration values measured from biospecimens with intra class correlation coefficients.	http://exposome-explorer.iarc.fr/ [32]
Madison-Qingdao Metabolomics Consortium Database	20,300 xenobiotics and endogenous metabolites, with chemical information	http://mmcd.nmrfam.wisc.edu/ [35]
Drugbank	10,513 drug entries with drug target information, part of the HMDB suite	https://www.drugbank.ca/ [36]
PubChem	93,977,784 compounds, xenobiotic and endogenous metabolites but also peptides, and chemically altered macromolecules. Data is derived from hundreds of sources.	https://pubchem.ncbi.nlm.nih.gov/ [37]
CompTox Chemistry Dashboard	758,000 xenobiotics with chemical information compiled from multiple sources; PubChem, and US EPA's DSSTox, ACToR, ToxCast, EDSP21, and CPCat.	https://comptox.epa.gov/dashboard [38]

important chemical information for identification. A few tools currently available for predicting phase I and II drug metabolism have been recently reviewed, along with the development of "DrugBug" which can predict xenobiotic metabolism by human gut microbiota enzymes [43]. Integration of such tools would facilitate exposome analysis.

The broad range of chemical classes that exist among the thousands of endogenous and environmentally derived metabolites contained within a biological sample has given rise to the need for analytical strategies that can separate and detect as much chemical diversity as possible from within the biological system under examination. The assessment of all metabolites present in a sample, untargeted metabolomics, is typically carried out using chromatography-based mass spectrometry and/or nuclear magnetic resonance spectroscopy, alongside bioinformatics that help understand the complex data generated [44]. Metabolomics research has undergone significant refocus over the past few years due to the improvements made in bioanalytical protocols and an evident shift towards the development of new chemoinformatic and bioinformatic tools [45]. These tools are designed to improve metabolite identification, particularly for microbial metabolites, and biological interpretation, which remain a major challenge for the field. For example, the mass spectrometry data generated in a metabolomics study have a high degree of degeneracy where the same metabolite can be represented as multiple signals [46]. Tools such as CAMERA [47], RAMClust [48], and "Credentialing" [49] have helped overcome this problem and improve peak annotation. Other notable tools include CSI:FingerID [50] which predicts the fragmentation of metabolites using an in silico

method, thus aiding in metabolite identification, and "integrated-omics" housed on XCMSOnline [51] (http://xcmsonline.scripps.edu/) which aids in both metabolite identification and biological interpretation. Excellent reviews on the technological advancements in this area can be found elsewhere [52–54]; in addition, an extensive list of all current metabolomics software and data analysis resources is available [55, 56]. For population-level studies, the application of metabolomics for the analysis of thousands of samples has been optimized and demonstrated [57, 58], but the field could still benefit from decades' worth of research and lessons learning in genetic epidemiology related to study design, statistical analyses, and reproducibility in large-scale population consortia.

Methodological challenges and considerations

Relevant and a priori formulated research questions and rigorous study designs and methods lay the foundation to perform a potentially successful piece of population-based research, after which replication is essential to confirm any associations, and to avoid the dissemination of potentially false research claims [59–61]. Prospective cohort studies follow a predefined population over time, capturing exposure information prior to occurrence of health events. This study design accommodates the appropriate temporal relationship between exposure and outcome, allows for testing of multiple risk factors and health outcomes, and permits collection of multiple pre-clinical biological specimens throughout the follow-up period. Although this is ideal from a metabolomics perspective, this study design often requires long follow-up durations and great expense. Case-control studies can be more efficient, and less expensive ways to test associations, but they

Rattray *et al. Human Genomics* (2018) 12:4

Page 5 of 14

lack the temporality criterion for causality, and metabolic profiles may be influenced by disease status. The use of nested case-control studies offers an efficient approach with the appropriate temporality between exposure and outcome. "Meet-in-the-middle" approaches, which involve linking intermediate biomarkers to both the exposure and outcome within cohort and nested case-control studies, are gaining popularity for their ability to reveal important linkages along the exposure-outcome pathway [62, 63].

While systems-level approaches hold great promise, they also pose challenges in the analysis of high-dimensional, complex data structure. The use of appropriate statistical tests within genomics, metabolomics, and epidemiology is dictated by the study design and the number of dimensions of data under investigation, with the application of univariate or multivariate techniques being applied to low-dimensional and high-dimensional datasets, respectively. Incorrect analytical decisions and interpretations that are made when conducting a study are a direct threat to reproducibility [64]. Table 2 [65–87] provides a list of some of the most commonly used statistical methods and tests in the interface of epidemiology, genetics, and metabolomics.

Many analyses in metabolomics involve the use of null hypothesis significance testing (NHST) and the reporting of p values. The p value, one of the most misused statistics in science [88], has not escaped the focus of members of the fields of epidemiology [89], metabolomics [90], and general biomedicine [91]. Poor application has contributed to the irreproducible nature of many studies, so much that the American Statistical Association felt moved to release a statement highlighting six underlying principles to dictate the proper use and interpretation of the p value [92, 93]. One should examine in each application whether NHST is best suited as an inferential tool or whether alternative approaches, such as the use of Bayesian methods or false discovery rates (FDR), are preferable [90, 94–96]. If p values are still used in multidimensional experiments, proper correction for multiplicity is important. There are numerous methods for accommodating family-wise error rates [90]. There are also some standard thresholds that can be used in specific settings, e.g., genome-wide significance $p < 5 \times 10^{-8}$ for genome-wide analyses. Some multiplicity corrections are more conservative than others; for instance, the Bonferroni correction (dividing the p value threshold required for significance by the number of tests performed) may be too conservative [97]. FDR and variants of FDR may be better suited [96] and can accommodate correlation structures between the multiple tested variables [98, 99].

Several methods are available that can help reduce complexity, detect trends, and generate predictive models within multidimensional datasets (Table 2) such as those generated by NGS and mass spectrometry when target genes or metabolites are not known. Unsupervised methods such as principal component analysis (PCA) provide an initial step to help reduce the complexity and indicate variables of interest by determining discriminant features linked to the "loadings" of different clusters. These loadings can be considered as the impact that a certain variable has on measured variance, so a high-level loading value displays a strong influence on clustered groups [100]. There also exist several extensions of the PCA architecture such as multiblock PCA, consensus PCA, or ANOVA-PCA that enable the user to control for underlying influential factors within datasets such as the intra-patient variability or other experimental confounders [65]. These approaches have been used for metabolomics and genetics analyses and also lend themselves to other cross-validation methods [66]. Supervised methods apply grouping stratification to the data based on some already known outcome variable(s). They aim to develop models that can accurately predict the correct grouping based on the input and identify genes, metabolites, or other statistical associations that underlie the grouping. The most commonly used methods are variants of regression tools (Table 2). Regression modeling can identify associations relevant to the disease [101], can predict association within gene expression patterns [102], and in metabolomics [103] can generate sample classification. However, as these tests are supervised, one of the issues with multivariate regression is that it tends to over-fit the data. Therefore, cross-validation (in the same dataset) and external validation (in additional datasets) are essential.

