Multi-omic integration of microbiome data for identifying disease-associated modules

Efrat Muller[1], Itamar Shiryan[1], Elhanan Borenstein[1,2,3,*]

[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel;

[2] Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel;

[3] Santa Fe Institute, Santa Fe, NM, USA;

* Corresponding author; elbo@tauex.tau.ac.il

Abstract

The human gut microbiome is a complex ecosystem with profound implications for health and disease. This recognition has led to a surge in multi-omic microbiome studies, employing various molecular assays to elucidate the microbiome's role in diseases across multiple functional layers. However, despite the clear value of these multi-omic datasets, rigorous integrative analysis of such data poses significant challenges, hindering a comprehensive understanding of microbiome-disease interactions.

Perhaps most notably, multiple approaches, including univariate and multivariate analyses, as well as machine learning, have been applied to such data to identify disease-associated markers, namely, specific features (e.g., species, pathways, metabolites) that are significantly altered in disease state. These methods, however, often yield extensive lists of features associated with the disease without effectively capturing the multi-layered structure of multi-omic data or offering clear, interpretable hypotheses about underlying microbiome-disease mechanisms.

Here, we address this challenge by introducing *MintTea* – an intermediate integration-based method for analyzing multi-omic microbiome data. MintTea combines a canonical correlation analysis (CCA) extension, consensus analysis, and an evaluation protocol to robustly identify disease-associated multi-omic modules. Each such module consists of a set of features from the various omics that both shift in concord, and collectively associate with the disease.

Applying MintTea to diverse case-control cohorts with multi-omic data, we show that this framework is able to capture modules with high predictive power for disease, significant cross-omic correlations, and alignment with known microbiome-disease associations. For example, analyzing samples from a metabolic syndrome (MS) study, we found a MS-associated module comprising of a highly correlated cluster of serum glutamate- and TCA cycle-related metabolites, as well as bacterial species previously implicated in insulin resistance. In another cohort, we identified a module associated with late-stage colorectal cancer, featuring *Peptostreptococcus* and *Gemella* species and several fecal amino acids, in agreement with these species' reported role in the metabolism of these amino acids and their coordinated increase in abundance during disease development. Finally, comparing modules identified in different datasets, we detected multiple significant overlaps, suggesting common interactions between microbiome features.

Combined, this work serves as a proof of concept for the potential benefits of advanced integration methods in generating integrated multi-omic hypotheses underlying microbiome-disease interactions and a promising avenue for researchers seeking systems-level insights into coherent mechanisms governing microbiome-related diseases.

Introduction

The human gut microbiome is an incredibly complex ecosystem that has a marked, multi-faceted impact on our health[1–3], and a well-established role in the pathogenesis of numerous diseases[4–8]. This appreciation for the importance and complexity of the gut microbiome has prompted a proliferation of multi-omic microbiome studies, which apply several molecular assays to the same set of samples, in the hope of capturing multiple layers of information concerning the involvement of the microbiome in disease. Among such multi-omic studies, one increasingly popular study design, for example, relies on collecting paired high-throughput microbiome and metabolome profiles[9–13] (e.g., coupling shotgun metagenomics sequencing and mass-spectrometry). Unfortunately, however, while such multi-omic data clearly offer an exciting opportunity to study the microbiome and its role in human health, rigorous integrative analysis of such data remains highly challenging, as does using such data to gain systems-level understanding of the microbiome[14–16].

A common aim of microbiome studies, and accordingly a key goal of many multi-omic microbiome analyses, is to identify disease-associated markers – specific features from the various omics (e.g., certain species, pathways, or metabolites) whose measured abundances are strongly associated with the disease in question[17]. Such markers can later be used to predict disease, informing, for example, microbiome-based diagnosis[18–20], or to suggest novel hypotheses concerning specific microbiome components or mechanisms that are at play, guiding future experimental and clinical studies. Given disease and control samples (or, more generally, some phenotypic measure of interest), disease-associated markers can be identified using a variety of statistical approaches, ranging from univariate statistical tests that are applied independently to each feature from each omic, to multivariate statistical methods that consider potential statistical dependencies between all features, such as partial least squares (PLS) or linear regression[10,21–24]. More recently, machine learning (ML) methods have also been applied widely for this purpose by first training some ML model for predicting the phenotype based on all features (either from a single omic or from multiple omics combined), and then using various model explainability methods for detecting informative features in this model[11,12,25–28]. Ultimately, however, such analyses generally result in long lists of features (from one or more omics) that are associated with the disease, without leveraging the multi-layered structure of multi-omic data and without offering clear, interpretable hypotheses concerning specific and coherent mechanisms underlying microbiome-disease interactions.

Several preliminary attempts directed at addressing this challenge and considering both cross-omic dependencies and associations with a disease state have been introduced. For example, some microbiome studies took a two-step approach by first identifying features that are associated with the disease, and then clustering these features based on pairwise correlations between features[29]. This approach, however, may fail to identify features that are not sufficiently informative by themselves but can be incorporated into larger modules that, as a whole, are strongly predictive of the disease. Other studies aiming to tackle a similar challenge constructed cross-omic correlation networks while also including phenotypes of interest in the network[30], or alternatively constructed separate networks for each host condition and then explored the differences between the networks to identify condition-specific motifs[31–33]. These approaches, however, are often hard to interpret given the high dimensionality of omic data, and the massive correlation networks that are accordingly produced.

A particularly intriguing approach that may be utilized for addressing such analyses originates from the domain of multi-view learning, and is known as "intermediate integration"[34,35]. Unlike traditional multi-view approaches that directly combine raw features, either by naïvely concatenating different omics into a single joint table (i.e., “early integration”) or by modeling each omic independently and then creating a final ensemble model (i.e., “late integration”), intermediate integration seeks to combine features from the various omics (or ‘views’) into an intermediary representation level before utilizing them for downstream tasks such as classification[34] (Supplementary Figure S1). This approach thus captures dependencies between omics and could accordingly be beneficial for generating multifaceted biological hypotheses. Canonical correlation analysis (CCA), for example, is a popular intermediate integration method that receives two feature tables, and outputs a linear transformation per table so that the resulting latent variables are maximally correlated[36]. CCA extensions that may be particularly relevant for microbiome mulit-omic data include sparse CCA (sCCA) – a CCA extension that includes sparsity constraints to deal with large numbers of features, and sparse generalized CCA (sGCCA), which generalizes sCCA to further support more than two views[36,37]. Indeed, CCA and its extensions have been previously applied to microbiome-related multi-omic data with insightful results[38–43]. One recent study, for example, applied sCCA to microbiome taxonomic profiles and host transcriptomics data from patients with either irritable bowel syndrome, inflammatory bowel disease (IBD), or colorectal cancer (CRC), identifying host-microbiome associations that are shared across these cohorts and others that are disease-specific[38]. In another study, sCCA was used to explore the association between early life gut microbiome and metabolome, finding that the specific taxa and metabolites driving these

