Scaling-up metabolomics: Current state and perspectives

Ghina Hajjar, Millena Cristina Barros Santos, Justine Bertrand-Michel, Cécile Canlet, Florence A Castelli, Nicolas Creusot, Sylvain Dechaumet, Binta Diémé, Franck Giacomoni, Patrick Giraudeau, et al.

Trends in Analytical Chemistry 167 (2023) 117225

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

journal homepage: www.elsevier.com/locate/trac

Scaling-up metabolomics: Current state and perspectives

Ghina Hajjar [a,1], Millena C. Barros Santos [b,1], Justine Bertrand-Michel [c], Cécile Canlet [d], Florence Castelli [e], Nicolas Creusot [f], Sylvain Dechaumet [e], Binta Diémé [a], Franck Giacomoni [a], Patrick Giraudeau [g], Yann Guitton [h], Etienne Thévenot [e], Marie Tremblay-Franco [d], Christophe Junot [e], Fabien Jourdan [d], François Fenaille [e], Blandine Comte [a,**], Pierre Pétriacq [b,***], Estelle Pujos-Guillot [a,*]

[a] Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand, France
[b] Université de Bordeaux, INRAE, Biologie du Fruit et Pathologie, UMR 1332, Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, 71 av E. Bourlaux, 33140, Villenave d'Ornon, France
[c] I2MC, Université de Toulouse, Inserm, Université Toulouse III – Paul Sabatier (UPS), MetaboHUB, 31432, Toulouse, France
[d] Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Plateforme Metatoul-AXIOM, MetaboHUB, 31300, Toulouse, France
[e] Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), MetaboHUB, 91191, Gif-sur-Yvette, France
[f] INRAE, UR EABX, Bordeaux Metabolome, 50 avenue de Verdun Gazinet, F-33612, Cestas, France
[g] Nantes Université, CNRS, CEISAM, UMR, 6230, Nantes, France
[h] Oniris, INRAE, LABERCA, 44300, Nantes, France

ARTICLE INFO

Keywords:
Metabolomics
Large-scale
Cohort
Mass spectrometry
Nuclear magnetic resonance
Data science
Interoperability
One-health

ABSTRACT

Metabolomics is now a mature phenotyping tool that provides substantial results within various scientific communities. Its application at large-scale, i.e. on large populations and/or samples, has shown its power for research activities from plant science to human epidemiology and medicine, but it still needs key methodological developments for its routine application. Here, we review the current state of large-scale metabolomics applications, providing recent examples of large cohort studies in human and plant/environment research, and present the remaining scientific challenges of both fields. Then, we address the key common methodological issues, from analytics to data science, to fulfil these objectives and go towards a more comprehensive and interoperable large-scale metabolomics, making it a new key actor in the frame of the One-Health future research.

1. Introduction

Exploration of metabolites, reflecting a series of biological processes modulated by genetic and environmental changes, is of major interest to fill the gaps between genotypes and phenotypes. Metabolic profiling was first applied in the fields of drug discovery and chemotaxonomy [1–3]. Metabolomics, defined as the comprehensive analysis of the small molecular weight compounds present in a biological system, emerged in the late 1990s and began to spread in various fields such as plant sciences, nutrition, pharmacology, medicine, and more recently environmental research [4]. As the metabolome is most closely linked to the phenotype, metabolomics is considered the ultimate strategy to decipher responses to internal and external stimuli, and thus discover new associated biomarkers. Today, it has increased in maturity and applicability, providing a phenotyping tool and systems biology approach [5–9].

Large-scale metabolomics, i.e. when applied to large populations and/or large numbers of samples (>1000), has shown its ability to provide substantial results within diverse scientific communities over the whole spectrum of life science (from unicellular to multicellular organisms). It successfully allowed defining individual phenotypes and their changes, elucidating the effects of factors (e.g. genetic, environment, intervention, senescence/ageing...), discovering biomarkers and

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: blandine.comte@inrae.fr (B. Comte), pierre.petriacq@inrae.fr (P. Pétriacq), estelle.pujos-guillot@inrae.fr (E. Pujos-Guillot).
[1] They should be considered as co-first authors.

https://doi.org/10.1016/j.trac.2023.117225
Received 22 March 2023; Received in revised form 19 July 2023; Accepted 7 August 2023
Available online 12 August 2023

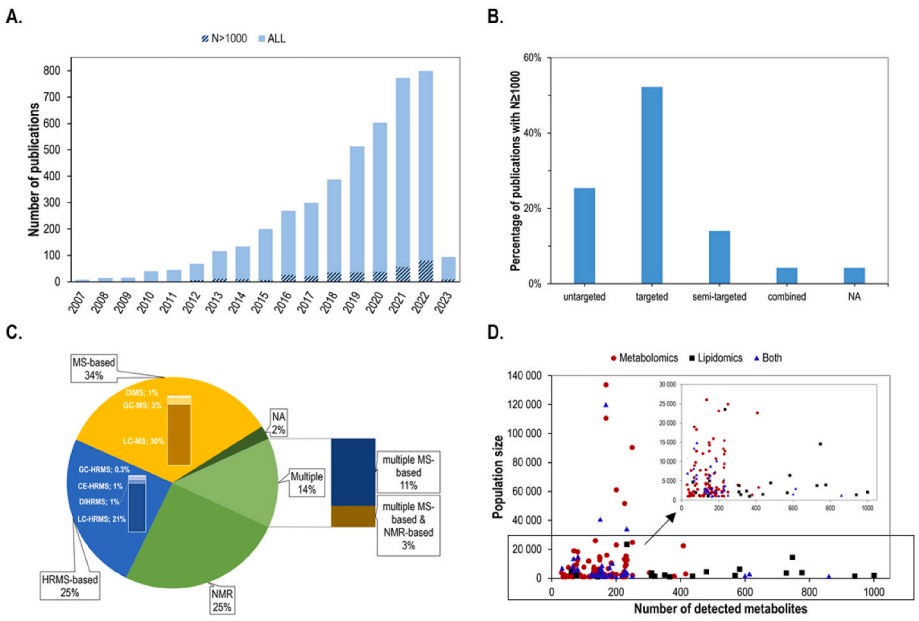

Fig. 1. A. Evolution of publication number, including metabolomics analyses of human cohorts (N = 4385), of which are studies involving over 1000 individuals (N = 335). **B.** Distribution of metabolomics approaches among studies involving over 1000 individuals. **C.** Distribution of analytical platforms used in metabolomics/lipidomics studies involving over 1000 individuals. NA indicates that information was not available in the paper. 'Combined' approach refers to studies that use more than one approach, for example targeted and untargeted methods. **D.** Number of detected metabolites in terms of population size analysed by targeted either metabolomics, lipidomics or both (N = 157).

validating metabolite patterns, that are characteristic of particular biological states [3,10,11]. A literature search (see section 2) showed that large-scale metabolomics studies are on the rise and currently represent 8% of the 4385 publications involving human cohorts (from PubMed, Feb. 2023), and ca. 9% of the 2174 in plant and environmental sciences (from WebOfScience, Feb. 2023).

Due to challenges associated with the analysis of metabolites of great chemical diversity, a stepwise strategy from exploration to validation is currently used to ensure high-quality data. First, metabolite profiling is commonly performed on a "limited" number (<1000) of individuals/samples and seeks to detect as many metabolites as possible in those samples. Then, some potential biomarkers are selected for validation in larger populations (from independent/multi-centre cohorts) for later use in routine practices, often using quantitative approaches. Generally, metabolite profiling is based on untargeted and/or semi-targeted approaches, applied to hundreds to thousands of samples and implies various methods specific to each laboratory, analytical platform and/or matrix. Alternatively, large-scale targeted approaches often relying on commercial solutions/kits (*e.g.*, Metabolon, Inc. (Morrisville, NC, USA), Biocrates® (Biocrates Life Sciences AG, Innsbruck, Austria)) can be used up to hundreds of thousands of individuals, which offer standardised time-effective procedures, yet on a restricted set of known metabolites.

However, even though metabolomics studies are constantly increasing and improving in plant and environmental sciences, as well as in system epidemiology/medicine, large-scale applications are still limited due to technical bottlenecks. Despite specific dedicated and standardised operating protocols [12,13] as well as recommendations for minimum reporting standards for chemical analysis [14], the limited interoperability and the lack of replicability between experiments and facilities prevent large intercomparable studies [15,16]. More precisely, the main requirements for advancing the field are comprehensive metabolite coverage and confident metabolite identification, reproducible, interoperable and robust data production workflows, throughput matching demand, accessible methods to users in terms of costs and results (user-friendly data visualisation and interpretation), and FAIR (Findable-Accessible-Interoperable-Reusable) data.

Beyond these common methodological issues and the mutual technical benefit of leveraging these bottlenecks, there is a crucial scientific interest in developing a more interdisciplinary analysis of the current biological challenges within transversal projects in the context of the One-Health concept. Indeed, this concept acknowledges that human health is interconnected with the health of ecosystems, including microbes and plants, in which they coexist. In this context, metabolomics has the potential of being a key actor in improving and preserving human health, through optimising crop quality, identifying plant-derived bioactive compounds, and increasing the efficacy of prevention and treatments *via* more personalised approaches.

Here, we review the current state of large-scale metabolomics applications, providing recent examples of large cohort studies in human and plant research, and present the remaining scientific challenges of both fields. Then, we address the key common methodological issues, from analytics to data science, to fulfil these objectives and achieve more comprehensive and interoperable large-scale metabolomics.

2. Large-scale metabolomics: definition and alternative approaches

Recent advances in analytical techniques, particularly in mass spectrometry (MS) and nuclear magnetic resonance (NMR), allowed increasing data quality in metabolomics in terms of sensitivity and robustness, opening the door to its large-scale application. In literature, the term 'large-scale' is still ambiguous as it is used to describe two different concepts, either related to the population size, or sometimes to the metabolome coverage (*i.e.*, number of metabolites detected). However, the present review focuses on its application to large cohorts and/or series of samples (>1000).

Because of challenges associated with analysing metabolites of a great chemical diversity present in wide concentration ranges, different alternative complementary techniques/approaches do coexist in cohort studies, and are referred to under different terms. As an example, lipidomics can be described as a subsection of metabolomics dedicated to lipid analysis, even if there is a continuum of polarity between lipophilic and hydrophilic metabolites. Some of their synonyms, such as 'metabolic/lipid profiles', refer either to biochemical measurements (*e.g.*, glucose, HDL-cholesterol, total triglycerides …) or to the overall metabolomics/lipidomics scientific field itself.

Consequently, a literature search aiming at evaluating the input of large-scale metabolomics/lipidomics in the various scientific fields was designed using a request combining words and expressions for three conceptual groups ("Metabolomics/lipidomics", "cohort" and "human/plant and environmental sciences"; see Supplemental Material 1).

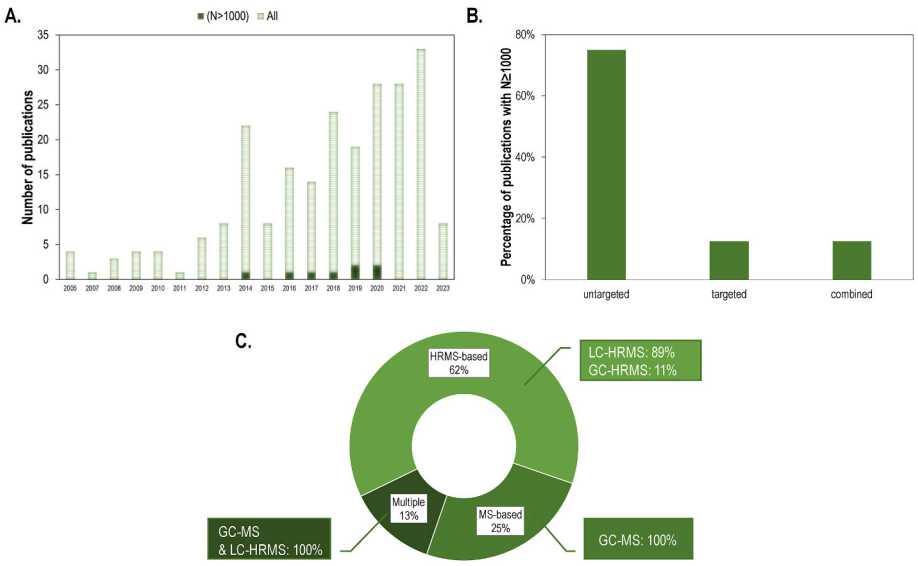

Fig. 2. A. Evolution of publication number, including large-scale plant science metabolomics analyses (N = 226), of which are studies involving more than 1000 biological samples (N = 8). **B.** Distribution of studies involving metabolomic approaches on more than 1000 biological samples (N = 45). **C.** Distribution of analytical platforms used in metabolomics/lipidomics studies involving more than 1000 biological samples).