Perhaps, the biggest challenge yet for exposome researchers is integration of the multiple types of data generated from systems-level analyses and assessing the role of one versus multiple exposures on the phenotype. Currently, there are platforms that enable biochemical pathway analysis and integration of systems-level data, and these platforms can identify pathways and networks that are related to a known exposure or health outcome (such as disease). Dissection of pathways may help direct mechanistic studies into causality. The most useful to date for untargeted metabolomics data is "mummichog," which uses computational algorithms to predict metabolic pathway effects directly from spectral feature tables without prior identification of metabolites [104]. Mummichog was recently integrated onto the XCMSOnline platform, with an added function to upload transcriptomic and proteomic data, for integrated pathway analysis [51]. Other notable software includes MarVis-Pathway [105], InCroMAP [106], GAM [107], and MetaCore™ (Thomson Reuters Corporation, Toronto, Canada) that can integrate multiple types of systems-level data for pathway interrogation. Combining this type of data with multiple measurements of xenobiotics has not yet been demonstrated, but tools are under development. Up to now, studies have

Rattray *et al. Human Genomics* (2018) 12:4

Page 6 of 14

Table 2 Common statistical methods and tests used in epidemiology, genetics, and metabolomics, with reference link to descriptive articles on appropriate general use

Class of test	Type of test	Application/description	Refs
Descriptive	Mean Median Mode	The simplest of tests used to describe basic features within data.	Covered in all general statistical textbooks and used in most if not all scientific disciplines. [67–69]
	Range, variance, SD	Describe spreads of data within a population	
Inferential	z test, t test, chi-square	Predicts/infers an observed mean, frequency, or proportion to a predetermined value, respectively.	
	ANOVA	Parametric method that tests the hypothesis that the means of two or more populations are equal. Frequently used to compare variance among groups relative to variance within groups	
	Kruskal-Wallis	Non-parametric method to rank statistical significant differences between two or more groups of an independent variable on a continuous/ordinal variable	
Scaling	Centering, auto, pareto, log, MD	Data pretreatment methods aim at reducing biological and analytical bias	[70, 71]
Principal component	PCA	Unsupervised dimensional reduction procedure used to explain the maximum variance within complex datasets.	[72–74]
	Multiblock PCA	PCA extension designed to find the underlying relationships between sets of related data	[65, 66, 75]
	ANOVA-PCA	Uses PC dimensional reduction to determines the effect of the experimental factors on multiple dependent variables	[65, 76]
	PC-DFA	Supervised test that summarizes the differentiation between groups while overlooking within-group variation.	[65, 77, 78]
Regression	Linear	Summarizes and quantifies the relationship between two continuous variables	[72, 79]
	PLS	Used to predict a set of dependent variables from a large set of independent variables	[73, 77, 80–82]
	O-PLS	orthogonal signal correction on PLS that maximizes the explained covariance on the first latent variable	[77, 81, 83]
	PLS-R	Combination of the predictive power of regression alongside the ability to deal with high dimensionality and multicollinearity of variables.	[77, 84]
	PLS-DA	Supervised approach to prediction on discrete variables	[77, 79, 83]
	LASSO	Parsimonious approach to variable selection and regularization in order to enhance interpretability and reduce noise	[79, 80, 85–87]
	Elastic net	Variable reduction approach where strongly correlated predictors coalesce in or out of the model together	[79, 80, 85, 87, 167]

Definitions: *SD* standard deviation, *MD* median, *PCA* principal component analysis, *ANOVA* analysis of variance, *PC-DFA* principal component discriminant function analysis, *PLS* partial least squares (also known as projection of latent structures), *O-PLS* orthogonal PLS, *PLS-R* PLS regression, *LASSO* least absolute shrinkage and selection operator

primarily assessed the effect of individual exposures and have combined multiple systems-level approaches to assess biological response (i.e., benzene exposure and toxicity, susceptibility genes, mRNA and DNA methylation) [108]. Phenome data has also been integrated into studies to account for population variability and reduce false positives [22]. A recent example, from the analysis of preterm birth in the Rhea mother-child cohort study, selected those metabolites that had significant association with birth outcomes in logistic regression models and significant correlation coefficients with metabolic syndrome traits to construct odds ratios (BMI, blood pressure, blood glucose) [109]. Moreover, new tools are being specifically designed with the exposome in mind; xMWAS can integrate metabolomics data with that derived from the transcriptome [110], microbiome [111], and cytokine [112] and can be used for genome, epigenome, proteome, and other integrated omics analyses. However, modeling the effect of combined exposures is extremely complex. Co-exposures can be linked and cause an additive effect on the biological outcome, but it is not possible to know beforehand which combinations of exposures may have the largest biological effect. A recent novel method was developed that first estimates the correlation between

Rattray *et al. Human Genomics* (2018) 12:4

Page 7 of 14

pairs of exposures, then groups the highly correlated exposures by unsupervised machine learning [26], and identifies co-occurring exposure networks. This technique reduces the total number of combinations of exposures to "prevalent co-occurring combinations"; however, integration with other systems-level data still remains very complex. The additional challenges associated with integrating exposome data with metabolomics, genomics, and proteomics have been recently reviewed [27] and were also highlighted in a recent symposium report [113].

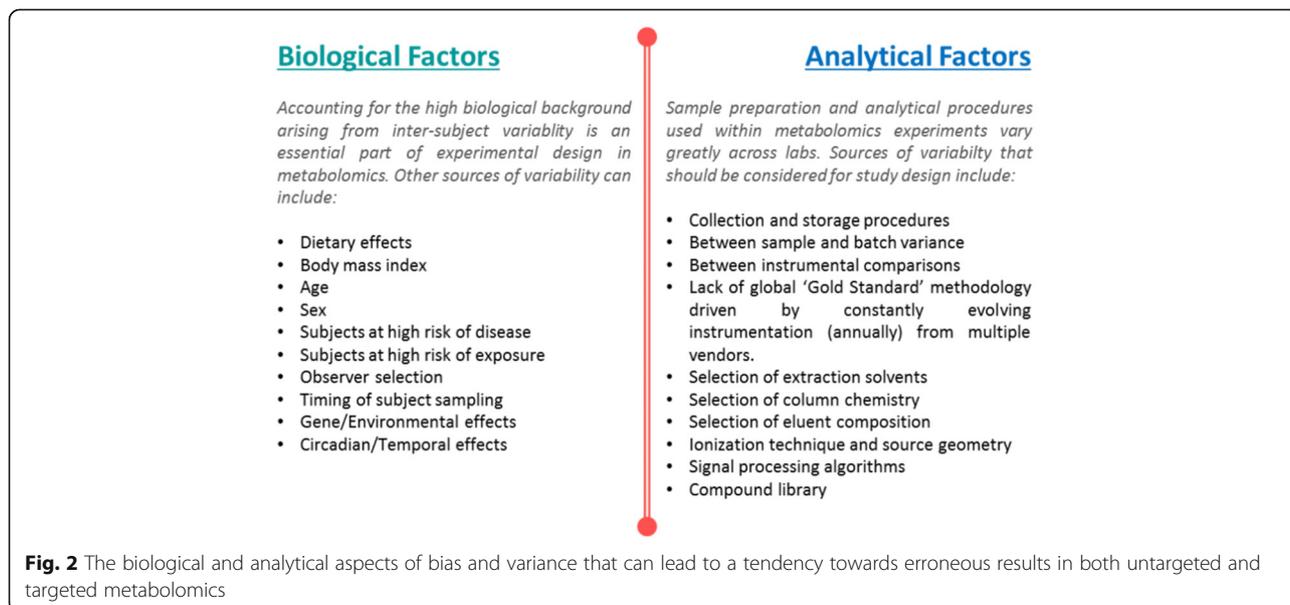## Analytical bias and biological variance in metabolomics analyses for epidemiologic studies

