nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-018-0064-z

A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links

Helle Krogh Pedersen[1,12], Sofia K. Forslund[2,3,12], Valborg Gudmundsdottir[4], Anders Østergaard Petersen[4], Falk Hildebrand[3], Tuulia Hyötyläinen[5], Trine Nielsen[1], Torben Hansen[1], Peer Bork[3], S. Dusko Ehrlich[6,7], Søren Brunak[4,8], Matej Oresic[9,10], Oluf Pedersen[1*], Henrik Bjørn Nielsen[11]

We recently presented a three-pronged association study that integrated human intestinal microbiome data derived from shotgun-based sequencing with untargeted serum metabolome data and measures of host physiology. Metabolome and microbiome data are high dimensional, posing a major challenge for data integration. Here, we present a step-by-step computational protocol that details and discusses the dimensionality-reduction techniques used and methods for subsequent integration and interpretation of such heterogeneous types of data. Dimensionality reduction was achieved through a combination of data normalization approaches, binning of co-abundant genes and metabolites, and integration of prior biological knowledge. The use of prior knowledge to overcome functional redundancy across microbiome species is one central advance of our method over available alternative approaches. Applying this framework, other investigators can integrate various '-omics' readouts with variables of host physiology or any other phenotype of interest (e.g., connecting host and microbiome readouts to disease severity or treatment outcome in a clinical cohort) in a three-pronged association analysis to identify potential mechanistic links to be tested in experimental settings. Although we originally developed the framework for a human metabolome–microbiome study, it is generalizable to other organisms and environmental metagenomes, as well as to studies including other -omics domains such as transcriptomics and proteomics. The provided R code runs in ~1 h on a standard PC.

Introduction

Common metabolic disorders are multifactorial, with risk factors including host genetics and multiple environmental exposures, such as lifestyle, alongside action of gut microbial symbionts and pathogens[1–4]. Circulating metabolite levels often act as intermediaries between states of the microbial ecosystem in the gastrointestinal tract and host biology[5–7]. Therefore, characterization of microbial community composition and functional potential[8] must be coordinated with targeted or untargeted metabolomic analysis of various biological compartments of the host[9]. Systems medicine–based approaches can subsequently be taken to mine these high-dimensional data for simple or complex associations, yielding insights into the biology of human health and the pathogenesis of common multifactorial disorders. However, few studies have thus far linked multiple high-dimensional biological feature spaces together in the same study sample, an effort in which additional challenges arise with regard to data integration and interpretation of analytical outcome.

We recently reported results from a study combining measures of host phenotype, gut metagenome and fasting serum metabolome[10], in which we developed a computational framework for a

[1]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [2]Experimental and Clinical Research Centre, a joint center of Max Delbrück Centre for Molecular Medicine & Charité University Hospital, Berlin, Germany. [3]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [4]Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark. [5]MTM Research Centre, School of Science and Technology, Örebro University, Örebro, Sweden. [6]MetaGénoPolis (MGP), INRA, Université Paris-Saclay, Jouy-en-Josas, France. [7]Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, King's College London, London, UK. [8]Novo Nordisk Foundation Center for Protein Research, Disease Systems Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [9]School of Medical Sciences, Örebro University, Örebro, Sweden. [10]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland. [11]Clinical Microbiomics A/S, Copenhagen, Denmark. [12]These authors contributed equally: Helle Krogh Pedersen, Sofia K. Forslund
*e-mails: oluf@sund.ku.dk; hbjorn@clinical-microbiomics.com

Box 1 | Pre-processing of raw data and input files for the protocol

Before the integrative analysis, the raw data must be processed. For microbiome data, this includes (i) pre-processing of sequencing reads (including quality control and filtering out of host human reads), (ii) generation of reference gene catalog (or selection of an existing one), (iii) mapping of reads, (iv) downsizing/rarefying (thus normalizing the data so as to be comparable between samples despite any differences in read depth) and (v) binning of genes to MGS entities.

For metabolome data, the typical preprocessing workflow includes (i) raw file import, (ii) detection of peaks, (iii) filtering/smoothing, (iv) peak list de-isotoping, (v) alignment, (vi) gap filling, (vii) integration of peaks, (viii) normalization and finally (ix) peak/feature identification.

Although these procedures fall outside of the scope of the present protocol, we refer to our previous paper[10] and the Supplementary Methods for more detailed descriptions of the pre-processing of the microbiome and metabolome data that were used to generate the input files provided with this protocol.

An archive containing the demonstration data is available at the accompanying Git repository (https://bitbucket.org/hellekp/clinical-micro-meta-integration); it includes preprocessed microbiome and metabolome data and phenotype information for 397 individuals in addition to functional microbiome annotation, a set of MGSs with taxonomic annotation and a manually composed annotation of metabolite clusters, as described in the 'Materials' section. These files have already undergone the numerous pre-processing steps listed above. For new data, one must tailor the pre-processing toward the particular experimental protocol and analytic platforms used to generate the data.

top–down integration approach of the three data types. Indeed, this hypothesis-free, yet phenotype-anchored, approach revealed several findings of biological relevance—and as such serves as a showcase for a successful data-integration framework. Although there are alternative metagenome analysis frameworks that could provide input equivalent to that of our core method, the analysis we conducted builds on previous developments from the MetaHIT consortium, notably, its microbial reference gene catalog construction[11] and metagenomic species (MGS) framework[12]. The protocol is provided as a script pipeline for the R programming language, which can be used as a concrete starting point for researchers seeking to undertake similar projects.

The objective of this protocol, and the novel advance offered by the outlined approach, is focused on the integrative analysis of high-throughput data; the protocol could be equally applied to other forms of data so long as they are structured similarly (e.g., quantitative and between-sample comparable measurements, provided as data matrices and originating from transcriptomics or proteomics profiling, e.g., with relevant a priori groupings of features, such as gene families organized into functional modules) to those for which we designed the method. Thus, the underlying principles of the protocol should be seen as agnostic with regard to data type, and are applicable to various forms of host and microbiome quantitative measurements, so long as functionality can be attributed to the organisms. A general method for microbiome and metabolome study design and data generation is outside the scope of this protocol; we refer interested readers to the following papers for examples of how to undertake microbiome[13–17] or metabolome[18–23] studies. However, we do provide a detailed description of the specific methods applied to generate the data for the study by Pedersen et al.[10] (see also Box 1 and the Supplementary Methods), which are also used here as demonstration data for the bioinformatics protocol.

Applications

We originally applied the computational framework to a Danish sample of nondiabetic individuals and type 2 diabetes (T2D) patients with measurements of clinical phenotypes, gut metagenome and fasting serum metabolome (325 polar metabolites and 876 molecular lipids; here collectively termed 'metabolites')[10]. The aim of the three-pronged study was to elucidate mechanisms for the gut microbial influence on circulating metabolites in the context of insulin resistance of the host. Briefly, we found the fasting serum metabolome of insulin-resistant individuals to be characterized by increased levels of branched-chain amino acids (BCAAs), alongside a gut microbiome enriched in microbes possessing biosynthetic potential for BCAA biosynthesis and depleted of microbes with bacterial inward transporters for these amino acids. Furthermore, we found the association between BCAA biosynthetic potential and insulin resistance to be driven mainly by *Prevotella copri* and *Bacteroides vulgatus*. By contrast, the association between BCAA transport and improved insulin sensitivity was driven by multiple species that each had only minor effects. The outcome of the three-pronged association analyses was further tested in rodent experiments[10].