2.1. Human research

Our results provided a useful snapshot of the evolving trend for large-scale metabolomics (Fig. 1A) and showed that around 8% involved over 1000 subjects within human studies, revealing that large-scale metabolomics has been emerging and constantly raising since 2018. These publications involved either one exploration/discovery study (74%, N = 335) or alternatively, a stepwise strategy from exploration to replication/validation, in order to discover biomarkers and validate metabolite patterns in various populations. Very few of them (8%) are based on longitudinal approaches with the follow-up of individuals, in the objective of characterising the changes of their metabolic phenotypes over time.

These publications involved a wide diversity of metabolomic approaches and analytical methods (Fig. 1B and C), often insufficiently or not adequately described from a technical point of view, reflecting the unsatisfactory standardisation level of the field. However, high coverage MS- and/or NMR-based analytical methods are generally used (in 25% of the studies) as a discovery and hypothesis-generating approach, involving the differential analysis of phenotypes in a semi-quantitative way, using an untargeted (global) approach (implying the analysis of all the detectable metabolites in a sample, including chemical unknowns). Alternatively, targeted hypothesis-driven strategies (focused on the analytical detection of predefined metabolites) are used to obtain normalised or quantitative data (Fig. 1B). Due to the limitations of not getting absolute quantification of metabolites, recent developments try to mix various analytical approaches, tending to merge targeted and untargeted strategies [17] with the objective of obtaining both high metabolite coverage, robustness and lower costs, for large-scale applications. These approaches (used in 18% of the identified publications) are generally based on the (semi)quantification of multiple known compounds from a targeted acquisition, either using low-resolution mass spectrometry or data-dependent methods performed on high-resolution instruments [18]. Alternatively, it can involve targeting multiple metabolites from an untargeted high-resolution data acquisition (identified hereafter as semi-targeted approach, i.e. untargeted data acquisition combined with targeted data treatment).

All these methods generally enabled detecting less than 250 metabolites in human biofluids, except for lipidomics approaches where more than 500 lipid species can be profiled in large-populations (Fig. 1D). Within the Consortium of Metabolomics Studies (COMETS) that gathers metabolomics data from 72 international cohorts [19], it was observed that up to 88% of these cohort studies relied on four major analytical platforms providing standardised metabolomics service, namely Metabolon, Inc. (Morrisville, NC, USA), Biocrates® (Biocrates Life Sciences AG, Innsbruck, Austria), Broad Institute (Massachusetts Institute of Technology & Harvard University, USA), and Nightingale Health Ltd (Helsinki, Finland). Nearly 600 metabolites were identified as 'frequent', being measured in at least 15 cohorts. Most of these metabolites (91%) were measured for 10,000 to 50,000 participants, while only 7% (43 metabolites) were determined for more than 50,000 participants, meaning that metabolome coverage is still restricted to a limited number of metabolites, when applied in large populations.

2.2. Plant and environment research

The literature search showed that, although the number of studies is small, the trend has been increasing in recent years (Fig. 2A, Supplemental Material 1). Our results revealed that in this field, the term "large-scale" is associated not only with the concepts already mentioned but also with the concepts "large-scale ecosystem", "large-scale chronic pollution", "large-scale fermentation", "large-scale production", "large-scale bioreactors", "large-scale distribution", "large-scale multisite", and "large-scale climate-associated diversity gradients". The term "cohorts" among the studies was associated with human studies. Of the 226 selected articles (see Supplemental Material 1), 80% of the studies had sampling lower than 50 (considering biological replicates). This number is remarkably low, as compared to human cohorts, which draws attention to the lack of large-scale studies in the plant field, thus positioning plant cohort metabolomics as a promising strategy in the future. The variability and difficulty of obtaining a large plant sample remain such that N = 20 has been considered by many authors in our literature search as large-scale. Our research highlighted that it is essential to define a standard minimum number of samples for plant science studies to qualify as large-scale, and the importance of making the number of samples for the metabolomics study explicit. Besides, the concept of a plant sample needs to be better defined, as unlike the human domain, some large plant cohort studies use pools of several specimens to constitute a single sample. The remarkably low number of samples in plant cohort studies can also be explained by the fact that two plant samples can represent a very large metabolomic diversity of organs, species, environments considered (compared to human samples), and a

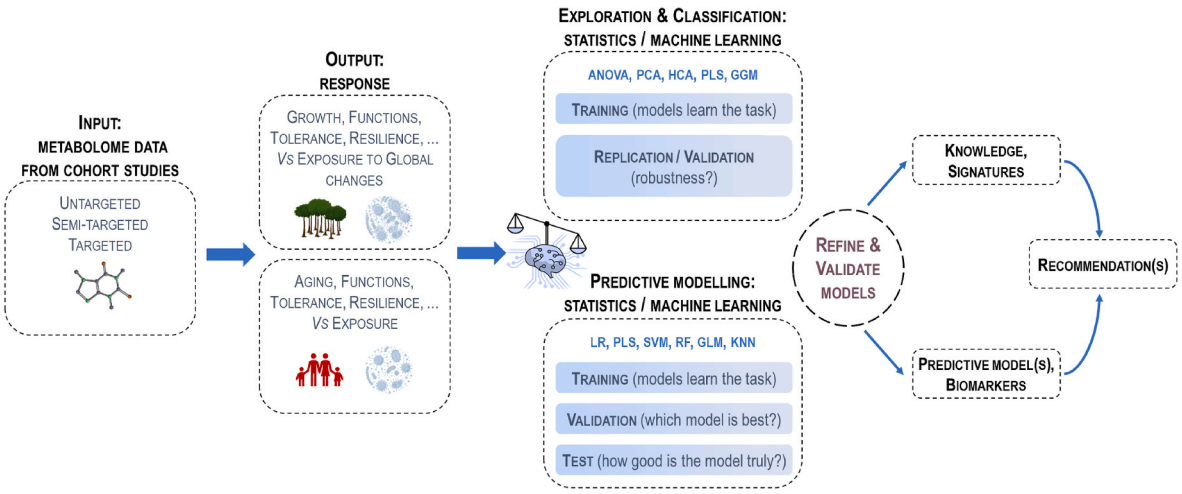

Fig. 3. Current large-scale metabolomic approaches aiming at analysing metabolic phenotypes and their complex interactions with intrinsic and extrinsic factors. They require statistical, artificial intelligence tools to enable top-down modelling in order to provide either novel biomarkers of status or treatment responses, or knowledge about mechanisms.

smaller number would be sufficient to encapsulate the metabolic shifts of interest. Some studies in this research had a sampling range between 100 and 1000 (considering biological replicates), representing 51% after the second screening. However, the number of studies that are effectively configured as large-scale, using a sampling greater than 1000 represented only 18% of the studies (N = 8).

As for human studies, we notice a variety of metabolomics strategies, but the untargeted approach prevails (76%) (Fig. 2B). The prevalence of untargeted approaches in plant science is directly related to the main objectives of these studies, namely molecular elucidation and the potential discovery of novel natural compounds for various species. Among the analytical methods used for large-scale plant studies, HRMS appears to be the most widely applied, regardless of the type of chromatography, as expected, to cover the vast metabolome of plants, given its broad chemical diversity (Fig. 2C).

In summary, the diversity of analytical approaches and platforms used for large cohort analysis, allowed large-scale metabolomics making a rapid ascent in the analysis of human, microbial and plant metabolisms. We will highlight some recent publications in both fields that illustrate its power in metabolic phenotype exploration, classification and prediction.

3. Large-scale metabolomics: from plant science to human epidemiology and medicine

3.1. Meta-metabolomics of plants and environments

In the context of global change, there is an increasing need to evaluate and predict the adaptation capacity of natural ecosystems facing these stresses in order to ensure their sustainability and associated services. To this end, the value of metabolomics is even more remarkable as the meta-metabolome (i.e. metabolome of multiple species) integrates information from the genome/transcriptome/proteome and the environmental influences, making ecosystem metabolomics a strategy of choice to better understand the impacts of climate change and chemical stress, for instance.

3.1.1. Predictive metabolomics for plant sciences

Metabolomics is experiencing an unprecedented boom in plant biology research, enabling the discovery of molecular mechanisms involved in areas as varied as development, stress response or chemical ecology [6,7,20]. However, the methods for deciphering these interactions are relatively scarce, and the lack of predictive models of intricate systems is a pressing issue. This is where the metabolic analysis of large plant cohorts could help remove methodological bottlenecks, particularly through predictive metabolomics. Predictive metabolomics combines artificial intelligence (Fig. 3) (*e.g.*, machine learning) and metabolomics to enable the top-down modelling (from phenotype to mechanism) of phenotype traits (*e.g.*, yield, quality or the response efficiency to environmental factors) that are predicted from metabolic data [21]. For the moment, such an approach is mainly limited to intraspecific diversity panels, like for single genotype tomato metabolome able to predict up to 87% of biotic resistance to various pathogens [22]. Nevertheless, predictive metabolomics also seems valid for species panels, as illustrated with eight fruit species where biomass composition is evaluated for the prediction of relative growth rate [23]. More strikingly, predictive metabolomics for 24 extremophiles harvested across multiple natural microenvironments in the Atacama desert showed that the meta-metabolome accurately predicts the plant environment, with plants harbouring a generic metabolic toolbox associated with extreme habitat resilience, notably redox and hormonal metabolites potentially involved in trade-offs [7]. Another elegant study of 416 vascular plants corroborated that phytochemical diversity revealed by metabolomics can predict the alpine habitat [23]. These results confirm that the metabolome contains valuable information that can be channelled into the phenotype, and that predictive metabolomics of large cohorts under environmental conditions is appropriate for discovering easily measurable traits (*e.g.*, resilience, plant environment). Furthermore, multi-species strategies would enable approaching the metabolic complexity of ecosystems. This appears particularly relevant for holobionts and microbial communities (Fig. 3).

3.1.2. Eco-metabolomics to explore the metabolome of microbial communities

(Meta)-metabolomics of microbial communities is a growing field of research since it can provide a comprehensive picture of chemical interactions into microbial communities and with their host (*e.g.*, plant-epiphyte interaction) but also their functioning (*i.e.*, microbial activity as an ecosystem phenotypic trait) and their response to the surrounding environment [24,25]. Despite its relevance, such an approach remains scarce in microbial ecology, chemical ecology and ecotoxicology. For instance, most of the knowledge about microbial, chemical interaction comes from metabolomic investigations of co-culture of only two (or a few) species that are not representative of actual interactions in complex natural communities [25]. The recent developments in microfluidics and robotics pave the way for the simultaneous implementation of

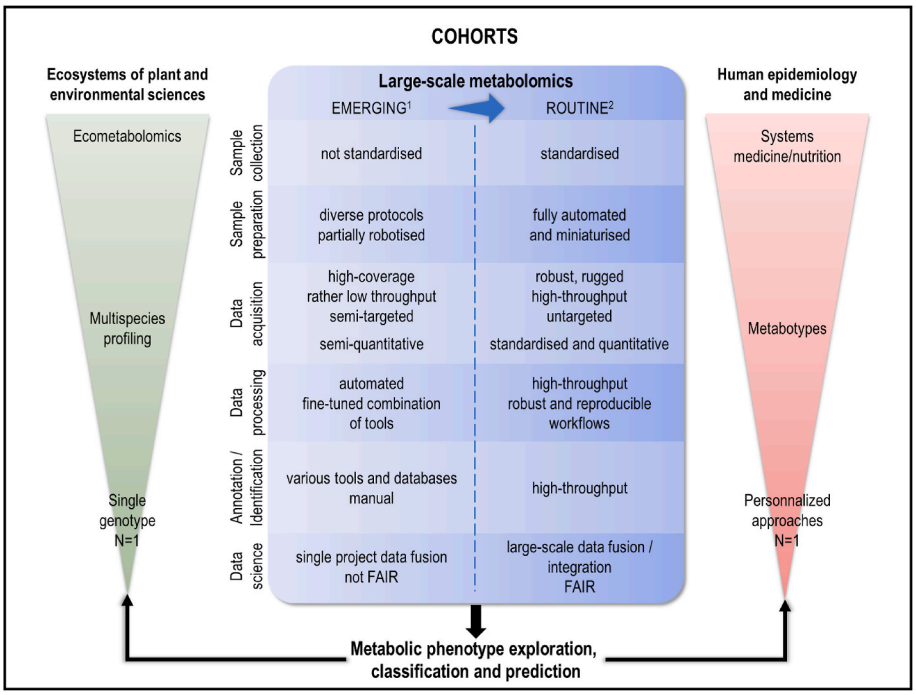

Fig. 4. Challenges towards large-scale metabolomics for application in plant and environmental sciences, and human epidemiology and medicine. 1: Represents the current state of large-scale metabolomics; 2: The required methodology for routine applications of large-scale metabolomics.