Metabolomics analyses in epidemiologic studies require additional consideration of sources of variability beyond traditional epidemiologic studies. There are a very large number of chemical features that can be detected by current highly sensitive mass spectrometers, and differences in metabolite recovery may arise from biological samples that are not collected under identical protocols. Additional batch variation can be introduced when handling large sample numbers [114], due to contaminant build-up and sample degradation [115].

Analytical bias in genomics and metabolomics can arise from practical laboratory aspects that, by their nature, favor the preselection of one type of variable (single nucleotide polymorphism (SNP) or chemical) over another. This is particularly evident when performing "untargeted" analyses in which the researcher is looking to maximize chemical coverage with a technology that cannot cover the full chemical space. With currently over 24 million SNPs having been documented within the human genome [116], the technology within SNP microarray chips has yet to catch up to this depth of coverage. The same issues are also present within metabolomics as no single technology can analyze the thousands of different metabolites within a sample. Therefore, pre-selecting approaches are commonly applied, be it using a gene-expression chip predefined for a subset of SNPs [117–120] or untargeted chromatography methods for metabolomics with a restricted spectrum of which metabolites can be captured [121]. These analytical biases are described in Fig. 2, but include the type of metabolite extraction method and column chemistry, which can enhance the analysis of some chemical functional groups and classes over others. For example, reversed-phase liquid chromatography (RPLC) can effectively analyze non-polar compounds such as lipids, carnitines, and bile acids, whereas hydrophilic interaction liquid chromatography (HILIC) is more suitable for the analysis of polar metabolites such as nucleotides, sugars, and amino acids. The two column chemistries have an analytical overlap of only 34%; thus, both column chemistries are needed if one wishes to obtain a relative quantification of the broadest chemical classes from a sample [122]. All types of study design need to consider inherent biological intra-individual variability as a potential source of variation (Fig. 2) as well as a source of discriminatory features. In addition to understanding and addressing potential methodological challenges and various sources of biases, open science practices are necessary to support the subsequent verification of research and use of the obtained data and results in subsequent secondary analyses and meta-analyses.

Moving from genome-wide association studies (GWAS) to metabolome-wide association studies (MWAS)

One of the most-used study approaches in big data genome research, first demonstrated in 2005, is GWAS

Biological Factors

Accounting for the high biological background arising from inter-subject variablity is an essential part of experimental design in metabolomics. Other sources of variability can include:

- Dietary effects
- Body mass index
- Age
- Sex
- Subjects at high risk of disease
- Subjects at high risk of exposure
- Observer selection
- Timing of subject sampling
- Gene/Environmental effects
- Circadian/Temporal effects

Analytical Factors

Sample preparation and analytical procedures used within metabolomics experiments vary greatly across labs. Sources of variabilty that should be considered for study design include:

- Collection and storage procedures
- Between sample and batch variance
- Between instrumental comparisons
- Lack of global 'Gold Standard' methodology driven by constantly evolving instrumentation (annually) from multiple vendors.
- Selection of extraction solvents
- Selection of column chemistry
- Selection of eluent composition
- Ionization technique and source geometry
- Signal processing algorithms
- Compound library

Fig. 2 The biological and analytical aspects of bias and variance that can lead to a tendency towards erroneous results in both untargeted and targeted metabolomics

Rattray *et al. Human Genomics* (2018) 12:4

Page 8 of 14

[123]. This technique examines genome-wide sets of genetic variants in samples of individuals to determine if any variants are associated with a trait and help pinpoint genes that may contribute to a person's risk for a certain disease or other phenotype of interest. GWAS can be described as an untargeted and sometimes a hypothesis-generating approach to associate genetic variants with specific phenotypes. GWAS and consortia-based meta-analyses have been conducted with increasing sample size [124], allowing for improved power [125] to detect genome-wide significant signals for what are typically very small effect sizes. Due to the analytical uniformity of sequencing, this is one area where genomic research has advanced more quickly than metabolomics.

Most of the early untargeted metabolomics experiments have had limited sample sizes ($n = 10–100$) often a result of technological, run-time, and statistical limitations. Given the large number of metabolic features that are typically generated by untargeted metabolomics (typically 1000s for liquid chromatography mass spectrometry), using such small sample sizes has led to over-fitting of data and spurious results [100]. Moreover, the highly collinear nature of metabolomics multivariate data [67] have not generally been properly factored in performing a priori power and sample size calculations, and there is no widely accepted method for sample size determination in metabolomics. In the absence of specific metabolic target hypothesis, one can use a data driven sample size determination (DSD) algorithm [126] where sample size estimation depends on the purpose of the study: whether it aims to find at least one statistically significant variation (biomarker discovery) or a maximum of statistically significant variations (metabolic exploration). Alternatively, one may adapt methods that have been developed for use with microarray gene expression(s) [127–129]. One common problem is that there is often high correlation between variables in one dataset, and in addition, not all variables have the same power. However, new more promising approaches have been generated using multivariate simulation to deal with this type of data structure [130].

Predictive power increases with sample size, and the current application of metabolomics to larger longitudinal cohort studies ($n > 1000$) is helping to give access to broader population data that can be linked to specific exposure such as alcohol [131, 132]. These types of studies are needed to improve biomarker discovery and inference of molecular mechanisms. Key issues continuously arise in the application of metabolomics to human subjects which can be overcome by putting metabolomics into epidemiological context. Common problems include causal and mechanistic claims based on differences between groups that have low numbers of individuals, lack of longitudinal data to avoid the possibility of reverse causation (a health outcome influencing pharmacokinetics and metabolite concentrations), limited information on lifestyle, socioeconomic and other influences, and the lack of multiple statistical tests and biological replication [133]. As metabolomics is incorporated into more population-level studies, it may be possible to more reliably model potential associations of metabolic profiles with phenotypes. The goal is to stratify metabolic data over exposure event data and ultimately determine the related disease risk. Confounding associations may still distort results and lead to erroneous conclusions. Yet it is more readily possible, with larger study numbers, and longitudinal testing, to control confounding by matching samples in to related sub-groups such as age, sex, or level-of-exposure.