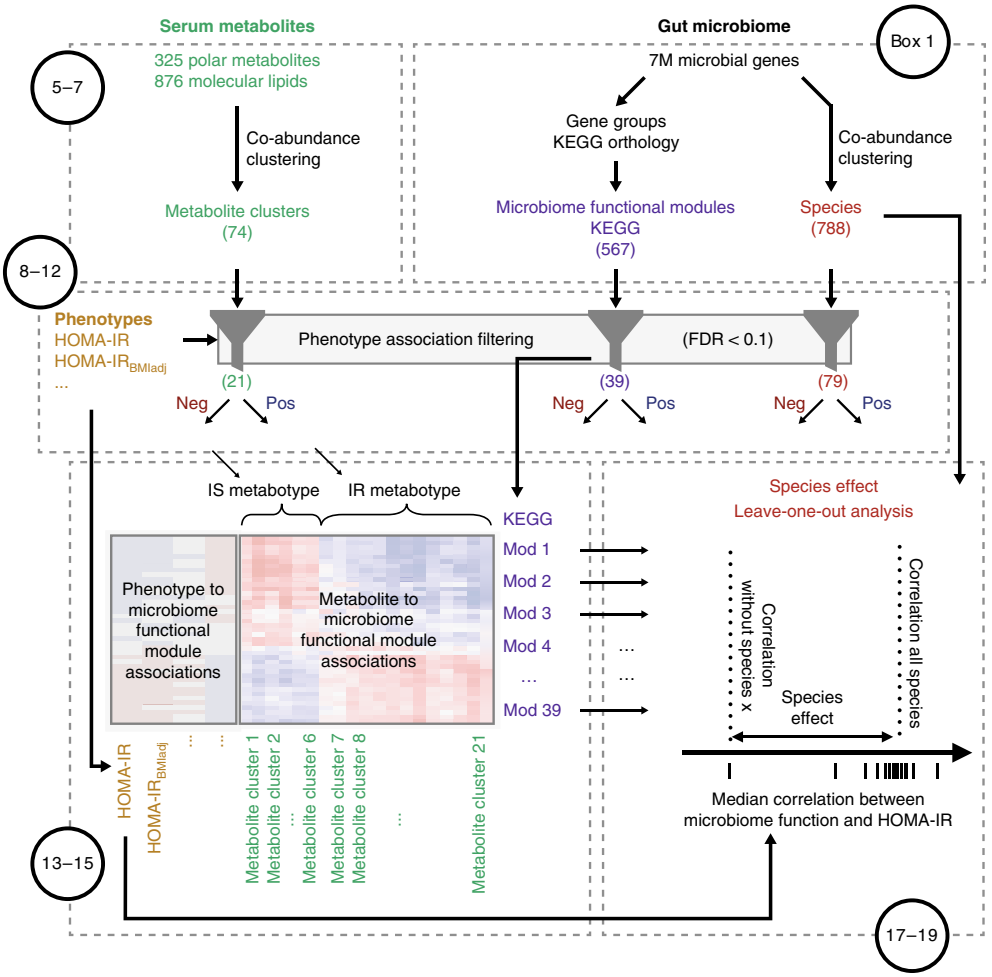

Fig. 1 | Overview of the protocol workflow integrating human phenotype, serum metabolome and gut microbiome data. After pre-processing raw microbiome and metabolome data, metabolites are summarized as co-abundance clusters (Steps 5–7), and KEGG module and species abundance profiles are extracted from gut microbiome data (Box 1). Next, in the phenotype-filtering step, features are filtered for statistically significant positive or negative associations with the phenotype of interest (here HOMA-IR or HOMA-IR$_{BMIadj}$) (Steps 8–12), and the resulting features are taken forward for cross-domain correlation/association analyses (Steps 13–15). Finally, microbial driver species for the KEGG module associations with HOMA-IR are identified by using the leave-one-MGS-out analysis (Steps 17–19). Numbers in circles refer to protocol steps. FDR, false-discovery rate; IR, insulin resistance; IS, insulin sensitivity; Mod, module; neg, negative; pos, positive. Adapted with permission from Pedersen et al.[10], Springer Nature.

In the present protocol, we describe and discuss in detail how such computational data integration can be carried out more generally, using our study on insulin resistance as a showcase. The protocol is generalizable to other phenotype study scenarios, in which the outcome of interest can be either a continuous variable, as shown here, or a discrete variable (in which case, univariate significance tests and effect size estimates simply would replace the correlation tests). Although the workflow described is for a combined metagenome taxonomy and function, metabolome and phenotype dataset, the general analysis framework can be adapted to other types of datasets, for example, by substituting the metabolomics data in the computational analysis with other types of -omics data, such as proteomics data, or completely different use cases, for example, transcriptome versus immune system versus treatment response, or environmental metagenome versus soil chemical screen versus crop yields.

Overview of the procedure

The overall workflow, with examples taken from Pedersen et al.[10], is shown in Fig. 1. The pre-processing steps of machine readouts taken before such integrative analyses are highlighted in Box 1.

Table 1 | Examples of different strategies for data-driven dimensionality reduction and resources that can be applied for knowledge-driven dimensionality reduction

	Data driven	Knowledge driven
Microbiome data	Binning of co-abundant genes, e.g., using the MGS framework[a] [12], MetaBat[50], MaxBin[51] and so on, or using single-copy phylogenetic marker genes (mOTU)[52], or including base composition information using, e.g., CONCOCT[53]. Reference genome–based methods, e.g., MetaPhlAn[54]	KEGG pathways[a] [55–57] MetaCyc[58] Clusters of orthologous groups (COGs)[59] Carbohydrate-active enzyme (CAZy) families (http://www.cazy.org)[60] Gut metabolic modules (GMMs)[61]
Metabolome data	Clustering of co-abundant metabolites, e.g., using the WGCNA framework[a] [26] or any other unsupervised clustering framework Principal component analysis (PCA)[62] Non-negative matrix factorization[63] (NMF)	KEGG pathways[55–57] ConsensusPathDB[64] (collection of pathways and metabolite sets from multiple databases, including KEGG)

[a]Indicates strategies and resources used in this protocol.

To address the high-dimensional feature space of the complex dataset, we collapsed co-abundant polar metabolites and molecular lipids into metabolite clusters (Fig. 1, Steps 5–7) and metagenomic reference genes into MGSs (Fig. 1, Box 1), respectively, alongside metagenomic functional potentials defined at the module level (Fig. 1, Box 1), thereby reducing dimensionality while retaining maximal resolution. We assessed the clinical phenotype relevance for the components of each type of -omics data (Fig. 1, Steps 8–12) and performed a cross-domain association between the metabolomics and microbiome functional profiles (Fig. 1, Steps 13–15). Associated microbes were further subjected to a so-called microbial driver species analysis that combined three data dimensions (MGSs, microbial functions and clinical phenotypes) in a leave-one-out analysis to identify MGSs that contributed particularly to the observed linkage between metagenomic functional potential and clinical phenotype (Fig. 1, Steps 17–19). Below, we provide a more in-depth discussion of the analysis framework.

Repeated dimensionality reduction to reduce heterogeneous data complexity and to 'sharpen' relevance by phenotype association filtering

One of the strengths of high-throughput technologies is also a main weakness: so many features may be measured at once that comprehensive statistical tests often yield apparently significant findings merely by chance[24]. This makes stringent corrections for multiple testing paramount, reducing statistical power to detect any differences. To circumvent the challenges posed by high data dimensionality, the data can be grouped in ways that are biologically meaningful to the application at hand. Such groupings can be identified by complementary, yet methodologically different, dimensionality-reduction approaches based on data-driven clustering or by utilizing biological prior knowledge (Table 1). Circulating metabolites and microbial genes in a metagenome sample each have innate correlation structures. Metabolite concentrations are often shaped by common pathways that are regulated in concert or by common precursors, whereas microbial genes from a chromosome show co-abundance across samples. In the latter case, MGSs can be identified, as previously reported[12] and described in the Supplementary Methods. Similarly, we can group metabolite features into clusters of both annotated and unannotated metabolites, as described in detail below. In parallel, dimensionality can be reduced by using prior biological knowledge. In the current protocol, this is done by grouping microbial genes into Kyoto Encyclopedia of Genes and Genomes (KEGG) functional modules based on sequence similarity to proteins with known functional characteristics. Analogously, one can also exploit predefined metabolic pathway maps as scaffolds for integrating metabolomics data and defining metabolite clusters, given a sufficient number of annotated metabolites to reliably cover (most of) the steps in the given pathways, which was not the case in our study. In this protocol, the feature space is further reduced by removing features that are too rare for their changes to reach statistical significance. In the second step, the dataset is filtered again to consider only features exhibiting significant association with the primary outcome variable (the clinical

phenotypes, here insulin resistance), while controlling for relevant confounding factors. This approach is designed to increase the effective power of a cross-omics association analysis.

Clustering of co-abundant metabolites to reduce the metabolomics feature space

Clustering of co-abundant metabolites is done here using the weighted gene co-expression network analysis (WGCNA) algorithm, originally developed for gene expression analysis[25–27] and recently reviewed in the context of proteomics and metabolomics data analysis[28]. Importantly, effective WGCNA performance depends on parameter settings for cluster reconstruction, which will vary to some extent between datasets. Establishing these for a dataset requires some parameter exploration and inspection of the resulting clustering. The full procedure for this lies beyond the scope of the present protocol, but readers are referred to the main WGCNA documentation[26]. We recommend using a signed network, effectively setting the link between negatively correlated metabolites (close) to zero; by contrast, using an unsigned network will group metabolites with high absolute correlation. It is further recommended to use biweighted mid-correlations as a robust alternative to Pearson correlations for calculating the similarity between any two metabolites. To preserve the continuous nature of the co-abundance patterns, so-called soft thresholding is often used to convert the similarity matrix to an abundance matrix; this requires choosing a value for the power parameter β. One common strategy is to choose the smallest value of β that satisfies the scale-free topology criteria[27]. Thus, for optimal performance, Steps 5 and 7 of the protocol should be performed for different values and the resulting curves inspected; thereafter, parameters should be specified as appropriate. The parameters used in our example correspond to optimal choices for that dataset and analysis[10]. One attractive feature of WGCNA is the use of a 'bin cluster' for unassigned metabolites (here named 'M_remaining' and 'L_remaining' for the polar metabolites and molecular lipids, respectively), rather than forcing all metabolites into a specified cluster, as is the case with many other clustering algorithms. Such unassigned metabolites can be analyzed as individual variables separately from the clusters at a later stage.

Choice of a representative value of composite features and performing of correlation analysis

The definition of features composed of multiple individual variables requires some consideration when selecting a single representative value for downstream analysis such as clinical phenotype and cross-omics associations; these definitions depend on the dimensionality-reduction approach employed. For co-abundance clusters defined by WGCNA, the default is to use the first principal component of the abundance matrix, which is basically a weighted average. By contrast, MGSs are defined as clusters of genes with extremely high correlation (Pearson's correlation coefficient >0.9), for which the median gene abundance of each MGS is considered an appropriate representative value[12]. For knowledge-based approaches, such as pathway or module grouping of microbial genes, it is similarly possible to use a median or mean abundance of genes in a given group, such as that implemented in the HUMAnN software[29], or, as implemented here, to employ a functional module enrichment analysis with a Mann–Whitney *U* (MWU) test of the correlation coefficients from a gene-level analysis.