thousands of co-culture combinations with increasing complexity [26]. Overall, as recently reviewed by Bauermeister, et al. [27], there are tremendous methodological/technical improvements (*e.g.*, molecular networking or machine learning-based annotation) in the field of MS-based metabolomics applied to microbiomes. At the computational level, emerging pipelines allowing the combination of meta-metabolomics and metagenomics are promising to enhance the prediction of microbial metabolic functions and microbial interaction (*i.e.*, Genome-scale metabolic models, GEMs) [28,29]. Nevertheless, there are still significant knowledge gaps about the spatiotemporal dynamics of microbial metabolomes, limiting the prediction of ecosystem functions (*e.g.*, biogeochemical cycles) across systems. To tackle this issue, Danczak, et al. [24] have recently proposed a theoretical paradigm named "meta-metabolome ecology" consisting of the application of ecological metrics to metabolome datasets funded on the hypothesis that metabolites assemblages are determined by stochastic or deterministic processes (as microbial taxa assemblage). Thus, identifying when, where and why metabolomes are directed by such processes would support a deeper understanding of the environmental factors (*e.g.*, chemical stress, temperature, pH) that shape (environmental) meta-metabolomes. The parallel implementation of *in situ* predictive metabolomics on microbial communities as described above in plants [7] also appears particularly relevant for such an end. This would further allow the discovery of metabolites as biomarkers able to detect and predict impaired ecosystem functions supported by microbial communities.

Despite these promising developments, Quinn, et al. [30] have recently highlighted that significant gaps still exist between analytical and microbial sciences to fulfil the above-mentioned objectives. In particular, microbial metabolomes' diversity and functional roles have only begun to be investigated since it requires both analytical capacity and microbiomes' scientific background. Thus, better communication between these two worlds would uncover the "dark matter" within microbiomes. One critical challenge in the following years will be our ability to unravel, in complex microbial communities (*e.g.*, biofilm) or holobionts, what is the contribution/role of each component but also to distinguish the metabolites contributing to the interaction from those involved in the functions.

3.2. *Human epidemiology and medicine*

With the advancement of analytical technologies and data science, but also a reduction in costs, the generation of metabolomics data from epidemiological studies has significantly increased in recent years. Indeed, since the technical possibility to analyze several hundred to thousand samples in the early 2010s, different outlying large-scale studies or consortia were published, involving prospective cohorts and case/control design [31]. The National Institute of Health (NIH) also established the Consortium of Metabolomics Studies (COMETS) to gather international cohorts with blood metabolomics data on samples collected between 1985 and 2017 [19] in order to advance the knowledge of metabolomes and to improve the understanding of disease aetiology, diagnosis and prognosis. At the beginning of 2022, COMETS already comprised 72 internationally distributed cohorts which together included measurements of 4647 metabolites in up to 134,742 participants (with a population size between 89 and 10,456 per dataset). More recently, metabolomics was used to provide a comprehensive readout of human physiology in the context of non-communicable disease (NCD) multimorbidity [32]. Indeed, this study was the first to perform a comprehensive metabolic profiling of plasma samples from a follow-up of 219,415 person.years, and to integrate it with deep phenotypic profiling, resulting in the identification of 420 metabolites shared between at least two NCDs. It also highlighted that the analysis of blood metabolome provides an integrated view of interactions between intrinsic (genes, sex …) and extrinsic (environment, nutrition, medical treatments, microbiota …) factors. In particular, with the objective of studying the human system as a holobiont, metabolomics was used to decipher the complex relationships between gut microbiota, diet and host metabolism [33,34] (Fig. 3). From fasting and postprandial serum metabolomics, it has been shown that large-scale phenotyping could potentially stratify the gut microbiome into different health status in subjects without clinically identified diseases. In terms of precision and personalised prevention, metabolomics was also applied at large-scale to stratify populations by the identification of metabotypes, which consists in grouping individuals based on the similarity of their metabolic phenotypes [35]. Finally, in medicine, metabolomics has shown its power to

provide clinicians with novel biomarkers for disease states and evaluate individual treatment responses [36]. However, although only targeted quantitative metabolomics methods have been translated into clinical practices until now, untargeted approaches open the door to a paradigm shift in the perception of diseases, by giving more complex signatures of metabolites than single-molecule disease biomarkers [31,36].

Finally, the future role of metabolomics in population-wide personalised medicine will require large, metabolomics-based screening programs to obtain comprehensive information across ethnicities, different environmental conditions and health status [37]. It will involve moving towards N = 1 clinical trials and having access to robust and high-throughput metabolomics methods requiring small sample volumes, for reliable assessment from repeated measurements of metabolic status. All these objectives will also require a more interoperable metabolomics, gathering adequate high-throughput and robust analytical methods that provide comprehensive and accessible phenotypic data. Besides, data science and semantics (ontologies, controlled vocabularies), will be essential to support the precise classification of patients for diagnosis, care management, and translational research [38] (Fig. 4).

4. Analytical tools and methods for large-scale metabolomics

4.1. Towards high-throughput analyses

Metabolomics approaches are constrained by the limits of the used analytical platforms and methods, on which depends the comprehensiveness of the metabolite landscape, especially in the case of large-scale studies. In the context of the development of high-throughput metabolomics, the need for speed without sacrificing quality is the most critical issue [5]. Metabolite identification is still a major bottleneck, and the validity of proposed identities is therefore, of deep concern. Nevertheless, discussing this aspect is out of the scope of the present review article. The interested reader can refer to the recent review of Theodoridis et al. [39].

4.1.1. High-throughput sample preparation

Firstly, sample preparation (*e.g.* extraction, filtration) has always been a time-consuming struggle and a crucial bottleneck in metabolomics. In terms of accessing high-throughput and standardisation, two approaches should be favoured: protocol simplification and automation. Protocol simplification is complex to implement, as it inevitably leads to a loss of metabolite diversity. For instance, metabolites such as oxidised lipid derivatives (*i.e.*, oxysterols, oxylipins) cannot be detected after simple protein precipitation. Similarly, plant-specific signalling metabolites (*e.g.,* redox compounds, phytohormones, metabolic intermediates) require sophisticated and stepwise extraction methods, which are incompatible with the high-throughput of cohort studies. Thus, protocol simplification entails compromising the coverage of the targeted metabolome. As such, protein precipitation by organic solvents (*e.g.*, acetonitrile, methanol, ethanol) is the most common approach [40]. Depending on the protocol, a standardisation method easy to implement should be provided to be compatible with large sample numbers. This step is well established for blood samples, but more complicated for urine or other tissues (*e.g.,* plants).

Besides, automation of sample preparation can be laborious and time-consuming for its implementation. Despite these difficulties, fully automated methods, including *e.g.*, weighing, extraction, filtration, quality control preparation, are today available on several robots [22,41]. However, partially automated protocols involving a simple dilution step [18] are the easiest to set up. Plant metabolomics has made tremendous progress in sample preparation, now allowing the standardised high-throughput extraction of about 400 samples in 3 h from a minimum of material (about 10 mg dry weight) [7,22]. The success of this implementation was achieved by designing separate robots to perform specific tasks, rather than highly versatile systems that stack multiple modules. The use of greener extraction solvents (e.g. ethanol) has also contributed to the success of robotised preparation of plant extracts. More prospectively, and owing to the advent of state-of-the-art facilities for automated plant phenotyping, there is a solid appetite for integrating rapid, sensitive and automated metabolomics to enable deep phenotyping [42]. In human metabolomics, blood sample preparation protocols based on single-step solvent-based protein precipitations have been easily automated for the analysis of polar metabolites, but this task is more complex in lipidomics, where most efficient biphasic extractions must be implemented [43,44].

Further developments in robotisation of sample preparation are still mandatory to scale-up metabolomics. These methodological advances will also involve miniaturization, in response to the demand for small sample volumes for longitudinal and less-invasive studies. Moreover, it will allow reducing reagents and consumables in the context of eco-responsibility. Besides, microsampling strategies, and in particular dried blot spots, are increasingly used within large cohort studies as they facilitate an easy and ultra-fast collection of blood sample, thus becoming an effective tool for epidemiological and medical research [45–47].

4.1.2. Fast data acquisition methods

In mass spectrometry-based methods, alternative strategies for high-throughput data production, and especially the need for liquid chromatographic separation prior to tandem MS analysis [48–50], are coexisting (Fig. 1C). Coupling LC to HRMS has become the most powerful MS-based approach for the analysis and profiling of both polar and non-polar metabolites and lipids. The most widely used high-resolution mass spectrometers for metabolomics analysis by LC-HRMS are Orbitrap and TOF-based systems having distinct but often complementary features [51]. LC-HRMS workflows often provide the broadest metabolome coverage due to their ability to resolve isobaric potentially and isomeric compounds, dereplicate complex biological extracts and their associated minimised ion suppression effects. However, the time associated with traditional LC-MS-based metabolomics can preclude its use for large-scale studies. Concurrently, direct introduction mass spectrometry (DIMS in low or high resolution) approaches with typical analysis time <1–3 min per sample have been proposed for high-throughput metabolomics [52], thus allowing the analysis of hundreds of samples per day [53] but at the expense of metabolome coverage. DIMS suffers from the inability to distinguish isomers or in-source fragments from true precursor ions, and is also directly impacted by (often strong) ion suppression effects. A recent comparative study showed that DIMS and LC-HRMS workflows highlighted shared discriminatory signals, with LC-HRMS providing (as expected) more comprehensive information in terms of metabolite identification [53]. This led the authors to recommend DIMS as a fast screening method for large sample batches and LC-HRMS for a more comprehensive analysis of selected samples. The introduction of ion mobility (IM) into metabolomics and lipidomics workflows, especially those involving DIMS, can allow the separation of some isobaric and isomeric compounds [54]. Fast LC-HRMS methods (<10–15 min analytical time) have been recently described and can represent platforms for high-throughput and high-confidence metabolome coverage [55]. This can be achieved by using shorter UHPLC columns (see below) or by increasing chromatographic flow rate, column temperature or by modifying the LC gradient while still using 2.1 mm × 150 mm UHPLC columns [56,57]. Interestingly, the National Phenome Centre's established platform combines both reversed phase chromatography and hydrophilic interaction liquid chromatography columns (both 2.1 mm × 150 mm) coupled to HRMS for the robust profiling (using standardised protocols and workflows, that have been made open) of more than 700 annotated biologically relevant metabolites in several hundreds of human biofluid samples [57]. Also, the recent development in nanoscale LC-ESI opens interesting perspectives regarding the need of miniaturization for improved detection sensitivity and enhancement of metabolome coverage [53],

but often at the expense of method robustness. Fitz et al. recently comparatively evaluated analytical (2.1 mm column i.d.), micro- (1.0 mm), and nano-flow (0.3 mm) LC systems coupled to HRMS (while maintaining injection volume, mobile and stationary phases, gradient and detection parameters constant) for their ability to robustly and sensitively detect more than 50 endogenous metabolites and xenobiotics in human plasma [58]. The authors concluded that micro-LC provided the best compromise between signal intensity, retention time stability and metabolome coverage [58]. By reducing both diameter and length of the UHPLC column to 1.0 × 50 mm provided an analytical time of 2.5 min/sample along with a 75% reduction in solvent consumption and improved batch reproducibility [59]. Such methodology offers the possibility of broad metabolic phenotyping for large sample sets at high throughput, as exemplified by the analysis of over 700 rat urine samples [59]. Finally, LC-IM-HRMS(/MS) workflows can present an unprecedented metabolome coverage. For instance, the recently introduced Parallel Accumulation Serial Fragmentation (PASEF, available on timsTOF Pro instruments), that synchronises Trapped Ion Mobility Spectrometry (TIMS) with MS/MS precursor selection and fragmentation, enabled the annotation of up to 1100 unique lipids in 1 μL of human plasma using two 30-min nanoLC-IM-HRMS/MS runs (one in positive and one in negative ion mode) [60].