Metabolome-wide association studies (MWAS) were first described in 2008 as the capture of "environmental and genomic influences to investigate the connections between phenotype variation and disease risk factors" [134, 135], thus helping reveal the complex gene-environment interactions on disease outcome. The method differs from conventional metabolomics in that high-throughput metabolomics is applied to large-scale epidemiologic studies at the population level and uses specialized algorithms to maximize the identification of biomarkers of disease risk [57]; for example, a recent algorithm was developed to correct for multiple testing using a permutation-based method to derive a metabolome-wide significance level controlling the family-wise error rate [136]. Initial studies showed that using high-throughput metabolomics, MWAS can be carried out on large population cohorts to provide individual metabolic phenotypes (metabotypes), and metabolic biomarkers correlated to exposures [137], and/or biological outcomes [138]. The proof-of-principle study used to coin the term MWAS identified discriminatory biomarkers of blood pressure and cardiovascular risk in 4630 individuals [138]. These types of studies may point to otherwise unknown features of the disease etiology or pathophysiology, which may be used to lead further mechanistic studies and potentially new avenues for therapeutic design, although the complexity of mechanisms makes such translation to therapeutic discovery very difficult. Comparison of metabotypes at the human population level can identify a signature of metabolites statistically correlated to disease risk and/or an exposure. Recent studies have shown the application of MWAS to identify metabolites correlated with cardiovascular events in a dietary intervention trial [139]. In another study, trimethylamine *N*-oxide (TMAO) was identified as a biomarker predictive of cardiovascular disease risk [140, 141] and was also shown to be involved in the production of atherosclerotic plaques. This discovery has resulted in a clinical test for TMAO, Cleveland HeartLab, and is the first to provide this blood test, and therapeutics are currently being designed to inhibit TMAO

Rattray *et al. Human Genomics* (2018) 12:4

Page 9 of 14

production as well as recommendations for dietary changes. Another application is to identify the enrichment of metabolites within specific biochemical pathways [142] to aid in the identification of genes and proteins/enzymes that may be related to the mechanism of disease. This method has gained traction within drug evaluation studies [143] trying to obtain more comprehensive understanding of individual responses to drug therapy [144, 145]. This application may be particularly useful for the design of immunotherapeutics where metabolites have been shown to modulate autoimmunity and can be targeted to improve the efficacy of these drugs [146, 147]. However, it should be acknowledged that therapeutic discovery or improvement in therapeutic management with known interventions has not yet been accomplished using metabolomics data; however, recent development in metabolomics technologies in both the bioanalytical and chemometric components is markedly improving, and thus, there is optimism for clinical translation as well.

Transparency, reproducibility, and open science

There is growing recognition of the need for improved transparency, reproducibility, and replication in the biomedical literature [64, 91, 148, 149]. With respect to multidimensional, big data analyses, transparency can be improved with the sharing of data, protocols, and analytical codes. Furthermore, the number of metabolomics studies that investigate reproducibility across multiple research centers are few in number, and ongoing interlaboratory efforts have struggled to generate metabolite data that is both accurate and reproducible across different labs [150]. Replication has been accepted as a sine qua non in certain disciplines, such as human genome epidemiology [149], and the same should apply across all multidimensional fields using big data. However, the research community is aware of this issue, and groups are convening to provide solutions to address this problem. For example, the European Centre for Ecotoxicology and Toxicology of Chemicals have provided a framework to facilitate the regulatory applicability and use of big data in chemical risk assessment [151, 152].

It is also important to protect inferences from data dredging/p-hacking (mining datasets prior to specifying a causal hypothesis), and unaccounted multiple comparisons in complex datasets that can lead to the inflation of false-positive rates. Therefore, to improve the reproducibility of metabolomics, it is necessary to understand certain methodological and statistical challenges, to protect against analytical biases and biological variance, and to promote transparency and open science. These open science practices, which include "the process of making the content and process of producing evidence and claims transparent and accessible to other researchers" [64], can increase the credibility of research. For

metabolomics in particular, both raw and metadata are essential to facilitate reproducibility, secondary analyses, and the synthesis of evidence by external metabolomics researchers [153]. Several measures can support the transparency and reproducibility of metabolomics. For maximal impact, the whole metabolomics research community should adopt and adhere to standards that promote the uniform preparation of study results. The metabolomics standards initiative (MSI), which was conceived in 2005 by the Metabolomics Society, highlights a range of minimum reporting standards covering biological [154], chemical [155], analytical, and data reporting methods [156] within the metabolomics experimental pipeline. However, ideally, metabolomics funders, reviewers, editors, and journals should require researchers to share their protocols, raw data, and analytical code. Broadly speaking, this does not happen (the Springer Journal *Metabolomics* (https://link.springer.com/journal/11306) and MDPI journal *Metabolites* (http://www.mdpi.com/journal/metabolites) being notable exceptions in which MSI compliance is asked for from authors and assessed by reviewers). Currently, most journals leave the suitability of metabolite submission data to reviewer and editor discretion.

Support is also beginning to appear from some funding bodies to help improve the reliability and efficiency of metabolomics. For example, the Data Repository and Coordination Center, which is part of the United States National Institutes of Health (NIH) Common Fund's Metabolomics Program, has created the Metabolomics Data Repository. All NIH Common Fund Metabolomics Program supported research projects which create metabolomics data as part of the funded research are required to submit all raw data (e.g., spectrometric, spectrographic, and chromatographic data) and metadata (e.g., details on how samples were obtained and the analytical methods that were used) to the repository [157]. In addition, the European Union funded data repository MetaboLights (http://www.ebi.ac.uk/metabolights/) has already assembled data from 317 metabolomics studies as of December 2017. Common data submission formats, such as *mzML/mzXML* for mass spectrometry, *nmrML* for NMR data, and ISA-Tab format for metadata, have helped to unify this process [158, 159]. But the research community must be careful to not generate an excess of unconnected data repositories. Multiple and potentially overlapping repositories could confuse researchers as to where they should submit their data and therefor limit the chance of uniform acceptance and adoption of standards. To this end, the COSMOS project (COordination of Standards in MetabOlomicS—http://www.cosmos-fp7.eu/) has been designed to address the challenges of e-infrastructure diversity in metabolomics by developing an interface that globally links community projects and output.

Rattray *et al. Human Genomics* (2018) 12:4

Page 10 of 14

The predominant reason behind the lack of data sharing in metabolomics is the complexity and lack of standardization in the data generated. For research areas such as genomics, transcriptomics, and, to a lesser extent, proteomics, the chemistry of the molecules under detection is highly symmetrical. Regardless of nucleobase-pair connectivity, DNA and RNA constructs can be detected and typed using highly reproducible sequencing chips that can work in a high-throughput manner. The sheer range of molecular chemistries available within the human metabolome demand a multitude of separation strategies when mass spectrometry is used as the detection technology. Consequently, different research groups align their experimental pipelines to one of the many instrument vendors (often dictated by geography and cost) leading to a multitude of protocols that cover all aspects of experimentation. Just within the confines of liquid chromatography mass spectrometry-based metabolomics, 84% use open source software and/or commercial software from instrument vendors, and within the open source software group, the majority use XCMS, and a smaller percentage use MZmine and MZmine 2. Therefore, variability in just the data processing limits integration of the MSI. One way to enable standardized data processing and biostatistics is to encourage the use of a universal workflow platform such as Galaxy (https://galaxyproject.org) [160]. In addition, the use of a standard reference material that can normalize and compare the detection levels from different instruments would be of value. A concerted effort is still needed by the community to enable broader reproducibility [161]. The lack of standardization and reporting is preventing the validation of metabolomics research [162].