Owing to the skewed distribution of metagenomic count data, we recommend the use of non-parametric statistical tests, such as Spearman's correlation, for downstream analysis. Similarly, we have chosen to use a partial Spearman correlation test, as implemented in the ppcor R package[30], when adjusting for the potentially confounding effects of a third variable (here, body mass index (BMI)). Although in this protocol we adjust for only a single variable, the analysis could be modified to adjust for multiple relevant variables, as seen in the paper by S. Kim[30], and could even be extended to perform a stepwise conditional analysis of -omics variables to identify independent signals.

Driver-species analysis to assess contributions to microbial guilds

Shotgun metagenomics as performed in the workflow outlined here can assess the functional potential of a microbiome directly from the gene sequences. Similar functional potential is often found across multiple species. In such cases, it is appropriate to group species with overlapping functional potential into so-called microbial guilds, that is, microbiome subsets varying in taxonomic composition and containing organisms that are not always closely related, but which together perform all or part of the same function. Direct quantification of functional potential from shotgun metagenomics identifies the presence and involvement of such microbial guilds. It then becomes relevant to ask which bacteria contribute most to the role that defines the guild.

Box 2 | Principle behind the driver-species analysis

The first step in the driver-species analysis is to establish a dataset-dependent baseline signal/extent of expected association (Fig. 2a,b). Previously, the gene abundance profiles for each sample were annotated to KOs (Supplementary Methods). For each KO, the abundance in each sample is computed by summarizing the abundances of all genes mapping to that KO. Each of the KOs in a given KEGG module is then correlated to the phenotypic variable of interest (here, HOMA-IR as a response variable in our analysis), resulting in a Spearman correlation coefficient (SCC) for each KO ($SCC_{KO\ with\ all\ genes}$) (Fig. 2a). To arrive at a single value for a given KEGG module–phenotype association, we finally collapse the signals over all KOs (that are part of the KEGG module) by taking the median $SCC_{KO\ with\ all\ genes}$ (Fig. 2b).

To evaluate the effect of a given MGS on a specific association, all genes forming that MGS are removed from the gene abundance table (across all samples) (Fig. 2c). The KO distributions are then recomputed and correlated to the response phenotype, and then a median $SCC_{KO\ without\ MGSi\ genes}$ is computed. The effect of the given MGS (MGS influence) is finally defined as the difference between the baseline SCC, which utilizes all genes in the sample (SCC_{module}), and the altered SCC, which utilizes all genes except those belonging to the given MGS ($SCC_{module\ without\ MGSi\ genes}$) (Fig. 2d). These steps are repeated, iterating over all MGSs containing at least one KO in the given KEGG module. More precisely, a large SCC reduction upon removal of an MGS's genes would indicate that this MGS plays a large role in the induction of a given microbial functional-phenotype association.

The microbiome functional correlations could potentially be driven either by one or a few prominent species or by many species that each contribute to the functional potential to different degrees. To identify the MGSs (or taxa identified under any other framework of resolution) that contributed the most to the association between a given functional KEGG module (hereafter termed 'KEGG module', i.e., a microbial guild) and a human phenotype, we developed an approach based on a variation of the leave-one-out principle, effectively evaluating the importance of each taxon by the difference in association strength that results from removing it from the analysis. Note that this driver-species analysis is different from a classic leave-one-out cross-validation analysis, in which the aim is to evaluate model performance rather than feature importance. The idea is simply to repeat the phenotype–KEGG module association analyses, removing all genes for a given MGS, one at a time, and see how much the association signal drops for each species that is removed, under the premise that the more the signal drops, the more important that species was for the originally observed association (Box 2 and Fig. 2). Although we use this approach for KEGG module–phenotype associations here, the overall approach for estimating the effect of a different taxon on a given association is easily generalizable to other functional definitions and/or other domains, e.g., microbial functional potential versus host metabolome associations.

Alternative methods

The computational workflow described here consists of multiple analytical steps developed during previous work[10], each of which could be implemented in other ways. We envision this protocol as providing a framework for scientists wishing to undertake an integrative analysis of a multi-omics dataset for which parts of the protocol can be adapted to other situations as needed, with some tools modularly switched out for other options, such as alternative approaches for dimensionality reduction listed in Table 1.

In the approach taken here, we use a stepwise approach to identify important features for each data type and subsequently evaluate the interrelationships. Which approach is most appropriate for a given multi-omics data-integration analysis depends largely on the intended outcome. For biomarker discoveries, one could use multivariate models to identify relevant features for an outcome of interest, such as those implemented in DIABLO as part of the versatile mixOmics package[31], which incorporates several multivariate projection-based methods, or by using any other supervised machine learning methodology. Alternatively, for identifying subgroups of individuals, one approach is a similarity network fusion[32], in which samples are clustered by shared data structure across multiple data types. Previous applications of software and statistical frameworks to microbiome data integrating a priori information include canonical correlations analyses with structure-constrained penalty functions[33], which makes use of phylogenetic similarity between amplicons for dimensionality reduction; and MIMOSA (model-based integration of metabolite observations and species abundances), which integrates genetic information with measured biomarker data to infer metabolic properties of bacteria as an extension of previously applied predictive relative metabolic turnover functions[34,35]. Thus, our approach is only one of many possible strategies that can be used for analysis of a multi-omics dataset. For further overview of strategies for integrating metabolomics with microbiome data, we refer to the recent review by Chong and Xia[35].

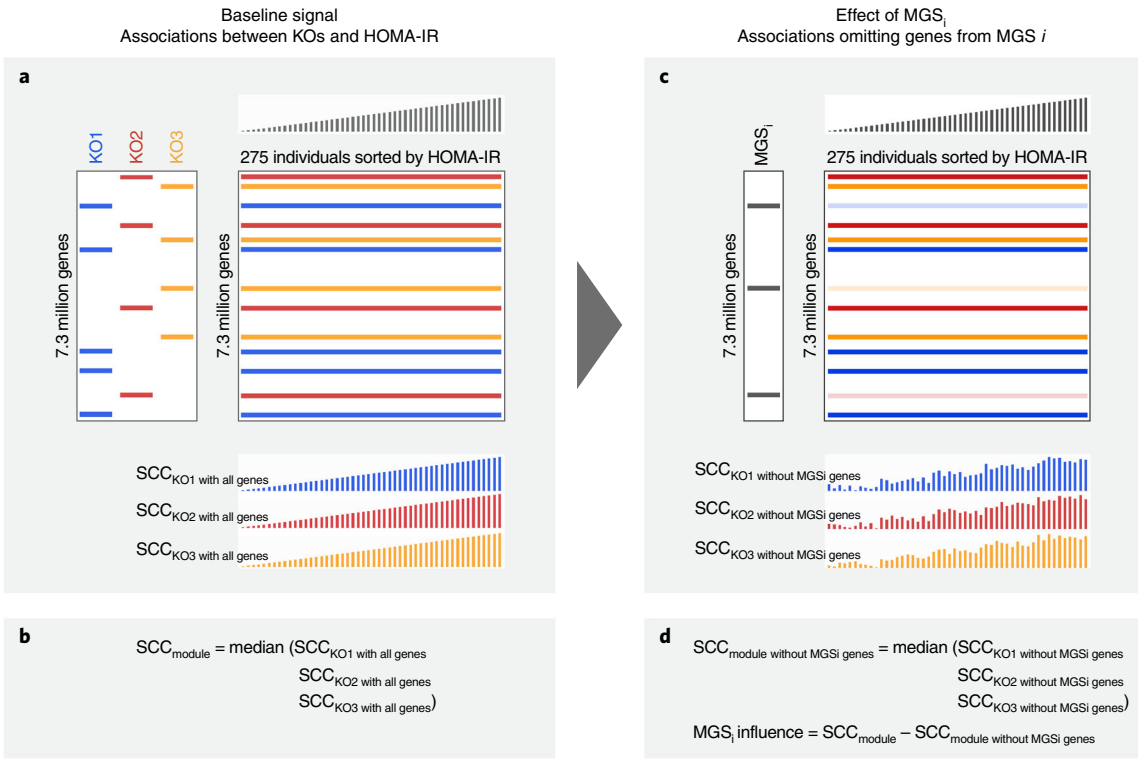

Fig. 2 | Schematic illustration of the leave-one-MGS-out approach in the driver-species analysis. For simplicity, the KEGG module here consists of three KOs (KO1, KO2 and KO3) and only one KEGG-module-HOMA-IR association is shown. **a**, The abundance of each of the three KOs is correlated to HOMA-IR, using Spearman correlation ($SCC_{KO\ with\ all\ genes}$). **b**, The median correlation for the three KOs is used to summarize the 'KEGG module signal' (SCC_{module}). **c**, Steps **a** and **b** are repeated, after all genes that constitute a given MGS (here denoted MGS_i) are removed. **d**, The influence of MGS *i* is estimated to be the difference between the 'KEGG module signal' with (SCC_{module}) and without ($SCC_{module\ without\ MGSi\ genes}$) the genes constituting this MGS. Finally, steps **c** and **d** are repeated for each MGS containing at least one KO that is part of the given KEGG module.This approach will identify species that are driving the association between HOMA-IR and the tested KEGG module. KO, KEGG orthology groups.