More informative and exhaustive data acquisition protocols [61,62] are also of great interest to the field. Regarding MS/MS investigations, conventional data-dependent acquisition (DDA) fragmentation methods are still the most widely used in metabolomics and lipidomics workflows. In this mode, precursor ions are selected using a small isolation window (typically 1 Da wide), which leads to high-quality and high-purity MS/MS spectra. However, selecting precursor ions is a (semi)stochastic event favouring the collection of the most abundant ions and sometimes of biologically irrelevant ones [63]. Therefore, specific rules have to be applied for the successful implementation of efficient DDA workflows, for instance with an iterative DDA script allowing to automatically remove pre-selected precursor and background ions of MS^2 acquisition by using repeated injections [64]. Besides, data-independent acquisition (DIA) workflows, including the Sequential Window Acquisition of all Theoretical fragment-ion spectra (SWATH) approach, are being increasingly used in metabolomics by enabling metabolite annotation and quantification through the acquisition of MS/MS spectra for all analytes in a single run [65–67]. A complementary approach to go towards high-throughput metabolomics analysis is the implementation of retrospective quantification from DDA or DIA data by using standardised internal standards (IS) broad mixture associated to bioinformatics tools able to identify the most relevant IS for each metabolite, as recently proposed in the chemical exposomics field [68].

In the case of NMR metabolomics, high-throughput data acquisition and processing workflows are well described in the literature [69]. Standard operating procedures rely on the acquisition of 1D NMR experiments with solvent signal presaturation, which are routinely applied on a broad variety of matrices (biofluids, extracts, or even tissues with HR-MAS spectroscopy) [70]. In recent years, standardised NMR hardware has been made commercially available, associated with databases and standard operating procedures (SOPs), so that large cohorts can be analysed on different sites equipped with the same analytical platform. Typical 1D NMR experiments include the nuclear Overhauser effect (NOESY) pulse sequence, which allows detecting signals from both metabolites and macromolecules, but also more selective experiments such as the CPMG (Carr-Purcell-Meiboom-Gill) experiments that filters out signals from macromolecules, or diffusion-edited pulse sequences that allow the selective observation of macromolecular signals. On typical NMR metabolomics hardware (600 MHz magnetic field), such experiments take between 5 and 30 min per sample, which confers them a high-throughput character. The latter is further ensured by well-defined SOPs as well as the use of refrigerated sample changers, which have been developed to facilitate the automated acquisition of large sample cohorts. These conditions typically make it possible to analyze up to 100 samples per day, making NMR well suited to the analysis of large sample cohorts (>1000). In addition to these routine detection methods, a great asset of NMR spectroscopy in metabolomics is its ability to provide a diversity of two-dimensional (2D) spectroscopy experiments that can be used to better spread overlapped peaks along an orthogonal dimension, thus offering the improved ability to discriminate between sample groups and identify potential biomarkers. While such 2D methods are relatively time-consuming, several fast acquisition approaches have been developed to make 2D NMR compatible with high-throughput metabolomics, resulting in data acquisition durations that do not exceed a few tens of minutes per sample [71].

4.2. Handling complex data

4.2.1. Fast data processing

With technological advances in the analytical platforms generating a huge amount of complex data, the development of data analytics is actually expanding with a great diversity of algorithms and tools with no real consensus.

On the MS side, there is a great diversity of processing algorithms and tools [72], some of which can be cited as the main drivers (e.g. R: XCMS, MS-Dial, mzMine, MetaboAnalyst…). Most of these tools follow a similar workflow from ion chromatogram building, peak detection, retention time alignment, to the correspondence of peaks across samples. Even if some of them have been wrapped in cloud environments, none has reached a complete consensus, and more importantly, none is yet ready to address all the high-throughput metabolomics challenges. Most of these tools can theoretically process hundreds of samples at the same time, but in practice, the end users face technical limitations of their computers with these locally installed software (e.g. amount of RAM, hard disk space). Even cloud based solutions have preset processing limitations usually made to limit huge calculation demand on shared servers (*e.g* uploads limited to 200 spectra files, disk quotas). Some alternative approaches have been recently developed, such as asari [73], which has delivered a new generation of computational performance, together with interesting linked and transparent data structures in all processing steps, contributing significantly to more reproducible data. In the same way, a new deep-learning based tool (i.e. NeatMS) has been recently proposed to handle peak alignment in large scale metabolomics [74]. Despite the increasing rate of MS technological development associated to ion characterisation (*e.g.*, ion mobility, new MS/MS fragmentation techniques), most of the features extracted from MS metabolomics studies are still defined by a unique pair of *m/z* value and retention time. Only a few tools like SLAW [75] are ready to handle high-throughput analysis with MS/MS dimension, and this fact can be explained by the computational complexity of concurrently handling a high number of samples and high MS dimensionality (MS/MS or IM).

In the case of NMR metabolomics, as described in the literature [69], untargeted metabolomics data processing can be achieved *via* several strategies. In the first one, NMR spectra are transformed into data matrices through bucketing, that is used to reduce the data dimensionality. Spectra are segmented into small buckets (fixed or variable size), and each bucket is integrated. The second approach is to work with full-resolution spectra, requiring specific algorithms for peak alignment. Several algorithms have been developed to achieve automated spectra processing and bucketing, so that cohort size is not an issue for the routine processing of NMR metabolomics data. In the case of very large cohorts, additional steps can be added in the processing workflow such as chemical shift alignment, removal of unwanted signals, normalisation and cohort or batch correction [76]. In a third approach, concentrations of all quantifiable metabolites in a biological sample are calculated using deconvolution tools. Several tools are available that support both semi-automated NMR data processing as well as automated or semi-automated small molecule identification and quantification in biofluids [77]. For instance, Buergel, et al. used a commercial tool to

quantify 168 markers in 117,981 NMR spectra and predict individual multi-disease outcomes [78]. In addition, several freely available academic programs can perform fully automated data processing and spectral deconvolution of 1D ^{1}H NMR spectra but are limited to analysing a specific biofluid, and the quantification is limited to 50–60 compounds. Note that most available tools are only available to deal with 1D spectra. However, recent applications of machine-learning methods, in particular of deep neural networks, have shown promising results in the ability to deconvolute both complex 1D and 2D NMR spectra [79]. Nevertheless, the analysis of 2D datasets recorded on large sample cohorts remains a challenge.

4.2.2. High-throughput data annotation

On the MS side, efficient tools are continuously released [62,72,80,81] in order to address high-confidence metabolite annotation, but despite their ability to correctly attribute compounds names to detected feature at least at level 2 [82,83] none of these tools is ready for high-throughput metabolite identification though. The main strategy for large-scale metabolomics in terms of MS-based compound identification is still based on targeted approaches [62]. From a global point of view, the annotation challenge is even more complex in large studies, where high-throughput analytical methods are generally used, resulting in a huge number of complex data files. Moreover, the associated advanced data protocols (*e.g.*, DDA, DIA …) are still limited to relatively small sample sequences and require specific dedicated processing tools [62,65]. In untargeted metabolomics, the issue of high-throughput data annotation remains one of the main bottlenecks [62,83], limiting its impact. Despite tremendous analytical and software evolution over the last years, the annotation step still leaves a large amount of unidentified or ambiguously identified compounds per dataset (>70% depending on matrices), limiting biological interpretability [72,84]. In both approaches, the huge number of data files and the requested amount of computational memory to query databases and resources are a limitation to the use of some software or comparison tools, initially designed for small datasets (*e.g.*, tools without batch query [72]). The actual growing number of bioinformatics tools and pipelines dedicated to annotation is in fact double-edged. Indeed, even if they give access to powerful MS data interpretation, they are based on algorithms with increasing complexity, and the lack of proper training can lead to a higher number of misinterpretations of the proposed annotations [62].

On the NMR side, software tools with specific spectral databases are also arising to facilitate the identification of metabolites in mixtures of 1D NMR spectra, but some of these metabolites are misidentified because of large overlapping signals, while a large number of compounds are still unidentified [77]. Two-dimensional spectra are often required to obtain more structural information and assign new metabolites. A promising tool has recently been developed, a web server for semi-automated 2D NMR analysis with peak fitting for quantification and database query for metabolite identification [85], opening the door to high-throughput annotation.

4.2.3. Data fusion and normalisation

The analysis of thousands of samples from large cohorts typically requires the acquisition of several batches (*e.g.*, to avoid clogging of the ionisation source in MS), sometimes spaced over several months (*e.g.*, in the case of limited patient inclusion over time), or even obtained on several instruments. The ability to merge and normalise these data for subsequent statistical analysis is a major challenge in metabolomics [86], and many methods and software tools have been proposed [87], such as WaveICA ([88] based on the biological samples), RUV ([89]; based on replicates), and SERRF ([90]; based on pooled quality controls QCs), which have been successfully applied to large cohort analysis.

One of the main challenges is to overcome the unwanted variability in the data unrelated to the factor(s) of interest (in particular the technical variability). This is why the identification of these sources of variation upstream of the experiment is essential to randomise the order of the samples in the analytical sequence and to position the controls to be included (*e.g.*, internal standards, pooled QCs, sample replicates, reference materials; [91]).

MS analytical drift within a batch can be corrected for each variable by modelling the values in the pooled QCs using local loess polynomial regression, splines, or machine learning [92]. Recently, it has been proposed to include the pooled QCs of variables with similar profiles in the model to obtain a more robust estimation of the drift [90,93]. Regarding the normalisation between acquisition batches, two strategies have been described: location-scale adjustment and matrix factorisation. The former seeks to harmonise the means (and variances) of the variables across the batches. Robust estimates of these parameters can be obtained by using an empirical Bayes approach (ComBat method; [94]). Matrix factorisation strategies focus on control variables (*i.e.*, not affected by the factor of interest) or on replicate samples to capture unwanted variation (*e.g.*, by singular value decomposition; [89]). A hierarchical strategy for intra- and inter-batch normalisation has been recently applied to a cohort study of more than 1000 human plasma samples [95]. This approach, which combines QC-based signal drift correction (*e.g.*, with loess) with replicate-based removal of unwanted variation, is of interest for large-scale untargeted studies, since it achieves robust and efficient results without the need to include additional (isotope-labelled) reference compounds. In NMR, the variation due to different cohorts or different time periods of analysis, can be more easily corrected using a mean-centering operation to remove batches differences in mean levels prior to statistical analysis [76].

Whatever the strategy, the use of complementary metrics (correlation coefficient between replicates, mixing of nearest-neighbour samples across batches, an increase of prediction performance) and visualisations (principal component analysis, sample intensities along run order, difference of intensities between replicates, etc.) is essential to control the quality of the normalisation [96]. Furthermore, the availability of data and analysis scripts is essential to reproduce the results [97]. When data were obtained with different analytical instruments or within different studies (*e.g.*, in meta-analyses), it is necessary to match features between the data sets. Various alignment methods have been proposed to model the drift in retention time and m/z, which either require the raw data or the processed data only [98].

4.2.4. Large-scale data contextualisation and reporting: F.A.I.R. principle implementation

Despite all the progress made in automating the analysis (analytical methods and bioinformatics) of metabolomic data in large cohort studies, many current approaches still involve manual curation by experts, to validate annotations, or in the choice of a chemical name or the addition of identifiers for further data contextualisation. Therefore, the production of knowledge is often isolated in a publication, or without coherent and machine-readable metadata, or even orphaned from their original raw data. The integration of existing metabolomics standards and FAIR (Findable, Accessible, Interoperable, Reproducible) [99] considerations for the processing and sharing of complex data often remains at the stage of recommendations or best practices. In order to move towards the routine application of large-scale metabolomics, it has become crucial to incorporate into management practices methods for addressing the ‘quantity’ and ‘quality’ dimensions of the data generated by these high-throughput approaches, as well as their integration with other omics fields, and the sharing (within and outside projects) and reuse of these data. We therefore inventoried the existing resources that meet the FAIR principles [100] and the remaining problems of large-scale data management. [99,100]. One of the main markers of the FAIR nature of data is the definition of unique and globally persistent identifiers. Regarding the specific problem of metabolite reporting, regular ambiguities are still observed in metabolite names, even though recent tools make it possible, for example, to normalise lipid names [101]. In addition, the compact hash code of the IUPAC International Chemical Identifier “InChIKey” seems to be an appropriate identifier

since it provides information on the molecular backbone, isomer identity or isotopes of identified Level 1 metabolites [14]. For compounds with incomplete structural annotation (e.g. unknown stereochemistry or unknown double bond position), other computational (e.g. SMILES, http://opensmiles.org/) or semantic (ChEBI, LipidMaps or PubChem) identifiers are possible alternatives. Concerning mass spectra, the hashed identifier "SPLASH" is an interesting solution for referencing spectra from the databases that identified a candidate metabolite [102]. A systematic data description is required with specific recommendations to improve (meta)data stewardship as recently observed in some fields [103]. The key point to address here is to find the minimum but also a good level of domain-specific descriptors (MSI [14], MERIT [104]) which should be precise enough for experts, and generic enough to be understood by other communities. Finally, finding data is also linked to the resources offered and consultable by the community. The diversity of metabolomics databases (spectral- or compounds-based) is an excellent point but the variability of "query engines" also limits the potential for data mining and federated searching necessary for large-scale contextualisation. Some large cohort studies have paved the way for the creation of open data repositories (comets-analytics.org [115], or https://omicscience.org/apps/mwasdisease/ [32]) but these resources have been built up independently of each other, and therefore remain community resources.