108 associations at 6 weeks of age differed from those at the age of 12 months[40]. Importantly, the
109 CCA transformation can further take the disease state of each sample into account, thus seeking
110 representations that highlight interactions both across different omics and between these omics
111 and the disease[36,44,45]. Galié et al., for example, studied the effects of a Mediterranean diet on
112 circulating metabolites and used an extension of sGCCA[44] to identify signatures of dietary
113 intervention reflected in both the gut microbiome and plasma metabolites in concert[46]. These
114 studies testify to the potential benefit of such CCA-based models, yet, their applicability to
115 diverse microbiome datasets, and the robustness of the obtained results, remain unclear.
116 Moreover, CCA and its extensions, as well as similar multivariate linear statistical methods,
117 are generally highly sensitive to small perturbations in the data, parameter choices (e.g. sparsity
118 constraint), atypical samples, and collinearity between variables[47–50], requiring careful
119 attention and interpretation when applied to complex microbiome data.

120 Setting out to address the challenges above and to allow researchers to gain systems-level
121 insights into coherent mechanisms underlying microbiome-disease interactions, here, we
122 introduce a comprehensive intermediate integration-based method (combining CCA extension,
123 consensus analysis, and a validation protocol), which we term *MintTea,* for analyzing multi-
124 omic microbiome data. We hypothesize that each such mechanism may involve various taxa,
125 functions, and metabolites which act in concert in disease, and thus MintTea aims to identifying
126 robust **disease-associated multi-omic modules,** each comprising a set of features from the
127 various omics that both exhibit coordinated variation across omics and, as a whole, associate
128 with the disease or phenotype of interest. We applied MintTea to 9 diverse case-control cohorts
129 with available shotgun metagenomics data (processed into both taxonomic and functional
130 profiles), 6 of which further included fecal or serum metabolomics data. We demonstrated that
131 MintTea was indeed able to capture modules that had high predictive power of the disease
132 (often comparable with that achieved by using *all* features), while also exhibiting significant
133 correlations between features from different omics. We also showed that some of the modules
134 identified by MintTea recapitulated previous observations concerning the role of the gut
135 microbiome in disease, and provided a catalog of identified multi-omic modules across all
136 datasets analyzed. This work thus serves as a proof of concept for the benefit of advanced
137 integration methods in generating integrated multi-omic biological hypotheses underlying
138 microbiome-disease associations.

Results

Introducing MintTea: A framework for identifying robust disease-associated microbiome multi-omic modules

Inspired by the potential of intermediate integration methods described above, we developed a novel multi-omic integration framework, termed *MintTea* (Multi-omic INTegration Tool for microbiomE Analysis) for identifying sets of features from multiple different omics that are strongly associated both with the disease and with each other (Figure 1A). MintTea is based on sparse generalized canonical correlation analysis (sGCCA) and other previously introduced related methods[36,44,51,52], and further applies repeated sampling, consensus analysis, and module evaluation to account for noisy data and to ensure robust and confident results.

Briefly, MintTea receives two or more feature tables describing different omics obtained for the same set of samples, as well as a label (e.g., healthy vs. disease) for each sample, as input. Following filtration of rare features and other pre-processing steps (see Methods), MintTea encodes the label as an additional omic (containing a single feature) as previously suggested[44,51,52], and then applies sGCCA, searching for a *sparse* linear transformation per feature table that yields maximal correlations between the respective latent variables, as well as between these variables and the label (Figure 1A top). This procedure yields a latent variable per feature table (omic), which is a sparse linear combination of the various features. We then record the set of features that were assigned non-zero coefficients across the various omics, and define this set as the first "*putative module*". sGCCA can then find additional sets of latent variables that are orthogonal to the previous ones, by deflation on those previous latent variables, with each iteration similarly providing a new putative module. Next, to identify specific modules of features that are robust to small changes in the input data, MintTea repeats the entire process multiple times on random data subsets (e.g., 90% of the samples) and records the resulting putative modules from each such iteration. It then constructs a co-occurrence network, where features are connected if they consistently co-occurred in the same putative module (e.g. over 80% of total iterations), and identifies "*consensus modules*" (i.e., connected subgraph; see Methods; Figure 1A middle). Importantly, as MintTea balances between two objectives, namely associations between omics and associations with the phenotype, the resulting consensus module may capture only one of these two objectives. Accordingly, to output robust, disease-associated, and cross-omic correlated modules, MintTea finally

170 evaluates each consensus module, discarding modules that do not adhere to the desired criteria.
171 To this end, it first filters out any module in which the average cross-omic correlation is not
172 higher than that obtained for random modules. Then, MintTea applies Principal Component
173 Analysis (PCA) to all the features included in each remaining module, uses the first principal
174 component (PC) of each module as the module's representative, quantifies how well this one
175 representative PC predicts the disease using the area under the receiver operating characteristic
176 curve (AUC), and preserves only modules for which this AUC is great than 0.7 and higher than
177 AUC obtained for random module (of the same size and distribution over omics; see Methods;
178 Figure 1A bottom). The final outcome of MintTea is therefore multiple *disease-associated,*
179 *multi-omic modules*, each capturing features from multiple omics that are highly correlated
180 across the different omics, and also with the disease.

181 In Supplementary note 1, we discuss the conceptual differences between MintTea and other
182 related approaches, and quantitatively compare MintTea to other sGCCA-based methods[36,44],
183 demonstrating that MintTea provides a good balance across multiple desired properties, while
184 achieving a substantially lower false discovery rate.