An important consideration in the selection of methods for integrative analysis such as that achieved using this protocol is the particular set of challenges following from the distributions and dynamics of the source dataset. Metabolomic data are structurally very different from microbiome data, which calls for other methods of clustering. Unlike microbiome data, metabolomic data are neither sparse nor particularly erratic in their abundance. The same applies to cross-species functional grouping on the basis of, e.g., KEGG orthology groups (KOs). Dominance by a few members reflects different properties and results in a less sparse dataset, as compared with microbiome datasets. In the exemplary analysis described here, correlations between phenotype and functional profiles, or between functional and metabolite profiles, are thus less sensitive to dominance by a small number of factors. Direct correlations between microbiome taxa and phenotypes are prone to false-positive associations. In particular, low-abundant taxa are prone to such errors. Such false discoveries may result from measurement error and biases. For this reason, our procedure for the identification of driver species was designed not to depend on correlations between single MGSs and phenotypes. For completeness, we include the computation of direct correlations between bacterial taxa and phenotypes in the protocol. Therefore, we suggest considering methods for detection and management of compositionality bias, such as SparCC[36,37] or the approaches reviewed in Gloor et al.[38]

Advantages and limitations

Metagenomics is an evolving field in which large and complex datasets are generated and downstream analysis remains challenging. With recent technological advances, the integration of metagenomics data with other complex -omics data will become increasingly common. The present protocol provides a step-by-step approach for the analysis of such large multi-omics datasets, thus

contributing to the development of standardized approaches in the field that will facilitate the comparison of such studies. A major advantage is that the analysis can be adapted to include different types of high-dimensional -omics data, can reflect domain-specific foreknowledge well and makes interpretation and design of validation tests straightforward. However, there are a number of biological and technical issues with current microbiome protocols that bioinformaticians should be aware of when interpreting results from this analytical approach; these are described below. In addition, the present protocol profiles only the cellular component of the microbiome (although we use the term 'microbiome', strictly speaking, we are ignoring the virome).

Correlations between gut microbiome and host serum metabolome data can be obscured by physiological processes

Spatial separation of the sampled body sites must be considered when integrating data, such as when joining gut microbiome data to serum metabolomics measures. Correlations between data from different body sites can be affected by physiological action; for example, the relationship between a gut microbe and serum metabolite could be obscured by the rate of metabolite uptake and/or breakdown by the liver. Another example might be when the correlation between serum and stool metabolites is affected by individual differences in intestinal epithelial filtering of a given metabolite. The connection between the gut microbiome and the blood circulatory system relies on nutrient translocation from the former to the latter. Not much is yet known about the exact permeability of specific metabolites through the gut wall, or whether changes in such permeability play a part in some diseases, as in the so-called 'leaky gut' model[39–41]. In summary, altered production of a given small molecule within the gut microbiota implies an altered presence of this molecule in circulation; however, this is not guaranteed, and caution must be taken in interpreting association as causation. How this issue is best addressed is outside the scope of the present protocol and calls for additional experiments, especially in the form of interventions including portal vein metabolomics[42].

What you measured might not be what you had hoped to measure

Correlations with microbiome data can be obscured by the fact that microbiome activities may take place in other parts of the gastrointestinal tract than the site that is sampled (typically stool). Assessments of gut microbiome functional activity through stool sample sequencing have inherent limitations in interpretation of more proximal sections. For example, the gut microbial composition is not uniform along the gastrointestinal tract; colonic microbial composition differs immensely from that of the small intestine[43]. Although samples (that we can measure noninvasively) from different study participants reflect this bias equally and thus remain comparable, care must be taken when inferring states in the upper gut from such data. Second, the functional redundancy of the microbiome can obscure associations more generally, something that can be addressed by systems biological analysis looking for biases in functional enrichment across species, or through insights from community-scale metabolic modeling. Third, metagenomic analyses measure the potential for expression of genes (e.g., relative abundance of genes) but not their actual expression (which can be measured directly from RNA and/or protein readouts). Therefore, metagenomics gives only indirect evidence for functional activity of the gut microbiome, making the importance of gene function especially difficult to access. Fourth, for many applications, using species-level (or broader taxonomical) resolution may blur the signal if, for example, only a subset of the subspecies populations is actually driving the investigated association. Future microbiome studies should utilize the potential of operating at increased taxonomical resolution; i.e., using subspecies population resolution or structure information. Fifth, although substantial effort has been made to standardize protocols for sample processing for gut metagenome analysis to minimize technical variation[44], for other types of microbiome samples, additional sources of error may arise, potentially necessitating experimental controls of various types to ensure high fidelity of data generation. Finally, standard approaches for sequencing-based microbiome profiling yield relative, not quantitative data. To circumvent this, one can adjust the sequencing data by the actual bacterial cell counts of the sample, as recently detailed by Vandeputte et al.[45].

Timescale of data and limitations of case–control versus longitudinal data

Correlations between high-throughput and clinical data can be obscured by the fact that our measurements are a snapshot, whereas the clinical variables may be long-term effects, and that other factors may play a large role (e.g., factors reflecting likelihood of recruitment as a cohort

participant). The microbiome composition reflects temporary factors such as diet and the transient colonization of microbial species. In many long-term progressive diseases such as diabetes or atherosclerosis, pathology builds up asymptomatically over decades before diagnosis is possible. Subtle effects probably accumulate over time to drive disease progression; these may or may not be the same effects that are visible at the clinical stage. Furthermore, comorbidities are hard to differentiate from the underlying cause (e.g., cardiovascular health in diabetes or in relation to obesity), confounder effects brought on by their co-occurrence, or effects of their treatment regimens post diagnosis. These factors call for caution in both exploration of confounding factors and interpretation of results.

Use of relative abundance

The present approach works with data, such as metagenome measurements, that are compositional, comprising relative abundance values. Such data are in principle vulnerable to, e.g., dilution effects and the effect of detection flaws interacting with diversity, as well as sampling depth differences between samples. In the analysis we previously reported, we chose to minimize bias resulting from such effects by downsampling all data matrices so that the samples have the same depth (other normalization approaches are reviewed in Weiss et al.[46]). This eliminates bias while also losing power, so an alternative approach may be to replace tests such as the Spearman correlation assessments with corresponding tests that explicitly model proportional data, such as the method suggested by Gloor et al.[47]. All other steps of the Procedure will remain the same. As discussed above, metabolomic data may be less susceptible to these issues.

Level of expertise required

Comprehending the step-by-step part of this protocol (to an extent where you can apply it to your own data) requires basic understanding of R, including how to install R packages. Good general introductions to R include the 'R Programming for Data Science' (https://bookdown.org/rdpeng/rprogdatascience/) by R. Peng and the Coursera course 'R Programming' offered by Johns Hopkins University (https://www.coursera.org/learn/r-programming).

Materials

Equipment

Software

- R (https://cran.r-project.org); we recommend the use of RStudio (https://www.rstudio.com/products/rstudio/download/). The analysis was tested using R 3.3.3. If packages are available under another version, it should run, but specifics of the implementation of each package may change the results slightly.
- Bioconductor (see https://www.bioconductor.org/install/ for install instructions and access) ▲ **CRITICAL** Ensure that the following packages (including their indirect dependencies in the form of other packages needed for their compilation and operation, some of which should be installed via Bioconductor) are installed and loaded correctly (in every case available via the built-in R and Bioconductor package managers).
- xlsx, for saving to a spreadsheet
- data.table, for fast read of large files into R
- WGCNA, clustering software. Our previously reported work[10] was done using v1.34.
- flashClust, clustering software
- ppcor, partial Spearman correlations, for confounder analysis. Our previously reported work[10] was done using v1.0.
- gplots, for plotting
- cowplot, for plotting; to arrange several plots on the same page
- ggplot2, for plotting
- plyr, for data transformations

Input data files: phenotype

- phenotypes.tab: file with clinical phenotypes (columns) per individual (rows). It is used to test for associations, or for confounder analysis. Individuals are labeled 'idv', followed by a number, e.g., 'idv001'. ▲ **CRITICAL** All demonstration files—both input data files and annotation files—are tab-delimited text files, but other formats would also work after modifying the respective file-import commands in the R script.

Input data files: metabolome

- metabolomic.tab/lipidomic.tab: input data matrix for abundance of 325 polar metabolites or 876 molecular lipids per individual. Note that no additional normalization is done in this script, so data are assumed to be comparable in these regards. Such different data types are eventually merged into a single set of metabolite cluster abundances. Individual metabolite/lipids are named M or L (for specifying a polar metabolite or molecular lipid, respectively), followed by a number and lastly the annotation or 'unknown' (in the case of unannotated metabolites/lipids), e.g., 'M_20_Valine'.