Historically in metabolomics, data providers make published resources available on the World-Wide Web using free, open and universal standardised communication protocols. A double challenge arises here: facilitating access to searchable data and facilitating large deposits. As "FAIRness" does not automatically mean openness, the addition of an Authentication and Authorisation Infrastructure (AAI) protocol is recommended in a FAIR data internet, although it should be as generic as possible or based on organisation-led solutions (*e.g.*, European AAI ELIXIR). Unfortunately, metabolomics resources offer many authentication solutions, poor documentation or access protocols and sometimes no programmatic access to data. Finally, metabolomics communities need to address the issue of (meta)data accessibility and persistence over time by adopting a targeted shared policy, but also by using a data management plan in line with FAIR principles.

Making dataset exchangeable and machine-readable means making it compatible with a (meta)data structuring model, *i.e.*, choosing a controlled vocabulary or a shared ontology. In metabolomics, the best-known example is the widely used spectral data exchange formats (mzML, nmrML, etc.). The ISA (Investigation/Study/Assay) data model [105] allows a complete representation of MS- or NMR-based studies and a description of associated experimental metadata. Choosing the 'best' ontology to find descriptors and the 'best' controlled vocabulary to define descriptors in any study remains a gamble and may depend on the application field. However, some initiatives, such as Ontologies 4Chem [106] and BioPortal web portal [107] provide a comprehensive overview of exploitable resources to access ontologies. Knowledge engineering also offers higher-level ontologies, such as the Semantic Science Integrated Ontology, for contextually enriching two linked data with qualified references. The benefit of annotations using ontology is also the potential it provides to create flexibility in the mapping between class of molecules and defined species (*e.g.*, automatically mapping class of lipids and individual species like with chain lengths [108]).

The next challenge is to reuse the deluge of data from the last decade, provided that the number and quality of labels attached to the data are sufficient. As it is easier to compare similar objects, studies complying with the minimum information of the Metabolomics Society Initiative (MSI) should be easily linkable. While metabolomic data repositories such as MetaboLights or Metabolomics Workbench offer such a "tagging service", they are still limited by the quality of the metadata linked to the deposited datasets and sometimes deemed far from the minimum level required by MSI. As high-throughput metabolomics is a big data provider and consumer, a clear policy on data licensing and use is needed. Attaching a license at the time of data generation is, therefore, mandatory to define the rights of owners and future users even after a public repository. FAIR spectral metabolomic databases make extensive use of Creative Commons type licenses, compatible with high-throughput approaches [15,109,110].

5. One step further towards large-scale metabolomics across studies

By adapting analytical methods for high-throughput robust profiling, the metabolomics community is actually establishing a framework for the rapid measurement and analysis of metabolites in large-scale studies. Progresses made in data science also enabled scalable processing workflows and e-resources. However, to go one step further towards large-scale metabolomics across studies, the question of the interoperability remains essential to be addressed, by first producing more standardised and confident, qualitative and quantitative data. Indeed, comparison and integration of data and results between studies are key in order to go towards a more knowledge-based and long-term strategy.

Today, even if the COMETS initiative gathered metabolomics data/results from several large-scale cohort studies to fulfil this objective, the assessment of metabolite overlap between the 5 widely used metabolomics platforms requested a huge work of curation of metabolite names using several metadata collected from the analytical platforms to propose a cross-references metabolite table. Moreover, the comparison of blood metabolite levels was limited because only 2 analytical platforms compared their metabolite measurements against one another and highlighted moderately intercorrelated metabolites values (median correlation of approximately 0.5) [19].

In this context, the metabolomics quality assurance and quality control consortium (mQACC) is actually working on the identification, development, prioritization, and promotion of suitable reference materials to be used in quality assurance and quality control to ensure standardisation of results obtained from data analysis, interpretation, to compare data within and across studies and across multiple laboratories [111]. In addition, the "metabolomic epidemiology" emerged as a growing area of research, a Task Group was recently created within the Metabolomics Society, to address the key challenges in order to advance the field [31]. It gathers experts in study design, data acquisition, data analysis and statistics, who identified a number of challenges mainly linked with standardisation. In particular, they highlighted the importance of establishing 1) standard protocols for sample collection and storage for large-scale metabolomic studies, 2) reporting standards associated with study design, sample collection and analysis, 3) standards for metabolomics data deposition together with metadata, 4) methods to enable causal inference links between metabolites and diseases. Successful translation also requires the adoption of SOPs, training in the interpretation of results, and adequate electronic infrastructural support [36].

Progress in the implementation of the FAIR principles in metabolomics in the various scientific communities is today imperative for successful comparisons across studies. Combined with data science efforts, it will enable more effective management of heterogeneous and complex data stored in large volumes. This evolution in practices cannot take place without the emergence of interoperability hubs. These meeting points between disciplines and areas of expertise, made possible by digital technologies and the web, are emerging as virtual research environments or fleets of application programming interfaces (APIs). Interoperability and data sharing are also currently progressing thanks to the development and adoption of universal standards, in particular those proposed by the World Wide Web Consortium. More specifically, the use of the extensible knowledge representation model (RDF) makes it possible to represent interconnected data on the web and to facilitate their exchange on the web as well. The Data Catalogue Vocabulary Format (DCAT) is designed to describe data sets and facilitate interoperability between data catalogues published on the web. These technologies, which are specific to the Semantic Web, make it possible to

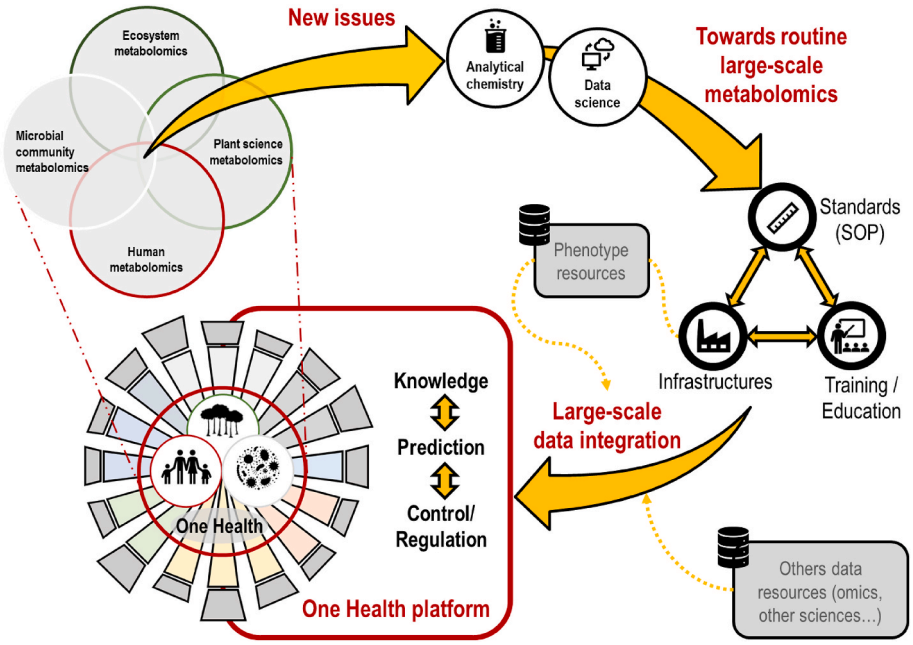

Fig. 5. Integration of metabolomics into the One-Health concept aiming an optimal health for humans and their ecosystems.

build increasingly massive inter-domain knowledge graphs [110], expanding the space of available life sciences data (or linked open data) and enabling to propose new hypotheses on health and environment issues.

By addressing all these previously mentioned challenges, the metabolomics field will be able to converge towards larger application across studies and disciplines, and ultimately to effectively integrate the One-Health concept. First, there are mutual benefits of integrating plant, environmental and human metabolomes. Indeed, it has been shown that plant metabolic engineering allows a better understanding of the plant metabolic pathways leading to the optimisation production of plant-derived metabolites that have beneficial effects on human health through nutrition and novel pharmaceutical compounds [2]. Integrating nutrimetabolomics into the One Health approach was recently proposed as a key element for personalised medicine advancement [112]. Through initiatives such as cross-field databases and resources, the understanding about how plant metabolites impact on health can be improved. In this context, the recent development of an integrated computational resource (the Aliment to Bodily Condition Knowledge-base) allows connecting plant products to health outcomes through their molecular mechanisms [109] for building informed nutritive hypotheses as the linking factor between dietary plants and human health [2]. Second, there are major opportunities for integrating metabolomes from the microbiomes (*i.e.*, as independent communities in soil and water or within plant and human holobionts), to plants and humans to study the interaction between these ecosystems and their contributions to systems health [113]. These perspectives call for the collaborative efforts of multiple disciplines working to attain optimal health for humans and their ecosystems (Fig. 5). They also represent great technological challenges both in analytical chemistry and data science within the metabolomics fields, as it will require its routine application at large-scale across studies and disciplines. It will require a shift from field-specific research tools to tools valid and useable in a variety of scientific domains. This will especially be the case concerning common SOPs from data production to treatment, the development of common infrastructures, able to run with highly interdisciplinary teams, operating on a daily basis with large-scale databases and multi-source data. Such transition will also require tailored education programs and continuous, complementary training for personnel as well as for systemic scientists [114].

Author contributions

Conceptualisation: GH, MBS, FF, BC, PP, EPG; Original draft Writing: Alternative approaches GH, MBS; Large-scale metabolomics applications: PP, NC, BC, EPG; Analytical approaches: JBM, FC, CC, PG, YG, FF, EPG; Data sciences: BD, SD, ET, MT, FG; One-Health: BD, FG, BC, EPG; Review & editing: GH, MBS, CJ, FJ, FF, BC, PP, EPG; Supervision: BC, PP, EPG; Project administration: BC, PP, EPG; Funding acquisition: CJ, FJ.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

We thank MetaboHUB (ANR-11-INBS-0010) and PHENOME (ANR-11-INBS-0012) projects for their financial support. PG acknowledges support from the European Research Council (grant no. 814747/SUMMIT).

Appendix ASupplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trac.2023.117225.

References

[1] E.W. Deutsch, Y. Perez-Riverol, J. Carver, S. Kawano, L. Mendoza, T. Van Den Bossche, R. Gabriels, P.A. Binz, B. Pullman, Z. Sun, J. Shofstahl, W. Bittremieux, T.D. Mak, J. Klein, Y. Zhu, H. Lam, J.A. Vizcaino, N. Bandeira, Universal spectrum identifier for mass spectra, Nat. Methods 18 (7) (2021) 768, https://doi.org/10.1038/s41592-021-01184-6.