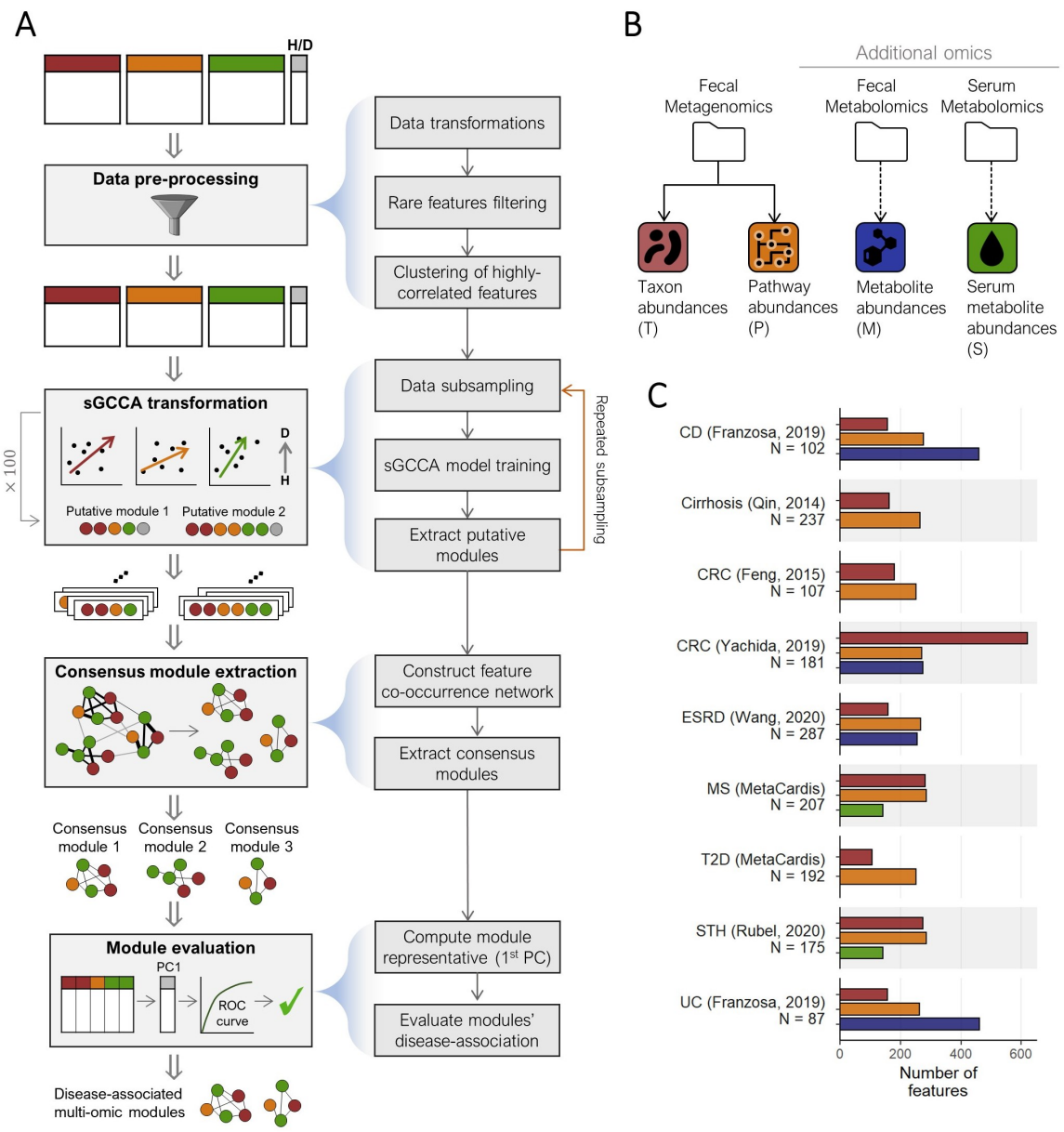


Figure 1: MintTea pipeline illustration and the multi-omic datasets analyzed in this study. (A) An illustration of the MintTea pipeline, including data preprocessing, repeated module discovery using sparse generalized CCA (sGCCA), consensus analysis, and evaluation of each module's association with the disease. See Methods. (B) Data types used in this analysis. Throughout the manuscript, taxonomic, pathway, fecal metabolite, and serum metabolite features are labeled with the letters 'T', 'P', 'M' and 'S', respectively. (C) The number of features available from each omic in each dataset after preprocessing. Sample sizes are noted under dataset names. CD: crohn's disease; CRC: colorectal cancer; ESRD: end-stage renal disease; MS: metabolic syndrome; T2D: type-2 diabetes; UC: ulcerative colitis;

Using MintTea for identifying multi-omic modules across cohorts and diseases

To examine MintTea's ability to identify disease-associated multi-omic modules in various settings, we obtained 9 different datasets of shotgun metagenomics sequencing. All datasets are from case-control studies, spanning multiple health conditions and diseases, including Crohn's disease (CD), ulcerative colitis (UC), colorectal cancer (CRC), metabolic syndrome (MS), type 2 diabetes (T2D) and others. For each dataset, we generated 2-3 different feature tables ('views') based on the available omic data. Specifically, in all datasets, metagenomic data was available and processed into taxonomic profiles ("T") and pathway-level functional profiles ("P") (see Methods). In addition, for 3 of the datasets, fecal metabolomics data ("M") were also available, and for 2 others datasets, fasting *serum* metabolomics data ("S") were available. Figure 1B-C summarizes the omics and datasets used for the main analysis (and see also Supplementary Table S1).

Applying MintTea to these 9 datasets, we found 2-5 consensus modules per dataset, each comprising 2 to 38 features from the various omics (average: 13.23; Figure 2A; Supplementary Table S2; Supplementary Table S3). Most of these modules captured correlations between features from different omics that significantly exceed those of randomly sampled modules (Figure 2B), with 1-2 modules per dataset also exhibiting strong association with the disease, as defined above (Figure 2C; Supplementary Tables S2 and S4). In the T2D (MetaCardis[13,53]) dataset, for example, we identified 5 consensus modules, 2 of which also passed our evaluation and were classified as disease-associated multi-omic modules, with AUCs reaching 0.86 (based on the 1st PC of the 21 features in this module) and 0.73 (based on only 2 features in this module), compared to 0.71 and 0.64 on average, respectively, when evaluating randomly selected modules. In the late stage CRC dataset (Yachida, 2019[11]), as another example, we found 2 consensus modules, one of which classified as disease-associated multi-omic module with an AUC of 0.72 compared to only 0.63 using equivalent, randomly selected modules (more on this module below). The entire list of identified modules, as well as detailed statistics about their association with disease and correlations between features across omics, is provided in Supplementary Tables S2-S4).

While most of our analyses below focus only on disease-associated multi-omic modules (i.e., modules that passed our evaluation phase and adhere to the criteria we defined), we sought to also evaluate the overall predictive power of *all* identified consensus modules in each dataset

217 in order to gain a better understanding of CCA-based methods' ability to identify informative
218 axes of variation in the data. To this end, we trained random forest (RF) models on the 1st PCs
219 of all modules in each dataset, and estimated their combined predictive accuracy (with cross-
220 validation). We further compared the AUC's of these models to RF models trained on the entire
221 concatenated multi-omic data (referred to as an "early integration" approach; see Methods;
222 Supplementary Table S5). We found that in most cases, these consensus modules by
223 themselves achieved an AUC comparable to that of the early-integration models, while
224 utilizing a substantially smaller number of features in total (Supplementary Figure S2). In the
225 UC (Franzosa, 2019[12]) dataset, for example, the 4 consensus modules identified by MintTea
226 (totaling 73 features) achieved a total AUC of 0.885 (SD: 0.12), where each module is
227 represented by its 1st PC as explained above. In comparison, the early integration approaches
228 achieved a similar AUC of 0.882 (SD: 0.14), but used 235 features on average for training the
229 RF model *after* feature selection. Similarly, the overall AUC of the CRC (Feng, 2015)[54]
230 modules was 0.839 (SD: 0.15), utilizing 36 features in just 3 modules (and hence based on just
231 3 values), compared to 0.824 (SD: 0.15) using early-integration based on 103 features.