Input data files: microbiome

- MGS_abundance.tab: file with the abundance (e.g., median gene abundance) of MGSs (columns) per individual (rows). These are assumed to have been rarefied to comparable depth or otherwise normalized. For historical reasons, the MGSs are labeled 'T2DCAG' followed by a number, e.g., 'T2DCAG00001'.
- KO_abundance.tab: file with the abundance of each KO (columns) per individual (rows). The data are assumed to be rarefied to comparable depth or otherwise normalized. For this, the software tool rtk[48] can be used.
- gene_abundance_sub.tab: file with the abundance of each catalog gene (subset version) in each individual assumed to be rarefied to comparable depth or otherwise normalized.

Annotation files: metabolome

- cluster_mapping_file.tab: input file with annotation for metabolite clusters, as available from curation of data in the specific dataset. The WGCNA clustering algorithm, by default, names the generated clusters with color codes. This mapping file simply facilitates renaming to more meaningful cluster descriptions. Here, the serum polar metabolite and serum molecular lipid clusters are labeled M01–M35 and L01–L39, respectively, and collectively termed 'metabolite clusters'.

Annotation files: microbiome

- MGS_taxonomy.tab: file with taxonomic annotation of the MGSs (rows) used in the analysis. Each row contains the following information: species_taxonomy, species_pct, genus_taxonomy, genus_pct, family_taxonomy, family_pct, order_taxonomy, order_pct, phylum_taxonomy, phylum_pct, where x_pct is the percentage of the MGS genes that can be annotated (by sequence similarity) to the taxonomy of the MGS. If no taxonomy can be assigned to the MGS, the value will be NA (not available).
- KEGG_modules.tab: file containing definition of KEGG gene functional modules (rows) specifying which KO gene groups constitute each KEGG module. The first two columns contain KEGG module entry (number) and name, and the third column lists all KOs, separated by semicolons. Any other functional annotation used analogously could be swapped in instead of the name. This file can be obtained by downloading KEGG modules from http://www.genome.jp/kegg-bin/get_htext?ko00002.keg; download htext and then run the provided script 'parse_kegg.pl' (after changing the input filename).
- KO_to_MGS.tab: file listing for each KO (rows) the MGSs (space-separated) of which it is a member. This is based on the KO annotation of the gene catalog (gene_to_KO.tab) and information regarding which gene from the gene catalog is within each MGS, as given in the list in the MGS_to_gene.tab file.
- gene_to_KO.tab: file containing the KO annotation (if any) per gene (rows) in the gene catalog (constituting 7,328,469 genes). Genes are labeled 'RefCat620' followed by a number (1, …, 7,328,469), e.g., 'RefCat620.1'. It is used to create the KO_to_MGS.tab file. Note, that for the purposes of this protocol and to considerably reduce the size of input data, only the subset of catalog genes with KO annotation ($n = 2{,}205{,}769$) are provided in files with gene abundance or annotation (gene_to_KO.tab, MGS_to_gene.tab and gene_abundance_sub.tab), as only those genes are used in the driver-species analysis.
- MGS_to_gene.tab: list of genes binned into a given MGS (rows). It is used to create the KO_to_MGS.tab file. Rather than gene names, the file contains the index value of the gene, i.e., the position of the gene in the gene catalog (subset version, thus 1, …, 2,205,769).

Procedure

▲ **CRITICAL** For full details of how to execute each step, see the enclosed source R code at the accompanying Git repository (https://bitbucket.org/hellekp/clinical-micro-meta-integration); the code can be performed step by step, e.g., in the RStudio interactive environment. For reference, we also provide examples of all output files generated by the script, including figures and tables (also at the accompanying Git repository).

All commands specified below are intended to be executed in R, within a single persistent environment, whether through bundling in a unified script or through consecutive commands in console or RStudio.

Stage I: starting an R session with all required packages and input files ● Timing ~32 min

1 *Set the working directory (10 min).* The code will look for input files and deposit output files in subdirectories relative to a main directory.

On your computer, create a directory for the analysis, to remain generic (and without assumptions about user operating system), we here call it 'top/'. Then create subdirectories to obtain the following directory hierarchy:

```
top/
  r-code/
  data/
  results/
```

Copy all R code from the Git repository to your 'top/r-code/' subdirectory (i.e., all R files). Then open the main R script, called 'protocol_main.R', in, e.g., RStudio.

Finally, modify the following command to set the working directory to your top/ directory:

```
setwd("~/top")
```

2 *Ensure availability of software packages and satisfaction of dependencies (10 min).* Use the following command to load the required libraries specified in the Software list above:

```
source("r-code/step2_load.libraries.R")
```

? TROUBLESHOOTING

3 *Import input files (10 min).* When using the demonstration data, the following input data are assumed to exist within the 'top/data/' subdirectory (the content and structure of the files are described in the 'Materials' section):

phenotypes.tab
metabolomic.tab/lipidomic.tab
cluster_mapping_file.tab
MGS_abundance.tab
MGS_taxonomy.tab
KEGG_modules.tab
KO_abundance.tab
KO_to_MGS.tab
gene_to_KO.tab
MGS_to_gene.tab
gene_abundance_sub.tab

These input files are available at the accompanying Git repository of this protocol article, as a demonstration example for the Procedure (https://bitbucket.org/hellekp/clinical-micro-meta-integration).

Download the 'example_input.zip file', unzip it and place all files in your 'top/data/' subdirectory.

Then use the following command to import the demonstration data files into R:

```
source("r-code/step3_import.input.files.R")
```

4 *Pre-processing/cleanup of loaded data for sparsity and domain limitation (2 min).* Use the following command to restrict the input data to account for method limitations:

```
source("r-code/step4_preprocessing.data.for.sparsity.R")
```

Stage II: co-abundant clustering of metabolome data ● Timing ~6 min

▲ **CRITICAL** In the following steps (Steps 5 and 7), WGCNA clustering is performed separately on lipidomic and metabolomic measurements to detect clusters of densely connected metabolites/lipids. The metabolite/lipid profiles constituting a given cluster are summarized by the first principal component of the metabolite/lipid abundance matrix ('Module Eigen-metabolite/lipid'), which is

basically a weighted average abundance profile. Before this, optimal parameters for WGCNA should be established for the dataset being analyzed.

5 *Identify clusters of polar metabolites (2 min).* Use the following command to generate WGCNA clusters for polar metabolites:

```
source("r-code/step5_identify.WGCNA.clusters.for.metabolites.R")
```

? TROUBLESHOOTING

6 *Link individual polar metabolites to the phenotype of interest (2 min).* The dimensionality reduction approach (detailed in Stage III) hinges on identifying features (e.g., metabolites and lipids clusters) that are associated with a host phenotype of interest. In the example work we describe here, this was insulin resistance as assessed by the homeostatic model assessment for insulin resistance (HOMA-IR) measurement.

In this step, for reference, the individual polar metabolites are linked with HOMA-IR, both directly and under adjustment for a potential confounder variable. In this case, such de-confounding was done for BMI. In the case of a binary phenotype variable, one can, for example, substitute the Spearman correlation test with a MWU test.

Use the following command to associate the individual polar metabolites with HOMA-IR:

```
source("r-code/step6_associate.metabolites.with.phenotype.R")
```

7 *Identify clusters of molecular lipids and link individual molecular lipids to the phenotype of interest (2 min).* To repeat Steps 5 and 6 for lipidomic data, use the following command to generate WGCNA clusters for molecular lipids:

```
source("r-code/step7a_identify.WGCNA.clusters.for.lipids.R")
```

Use the following command to associate the individual molecular lipids with HOMA-IR:

```
source("r-code/step7b_associate.lipids.with.phenotype.R")
```

The resulting metabolite and lipid clusters are thereafter combined into a single dataset, collectively termed 'metabolite clusters', for downstream analyses.

? TROUBLESHOOTING

Stage III: phenotype filtering ● Timing ~10 min

▲ CRITICAL This analysis stage generates associations between the dimensionality-reduced -omics data and a clinically interesting phenotype. In our previous work, this was insulin resistance (HOMA-IR measurement), but any phenotype is possible, as is checking against other -omics domains or overall -omics measurements such as gut diversity or enterotype. Furthermore, this analysis can be conducted while controlling for confounders (such as BMI in the analysis we previously reported), by performing tests with partial correlations, or can be extended to binary phenotype variables by substituting tests of Spearman correlation with, e.g., MWU tests.

8 *Link metabolite clusters to the phenotype of interest (2 min).* To repeat Step 6 for metabolite clusters, use the following command to associate metabolite clusters with HOMA-IR:

```
source("r-code/step8_associate.metabolite.clusters.with.pheno-
type.R")
```

9 *Link MGS metagenomic entities to the phenotype of interest (2 min).* To repeat Step 6 for metagenomic taxonomic data, use the following command to associate MGSs with HOMA-IR:

```
source("r-code/step9_associate.MGSs.with.phenotype.R")
```

10 *Link KEGG functions to the phenotype of interest (2 min).* This step is analogous in goal to Step 6 but is used for metagenomics functional data. Here, we use KEGG modules, but any other groupings of genes into functional modules could similarly be used (see examples in Table 1). Use the following command to associate KEGG modules with HOMA-IR:

```
source("r-code/step10_associate.KEGG.modules.with.phenotype.R")
```

▲ CRITICAL STEP It is important to note that each KEGG module is constituted of multiple KOs. Thus, to generate results on the level of modules, we test if correlations between the phenotype and the abundances of KOs in the module are significantly higher or lower (MWU test) for the module member KOs than for all other KOs, thus also considering module completeness beyond the single-gene level. In the case of a binary phenotype variable, the KOs can be ranked based on, e.g., Wald statistics (instead of Spearman correlation coefficients (SCCs)) for testing differentially abundant KOs with a negative binomial test with the DESeq2 R package[49] using non-rarefied gene counts.