[2] A.S. Marchev, L.V. Vasileva, K.M. Amirova, M.S. Savova, Z.P. Balcheva-Sivenova, M.I. Georgiev, Metabolomics and health: from nutritional crops and plant-based

pharmaceuticals to profiling of human biofluids, Cell. Mol. Life Sci. 78 (2021) 19–20, https://doi.org/10.1007/s00018-021-03918-3, 6487.
[3] W. Sun, Z. Chen, J. Hong, J. Shi, Promoting human nutrition and health through plant metabolomics: current status and challenges, Biology 10 (1) (2020) 20, https://doi.org/10.3390/biology10010020.
[4] S.G. Oliver, M.K. Winson, D.B. Kell, F. Baganz, Systematic functional analysis of the yeast genome, Trends Biotechnol. 16 (9) (1998) 373, https://doi.org/10.1016/s0167-7799(98)01214-1.
[5] F.A. Castelli, G. Rosati, C. Moguet, C. Fuentes, J. Marrugo-Ramírez, T. Lefebvre, H. Volland, A. Merkoçi, S. Simon, F. Fenaille, Metabolomics for personalized medicine: the input of analytical chemistry from biomarker discovery to point-of-care tests, Anal. Bioanal. Chem. 414 (2) (2021) 759, https://doi.org/10.1007/s00216-021-03586-z.
[6] E. Defossez, C. Pitteloud, P. Descombes, G. Glauser, P.M. Allard, T.W.N. Walker, P. Fernandez-Conradi, J.L. Wolfender, L. Pellissier, S. Rasmann, Spatial and evolutionary predictability of phytochemical diversity, Proc. Natl. Acad. Sci. U. S. A. 118 (3) (2021), e2013344118, https://doi.org/10.1073/pnas.2013344118.
[7] T. Dussarrat, S. Prigent, C. Latorre, S. Bernillon, A. Flandin, F.P. Diaz, C. Cassan, P. Van Delft, D. Jacob, K. Varala, J. Joubes, Y. Gibon, D. Rolin, R.A. Gutierrez, P. Petriacq, Predictive metabolomics of multiple Atacama plant species unveils a core set of generic metabolites for extreme climate resilience, New Phytol. 234 (5) (2022) 1614, https://doi.org/10.1111/nph.18095.
[8] J.K. Nicholson, A. Darzi, E. Holmes, J.C. Lindon, Metabolic Phenotyping in Personalized and Public Healthcare, Academic Press, 2016.
[9] D.K. Trivedi, K.A. Hollywood, R. Goodacre, Metabolomics for the masses: the future of metabolomics in a personalized world, New Horiz. Transl. Med. 3 (6) (2017) 294, https://doi.org/10.1016/j.nhtm.2017.06.001.
[10] G. Cao, Z. Song, Y. Hong, Z. Yang, Y. Song, Z. Chen, Z. Chen, Z. Cai, Large-scale targeted metabolomics method for metabolite profiling of human samples, Anal. Chim. Acta 1125 (2020) 144, https://doi.org/10.1016/j.aca.2020.05.053.
[11] A. Razzaq, D.S. Wishart, S.H. Wani, M.K. Hameed, M. Mubin, F. Saleem, Advances in metabolomics-driven diagnostic breeding and crop improvement, Metabolites 12 (6) (2022) 511, https://doi.org/10.3390/metabo12060511.
[12] R.C. De Vos, S. Moco, A. Lommen, J.J. Keurentjes, R.J. Bino, R.D. Hall, Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry, Nat. Protoc. 2 (4) (2007) 778, https://doi.org/10.1038/nprot.2007.95.
[13] W.B. Dunn, I.D. Wilson, A.W. Nicholls, D. Broadhurst, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, Bioanalysis 4 (18) (2012) 2249, https://doi.org/10.4155/bio.12.204.
[14] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M. D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI), Metabolomics 3 (3) (2007) 211, https://doi.org/10.1007/s11306-007-0082-2.
[15] S. Alseekh, A. Aharoni, Y. Brotman, K. Contrepois, J. D'Auria, J. Ewald, J. C. Ewald, P.D. Fraser, P. Giavalisco, R.D. Hall, Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices, Nat. Methods 18 (7) (2021) 747, https://doi.org/10.1038/s41592-021-01197-1.
[16] S. Alseekh, A.R. Fernie, Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J. 94 (6) (2018) 933, https://doi.org/10.1111/tpj.13950.
[17] T. Cajka, O. Fiehn, Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics, Anal. Chem. 88 (1) (2016) 524, https://doi.org/10.1021/acs.analchem.5b04491.
[18] J. Medina, R. Borreggine, T. Teav, L. Gao, S. Ji, J. Carrard, C. Jones, N. Blomberg, M. Jech, A. Atkins, C. Martins, A. Schmidt-Trucksass, M. Giera, A. Cazenave-Gassiot, H. Gallart-Ayala, J. Ivanisevic, Omic-scale quantitative HILIC-MS/MS approach for circulatory lipid phenotyping in clinical research, Anal. Chem. 95 (6) (2023) 3168, https://doi.org/10.1021/acs.analchem.2c02598.
[19] B. Yu, K.A. Zanetti, M. Temprosa, D. Albanes, N. Appel, C.B. Barrera, Y. Ben-Shlomo, E. Boerwinkle, J.P. Casas, C. Clish, C. Dale, A. Dehghan, A. Derkach, A. H. Eliassen, P. Elliott, E. Fahy, C. Gieger, M.J. Gunter, S. Harada, T. Harris, D. R. Herr, D. Herrington, J.N. Hirschhorn, E. Hoover, A.W. Hsing, M. Johansson, R. S. Kelly, C.M. Khoo, M. Kivimaki, B.S. Kristal, C. Langenberg, J. Lasky-Su, D. A. Lawlor, L.A. Lotta, M. Mangino, L. Le Marchand, E. Mathe, C.E. Matthews, C. Menni, L.A. Mucci, R. Murphy, M. Oresic, E. Orwoll, J. Ose, A.C. Pereira, M. C. Playdon, L. Poston, J. Price, Q. Qi, K. Rexrode, A. Risch, J. Sampson, W. J. Seow, H.D. Sesso, S.H. Shah, X.O. Shu, G.C.S. Smith, U. Sovio, V.L. Stevens, R. Stolzenberg-Solomon, T. Takebayashi, T. Tillin, R. Travis, I. Tzoulaki, C. M. Ulrich, R.S. Vasan, M. Verma, Y. Wang, N.J. Wareham, A. Wong, N. Younes, H. Zhao, W. Zheng, S.C. Moore, The Consortium of Metabolomics Studies (COMETS): metabolomics in 47 prospective cohort studies, Am. J. Epidemiol. 188 (6) (2019) 991, https://doi.org/10.1093/aje/kwz028.
[20] T. Tohge, A.R. Fernie, Metabolomics-inspired insight into developmental, environmental and genetic aspects of tomato fruit chemical composition and quality, Plant Cell Physiol. 56 (9) (2015) 1681, https://doi.org/10.1093/pcp/pcv093.
[21] O. Fernandez, E.J. Millet, R. Rincent, S. Prigent, P. Pétriacq, Y. Gibon, Plant Metabolomics and Breeding, Elsevier, 2021, p. 207.
[22] E. Luna, A. Flandin, C. Cassan, S. Prigent, C. Chevanne, C.F. Kadiri, Y. Gibon, P. Petriacq, Metabolomics to exploit the primed immune system of tomato fruit, Metabolites 10 (3) (2020) 96, https://doi.org/10.3390/metabo10030096.
[23] L. Roch, S. Prigent, H. Klose, C.B. Cakpo, B. Beauvoit, C. Deborde, L. Fouillen, P. van Delft, D. Jacob, B. Usadel, Z. Dai, M. Genard, G. Vercambre, S. Colombie, A. Moing, Y. Gibon, Biomass composition explains fruit relative growth rate and discriminates climacteric from non-climacteric species, J. Exp. Bot. 71 (19) (2020) 5823, https://doi.org/10.1093/jxb/eraa302.
[24] R.E. Danczak, R.K. Chu, S.J. Fansler, A.E. Goldman, E.B. Graham, M.M. Tfaily, J. Toyoda, J.C. Stegen, Using ecology to understand environmental metabolomes, Nat. Commun. 11 (1) (2020) 1, https://doi.org/10.1038/s41467-020-19989-y.
[25] A.E. Douglas, The microbial exometabolome: ecological resource and architect of microbial communities, Philos. Trans. R. Soc. Lond. B Biol. Sci. 375 (2020), 20190250, https://doi.org/10.1098/rstb.2019.0250.
[26] J. Kehe, A. Kulesa, A. Ortiz, C.M. Ackerman, S.G. Thakku, D. Sellers, S. Kuehn, J. Gore, J. Friedman, P.C. Blainey, Massively parallel screening of synthetic microbial communities, Proc. Natl. Acad. Sci. U. S. A. 116 (26) (2019), 12804, https://doi.org/10.1073/pnas.1900102116.
[27] A. Bauermeister, H. Mannochio-Russo, L.V. Costa-Lotufo, A.K. Jarmusch, P. C. Dorrestein, Mass spectrometry-based metabolomics in microbiome investigations, Nat. Rev. Microbiol. 20 (3) (2022) 143, https://doi.org/10.1038/s41579-021-00621-9.
[28] G. Daly, V. Ghini, A. Adessi, M. Fondi, A. Buchan, C. Viti, Towards a mechanistic understanding of microalgae–bacteria interactions: integration of metabolomic analysis and computational models, FEMS (Fed. Eur. Microbiol. Soc.) Microbiol. Rev. 46 (5) (2022) 1, https://doi.org/10.1093/femsre/fuac020.
[29] V. Mataigne, N. Vannier, P. Vandenkoornhuyse, S. Hacquard, Microbial systems ecology to understand cross-feeding in microbiomes, Front. Microbiol. 12 (2021), 780469, https://doi.org/10.3389/fmicb.2021.780469.
[30] R.A. Quinn, K.A. Hagiwara, K. Liu, M. Goudarzi, W. Pathmasiri, L.W. Sumner, T. O. Metz, Bridging the gap between analytical and microbial sciences in microbiome research, mSystems 6 (5) (2021), e0058521, https://doi.org/10.1128/mSystems.00585-21.
[31] J. Lasky-Su, R. Kelly, C.E. Wheelock, D. Broadhurst, A strategy for advancing for population-based scientific discovery using the metabolome: the establishment of the Metabolomics Society Metabolomic Epidemiology Task Group, Metabolomics 17 (5) (2021) 45, https://doi.org/10.1007/s11306-021-01789-0.
[32] M. Pietzner, I.D. Stewart, J. Raffler, K.T. Khaw, G.A. Michelotti, G. Kastenmuller, N.J. Wareham, C. Langenberg, Plasma metabolites to profile pathways in noncommunicable disease multimorbidity, Nat. Med. 27 (3) (2021) 471, https://doi.org/10.1038/s41591-021-01266-0.
[33] F. Asnicar, S.E. Berry, A.M. Valdes, L.H. Nguyen, G. Piccinno, D.A. Drew, E. Leeming, R. Gibson, C. Le Roy, H.A. Khatib, Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals, Nat. Med. 27 (2) (2021) 321, https://doi.org/10.1038/s41591-020-01183-8.
[34] B. Bana, F. Cabreiro, The microbiome and aging, Annu. Rev. Genet. 53 (2019) 239, https://doi.org/10.1146/annurev-genet-112618-043650.
[35] M. Palmnas, C. Brunius, L. Shi, A. Rostgaard-Hansen, N.E. Torres, R. Gonzalez-Dominguez, R. Zamora-Ros, Y.L. Ye, J. Halkjaer, A. Tjonneland, G. Riccardi, R. Giacco, G. Costabile, C. Vetrani, J. Nielsen, C. Andres-Lacueva, R. Landberg, Perspective: metabotyping-A potential personalized nutrition strategy for precision prevention of cardiometabolic disease, Adv. Nutr. 11 (3) (2020) 524, https://doi.org/10.1093/advances/nmz121.
[36] H. Ashrafian, V. Sounderajah, R. Glen, T. Ebbels, B.J. Blaise, D. Kalra, K. Kultima, O. Spjuth, L. Tenori, R.M. Salek, Metabolomics: the stethoscope for the twenty-first century, Med. Princ. Pract. 30 (4) (2021) 301, https://doi.org/10.1159/000513545.
[37] J. Van Der Greef, T. Hankemeier, R.N. McBurney, Metabolomics-based systems biology and personalized medicine: moving towards n=1 clinical trials? Pharmacogenomics 7 (7) (2006) 1087, https://doi.org/10.2217/14622416.7.7.1087.
[38] M.A. Haendel, C.G. Chute, P.N. Robinson, Classification, ontology, and precision medicine, N. Engl. J. Med. 379 (15) (2018) 1452, https://doi.org/10.1056/NEJMra1615014.
[39] G. Theodoridis, H. Gika, D. Raftery, R. Goodacre, R.S. Plumb, I.D. Wilson, Ensuring fact-based metabolite identification in liquid chromatography-mass spectrometry-based metabolomics, Anal. Chem. 95 (8) (2023) 3909, https://doi.org/10.1021/acs.analchem.2c05192.
[40] A.D. Southam, L.D. Haglington, L. Najdekr, A. Jankevics, R.J.M. Weber, W. B. Dunn, Assessment of human plasma and urine sample preparation for reproducible and high-throughput UHPLC-MS clinical metabolic phenotyping, Analyst 145 (20) (2020) 6511, https://doi.org/10.1039/D0AN01319F.
[41] J.M. Malinowska, T. Palosaari, J. Sund, D. Carpi, G.R. Lloyd, R.J. Weber, M. Whelan, M.R. Viant, Automated sample preparation and data collection workflow for high-throughput in vitro metabolomics, Metabolites 12 (1) (2022) 52, https://doi.org/10.3390/metabo12010052.
[42] R.D. Hall, J.C. D'Auria, A.C.S. Ferreira, Y. Gibon, D. Kruszka, P. Mishra, R. Van de Zedde, High-throughput plant phenotyping: a role for metabolomics? Trends Plant Sci. 27 (6) (2022) 549, https://doi.org/10.1016/j.tplants.2022.02.001.
[43] L. Lofgren, M. Stahlman, G.B. Forsberg, S. Saarinen, R. Nilsson, G.I. Hansson, The BUME method: a novel automated chloroform-free 96-well total lipid extraction method for blood plasma, J. Lipid Res. 53 (8) (2012) 1690, https://doi.org/10.1194/jlr.D023036.
[44] M.A. Surma, R. Herzog, A. Vasilj, C. Klose, N. Christinat, D. Morin-Rivron, K. Simons, M. Masoodi, J.L. Sampaio, An automated shotgun lipidomics platform for high throughput, comprehensive, and quantitative analysis of blood plasma intact lipids, Eur. J. Lipid Sci. Technol. 117 (10) (2015) 1540, https://doi.org/10.1002/ejlt.201500145.