Conclusions

Human populations are exposed to a complex mix of chemicals and toxicants, from multiple sources, for varying durations. These exposures are affecting the health of the global population dramatically, for example, over seven million premature deaths annually linked to air pollution exposure alone [163]. It is vital that a more comprehensive understanding of how these environmental exposures affect biology at the systems level to determine cause, effect, and susceptibilities. In doing so, a compound specific "exposotype" can be developed that accounts for the totality of the multileveled downstream biological changes that an individual exposure event produces [18]. To better understand these effects, metabolomics can be used to develop not only metabolic biomarkers of exposure but can also be used to build metabolic models that identify upstream genetic and enzymatic changes. This may complement GWAS studies as knowledge of a potential enzymatic mutation can narrows down the DNA search space needed to identify relevant SNPs linked to the exposure [144, 145].

In-depth biological data generated by metabolomics can be used to enhance exposure studies by supplying information not only on directly affected metabolic pathways but also on off-target metabolic effects. The value of metabolomics to identify gene-environment interactions lends itself to the study of the exposome and will be the most complex and important integration of metabolomics to date. Further characterization of gene variants associated with those metabolic pathways could help forecast disease prevalence by either using pre-diagnostic metabolic signatures (collections of metabolites that change prior to disease onset) and genetic risk data. Therefore, preventive measures may be tailored specifically for those individuals. The combination of metabolomics with genomics offers one tool that may prove helpful towards materializing precision medicine. Success in precision medicine has been difficult to achieve [164], but the recent US Food and Drug Administration approval of pembrolizumab, a "tumor-agnostic" therapeutic which targets any solid tumor with a specific genetic feature, shows that the field is starting to head in that direction [165]. Given recent evidence that non-genomic influences such as the microbiome can influence therapeutic response, metabolomics may be used in this context to identify factors that are related to non-responders and responders [166].

However, some of the caveats that still exist within conventional metabolomics and population studies are still present, such as accurate identification of new metabolites, controlling for multiple levels of confounders, and the integration of different forms of data from different analytical platforms. Further advancement can be made by routine application of appropriate statistical tools to metabolomics as well as the adoption and promotion of transparent and reproducible research practices. Reproducible, transparent advances may then be examined for their impact in changing outcomes in single patients and at the population level to judge their utility.

Rattray *et al. Human Genomics* (2018) 12:4

Page 11 of 14

Publisher's Note

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

Author details
[1]Department of Environmental Health Sciences, Yale School of Public Health, Yale University, New Haven, CT, USA. [2]Collaboration for Research Integrity and Transparency (CRIT), Yale Law School, New Haven, CT, USA. [3]Center for Outcomes Research and Evaluation (CORE), Yale-New Haven Health System, New Haven, CT, USA. [4]Department of Surgery, Section of Surgical Oncology, Yale University School of Medicine, New Haven, CT, USA. [5]Yale Cancer Center, Yale University School of Medicine, New Haven, CT, USA. [6]Stanford Prevention Research Center, Department of Medicine, Stanford University, Stanford, CA, USA. [7]Department of Health Research and Policy, Stanford University, Stanford, CA, USA. [8]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [9]Department of Statistics, Stanford University, Stanford, CA, USA. [10]Meta-Research Innovation Center at Stanford, Stanford University, Stanford, CA, USA.

References

1. Neel JV, Schull WJ. Human heredity. Chicago: Chicago Press; 1954.
2. DeWan AT. Five classic articles in genetic epidemiology. Yale J Biol Med. 2010;83:87–90.
3. Beaty TH, Khoury MJ. Interface of genetics and epidemiology. EpidemiolRev. 2000;22:120–5.
4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74:5463–7.
5. National Human Genome Research Institute. All about the Human Genome Project (HGP). 2014. Available from: http://www.genome.gov/10001772. Accessed 17 Jan 2018.
6. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52:413–35.
7. Hayden EC. The $1,000 genome. Nature. 2014;507:295.
8. Goldfeder RL, Wall DP, Khoury MJ, JPA I. Human genome sequencing at population scale: a primer on high throughput DNA sequencing and analysis. Am J Epidemiol. 2017;186:1000–9.
9. Goodwin S, JD MP, WR MC. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
10. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
11. Theodoratou E, Timofeeva M, Li X, Meng X, JPA I. Nature, nurture, and cancer risks: genetic and nutritional contributions to cancer. Annu Rev Nutr. 2017;21:293–320.
12. Willett WC. Balancing life-style and genomics research for disease prevention. Science (80-). 2002;296:695–8.
13. Rappaport SM, Smith MT. Environment and disease risks. Science (80-). 2010;330:460–1.
14. Wild CP. Complementing the genome with an "Exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers. 2005;14:1847. LP-1850
15. Patel CJ, Ioannidis JPA. Studying the elusive environment in large scale. JAMA. 2014;311:2173–4.
16. Romero R, Espinoza J, Gotsch F, Kusanovic JP, Friel LA, Erez O, et al. The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. BJOG. 2006;113:118–35.
17. Wilson SH. Disease-first: a new paradigm for environmental health science research. Environ Health Perspect. 2006;114:2006.
18. Rattray NJW, Charkoftaki G, Rattray Z, Hansen JE, Vasiliou V, Johnson CH. Environmental influences in the etiology of colorectal cancer: the premise of metabolomics. Curr Pharmacol Reports. 2017;3:114–25.
19. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012;13:263–9.
20. Ellis JK, Athersuch TJ, Thomas LD, Teichert F, Perez-Trujillo M, Svendsen C, et al. Metabolic profiling detects early effects of environmental and lifestyle exposure to cadmium in a human population. BMC Med. 2012;10:61.
21. Andra SS, Austin C, Wright RO, Arora M. Reconstructing pre-natal and early childhood exposure to multi-class organic chemicals using teeth: towards a retrospective temporal exposome. Environ Int. 2015;83:137–45.
22. Maitre L, Villanueva CM, Lewis MR, Ibarluzea J, Santa-Marina L, Vrijheid M, et al. Maternal urinary metabolic signatures of fetal growth and associated clinical and environmental factors in the INMA study. BMC Med. 2016;14:1–12.
23. Baker MG, Simpson CD, Lin YS, Shireman LM, Seixas N. Original article the use of metabolomics to identify biological signatures of manganese exposure. Ann Work Expo Heal. 2017;61:406–15.
24. Walker DI, Uppal K, Zhang L, Vermeulen R, Smith M, Hu W, et al. High-resolution metabolomics of occupational exposure to trichloroethylene. Int J Epidemiol. 2016;45:1517–27.
25. van Veldhoven K, Keski-Rahkonen P, Barupal DK, Villanueva CM, Font-Ribera L, Scalbert A, et al. Effects of exposure to water disinfection by-products in a swimming pool: a metabolome-wide association study. Environ Int Elsevier. 2018;111:60–70.
26. Patel CJ. Analytic complexity and challenges in identifying mixtures of exposures associated with phenotypes in the Exposome era. Curr Epidemiol Reports. 2017;4:22–30.
27. Patel CJ, Kerr J, Thomas DC, Mukherjee B, Ritz B, Chatterjee N, et al. Opportunities and challenges for environmental exposure assessment in population-based studies. Cancer Epidemiol Biomarkers Prev. 2017;26:cebp.0459.2017.
28. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: the toxic exposome database. Nucleic Acids Res. 2015;43:D928–34.
29. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, et al. T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res. 2009;38:781–6.
30. Warth B, Spangler S, Fang M, Johnson CH, Forsberg EM, Granados A, et al. Exposome-scale investigations guided by global metabolomics, pathway analysis, and cognitive computing. Anal Chem. 2017; In-Press
31. Richard AM, Williams CR. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutat Res Fundam Mol Mech Mutagen. 2002;499:27–52.
32. Neveu V, Moussy A, Rouaix H, Wedekind R, Pon A, Knox C, et al. Exposome-explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res. 2017;45:D979–84.
33. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the human metabolome database. Nucleic Acids Res. 2007;35:521–6.
34. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. A metabolite mass spectral database. Ther Drug Monit. 2005;27:747–51.
35. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, et al. Metabolite identification via the Madison Metabolomics Consortium Database [3]. Nat Biotechnol. 2008;26:162–4.
36. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34:D668–72.
37. Kaiser J. Chemists want NIH to curtail database. Science (80-). 2005;308:774.
38. Williams AJ, Grulke CM, Edwards J, AD ME, Mansouri K, Baker NC, et al. The CompTox chemistry dashboard: a community data resource for environmental chemistry. J Cheminform. 2017;9:61.
39. Beger RD, Dunn W, Schmidt MA, Gross SS, Kirwan JA, Cascante M, et al. Metabolomics enables precision medicine: "a white paper, community perspective". Metabolomics. 2016;12:149.
40. Chen C, Krausz KW, Idle JR, Gonzalez FJ. Identification of novel toxicity-associated metabolites by metabolomics and mass isotopomer analysis of acetaminophen metabolism in wild-type and Cyp2e1-null mice. J Biol Chem. 2008;283:4543–59.