11 *Save phenotype associations (2 min).* Use the following command to save the (BMI-corrected) HOMA-IR association of metabolite clusters, MGSs and KEGG modules calculated in Steps 8–10:

```
source("r-code/step11_save.phenotype.associations.R")
```

12 *Select features with significant differences (2 min).* Combine and integrate those functional, taxonomic and metabolomics features that reliably correspond to the host phenotype of interest, using the following command to select the subset of features significantly associated with HOMA-IR:

```
source("r-code/step12_select.significant.features.R")
```

? TROUBLESHOOTING

Stage IV: cross-domain association analyses ● Timing ~11 min

▲ CRITICAL This stage tests the association of the set of metabolite clusters associated with the phenotype of interest with the set of functional metagenome features likewise so associated (identified in Step 12). Here, once again, the complex nature of KEGG modules (consisting of multiple KOs) must be taken into account.

13 *Correlate metabolite clusters to functional metagenomic potentials (KOs) (5 min).* Use the following command to calculate correlations between each metabolite cluster and each KO:

```
source("r-code/step13_associate.metabolite.clusters.with.KOs.R")
```

14 *Associate metabolite clusters with functional metagenomic potentials (KEGG modules) (2 min).* To use the KO-level data generated in Step 13 to calculate module-level associations between a KEGG module and a metabolite cluster, execute the following command:

```
source("r-code/step14_associate.metabolite.clusters.with.KEGG.
modules.R")
```

15 *Plot metabolome–microbiome functional analysis results (2 min).* Use the following command to create a visual representation of the generated results:

```
source("r-code/step15_plot.metabolome.microbiome.functional.ana-
lysis.R")
```

This creates a plot similar to that shown in Fig. 3.

? TROUBLESHOOTING

16 (Optional) *Export metabolome–microbiome associations for network analysis (2 min).* Further exploration of such high-dimensional association data can be performed using software for network analysis, e.g., Cytoscape (http://www.cytoscape.org) or igraph (http://www.igraph.org). Here, we provide code for exporting an edge file with pairwise association scores and false discovery rate (FDR) values between metabolite clusters and KEGG modules, as well as a corresponding node attribute file (both files in .txt format). Use the following command to export an edge file and a corresponding node attribute file:

```
source("r-code/step16_export.edge.node.files.R")
```

▲ CRITICAL STEP Several tutorials for importing, visualizing and analyzing networks in Cytoscape can be found here: https://github.com/cytoscape/cytoscape-tutorials/wiki.

Stage V: driver-species analysis ● Timing ~64 min

▲ CRITICAL This stage of the analysis allows users to test which bacterial taxa (in the sense of MGS as defined from the metagenomic datasets themselves) are driving the functional effects seen, enabling assessment of the extent to which different taxa explain a functional potential association with a phenotype of interest, such as insulin resistance in the analysis we previously reported[10]. It is done by testing for each functional feature (here: KEGG modules) to what extent leaving out each MGS and the genes it contains causes a change in the association between those modules and the target phenotype (Box 2).

17 *Leave-one-MGS-out analysis (~1 h).* First, specify which KEGG modules and taxa to include in the leave-one-MGS-out analysis. Then compute the contribution per taxon to each KEGG module for all KEGG module–taxa combinations. Use the following command to perform the leave-one-MGS-out analysis:

```
source("r-code/step17_leave.one.MGS.out.analysis.R")
```

? TROUBLESHOOTING

18 *Extraction of the top driver species from leave-one-MGS-out analysis for each microbiome functional module (2 min).* Use the following command to extract the top five driver taxa for each KEGG module for interpretation:

```
source("r-code/step18_extract.top.driver.species.R")
```

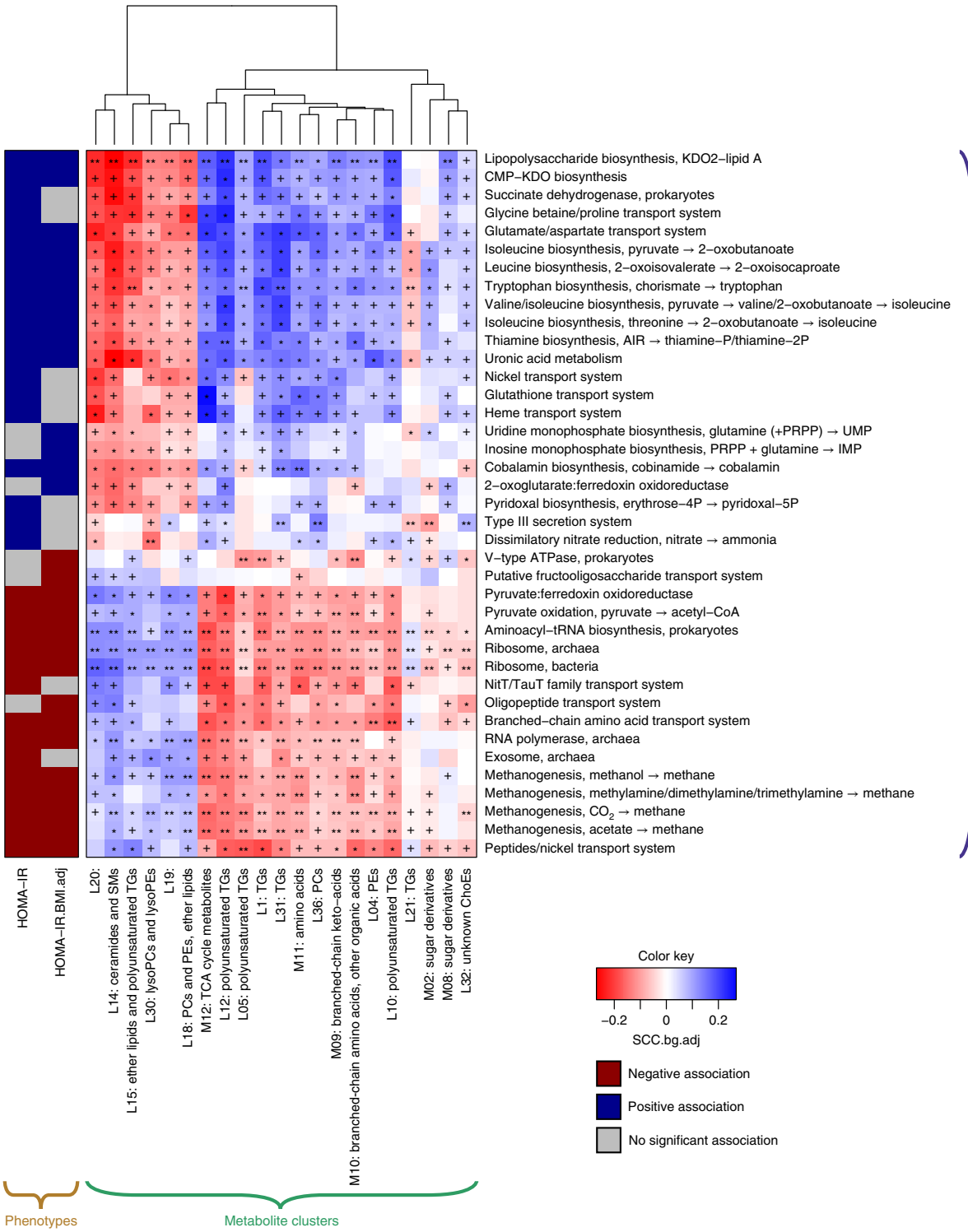

Fig. 3 | Sample plot produced by the Procedure in Step 15. Association map of the phenome, the gut microbiome and the fasting serum metabolome. The left panel shows significant associations (Mann–Whitney *U*-test FDR < 0.1) between KEGG modules and the indicated phenotypes; coloring indicates the direction of association (red: negative; blue: positive; gray: not significant). The right panel shows associations between the same KEGG modules and serum metabolite clusters. Coloring represents the median Spearman correlation coefficient between metabolite clusters and the indicated KEGG modules, corrected for background distribution (SCCbg.adj.), where MWU FDRs are denoted: $^{+}$FDR < 0.1; *FDR < 0.01; **FDR < 0.001. Module names are shown as designated by KEGG (https://www.genome.jp/kegg/module.html). Interpreting the figure, it is apparent that both the KEGG modules and metabolite clusters segregate into two overall groups: metabolically favorable and unfavorable (here assessed by insulin resistance), giving rise to extensive—either positive or negative—cross-omics associations; a pattern we often observe in such inter-domain association analyses. Essentially, features that participate in the same dimension of variability in an aspect, e.g., health/sickness, will also intercorrelate in the expected manner, if associations are robust.