232 It should also be noted that the features included in the disease-associated multi-omic modules
233 reported above, did not necessarily overlap with those ranked highest by the early integrated
234 multi-omic models. Specifically, on average, only 63.9% (SD: 24.8%) of the features
235 comprising each disease-associated module (as defined above), were also significant
236 contributors to the multi-omic early-integration models (see Methods, Supplementary Table
237 S6). This observation may attest to the ability of the intermediate integration approach to
238 consider features in the context of their associations with other features and not only their
239 association with disease.

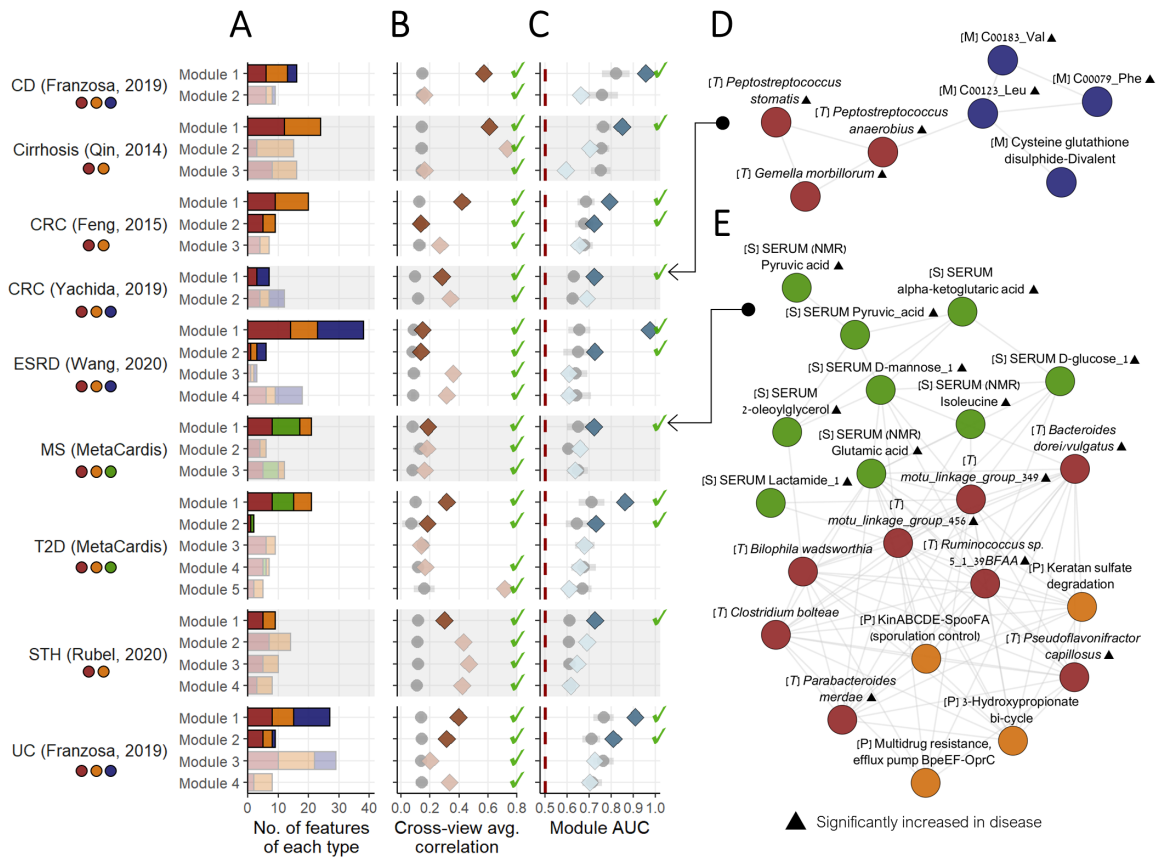


Figure 2: **Identifying disease-associated multi-omic modules via MintTea intermediate integration framework.** (A-C) Properties of the multi-omic consensus modules obtained by applying MintTea to several key datasets, including (A) the number of features in each module, stratified by feature type, (B) cross-omic correlation per module, calculated as the average pairwise Spearman correlation between features of different omics (with grey points and lines indicating the average and standard deviation of correlations obtained from random modules of the same size), and (C) AUC per module, calculated using the first principal component values against disease label of each sample (again compared to random modules, in grey). Modules that only included features from a single omic were discarded from our analyses but are listed in Supplementary Tables S2 and S3. Modules that exhibit between-omic correlations (higher than random modules) and that are disease-associated (AUC > 0.7 and above that of random modules) are shown in darker color. Circle colors under each dataset name indicate which omics were available for this dataset. (D) A late stage CRC-associated multi-omic module. Node colors represent feature types (see Figure 1B). Edges connect features that appeared together in a sGCCA putative module in >80% of data

subsampling iterations. Correlations between pairs of features within the module can be viewed in Supplementary Table S4. (E) A MS-associated multi-omic.

240

241 Disease-associated multi-omic modules reveal multifaceted biological 242 mechanisms

243 Figure 2D presents one example of a MintTea module identified in the late-stage CRC 244 dataset[11]. This module included 7 features in total; 3 bacterial species, namely 245 *Peptostreptococcus stomatis*, *Peptostreptococcus anaerobius*, and *Gemella morbillorum*, and 246 4 fecal metabolites: 2 branched-chain amino acids (BCAA's), namely valine (Val) and leucine 247 (Leu), an aromatic amino acid - phenylalanine (Phe), and cysteine-glutathione disulfide 248 divalent. The 1st PC of the 7 features included in this module, by itself, yielded an AUC of 0.72 249 in classifying disease state (Figure 2C), and the average Spearman correlation between features 250 from different omics was 0.28 (all correlations had an FDR-corrected p-value < 0.05, 251 Supplementary Table S4). All features in this module, except for the last, were significantly 252 elevated in disease (Mann-Whitney tests, FDR < 0.05), and were also reported as disease 253 biomarkers in multiple other independent studies[55–59]. Specifically, in the original paper by 254 Yachida et al.[11], all three species were overrepresented consistently across *all disease stages*, 255 as opposed to other species that were elevated in specific stages only. A similar trend was 256 reported for the three amino acids (Val, Leu and Phe), indicating a coordinated pattern with the 257 above taxa[11]. *P. stomatis* and *G. morbillorum* were also found to have significantly high 258 replication rates compared to controls across all cancer stages in that study, hinting to their 259 increased metabolic activity during cancer development. One possible explanation for 260 *Peptostreptococcus* species appearing together with the BCAA's is their involvement in BCAA 261 metabolism. Interestingly, *Peptostreptococci* species were specifically identified as major 262 fermenters of Phe and Leu, while the most prominent CRC-associated species, *Fusobacterium* 263 *nucleatum*, for comparison, was reported to prefer different amino acid substrates in the same 264 review study[60]. Phenylalanine was also significantly and positively correlated with both *P.* 265 *stomatis* and *G. morbillorum* in a recent meta-analysis[61]. Lastly, though we could not recover 266 a mechanistic link between cysteine-glutathione disulfide and the other features, we note that 267 the correlations between this metabolite and the other 3 amino acids were extremely high, even