Rattray *et al. Human Genomics* (2018) 12:4

Page 12 of 14

41. Johnson CH, Krausz KW, Kang DW, Patterson AD, Kim J, Luecke H, et al. Novel metabolites and roles for a-tocopherol in humans and mice discovered by mass spectrometry-based metabolomics 1–5. Am J Clin Nutr. 2012;96:818–30.

42. Chen C, Ma X, Malfatti MA, Krausz KW, Kimura S, Felton JS, et al. A comprehensive investigation of 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) metabolism in the mouse using a multivariate data analysis approach. Chem Res Toxicol. 2007;20:531–42.

43. Sharma AK, Jaiswal SK, Chaudhary N, Sharma VK. A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota. Sci Rep. 2017;7:1–13.

44. Gavaghan CL, Holmes E, Lenz E, Wilson ID, Nicholson JK. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. FEBS Lett 2000;484:169–174.

45. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 2016;17:451–9.

46. Mahieu NG, Patti GJ. Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites. Anal Chem. 2017;89:10397–406.

47. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012;84:283–9.

48. Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE. RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. Anal Chem. 2014;86:6812–7.

49. Mahieu NG, Huang X, Chen YJ, Patti GJ. Credentialing features: a platform to benchmark and optimize untargeted metabolomic methods. Anal Chem. 2014;86:9583–9.

50. da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. Proc Natl Acad Sci. 2015;112:12549–50.

51. Huan T, Forsberg EM, Rinehart D, Johnson CH, Ivanisevic J, Benton HP, et al. Systems biology guided by XCMS online metabolomics. Nat Methods. 2017;14:461–2.

52. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal Chem. 2009;81:6656–67.

53. Lankadurai BP, Nagato EG, Simpson MJ. Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. Environ Rev. 2013;21:180–205.

54. Johnson CH, Ivanisevic J, Benton HP, Siuzdak G. Bioinformatics: the next frontier of metabolomics. Anal Chem. 2015;87:147–56.

55. Misra BB, van der Hooft JJJ. Updates in metabolomics tools and resources: 2014-2015. Electrophoresis. 2016;37:86–110.

56. Misra BB, Fahrmann JF, Grapov D. Review of emerging metabolomic tools and resources: 2015–2016. Electrophoresis. 2017;38:2257–74.

57. Chan Q, Loo R, Ebbels T, Van Horn L, Daviglus M, Stamler J, et al. Metabolic phenotyping for discovery of urinary biomarkers of diet, xenobiotics and blood pressure in the INTERMAP study: an overview. Hypertens Res. 2016;40:1–10.

58. Karaman I, Ferreira DLS, Boulangé CL, Kaluarachchi MR, Herrington D, Dona AC, et al. Workflow for integrated processing of multicohort untargeted ^{1}H NMR metabolomics data in large-scale metabolic epidemiology. J Proteome Res. 2016;15:4188–94.

59. Ioannidis J, Allison D, Ball C, Coulibaly I, Cui X, Culhane A, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009;41:149–204.

60. Kraft P, Zeggini E, Ioannidis J. Replication in genome-wide association studies. Stat Sci. 2010;24:561–73.

61. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:0696–701.

62. Chadeau-Hyam M, Athersuch TJ, Keun HC, De Iorio M, TMD E, Jenab M, et al. Meeting-in-the-middle using metabolic profiling—a strategy for the identification of intermediate biomarkers in cohort studies. Biomarkers. 2011;16:83–8.

63. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomark Prev. 2007;16:1954–65.

64. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1:1–9.

65. Xu Y, Goodacre R. Multiblock principal component analysis: an efficient tool for analyzing metabolomics data which contain two influential factors. Metabolomics. 2012;8:37–51.

66. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. Wiley Interdiscip Rev Comput Stat. 2013;5:149–79.

67. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. Meta. 2012;2:775–95.

68. Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36:666–76.

69. Szklo M, Nieto FJ. Epidemiology: beyond the basics. 3rd Ed. Aspen: Jones & Bartlett Learning; 2000.

70. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.

71. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011;25:534–55.

72. Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med. 1997;29:221–39.

73. Eriksson L, Antti H, Gottfries J, Holmes E, Johansson E, Lindgren F, et al. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). Anal Bioanal Chem. 2004;380:419–29.

74. DiBello JR, Kraft P, ST MG, Goldberg R, Campos H, Baylin A. Comparison of 3 methods for identifying dietary patterns associated with risk of disease. Am J Epidemiol. 2008;168:1433–43.

75. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. J Chemom. 1998;12:301–21.

76. Zwanenburg G, Huub CJ, Westerhuis JA, Jansen JJ, Smilde AK. ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. J Chemom. 2011;25:561–7.

77. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: metabolomics and partial least squares-discriminant analysis—a marriage of convenience or a shotgun wedding. Anal Chim Acta. 2015;879:10–23.

78. Jombart T, Devillard S, Balloux F, Falush D, Stephens M, Pritchard J, et al. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11:94.