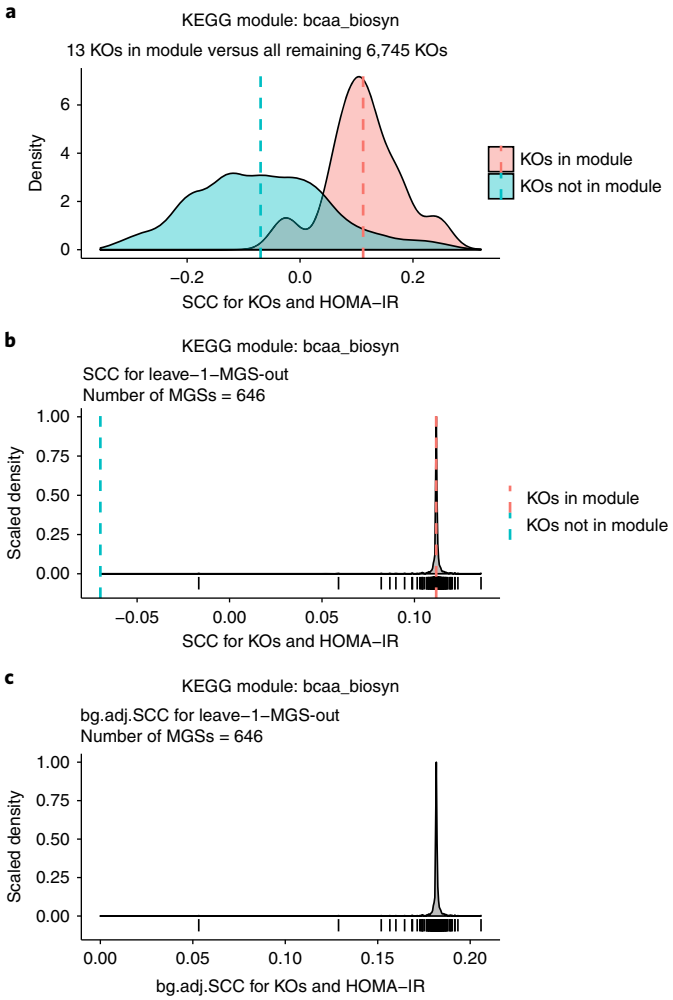

Fig. 4 | Sample plot produced by the Procedure in Step 19 for the leave-one-MGS-out analysis, here shown for the combined BCAA-biosynthesis module (KEGG modules M00019, M00570, M00535 and M00432; together constituting 13 KOs). a, Distribution of SCC for KOs in the combined BCAA-biosynthesis modules (red) and all other KOs (blue). **b**, Median SCC (i.e., 'MGS influence' in Fig. 2) between HOMA-IR and the combined BCAA-biosynthesis module when a given MGS has been excluded from the analysis, shown as density (top) and rug (bottom) plots. **c**, Background-adjusted median SCC (i.e., 'MGS influence' in Fig. 2) between HOMA-IR and the combined BCAA-biosynthesis module when a given MGS has been excluded from the analysis, shown as density (top) and rug (bottom) plots. The median SCC for KOs within a module (red) and all other remaining KOs (blue) are indicated in **a**,**b** by dotted lines. In the rug plots (in **b** and **c**), each vertical line corresponds to an MGS (i.e., the (background-adjusted) median SCC between the 13 KOs and HOMA-IR when the respective MGS is left out). The majority of MGSs show no or only minor effects, as seen by the many overlapping lines around the SCC based on all MGSs (indicated by the dotted red line); this is even clearer in the corresponding density plots. A few MGSs show noticeable effects in driving the HOMA-IR-KEGG module association; the closer to 0 in **c**, the larger effect. **c** adapted with permission from Pedersen et al.[10], Springer Nature.

19 *Plotting leave-one-MGS-out results (2 min).* Use the following command to plot the top driver taxa for the phenotype-associated gene functions:

```
source("r-code/step19_plot.leave.one.MGS.out.results.R")
```

This creates a plot similar to those shown in Fig. 4, with the following subplots:
Distribution/density plots of SCC for KOs in a KEGG module versus all other KOs
Distribution of SCC when leave-one-MGS-out is applied
Distribution of SCC.bg.adj when leave-one-MGS-out (median SCC for KOs within the respective module minus the median SCC for all other KOs not in the module) is applied (i.e., what is shown in Fig. 3c,d in Pedersen et al.[10]).

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting table

Step	Problem	Possible reason	Solution
2	Problems installing packages, or not able to reproduce sample results from sample input files	Missing dependency packages; using different versions of packages or R	Make sure that all dependency packages are installed, some of which need to be installed via Bioconductor (see source code). Change to the same versions of packages and R that were used to test the code, as listed at the end of the code
5, 7	No clusters observed in -omics data	-Omics data coverage and/or quality is low	Investigate the quality of the -omics data. If data coverage is low, i.e., few variables are measured, it may be unnecessary to perform the clustering steps. Consult the FAQ on the main WGCNA webpage (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/) for further guidelines on parameter settings, pre-processing of data and troubleshooting
	Many small clusters, where members have small kIN (intramodular connectivity) and kME (eigengene-based connectivity) values (for further explanation, see Langfelder and Horvath[26])	This can be a sign of a lack of structure within the data. It can occur if the biological space is sampled very sparsely or tested with randomly permuted data	Investigate the quality of the -omics data. If data coverage is low, i.e., few variables are measured, it may be unnecessary to perform the clustering steps
12	No significant features are obtained	There may not be a relationship between the phenotype and the -omics data tested; phenotype is inappropriately defined; data quality is low; sample size is too small	Make sure that ordering of individuals is the same in the two domains when testing for associations. Investigate data quality and pre-processing steps. Reconsider the outcome phenotype; if using a binary outcome, then investigating a related continuous variable might increase statistical power. Increase the sample size if possible
15	Error with 'issig = rowSums (tmp, na.rm = T)' or 'issig = colSums (tmp, na.rm=T)'	No significant metabolite clusters and/or KEGG functional modules; consequently, the cross-domain heatmap cannot be defined and plotted	See solutions for Step 12
17	Out-of-memory issues	The analysis of the given dataset is too computationally expensive	Rerun the analysis on a subset of the MGSs/modules (see source code)
	Strange or unexpected results	Batch effects or other confounders in data	Make sure to use appropriate normalization methods for your data and check for effects of technical covariates, such as batch/run status, which should then be accounted for in downstream analysis. In the case of such effects, strategies forward include (i) using residuals after regressing out the confounding factors and (ii) replacing univariate tests with variants that can handle multiple covariates (e.g., ppcor or nested model comparisons). For an example of practical application of the latter approach, see Forslund et al.[65]

Timing

The times listed below are computation time estimates for running the source code on the provided input files, using a MacBook Pro (2.9-GHz quad-core seventh-generation Intel Core i7 processor, 16-GB 2133-MHz LPDDR3 memory), while also including time for reading the procedural steps and copying/pasting the commands into R (the computation time itself is ~1 h). The run time may change considerably with the use of other input data. In addition, one should allocate time for ensuring full comprehension of the protocol and code and assessment of optimal parameters for the given input data, which probably will add up to a few days, depending on prior expertise.

Stage I, Steps 1–4, starting an R session with all required packages and input files: ~32 min
Stage II, Steps 5–7, co-abundant clustering of metabolome data: ~6 min
Stage III, Steps 8–12, phenotype filtering: ~10 min

Stage IV, Steps 13–16, cross-domain association analyses: ~11 min
Stage V, Steps 17–19, driver-species analysis: ~64 min

Anticipated results

If successful, the provided R code will generate the following ten files, which are also, for reference, provided in the example_output directory in the accompanying Git repository (https://bitbucket.org/hellekp/clinical-micro-meta-integration):

individual_metabolites.txt (made in Step 6): overview of all serum polar metabolites mapped to their corresponding 35 co-abundance clusters, including their individual associations with HOMA-IR and HOMA-IR_{BMIadj} (HOMA-IR adjusted for BMI). The resulting 35 clusters are named by both module label (column: 'module'; i.e., a color as is the default output by the WGCNA algorithm) and our cluster annotation (column: 'cluster_name', i.e., MXX). Be aware that the 'gray' cluster ('M_remaining') is not a real or meaningful cluster but rather a collection of all the metabolites that could not be reliably assigned to any of the resulting 35 clusters and, consequently, should be excluded from downstream analysis. The columns kIN and kME specify the within-module connectivity, determined by summing connectivity with all other metabolites in the given cluster and bicor correlation between the metabolite profile and module eigenvector, respectively; both measures of intramodular hubmetabolite status. For association with HOMA-IR and HOMA-IR_{BMIadj}, the resulting (partial) SCC coefficients (column: 'XX_estimate') are reported together with both nominal and FDR-adjusted *P* values (columns: 'XX_p.value' and 'XX_p.adjust', respectively).

individual_lipids.txt (made in Step 7): Overview of all serum molecular lipids mapped to their corresponding 39 co-abundance clusters, including their individual associations with HOMA-IR and HOMA-IR_{BMIadj}. The file is similar in structure to that of *individual_metabolites.txt* explained above, except the column 'cluster_name' is now in the format of LXX.

MEs_metabolite_clusters.txt (made in Step 7): Matrix with cluster 'eigen-metabolite' values for each individual (rows) and metabolite cluster (columns). This file is the result of the dimensionality-reduction step performed using the WGCNA algorithm, effectively reducing the 325 polar metabolites to 35 entities and the 876 molecular lipids to 39 entities.