268 when controlling for disease state (with all Spearman and partial Spearman correlations $\rho >$
269 0.47 and FDR $< 1 \cdot 10^{-10}$).

270 Another example, illustrated in Figure 2E, is a module identified in the metabolic syndrome
271 (MS) cohort of MetaCardis (these are subjects with MS according to International Diabetes
272 Federation criteria, without type 2 diabetes (T2D) or coronary artery disease). This module
273 included 8 species (2 of which lacking taxonomic annotation), 4 pathways, and 9 serum
274 metabolites (measured by either NMR or mass spectrometry), with the 1st PC of these features
275 achieving an AUC of 0.72 for predicting MS. Notably, the interplay between the gut
276 microbiome and serum metabolites is significantly less straightforward compared to that
277 observed with fecal metabolites, as it is likely overshadowed by other physiological processes
278 and systemic metabolic activities[22,62,63]. Indeed, the average correlation between features from
279 different omics in this model was lower than those in modules of other datasets, but nonetheless
280 higher than randomly sampled modules (Figure 2B). Specifically, 19 of 108 correlations
281 between serum metabolites and gut microbiome features in this module were significantly
282 positive (FDR < 0.05; for comparison, in random modules of the same size, less than 1
283 metabolite-microbiome feature pair was significantly correlated). Of the serum metabolites
284 included in this module, several are well-known signatures of metabolic disorders, including
285 increased glucose, mannose[64], TCA cycle metabolites (e.g. pyruvic acid and alpha-ketoglutaric
286 acid)[65], and glutamate[66]. Circulating levels of isoleucine (as well as other BCAA's and BCAA-
287 related metabolites) have also been repeatedly shown to play a role in glucose homeostasis and
288 metabolic syndrome risk[65,67,68]. Though some of the other module's metabolites have not been
289 specifically discussed in the context of metabolic disorders, they are all significantly positively
290 correlated with one another (32 of 36 metabolite pairs are positively correlated with an FDR <
291 0.05), effectively representing a cohesive cluster of metabolites coordinately elevated in MS.
292 Interestingly, in a recent study, several of these metabolites (pyruvic acid, glucose, glutamic
293 acid) were found to be significantly well-predicted by the gut microbiome when measured in
294 feces, but only marginally well-predicted in blood[63], further strengthening the assumption that
295 the gut microbiome's association with these serum metabolites is likely obscured by additional
296 processes, but nonetheless existent.

297 Moreover, the inclusion of certain bacterial species in this module along with these metabolites
298 is likely related to their previously reported role in glutamate metabolism and/or BCAA
299 metabolism. *Bacteroides dorei/vulgatus*, for example, though only marginally significantly
300 increased in MS patients in our dataset (FDR = 0.07), have been shown to impact serum

301 isoleucine levels[69]. *B. vulgatus* specifically was also reported to take part in glutamate
302 metabolism[70] and further recognized as one of the main species driving the association between
303 biosynthesis of BCAAs in the gut and insulin resistance[7,71]. The exact role of *B. vulgatus* in the
304 development of metabolic disorders, however, remains unclear and is probably context- and
305 strain- dependent, as it has been shown to confer health benefits (including metabolic related)
306 in some studies, while positively correlating with insulin resistance and body fat in others[7,69,72–
307 75]. Interestingly, *B. vulgatus* was also reported to induce compositional shifts when
308 administered to rats, specifically promoting an increase in *Parabacteroides* species that is also
309 included in this module[72]. *Parabacteroides mardeae*, another species in this module, has also
310 been shown to enhance BCAA catabolism, again with positive health benefits[76], albeit
311 significantly increased in MS patients in our dataset. Lastly, we note that *Bilophila*
312 *wadsworthia* has been mostly discussed in the literature in the context of inflammatory bowel
313 disease and colon inflammation[77], but has also been implicated in metabolic dysfunctions[78]. It
314 was also found to cluster with *Clostridium bolteae*, another species in the module, in terms of
315 their responses to short and long term diet[79], possibly explaining their co-occurrence in this
316 module. As for the pathway features included in this module, likely not all represent specific
317 metabolic functions relevant to MS, but nonetheless exhibit strong correlations with some of
318 the modules' species, suggesting potential dependencies. Overall, we find multiple reports
319 supporting specific links within the module, but the shared role of these bacterial species in
320 MS phenotypes and circulating levels of MS-related metabolites specifically remains elusive.

321 Importantly, even when analyzing metagenomic-derived features alone (i.e. in the absence of
322 additional omics), MintTea may highlight interesting insights related to associated taxonomic
323 and functional disease biomarkers. In the liver cirrhosis dataset[32], for example, we found a
324 module (Supplementary Figure S3A) that included multiple species typically found in the
325 human oral cavity, including several *Streptococcus*, *Veillonella*, and *Megasphaera* species
326 (both commensal and opportunistic pathogens)[80], all of which previously found to be increased
327 in stool samples of individuals with liver diseases[81–83], as well as other gastrointestinal
328 disorders[84–86]. The co-occurrence of multiple *Streptococcus* and *Veillonella* species in this
329 module is in line with their strong co-abundance patterns, confirmed across multiple diverse
330 large-scale cohorts[87], and also with known metabolic interactions between them[88].
331 Furthermore, several pathways in the module relate to vitamin K metabolism (menaquinol-8
332 biosynthesis I, menaquinol-11 biosynthesis, 2-carboxy-1,4-naphthoquinol biosynthesis,
333 phylloquinol biosynthesis, geranylgeranyl diphosphate biosynthesis), also relevant to liver
334 health[89,90]. Indeed, several vitamin K forms (including menaquinone) are known to be

335 bacterially-produced in the human intestines, with *Veillonella parvula* specifically reported to
336 produce menaquinones in vitro[91]. In another study, menaquinone biosynthesis was found to
337 associate with *Streptococcus* overgrowth in IBD[87]. More generally, fecal menaquinone profiles
338 have been shown to strongly associate with overall gut microbiome composition even on a
339 controlled diet[92]. Additional pathways in the module, e.g. those related to L-metionine and
340 heme b biosynthesis, are also possibly tied to menaquinone metabolism and to *Veillonella*'s
341 specific metabolic activities. L-methionine's role, for example, as a methyl group donor for
342 certain menaquinones biosynthesis has been confirmed in several bacterial species[93], though
343 not specifically in the species of this module. Also, previous genomic analysis of *Veillonella*
344 species showed that they harbor complete gene sets for heme biosynthesis[94]. Overall, this
345 module suggests a potential multi-layer vitamin k-related and oral bacteria-driven mechanism
346 that is highly associated with liver damage. Supplementary Figure S3 further illustrates two
347 additional multi-omic modules associated with IBD, as additional examples.