79. Ogutu JO, Schulz-Streeck T, Piepho H-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 2012;6:S10.

80. Acharjee A, Finkers R, Visser RG, Maliepaard C. Comparison of regularized regression methods for ~omics data. Metabolomics. 2013;3:126.

81. Tzoulaki I, Ebbels TMD, Valdes A, Elliott P, JPA I. Design and analysis of metabolomics studies in epidemiologic research: a primer on-omic technologies. Am J Epidemiol. 2014;180:129–39.

82. Abdi H. Partial least squares (PLS) regression. Encycl Res Methods Soc Sci. 2003;2003:792–5.

83. Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom. 2006;20:3541–351.

84. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst. 2001;58:109–30.

85. Waldron L, Pintilie M, Tsao MS, Shepherd FA, Huttenhower C, Jurisica I. Optimized application of penalized regression methods to diverse genomic data. Bioinformatics. 2011;27:3399–406.

86. Tibshirani R. Regression selection and shrinkage via the lasso. J R Stat Soc B. 1996;58:267–88.

87. Vaarhorst AAM, Verhoeven A, Weller CM, Böhringer S, Göraler S, Meissner A, et al. A metabolomic profile is associated with the risk of incident coronary heart disease. Am Heart J. 2014;168:45–52. e7

88. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016;531:151.

89. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol Springer Netherlands. 2016;31:337–50.

90. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006;2:171–96.

Rattray *et al. Human Genomics* (2018) 12:4

Page 13 of 14

91. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA, Gigerenzer G, Berlin J, et al. Evolution of reporting P values in the biomedical literature, 1990–2015. JAMA. 2016;315:1141.

92. The American Statistical Association. Statement on statistical significance and P-values. 2016;

93. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70:129–33.

94. Chong EY, Huang Y, Wu H, Ghasemzadeh N, Uppal K, Quyyumi AA, et al. Local false discovery rate estimation using feature reliability in LC/MS metabolomics data. Sci Rep. 2015;5:17221.

95. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. Curr Bioinforma. 2012;7:96–108.

96. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995:289–300.

97. McDonald JH. Handbook of biological statistics. Baltimore: Sparky House Publishing; 2015.

98. Efron B. Size, power and false discovery rates. Ann Stat. 2007;35:1351–77.

99. Efron B. Microarrays, empirical Bayes and the two-groups model. Stat Sci. 2008;23:1–22.

100. Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. Comput Struct Biotechnol J. 2013;4: e201301009

101. Lewis FI, Ward MP. Improving epidemiologic data analyses through multivariate regression modelling. Emerg Themes Epidemiol. 2013;10:2–11.

102. Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci U S A. 2006;103:19430–5.

103. Saccenti E, Hoefsloot HCJ, Smilde AK, Westerhuis JA, MMWB H. Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics. 2014;10:361–74.

104. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting network activity from high throughput metabolomics. PLoS Comput Biol. 2013;9

105. Kaever A, Landesfeind M, Feussner K, Mosblech A, Heilmann I, Morgenstern B, et al. MarVis-pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. Metabolomics. 2015;11:764–77.

106. Wrzodek C, Eichner J, Büchel F, Zell A. InCroMAP: integrated analysis of cross-platform microarray and pathway data. Bioinformatics. 2013;29:506–8.

107. Sergushichev AA, Loboda AA, Jha AK, Vincent EE, Driggers EM, Jones RG, et al. GAM: a web-service for integrated transcriptional and metabolic network analysis. Nucleic Acids Res. 2016;44:W194–200.

108. Zhang L, CM MH, Rothman N, Li G, Ji Z, Vermeulen R, et al. Systems biology of human benzene exposure. Chem Biol Interact. 2010;184:86–93.

109. Maitre L, Fthenou E, Athersuch T, Coen M, Toledano MB, Holmes E, et al. Urinary metabolic profiles in early pregnancy are associated with preterm birth and fetal growth restriction in the Rhea mother-child cohort study. BMC Med. 2014;12:1–14.

110. Roede JR, Uppal K, Park Y, Tran VL, Jones DP. Transcriptome-metabolome wide association study (TMWAS) of maneb and paraquat neurotoxicity reveals network level interactions in toxicologic mechanism. Toxicol Rep. 2014;1:435–44.

111. Cribbs SK, Uppal K, Li S, Jones DP, Huang L, Tipton L, et al. Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. Microbiome. 2016;4:1–11.

112. Chandler JD, Hu X, Ko E-J, Park S, Lee Y-T, Orr ML, et al. Metabolic pathways of lung inflammation revealed by high-resolution metabolomics (HRM) of H1N1 influenza virus infection in mice. Am J Physiol Regul Integr Comp Physiol. [Internet]. 2016;ajpregu.00298.2016.

113. Johnson CH, Athersuch TJ, Collman GW, Dhungana S, Grant DF, Jones DP, et al. Yale school of public health symposium on lifetime exposures and human health: the exposome; summary and future reflections. Hum Genomics. 2017;11:32.

114. Wang SY, Kuo CH, Tseng YJ. Batch normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration met. Anal Chem. 2013;85:1037–46.

115. Reisetter AC, Muehlbauer MJ, Bain JR, Nodzenski M, Stevens RD, Ilkayeva O, et al. Mixture model normalization for non-targeted gas chromatography/ mass spectrometry metabolomics data. BMC Bioinformatics. 2017;18:84.

116. NCBI dbSNP Database - www.ncbi.nlm.nih.gov/projects/SNP. Accessed 6 Nov 2017.

117. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.

118. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 2008;5:1320–32.

119. Stefano GB. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res. 2014;20:138–42.

120. Siddiqui AS, Delaney AD, Schnerch A, Griffith OL, Jones SJM, Marra MA. Sequence biases in large scale gene expression profiling data. Nucleic Acids Res. 2006;34:e84.

121. Büscher JM, Czernik D, Ewald JC, Sauer U, Zamboni N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. Anal Chem. 2009;81:2135–43.

122. Ivanisevic J, Zhu ZJ, Plate L, Tautenhahn R, Chen S, O'Brien PJ, et al. Toward 'omic scale metabolite profiling: a dual separation-mass spectrometry approach for coverage of lipid and central carbon metabolism. Anal Chem. 2013;85:6876–84.

123. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science (80-.). 2005;308:385–9.

124. Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA. The power of meta-analysis in genome-wide association studies. Annu Rev Genomics Hum Genet. 2013;14:441–65.

125. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform. 2012;10:117–22.

126. Blaise BJ. Data-driven sample size determination for metabolic phenotyping studies. Anal Chem. 2013;85:8943–50.

127. Van Iterson M, 't Hoen PAC, Pedotti P, Hooiveld GJ, Den Dunnen JT, van Ommen GJ, et al. Relative power and sample size analysis on gene expression profiling data. BMC Genomics. 2009;10:439.

128. Ferreia JA, Zwinderman A. Approximate power and sample size calculations with the Benjamini-Hochberg method. Int J Biostat. 2006;2:1–36.