HOMA.IR_associations.xlsx (made in Step 11): Association of metabolite clusters (sheet: metlip), MGSs (sheet: MGSs) and KEGG modules (sheet: keggmodules) with HOMA-IR and HOMA-IR_{BMIadj} (adjusted for BMI). Both nominal and FDR-adjusted *P* values are shown. For metabolite clusters and MGSs, 'estimate' refers to SCC, whereas for KEGG modules, it refers to SCC.bg.adj., defined as the median SCC for KOs within the respective module minus the median SCC for all other KOs not in the module. By subsetting the file to show only significant FDR-adjusted *P* values, one can quickly get an overview of which metabolite clusters, MGSs and KEGG modules are associated with HOMA-IR and/or HOMA-$IR_{BMIadj.}$

heatmap_KEGG_vs_metabolite_clusters.pdf (made in Step 15): The plot shown in Fig. 3.

edge_file.txt (made in (optional) Step 16): Exported edge file; can be used to explore pairwise relationships in, e.g., Cytoscape. It contains the following four columns: KEGG_module, Metabolite_cluster, association estimate (SCC.bg.adj) and FDR-adjusted *P* values.

node_file.txt (made in (optional) Step 16): Exported node annotation file; can be used to explore pairwise relationships in, e.g., Cytoscape. It contains the following three columns: Node_name (corresponds to names in the edge file), Description (where existing) and Examples (example metabolites, only for metabolite clusters).

delta_SCC_per_MGS.RData (made in Step 17): output from the driver-species analysis; a list with 40 entries, one for each tested KEGG module. Each entry contains a data frame with one row for each of the tested MGSs that contain at least one KO, constituting the given KEGG module (i.e., the number of resulting MGSs varies for different KEGG modules); the MGSs are sorted by decreasing importance in driving the HOMA-IR-KEGG module association. Thus, the analogous output for another application would provide a similar ranking.

top_driver_species.txt (made in Step 18): For each KEGG module, the five most important species driving the association between the microbial module and HOMA-IR are listed. For each of these species, the following information is provided: DeltaMGS_SCC (the change in median SCC between KOs and HOMA-IR when the respective MGS is left out) and pctSCCeffect_bg.adj (the percentage change compared with the original background-adjusted median SCC). The larger the DeltaMGS_SCC and the pctSCCeffect_bg.adj, the more important the species is in driving the association between the given KEGG module and HOMA-IR.

density_plot_SCC_HOMA.IR.pdf (made in Step 19): The plots shown in Fig. 4.

Data and code availability

An archive containing all code and the demonstration data required to run the procedural part of the protocol is available in the Supplementary Data as well as at the accompanying Git repository (https://bitbucket.org/hellekp/clinical-micro-meta-integration). It includes preprocessed microbiome and metabolome data and phenotype information. For all demonstration data, pseudonymized sample names were re-randomized to generate anonymized data.

References

1. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
2. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
3. Karlsson, F. H. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
4. Forslund, K. et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
5. Sharon, G. et al. Specialized metabolites from the microbiome in health and disease. *Cell Metab.* **20**, 719–730 (2014).
6. Antharam, V. C. et al. An integrated metabolomic and microbiome analysis identified specific gut microbiota associated with fecal cholesterol and coprostanol in *Clostridium difficile* infection. *PLoS ONE* **11**, 1–23 (2016).
7. Musso, G., Gambino, R. & Cassader, M. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. *Annu. Rev. Med.* **62**, 361–380 (2011).
8. Escobar-Zepeda, A., De León, A. V. P. & Sanchez-Flores, A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Front. Genet.* **6**, 1–15 (2015).
9. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
10. Pedersen, H. K. et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
11. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
12. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
13. Methé, B. A. et al. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
14. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
15. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Appl. Environ. Microbiol.* https://doi.org/10.1128/AEM.02627-17 (2018).
16. Mallick, H. et al. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol.* **18**, 228 (2017).
17. Knight, R. et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
18. Nygren, H., Seppänen-Laakso, T., Castillo, S., Hyötyläinen, T. & Orešič, M. Liquid chromatography-mass spectrometry (LC-MS)-based lipidomics for studies of body fluids and tissues. *Methods Mol. Biol.* **708**, 247–257 (2011).
19. Considine, E. C., Thomas, G., Boulesteix, A. L., Khashan, A. S. & Kenny, L. C. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics* **14**, 7 (2018).
20. Castillo, S., Mattila, I., Miettinen, J., Orešič, M. & Hyötyläinen, T. Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. *Anal. Chem.* **83**, 3058–3067 (2011).
21. Hyötyläinen, T. & Orešič, M. Optimizing the lipidomics workflow for clinical studies—practical considerations. *Anal. Bioanal. Chem.* **407**, 4973–4993 (2015).
22. Cajka, T. & Fiehn, O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Anal. Chem.* **88**, 524–545 (2016).
23. Begley, P. et al. Development and performance of a gas chromatography-time-of-flight mass spectrometry analysis for large-scale nontargeted metabolomic studies of human serum. *Anal. Chem.* **81**, 7038–7046 (2009).
24. Tukey, J. Some thoughts on clinical trials, especially problems of multiplicity. *Science* **198**, 679–684 (1977).
25. Zhao, W. et al. Weighted gene coexpression network analysis: state of the art. *J. Biopharm. Stat.* **20**, 281–300 (2010).
26. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
27. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
28. Pei, G., Chen, L. & Zhang, W. WGCNA application to proteomic and metabolomic data analysis. *Methods Enzymol.* **585**, 135–158 (2017).

29. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
30. Kim, S. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods* **22**, 665–674 (2015).
31. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration'. *PLoS Comput. Biol.* **13**, e1005752 (2017).
32. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
33. Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. & Li, H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**, 244–258 (2013).
34. Noecker, C. et al. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* **1**, e00013–e00015 (2016).
35. Chong, J. & Xia, J. Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites* **7**, E62 (2017).
36. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, 1–11 (2012).
37. Weiss, S. et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**, 1669–1681 (2016).
38. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
39. Quigley, E. M. M. Leaky gut-concept or clinical entity? *Curr. Opin. Gastroenterol.* **32**, 74–79 (2016).
40. Kelly, J. R. et al. Breaking down the barriers: the gut microbiome, intestinal permeability and stress-related psychiatric disorders. *Front. Cell. Neurosci.* **9**, 392 (2015).
41. Mu, Q., Kirby, J., Reilly, C. M. & Luo, X. M. Leaky gut as a danger signal for autoimmune diseases. *Front. Immunol.* **8**, 1–10 (2017).
42. Meijnikman, A. S., Gerdes, V. E., Nieuwdorp, M. & Herrema, H. Evaluating causality of gut microbiota in obesity and diabetes in humans. *Endocr. Rev.* **39**, 133–153 (2018).
43. Walter, J. & Ley, R. The human gut microbiome: ecology and recent evolutionary changes. *Annu. Rev. Microbiol.* **65**, 411–429 (2011).
44. Costea, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
45. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
46. Weiss, S. et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* https://doi.org/10.1186/s40168-017-0237-y (2017).
47. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* **62**, 692–703 (2016).
48. Saary, P., Forslund, K., Bork, P. & Hildebrand, F. RTK: efficient rarefaction analysis of large datasets. *Bioinformatics* **33**, 2594–2595 (2017).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
50. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
51. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
52. Sunagawa, S. et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
53. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
54. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
55. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
56. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
57. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
58. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).
59. Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
60. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, 490–495 (2014).
61. Vieira-Silva, S. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* **1**, 16088 (2016).

62. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).
63. Devarajan, K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.* **4**, e1000029 (2008).
64. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41**, D793–D800 (2013).
65. Forslund, K. et al. Country-specific antibiotic use practices impact the human gut resistome. *Genome Res.* **23**, 1163–1169 (2013).

Acknowledgements

This research received funding from the European Community's Seventh Framework Programme (FP7/2007–2013): MetaHIT, grant agreement HEALTH-F4-2007-201052 and MetaCardis, grant agreement HEALTH-2012-305312. The Department of Bio and Health Informatics, Technical University of Denmark, and the Novo Nordisk Foundation Center for Basic Metabolic Research have in addition received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115317 (DIRECT), the resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. The Novo Nordisk Foundation Center for Protein Research received funding from the Novo Nordisk Foundation (grant agreement NNF14CC0001). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent research center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (http://www.metabol.ku.dk). A.Ø.P. received funding from the Lundbeck Foundation (grant R218-2016-1367) and S.D.E. received funding from Agence Nationale de la Recherche MetaGenoPolis grant 'Investissements d'avenir' ANR-11-DPBS-0001.

Author contributions

The protocol was written by S.K.F., H.K.P., V.G., A.Ø.P., F.H., T. Hyötyläinen., T.N. and H.B.N., together with T. Hansen, S.D.E., S.B., M.O., P.B. and O.P., with reusable code and example data compiled and tested by H.K.P. and V.G.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-018-0064-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to O.P. or H.B.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 31 October 2018

Related link
Key references using this protocol
Pedersen, H. K. et al. *Nature* **535**, 376–381 (2016): https://doi.org/10.1038/nature18646