348 Multi-omic modules across diverse cohorts and disease states

349 Having shown that MintTea can identify multi-omic modules exhibiting associations both
350 across omics and with a disease, we finally sought to detect prevalent and recurring multi-omic
351 modules shared across diverse datasets. Due to the substantial technical and methodological
352 differences between the multi-omic datasets analyzed, which render cross-study comparisons
353 impractical, here we focused on shotgun data only, where both taxonomic and functional
354 (pathway-level) profiles were generated uniformly by the curatedMetagneomicData resource[95].
355 Specifically, we utilized the 3 datasets, previously presented, from the curatedMetagonmicData
356 resource (i.e., the CRC[54], Cirrhosis[96], and STH[97] datasets; Figure 1C), as well as 11 additional
357 datasets from the same resource (Supplementary Table S1; Methods), totaling 14 uniformly
358 processed datasets. After applying MintTea to these datasets (Supplementary Figure S4;
359 Supplementary Tables S3 and S4), we quantified the overlaps between modules from the
360 different datasets and disease states, and identified significant overlaps (Figure 3A;
361 Supplementary Table S7; Methods). Of the 43 consensus module-pairs with an overlap of at
362 least 2 features, 10 exhibited a statistically significant overlap (Fisher's exact test, FDR-
363 corrected P value < 0.1), suggesting prevalent multi-omic associations between microbiome
364 features.

365 We additionally identified specific bacterial species that appeared in multiple modules of
366 different datasets. *Bacteroides uniformis,* for example – a gram-negative, fiber-degrading,

367 highly abundant commensal in the human gastrointestinal tract[98–100] – appeared in modules of
368 5 different datasets (Figure 3B). Other species commonly appearing with *B. uniformis* in
369 MintTea modules were also from the *Bacteroides* genus, including *B. xylanivsolvens*, *B. ovatus*,
370 and *B. thetaiotaomicron*. Interestingly, *B. uniformis* tended to appear with vitamin B related
371 pathways such as pyridoxal 5'-phosphate biosynthesis pathways (the active form of vitamin
372 B6) and thiamine (vitamin B1) diphosphate synthesis. Indeed, *B. uniformis* genomes as well as
373 the other Bacteroides genomes, are known to have a relatively high coverage of biosynthetic
374 pathways for multiple vitamins, including B-vitamins[101,102]. Examining the taxon-stratified
375 outputs of HUMAnN3, we observed that *B. uniformis* was one of the top contributors of reads
376 assigned to these vitamin B-related pathways, further explaining their co-occurrence in the
377 same modules. *B. uniformis* was also a top contributor to other pathways that co-appeared with
378 it across modules from diverse datasets, including 6-hydroxymethyl-dihydropterin diphosphate
379 biosynthesis I and lipid IV_A biosynthesis (which is part of the biosynthesis process of
380 lipopolysaccharide (LPS), a crucial component of bacterial outer membrane). Taken together,
381 these modules suggest a common axis of variation shared between multiple cohorts and
382 characterized mainly by *Bacteroides* species and metabolic processes essential for cell survival
383 including biosynthesis of several cofactors and LPS.

384 Noticing the tendency described above of species from the *Bacteroides* genus to co-occur in
385 modules, we conducted a permutation-based analysis to test whether species in MintTea
386 modules are indeed significantly more likely to appear with other species from their respective
387 genus, compared to random modules. We confirmed this was indeed a general trend (22.8%
388 compared to 12.7% [7.7%-18.6%], p-value: 0.001). This observation is in line with the
389 hypothesis that phylogenetically related species are more likely to compete for a shared niche,
390 and accordingly present stronger co-occurrence patterns.

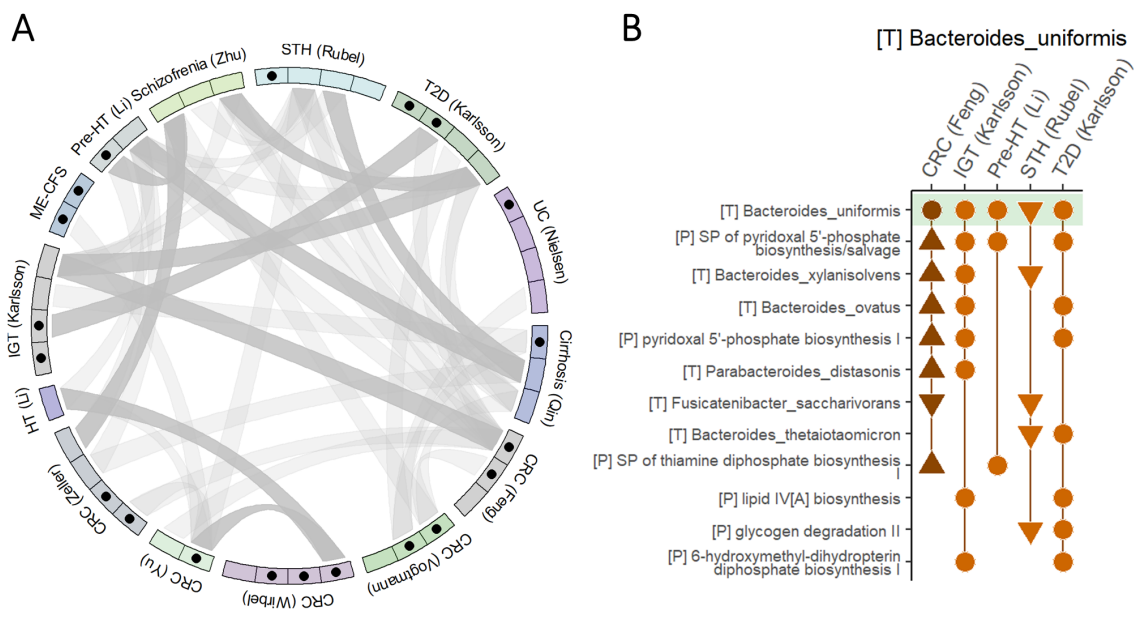


Figure 3: Multi-omic modules across datasets and disease states. (A) Overlaps between multi-omic modules of different datasets. Each sector in the circus plot represents a consensus module, grouped by the datasets to which they belong. Black dots represent disease-associated multi-omic modules (as previously defined). Links between modules indicate overlaps of at least 2 features, with darker links indicating statistically significant dependences (Fisher's exact test FDR < 0.1). (B) Multi-omic modules from multiple datasets that include *B. uniformis*, and additional overlapping features. For each module (from the dataset listed on top), all module features that appear in at least one other module are presented. Triangles pointing up indicate that the feature level was significantly increased in disease in that dataset, while triangles pointing down indicate an opposite trend (Mann Whitney tests, FDR < 0.1). Circles indicate no significant difference between study groups. The module in dark orange (CRC, Feng) was also associated with disease state.