[45] X. Guo, L. Zhou, Y. Wang, F. Suo, C. Wang, W. Zhou, L. Gou, M. Gu, G. Xu, Development of a fast and robust liquid chromatography-mass spectrometry-based metabolomics analysis method for neonatal dried blood spots, J. Pharm. Biomed. Anal. 230 (2023), 115383, https://doi.org/10.1016/j.jpba.2023.115383.
[46] H.B. Ferreira, I.M.S. Guerra, T. Melo, H. Rocha, A.S.P. Moreira, A. Paiva, M. R. Domingues, Dried blood spots in clinical lipidomics: optimization and recent findings, Anal. Bioanal. Chem. 414 (24) (2022) 7085, https://doi.org/10.1007/s00216-022-04221-1.
[47] K. Li, J.C. Naviaux, J.M. Monk, L. Wang, R.K. Naviaux, Improved dried blood spot-based metabolomics: a targeted, broad-spectrum, single-injection method, Metabolites 10 (2020) 3, https://doi.org/10.3390/metabo10030082.
[48] S. Bravo-Veyrat, G. Hopfgartner, Mass spectrometry based high-throughput bioanalysis of low molecular weight compounds: are we ready to support personalized medicine? Anal. Bioanal. Chem. 414 (1) (2022) 181, https://doi.org/10.1007/s00216-021-03583-2.
[49] B. Habchi, S. Alves, A. Paris, D.N. Rutledge, E. Rathahao-Paris, How to really perform high throughput metabolomic analyses efficiently? TrAC, Trends Anal. Chem. 85 (2016) 128, https://doi.org/10.1016/j.trac.2016.09.005.
[50] S. Wernisch, S. Pennathur, Application of differential mobility-mass spectrometry for untargeted human plasma metabolomic analysis, Anal. Bioanal. Chem. 411 (24) (2019) 6297, https://doi.org/10.1007/s00216-019-01719-z.
[51] L. Perez de Souza, S. Alseekh, F. Scossa, A.R. Fernie, Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research, Nat. Methods 18 (7) (2021) 733, https://doi.org/10.1038/s41592-021-01116-4.
[52] A.D. Southam, R.J. Weber, J. Engel, M.R. Jones, M.R. Viant, A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics, Nat. Protoc. 12 (2) (2017) 310, https://doi.org/10.1038/nprot.2016.156.
[53] E. Chekmeneva, G. dos Santos Correia, Q. Chan, A. Wijeyesekera, A. Tin, J. H. Young, P. Elliott, J.K. Nicholson, E. Holmes, Optimization and application of direct infusion nanoelectrospray HRMS method for large-scale urinary metabolic phenotyping in molecular epidemiology, J. Prot. Res. 16 (4) (2017) 1646, https://doi.org/10.1021/acs.jproteome.6b01003.
[54] A. Delvaux, E. Rathahao-Paris, S. Alves, Different ion mobility-mass spectrometry coupling techniques to promote metabolomics, Mass Spectrom. Rev. 41 (5) (2022) 695, https://doi.org/10.1002/mas.21685.
[55] R.S. Plumb, L.A. Gethings, P.D. Rainville, G. Isaac, R. Trengove, A.M. King, I. D. Wilson, Advances in high throughput LC/MS based metabolomics: a review, TrAC, Trends Anal. Chem. 160 (2023), 116954, https://doi.org/10.1016/j.trac.2023.116954.
[56] M.R. Lewis, J.T. Pearce, K. Spagou, M. Green, A.C. Dona, A.H. Yuen, M. David, D. J. Berry, K. Chappell, V. Horneffer-van der Sluis, R. Shaw, S. Lovestone, P. Elliott, J. Shockcor, J.C. Lindon, O. Cloarec, Z. Takats, E. Holmes, J.K. Nicholson, Development and application of ultra-performance liquid chromatography-TOF MS for precision large scale urinary metabolic phenotyping, Anal. Chem. 88 (18) (2016) 9004, https://doi.org/10.1021/acs.analchem.6b01481.
[57] M.R. Lewis, E. Chekmeneva, S. Camuzeaux, C. Sands, A. Yuen, M. David, A. Salam, K. Chappell, B. Cooper, G. Haggart, L. Maslen, M. Gómez-Romero, V. Horneffer-van der Sluis, G. Correia, Z. Takats, An open platform for large scale LC-MS-based metabolomics, ChemRixv (2022), https://doi.org/10.26434/chemrxiv-2022-nq9k0.
[58] V. Fitz, Y. El Abiead, D. Berger, G. Koellensperger, Systematic investigation of LC miniaturization to increase sensitivity in wide-target LC-MS-based trace bioanalysis of small molecules, Front. Mol. Biosci. 9 (2022), 857505, https://doi.org/10.3389/fmolb.2022.857505.
[59] N. Gray, K. Adesina-Georgiadis, E. Chekmeneva, R.S. Plumb, I.D. Wilson, J. K. Nicholson, Development of a rapid microbore metabolic profiling ultraperformance liquid chromatography-mass spectrometry approach for high-throughput phenotyping studies, Anal. Chem. 88 (11) (2016) 5742, https://doi.org/10.1021/acs.analchem.6b00038.
[60] C.G. Vasilopoulou, K. Sulek, A.D. Brunner, N.S. Meitei, U. Schweiger-Hufnagel, S. W. Meyer, A. Barsch, M. Mann, F. Meier, Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts, Nat. Commun. 11 (1) (2020) 331, https://doi.org/10.1038/s41467-019-14044-x.
[61] G. Paglia, A.J. Smith, G. Astarita, Ion mobility mass spectrometry in the omics era: challenges and opportunities for metabolomics and lipidomics, Mass Spectrom. Rev. 41 (5) (2022) 722, https://doi.org/10.1002/mas.21686.
[62] E. Rampler, Y.E. Abiead, H. Schoeny, M. Rusz, F. Hildebrand, V. Fitz, G. Koellensperger, Recurrent topics in mass spectrometry-based metabolomics and lipidomics—standardization, coverage, and throughput, Anal. Chem. 93 (1) (2021) 519, https://doi.org/10.1021/acs.analchem.0c04698.
[63] F. Fenaille, P. Barbier Saint-Hilaire, K. Rousseau, C. Junot, Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: where do we stand? J. Chromatogr. A 1526 (2017) 1, https://doi.org/10.1016/j.chroma.2017.10.043.
[64] E. Defossez, J. Bourquin, S. von Reuss, S. Rasmann, G. Glauser, Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics, Mass Spectrom. Rev. 42 (1) (2023) 131, https://doi.org/10.1002/mas.21715.
[65] P. Barbier Saint Hilaire, K. Rousseau, A. Seyer, S. Dechaumet, A. Damont, C. Junot, F. Fenaille, Comparative evaluation of data dependent and data independent acquisition workflows implemented on an orbitrap fusion for untargeted metabolomics, Metabolites 10 (4) (2020) 158, https://doi.org/10.3390/metabo10040158.
[66] M. Raetz, R. Bonner, G. Hopfgartner, SWATH-MS for metabolomics and lipidomics: critical aspects of qualitative and quantitative analysis, Metabolomics 16 (6) (2020) 71, https://doi.org/10.1007/s11306-020-01692-0.
[67] T. van der Laan, I. Boom, J. Maliepaard, A.C. Dubbelman, A.C. Harms, T. Hankemeier, Data-independent acquisition for the quantification and identification of metabolites in plasma, Metabolites 10 (12) (2020) 514, https://doi.org/10.3390/metabo10120514.
[68] K. Kiefer, A. Muller, H. Singer, J. Hollender, New relevant pesticide transformation products in groundwater detected using target and suspect screening for agricultural and urban micropollutants with LC-HRMS, Water Res. 165 (2019), 114972, https://doi.org/10.1016/j.watres.2019.114972.
[69] A. Vignoli, V. Ghini, G. Meoni, C. Licari, P.G. Takis, L. Tenori, P. Turano, C. Luchinat, High-throughput metabolomics by 1D NMR, Angew Chem. Int. Ed. Engl. 58 (4) (2019) 968, https://doi.org/10.1002/anie.201804736.
[70] G.N. Gowda, D. Raftery, NMR-Based Metabolomics: Methods and Protocols, Springer, 2019.
[71] E. Martineau, J.N. Dumez, P. Giraudeau, Fast quantitative 2D NMR for metabolomics and lipidomics: a tutorial, Magn. Reson. Chem. 58 (5) (2020) 390, https://doi.org/10.1002/mrc.4899.
[72] B.B. Misra, New software tools, databases, and resources in metabolomics: updates from 2020, Metabolomics 17 (5) (2021) 49, https://doi.org/10.1007/s11306-021-01796-1.
[73] S. Li, A. Siddiqa, M. Thapa, S. Zheng, Trackable and scalable LC-MS metabolomics data processing using asari, bioRxiv 2022 (6) (2022), https://doi.org/10.1101/2022.06.10.495665, 10.495665.
[74] Y. Gloaguen, J.A. Kirwan, D. Beule, Deep learning-assisted peak curation for large-scale LC-MS metabolomics, Anal. Chem. 94 (12) (2022) 4930, https://doi.org/10.1021/acs.analchem.1c02220.
[75] A. Delabriere, P. Warmer, V. Brennsteiner, N. Zamboni, SLAW: a scalable and self-optimizing processing workflow for untargeted LC-MS, Anal. Chem. 93 (45) (2021), 15024, https://doi.org/10.1021/acs.analchem.1c02687.
[76] I. Karaman, D.L. Ferreira, C.L. Boulange, M.R. Kaluarachchi, D. Herrington, A. C. Dona, R. Castagne, A. Moayyeri, B. Lehne, M. Loh, P.S. de Vries, A. Dehghan, O.H. Franco, A. Hofman, E. Evangelou, I. Tzoulaki, P. Elliott, J.C. Lindon, T. M. Ebbels, Workflow for integrated processing of multicohort untargeted 1H NMR metabolomics data in large-scale metabolic epidemiology, J. Proteome Res. 15 (12) (2016), https://doi.org/10.1021/acs.jproteome.6b00125, 4188.
[77] D.S. Wishart, L.L. Cheng, V. Copié, A.S. Edison, H.R. Eghbalnia, J.C. Hoch, G. J. Gouveia, W. Pathmasiri, R. Powers, T.B. Schock, NMR and metabolomics—a roadmap for the future, Metabolites 12 (8) (2022) 678, https://doi.org/10.3390/metabo12080678.
[78] T. Buergel, J. Steinfeldt, G. Ruyoga, M. Pietzner, D. Bizzarri, D. Vojinovic, J. Upmeier Zu Belzen, L. Loock, P. Kittner, L. Christmann, N. Hollmann, H. Strangalies, J.M. Braunger, B. Wild, S.T. Chiesa, J. Spranger, F. Klostermann, E.B. van den Akker, S. Trompet, S.P. Mooijaart, N. Sattar, J.W. Jukema, B. Lavrijssen, M. Kavousi, M. Ghanbari, M.A. Ikram, E. Slagboom, M. Kivimaki, C. Langenberg, J. Deanfield, R. Eils, U. Landmesser, Metabolomic profiles predict individual multidisease outcomes, Nat. Med. 28 (11) (2022) 2309, https://doi.org/10.1038/s41591-022-01980-3.
[79] D.W. Li, A.L. Hansen, C. Yuan, L. Bruschweiler-Li, R. Bruschweiler, DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra, Nat. Commun. 12 (2021) 5229, https://doi.org/10.1038/s41467-021-25496-5.
[80] S.M. Colby, J.R. Nuñez, N.O. Hodas, C.D. Corley, R.R. Renslow, Deep learning to generate in silico chemical property libraries and candidate molecules for small molecule identification in complex samples, Anal. Chem. 92 (2) (2019) 1720, https://doi.org/10.1021/acs.analchem.9b02348.
[81] H. Ji, Y. Xu, H. Lu, Z. Zhang, Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification, Anal. Chem. 91 (9) (2019) 5629, https://doi.org/10.1021/acs.analchem.8b05405.
[82] W. Bittremieux, L.I. Levitsky, M. Pilz, T. Sachsenberg, M. Wang, P.C. Dorrestein, Unified and standardized mass spectrometry data processing in Python using spectrum_utils, J. Prot. Res. 22 (2) (2023) 265, https://doi.org/10.1021/acs.jproteome.2c00632.
[83] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying small molecules via high resolution mass spectrometry: communicating confidence, Environ. Sci. Technol. 48 (4) (2014), https://doi.org/10.1021/es5002105, 2097.
[84] Y. Cai, Z. Zhou, Z.-J. Zhu, Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics, Trac. Trends Anal. Chem. 158 (2023), 116903, https://doi.org/10.1016/j.trac.2022.116903.
[85] D.W. Li, A. Leggett, L. Bruschweiler-Li, R. Bruschweiler, COLMARq: a web server for 2D NMR peak picking and quantitative comparative analysis of cohorts of metabolomics samples, Anal. Chem. 94 (24) (2022) 8674, https://doi.org/10.1021/acs.analchem.2c00891.
[86] E. Stancliffe, M. Schwaiger-Haber, M. Sindelar, M.J. Murphy, M. Soerensen, G. J. Patti, An untargeted metabolomics workflow that scales to thousands of samples for population-based studies, Anal. Chem. 94 (50) (2022), 17370, https://doi.org/10.1021/acs.analchem.2c01270.
[87] W. Han, L. Li, Evaluating and minimizing batch effects in metabolomics, Mass Spectrom. Rev. 41 (3) (2022) 421, https://doi.org/10.1002/mas.21672.
[88] K. Deng, F. Zhang, Q. Tan, Y. Huang, W. Song, Z. Rong, Z.J. Zhu, K. Li, Z. Li, WaveICA: a novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis, Anal. Chim. Acta 1061 (2019) 60, https://doi.org/10.1016/j.aca.2019.02.010.