129. Langaas M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J R Stat Soc Ser B Stat Methodol. 2005;67:555–72.

130. Blaise BJ, Correia G, Tin A, Young JH, Vergnaud AC, Lewis M, et al. Power analysis and sample size determination in metabolic phenotyping. Anal Chem. 2016;88:5179–88.

131. Jaremek M, Yu Z, Mangino M, Mittelstrass K, Prehn C, Singmann P, et al. Alcohol-induced metabolomic differences in humans. Transl Psychiatry. 2013;3:e276.

132. Homuth G, Teumer A, Völker U, Nauck M. A description of large-scale metabolomics studies: increasing value by combining metabolomics with genome-wide SNP genotyping and transcriptional profiling. J Endocrinol. 2012;215:17–28.

133. Mäkinen V-P, Ala-Korpela M. Metabolomics of aging requires large-scale longitudinal studies with replication. Proc Natl Acad Sci 2016; 113:E3470–3470.

134. Nicholson JK, Holmes E, Elliott P. The metabolome-wide association study: a new look at human disease risk factors. J Proteome Res. 2008;7:3637–8.

135. Chadeau-Hyam M, Ebbels TM, Brown IJ, Chan Q, Stamler J, et al. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. J Proteome Res. 2010;9:4620-7.

136. Castagné R, Boulangé CL, Karaman I, Campanella G, Santos Ferreira DL, Kaluarachchi MR, et al. Improving visualization and interpretation of metabolome-wide association studies: an application in a population-based cohort using untargeted 1 H NMR metabolic profiling. J Proteome Res. 2017;16:3623–33.

137. Walker DI, Pennell KD, Uppal K, Xia X, Hopke PK, Utell MJ, et al. Pilot Metabolome-Wide Association Study of Benzo(a)pyrene in Serum From Military Personnel. J Occup Environ Med. 2016;58:S44-52.

138. Bictash M, Ebbels TM, Chan Q, Loo RL, Yap IKS, Brown IJ, et al. Opening up the "black box": metabolic phenotyping and metabolome-wide association studies in epidemiology. J Clin Epidemiol Elsevier Inc. 2010;63:970–9.

139. Toledo E, Wang DD, Ruiz-Canela opez M, Clish CB, Razquin C, Zheng Y, et al. Plasma lipidomic profiles and cardiovascular events in a randomized intervention trial with the Mediterranean diet. Am J Clin Nutr. 2017;106:973–83.

140. Li XS, Obeid S, Klingenberg R, Gencer B, Mach F, Räber L, et al. Gut microbiota-dependent trimethylamine N-oxide in acute coronary syndromes: a prognostic marker for incident cardiovascular events beyond traditional risk factors. Eur Heart J. 2017;14:814–24.

Rattray *et al. Human Genomics* (2018) 12:4

Page 14 of 14

141. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. 2011;472:57–63.

142. Igari M, Alexander JC, Ji Y, Qi XL, Papke RL, Bruijnzeel AW. Varenicline and cytisine diminish the dysphoric-like state associated with spontaneous nicotine withdrawal in rats. Neuropsychopharmacology. 2014;39:445–55.

143. Renier N, Adams EL, Kirst C, Wu Z, Azevedo R, Kohl J, et al. Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. Cell. 2016;165:1789–802.

144. Gupta M, Neavin D, Liu D, Biernacka J, Hall-Flavin D, Bobo WV, et al. TSPAN5, ERICH3 and selective serotonin reuptake inhibitors in major depressive disorder: pharmacometabolomics-informed pharmacogenomics. Mol Psychiatry. 2016;21:1717–25.

145. Ji Y, Hebbring S, Zhu H, Jenkins GD, Biernacka J, Snyder K, et al. Glycine and a glycine dehydrogenase (GLDC) SNP as citalopram/escitalopram response biomarkers in depression: pharmacometabolomics-informed pharmacogenomics. Clin Pharmacol Ther. 2011;89:97–104.

146. Kepp O, Loos F, Liu P, Kroemer G. Extracellular nucleosides and nucleotides as immunomodulators. Immunol Rev. 2017;280:83–92.

147. Johnson CH, Spilker ME, Goetz L, Peterson SN, Siuzdak G. Metabolite and microbiome interplay in cancer immunotherapy. Cancer Res. 2016;76:6146–52.

148. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383:166–75.

149. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, JPA I. Reproducible research practices and transparency across the biomedical literature. PLoS Biol. 2016;14:1–13.

150. Siskos AP, Jain P, Römisch-Margl W, Bennett M, Achaintre D, Asad Y, et al. Interlaboratory reproducibility of a targeted Metabolomics platform for analysis of human serum and plasma. Anal Chem. 2017;89:656–65.

151. Buesen R, Chorley BN, da Silva Lima B, Daston G, Deferme L, Ebbels T, et al. Applying 'omics technologies in chemicals risk assessment: report of an ECETOC workshop. Regul Toxicol Pharmacol. 2017:1–11.

152. Kauffmann HM, Kamp H, Fuchs R, Chorley BN, Deferme L, Ebbels T, et al. Framework for the quality assurance of 'omics technologies considering GLP requirements. Regul Toxicol Pharmacol. 2017;91:1–9.

153. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016;44:D463–70.

154. Griffin JL, Nicholls AW, Daykin CA, Heald S, Keun HC, Schuppe-Koistinen I, et al. Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments. Metabolomics. 2007;3:179–88.

155. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis: chemical analysis working group (CAWG) metabolomics standards initiative (MSI). Metabolomics. 2007;3:211–21.

156. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, et al. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics. 2007;3:231–41.

157. Metabolomics workbench - www.metabolomicsworkbench.org/. Accessed 17 Jan 2018.

158. Schober D, Jacob D, Wilson M, Cruz JA, Marcu A, Grant JR, et al. nmrML: a community supported open data standard for the description, storage, and exchange of NMR data. Anal Chem. 2017. In-Press

159. Rocca-Serra P, Salek RM, Arita M, Correa E, Dayalan S, Gonzalez-Beltran A, et al. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics. 2016;12:1–13.

160. Weber RJM, Lawson TN, Salek RM, Ebbels TMD, Glen RC, Goodacre R, et al. Computational tools and workflows in metabolomics: an international survey highlights the opportunity for harmonisation through galaxy. Metabolomics. 2017;13:1–5.

161. Salek RM, Steinbeck C, Viant MR, Goodacre R, Dunn WB. The role of reporting standards for metabolite annotation and identification in metabolomic studies. Gigascience. 2013;2:13.

162. van Rijswijk M, Beirnaert C, Caron C, Cascante M, Dominguez V, Dunn WB, et al. The future of metabolomics in ELIXIR. F1000Research. 2017;6:1649.

163. WHO. 7 million premature deaths annually linked to air pollution - http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/. Accessed 17 Jan 2018.

164. Shin SH, Bode AM, Dong Z. Precision medicine: the foundation of future cancer therapeutics. Precis Oncol. 2017;1:12.

165. FDA approves first cancer treatment for any solid tumor with a specific genetic feature - https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm560167.htm. Accessed 17 Jan 2018.

166. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016;535:94–103.

167. Zou H, Hastie T. Regularization and variable selection via the elastic-net. J R Stat Soc. 2005;67:301–20.