STH: soil-transmitted helminths; HT: hypertension; ME-CFS: myalgic encephalomyelitis/ chronic fatigue syndrome; IGT: impaired glucose tolerance; SP: super-pathway; Also see abbreviations in the legend of Figure 1.

391

Methods

Dataset collection

In this work we collected, processed, and analyzed data from a few different multi-omic studies, as well as from the curatedMetagenomicData[95] resource. Below are details describing the data obtained from each of these sources, and further processing we performed. We first obtained data from 3 case-control studies that profiled both stool shotgun metagenomics and metabolomics, including an inflammatory bowel disease (IBD) cohort (Franzosa et al., 2019[12]), a colorectal cancer (CRC) cohort (Yachida et al., 2019[11]), and an end stage renal disease (ESRD) cohort (Wang et al., 2020[32]). For each of these studies we ran HUMAnN3[103] (version v3.0.1) on the raw metagenomic sequencing files to obtain MetaCyc[104] pathway profiles. MetaPhlAn3[103] (version 3.0.14) was used to obtain taxonomic profiles at the species level. Processed metabolomics data was obtained from the supplementary information of each of the original publications or online repositories. Only identified metabolites (i.e., annotated with metabolite name by the original study's authors) were included in the analysis, to allow downstream interpretation. Metadata per subject was obtained from the supplementary information. When distinct disease subtypes were present, we split the data into separate datasets per disease subtype (or kept only a specific subtype), taking each subtype as the "case" patients, and the same set of control subjects as shared controls. Specifically, the IBD cohort[12] were split into Chron's disease (CD) and ulcerative colitis (UC), and in the CRC cohort[11] we kept only late-stage cancer patients (Supplementary Table S1). In both of these cases, the original publications indicated unique microbiome signatures for each disease subtype/stage and performed separate analyses accordingly, as we do here.

We additionally included data from the European MetaCardis cohort[13,53], in which participants underwent deep clinical phenotyping combined with gut microbiome and serum metabolome profiling. Pre-processed taxonomic, functional, and serum metabolome profiles were downloaded from zenodo.org/records/6242715. Notably, taxonomic profiles were obtained using the mOTU[105] approach, and functional profiles at the KEGG-module level were generated using a custom pipeline. Full data processing details are described in the original publication[53]. As the MetaCardis cohort is inherently divided into several study groups based on metabolic and cardiac phenotypes, we focused on two specific conditions, T2D and metabolic syndrome (without T2D or coronary artery disease). Subjects with antibiotics indications ('ANTIBIOTICS_TOTAL' > 0) or multiple drugs intake ('DRUGTOTAL' >= 3) were excluded. As before, the same set of control samples were used for both comparisons.

425 Lastly, 14 processed shotgun metagenomics datasets were extracted from the
426 *curatedMetagenomicData* R package[95], version 3.2.3, together with the corresponding sample
427 metadata. Specifically, MetaPhlAn3[103] species-levels abundances and HUMAnN3[103]
428 unstratified gene-level (UniRef) and pathway-level (MetaCyc) abundances were obtained. We
429 only considered datasets with data from at least 40 stool samples from different subjects in
430 each study group (as defined by the "study_condition" variable). Longitudinal studies were
431 excluded. Additional details for each dataset are available in Supplementary Table S1.

432 In total, data analyzed in this study included 2,629 samples, obtained from 16 studies, and
433 partitioned into 20 case-control datasets, covering 13 disease states. We preliminary focused
434 on 9 datasets (all multi-omic datasets and 3 additional shotgun-only datasets), and in the last
435 section expand the analysis to all datasets.

436 *Shared data preprocessing*

437 For each of the datasets described above, we obtained sample metadata, taxonomy abundances
438 ('T'), pathway abundances ('P'), and, when available, fecal/serum metabolite abundances
439 ('M'/'S') (Figure 1B). For all datasets, T and P profiles were converted into relative
440 abundances (by total sum scaling), metabolite values were log-transformed, and constant and
441 rare features (defined as those with <15% non-zero abundance values, or a mean abundance
442 <0.005% for taxonomy/pathways) were discarded. Moreover, *extremely* highly correlated
443 features, which likely represent technical artifacts rather than biological phenomena, were
444 clustered (to avoid misleading interpretation downstream) such that absolute correlations
445 between any two cluster members is higher than 0.99 and using one random feature as cluster
446 representative (Supplementary Table S8). Lastly, non-bacterial MetaCyc pathways were
447 removed from the P profiles, as annotated in the MetaCyc database within the "Expected
448 Taxonomic Range" field.

449 *Identifying disease-associated multi-omic modules with "MintTea"*

450 To identify multi-omic sets of features that are both tightly associated with each other and
451 collectively associated with the disease, we developed "MintTea" (Multi-omic INTegration
452 Tool for microbiomE Analysis). MintTea is based on sparse generalized canonical correlation
453 analysis (sGCCA)[36,37] and additional methods that extend sGCCA to a supervised scenario by
454 encoding the disease label of each subject as an additional single-variable omic[44,51,52,106].
455 Briefly, sGCCA searches for a *sparse* linear transformation per feature table that yields
456 maximal correlations between the respective latent variables, as well as between these variables
457 and the label (see also Supplementary Note 2). Herein, we define the set of features across all

458 omics that were assigned a non-zero loading by sGCCA, as a "putative module". SGCCA can
459 further find additional linear transformations, orthogonal to the previous ones, by a process
460 called matrix deflation[37]. Each such additional set of omic transformations yields additional
461 "putative module".