[89] R. Molania, J.A. Gagnon-Bartsch, A. Dobrovic, T.P. Speed, A new normalization for Nanostring nCounter gene expression data, Nucleic Acids Res. 47 (12) (2019) 6073, https://doi.org/10.1093/nar/gkz433.
[90] S. Fan, T. Kind, T. Cajka, S.L. Hazen, W.W. Tang, R. Kaddurah-Daouk, M.R. Irvin, D.K. Arnett, D.K. Barupal, O. Fiehn, Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data, Anal. Chem. 91 (5) (2019) 3590, https://doi.org/10.1021/acs.analchem.8b05592.
[91] J.A. Kirwan, H. Gika, R.D. Beger, D. Bearden, W.B. Dunn, R. Goodacre, G. Theodoridis, M. Witting, L.R. Yu, I.D. Wilson, A. metabolomics Quality, C. Quality Control, Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management, Metabolomics 18 (2022) 70, https://doi.org/10.1007/s11306-022-01926-3.
[92] H. Luan, F. Ji, Y. Chen, Z. Cai, statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data, Anal. Chim. Acta 1036 (2018) 66, https://doi.org/10.1016/j.aca.2018.08.002.
[93] C. Brunius, L. Shi, R. Landberg, Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction, Metabolomics 12 (11) (2016) 173, https://doi.org/10.1007/s11306-016-1124-4.
[94] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (6) (2012) 882, https://doi.org/10.1093/bioinformatics/bts034.
[95] T. Kim, O. Tang, S.T. Vernon, K.A. Kott, Y.C. Koay, J. Park, D.E. James, S. M. Grieve, T.P. Speed, P. Yang, G.A. Figtree, J.F. O'Sullivan, J.Y.H. Yang, A hierarchical approach to removal of unwanted variation for large-scale metabolomics data, Nat. Commun. 12 (2021) 4992, https://doi.org/10.1038/s41467-021-25210-5.
[96] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D.Y. Weiss-Solís, R. Duque, H. Bersini, A. Nowé, Batch effect removal methods for microarray gene expression data integration: a survey, Briefings Bioinf. 14 (4) (2013) 469, https://doi.org/10.1093/bib/bbs037.
[97] A. Imbert, M. Rompais, M. Selloum, F. Castelli, E. Mouton-Barbosa, M. Brandolini-Bunlon, E. Chu-Van, C. Joly, A. Hirschler, P. Roger, T. Burger, S. Leblanc, T. Sorg, S. Ouzia, Y. Vandenbrouck, C. Medigue, C. Junot, M. Ferro, E. Pujos-Guillot, A.G. de Peredo, F. Fenaille, C. Carapito, Y. Herault, E. A. Thevenot, ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis, Sci. Data 8 (1) (2021) 311, https://doi.org/10.1038/s41597-021-01095-3.
[98] R. Climaco Pinto, I. Karaman, M.R. Lewis, J. Hallqvist, M. Kaluarachchi, G. Graca, E. Chekmeneva, B. Durainayagam, M. Ghanbari, M.A. Ikram, H. Zetterberg, J. Griffin, P. Elliott, I. Tzoulaki, A. Dehghan, D. Herrington, T. Ebbels, Finding correspondence between metabolomic features in untargeted liquid chromatography–mass spectrometry metabolomics datasets, Anal. Chem. 94 (14) (2022) 5493, https://doi.org/10.1021/acs.analchem.1c03592.
[99] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M. E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016), 160018, https://doi.org/10.1038/sdata.2016.18.
[100] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C.T. Evelo, FAIR principles: interpretations and implementation considerations, Data Intell. 2 (1–2) (2020) 10, https://doi.org/10.1162/dint_r_00024.
[101] D. Kopczynski, N. Hoffmann, B. Peng, G. Liebisch, F. Spener, R. Ahrends, Goslin 2.0 implements the recent lipid shorthand nomenclature for MS-derived lipid structures, Anal. Chem. 94 (16) (2022) 6097, https://doi.org/10.1021/acs.analchem.1c05430.
[102] G. Wohlgemuth, S.S. Mehta, R.F. Mejia, S. Neumann, D. Pedrosa, T. Pluskal, E. L. Schymanski, E.L. Willighagen, M. Wilson, D.S. Wishart, M. Arita, P. C. Dorrestein, N. Bandeira, M. Wang, T. Schulze, R.M. Salek, C. Steinbeck, V. C. Nainala, R. Mistrik, T. Nishioka, O. Fiehn, SPLASH, a hashed identifier for mass spectra, Nat. Biotechnol. 34 (11) (2016) 1099, https://doi.org/10.1038/nbt.3689.
[103] S. Savoi, P. Arapitsas, E. Duchene, M. Nikolantonaki, I. Ontanon, S. Carlin, F. Schwander, R.D. Gougeon, A.C.S. Ferreira, G. Theodoridis, R. Topfer, U. Vrhovsek, A.F. Adam-Blondon, M. Pezzotti, F. Mattivi, Grapevine and wine metabolomics-based guidelines for fair data and metadata management, Metabolites 11 (11) (2021) 757, https://doi.org/10.3390/metabo11110757.
[104] M.R. Viant, T.M.D. Ebbels, R.D. Beger, D.R. Ekman, D.J.T. Epps, H. Kamp, P.E. G. Leonards, G.D. Loizou, J.I. MacRae, B. van Ravenzwaay, P. Rocca-Serra, R. M. Salek, T. Walk, R.J.M. Weber, Use cases, best practice and reporting standards for metabolomics in regulatory toxicology, Nat. Commun. 10 (1) (2019) 3041, https://doi.org/10.1038/s41467-019-10900-y.
[105] S.A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C.T. Evelo, M.J. Forster, P. Gaudet, J. Gilbert, C. Goble, J.L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S.J. Ho Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C.E. Shamu, C.A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, W. Hide, Toward interoperable bioscience data, Nat. Genet. 44 (2) (2012) 121, https://doi.org/10.1038/ng.1054.
[106] P. Strömert, J. Hunold, A. Castro, S. Neumann, O. Koepler, Ontologies4Chem: the landscape of ontologies in chemistry, Pure Appl. Chem. 94 (6) (2022) 605, https://doi.org/10.1515/pac-2021-2007.
[107] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, Nucleic Acids Res. 39 (2011) W541, https://doi.org/10.1093/nar/gkr469, suppl_2.
[108] N. Poupin, F. Vinson, A. Moreau, A. Batut, M. Chazalviel, B. Colsch, L. Fouillen, S. Guez, S. Khoury, J. Dalloux-Chioccioli, A. Tournadre, P. Le Faouder, C. Pouyet, P. Van Delft, F. Viars, J. Bertrand-Michel, F. Jourdan, Improving lipid mapping in genome scale metabolic networks using ontologies, Metabolomics 16 (4) (2020) 44, https://doi.org/10.1007/s11306-020-01663-5.
[109] A. Trautman, R. Linchangco, R. Walstead, J.J. Jay, C. Brouwer, The Aliment to Bodily Condition knowledgebase (ABCkb): a database connecting plants and human health, BMC Res. Notes 14 (1) (2021) 433, https://doi.org/10.1186/s13104-021-05835-x.
[110] M. Delmas, O. Filangi, N. Paulhe, F. Vinson, C. Duperier, W. Garrier, P.E. Saunier, Y. Pitarch, F. Jourdan, F. Giacomoni, C. Frainay, FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases, Bioinformatics 37 (21) (2021) 3896, https://doi.org/10.1093/bioinformatics/btab627.
[111] K.A. Lippa, J.J. Aristizabal-Henao, R.D. Beger, J.A. Bowden, C. Broeckling, C. Beecher, W. Clay Davis, W.B. Dunn, R. Flores, R. Goodacre, G.J. Gouveia, A. C. Harms, T. Hartung, C.M. Jones, M.R. Lewis, I. Ntai, A.J. Percy, D. Raftery, T. B. Schock, J. Sun, G. Theodoridis, F. Tayyari, F. Torta, C.Z. Ulmer, I. Wilson, B. K. Ubhi, Reference materials for MS-based untargeted metabolomics and lipidomics: a review by the metabolomics quality assurance and quality control consortium (mQACC), Metabolomics 18 (4) (2022) 29, https://doi.org/10.1007/s11306-021-01848-6.
[112] I. Hotea, C. Sirbu, A.M. Plotuna, E. Tîrziu, C. Badea, A. Berbecea, M. Dragomirescu, I. Radulov, Integrating (Nutri-)Metabolomics into the one health tendency—the key for personalized medicine advancement, Metabolites 13 (7) (2023) 800, https://doi.org/10.3390/metabo13070800.
[113] K. Schlaeppi, J.J. Gross, S. Hapfelmeier, M. Erb, Plant chemistry and food web health, New Phytol. 231 (3) (2021) 957, https://doi.org/10.1111/nph.17385.
[114] B. Comte, J. Baumbach, A. Benis, J. Basilio, N. Debeljak, A. Flobak, C. Franken, N. Harel, F. He, M. Kuiper, J.A. Mendez Perez, E. Pujos-Guillot, T. Rezen, D. Rozman, J.A. Schmid, J. Scerri, P. Tieri, K. Van Steen, S. Vasudevan, S. Watterson, H. Schmidt, Network and systems medicine: position paper of the European collaboration on science and Technology action on open multiscale systems medicine, Netw. Syst. Med. 3 (1) (2020) 67, https://doi.org/10.1089/nsm.2020.0004.
[115] M Temprosa, SC Moore, KA Zanetti, N Appel, D Ruggieri, KM Mazzilli, KL Chen, RS Kelly, JA Lasky-Su, E Loftfield, K McClain, B Park, L Trijsburg, OA Zeleznik, EA Mathé, COMETS Analytics: An Online Tool for Analyzing and Meta-Analyzing Metabolomics Data in Large Research Consortia, Am J Epidemiol 191 (1) (2022 Jan 1) 147–158, https://doi.org/10.1093/aje/kwab120. PMID: 33889934; PMCID: PMC8897993.