462 MintTea pipeline, as also illustrated in Figure 1A, includes the following steps: After data pre-
463 processing described above, MintTea performs a repeated sub-sampling procedure so that for
464 each subsample of the data (80% of the samples, 100 iterations) sGCCA is applied and the first
465 5 resulting putative modules are recorded. Next, MintTea identifies "consensus modules", by
466 constructing a network of multi-omic features where 2 features are connected by an edge if
467 they appeared together in a putative module in at least 80% of the iterations. Connected
468 components in the graph of size ≥ 2 are then considered "consensus modules". Lastly, to assess
469 the extent to which each identified consensus module adheres to the desired properties
470 described above, MintTea further evaluates each such consensus module preserving only those
471 that meet certain criteria. Specifically, to evaluate whether a given consensus module is
472 associated with the disease state, MintTea applies principal component analysis (PCA) on all
473 features included in that module (scaled and centered) and takes the first principal component
474 (PC) as a single representative of the module. It then computes the area under the receiver
475 operating characteristic curve (AUC) for this first PC of each module, and compares it to "null"
476 modules where the same number of features, from the same omics, are randomly selected.
477 MintTea further evaluates, per module, how well it captures associations between different
478 omics, by calculating the average Spearman correlation between pairs of features from different
479 omics within that module (Supplementary Table S4). Consensus modules with an AUC>0.7
480 and above that of random modules, and an average Spearman correlation between features from
481 different omics that was also above that of random modules, are considered as '*disease-*
482 *associated multi-omic modules*'. Importantly, all parameters mentioned above can be adjusted
483 by the user (see below). Supplementary Note 1 further describes a conceptual and quantitative
484 comparison between MintTea and related methods.

485 The pipeline was developed in R using the packages "mixOmics"[106] (version 6.18.1,
486 "block.splsda" function), and "igraph"[107] (version 1.3.4) for graph operations. Classification
487 performances were generally estimated using repeated 10-fold cross-validation and averaged
488 over folds and repeats.

MintTea pipeline parameters and sensitivity analysis

The MintTea pipeline has several user-adjustable parameters required for tuning the process to specific datasets. Different parameters may be favored based on the size of the outputted modules and downstream analyses, correlation structures in the data, dimensionality of the data, etc. Briefly, these parameters include: '*keep'* – the number of features to select from each omic as provided to sGCCA's computation; '*des*' – the value to set as design matrix default (see Supplementary Note 2); '*nrep*' – number of repetitions; '*nfol*' – number of folds to which the entire data is split in order to run sGCCA on *nfol*-1 folds at a time (i.e., subsampling ratio); '*ncom'* – number of orthogonal sGCCA components; '*edge*' – the threshold defining how often should a pair of features co-appear in a putative module in order for them to co-appear in a consensus module. Exact parameters used per dataset in our analysis are listed in Supplementary Table S2. Supplementary Figure S5 shows an example of a sensitivity analysis applied to a specific module from a specific dataset, assessing that module's sensitivity to different parameter choices. Overall, and as demonstrated in Supplementary Figure S5, modules are generally robust to choice of parameters.

Modules overlap analysis

To identify significantly overlapping modules from different datasets, we first narrowed the consensus modules to the set of features available in both datasets, and then ran Fisher's exact test to determine odds ratios and *P* values. Results are listed in Supplementary Table S7.

Predicting disease state using simple, early-integration random forest classifiers

To provide a reference value for the AUC's estimations of MintTea's multi-omic modules, we implemented a standard ML pipeline to assess disease predictability in each dataset. Input data was a simple concatenation of all features from all omics (AKA an "early integration" approach). Training was performed with a repeated 5-fold cross validation (10 repeats) procedure and included reducing dimensionality with a feature selection step (using Boruta[108]), training a random forest (RF) classifier, evaluating model performance on out-of-fold data using AUC, and evaluating feature importance (as described below).

Final model performances, for each dataset, were determined by averaging AUC over the 5 repeats and 10 folds (i.e. a total of 50 trained models).

We used the 'ranger' R package (version 0.14.1)[109] for RF modelling with default parameters. We opted for RF due to its simplicity on the one hand and its tendency to perform well on microbiome abundance data as consistently shown in previous studies[25,110–114]. Hyper-

521 parameter tuning for the RF classifiers with a nested cross-validation approach did not result
522 in significantly improved performance for most datasets and was therefore discarded. 'Boruta'
523 R package (version 7.0.0)[108] was used for feature selection. All results from RF modelling are
524 provided in Supplementary Table S5.

525 *Random forest's feature importance analysis*

526 We examined the features importance in each model that achieved an AUC > 0.7, as follows:
527 Within each fold, we calculated permutation-based feature importance scores implemented in
528 'ranger' package[109]. We additionally applied the method introduced by Altmann et al.[115] to
529 assign a *P* value for each feature in each model. Final feature importance per feature was
530 averaged over repeats and folds. Final *P* values were FDR-corrected over all features in each
531 feature set. We refer to features that were selected by the feature selection method in at least
532 50% of repeats/folds and obtained a final Altmann FDR < 0.1 as "contributors". Feature
533 importance results are provided in Supplementary Table S6.

534 Discussion

535 In this work, we set out to enhance the analysis of microbiome multi-omic data, allowing
536 researchers to identify robust, disease-associated microbiome signatures across multiple
537 molecular processes and to generate multi-faceted hypotheses concerning the role of the
538 microbiome in disease. We specifically utilized an "intermediate integration" approach to
539 analyze multi-view (multi-omic) human gut microbiome data, presenting a novel framework
540 (termed "MintTea"), for simultaneously identifying sets of features from different omics that
541 are both highly associated with one another and jointly associated with the studied disease. Our
542 "views" included taxonomic and pathway functional profiles derived from shotgun
543 metagenomics data, as well as metabolomic profiles (from either feces or serum). Applied to
544 multiple diverse case-control datasets, MintTea was able to pinpoint cohesive multi-omic
545 modules of features that reflected coordinated shifts in the data across omics, while also being
546 highly predictive of disease.

547 Using MintTea, we identified, for example, a colorectal cancer-associated module driven by
548 *Peptostreptococcus* and *Gemella* species and the metabolism of specific amino acids. In
549 another cohort of patients with metabolic syndrome, we found a module that included a highly
550 correlated cluster of fasting serum metabolites (mostly glutamate- and TCA cycle-related
551 metabolites) as well as bacterial species that were previously implicated in insulin

552 resistance/metabolic disorders as well as in isoleucine metabolism. Applying MintTea to a
553 large collection of shotgun datasets, and comparing the modules identified in each dataset, we
554 further identified multiple modules with a significantly high overlap, suggesting common axes
555 of variation in the gut's microbiome composition and function, shared across multiple cohorts.
556 In a way, these modules can thus be thought of as fundamental building blocks of the
557 microbiome or "modes" of operation.

558 Intermediate integration methods, in general, can serve as an alternative (or complementary)
559 analytical approach to early integration, in which different omics are simply concatenated
560 together prior to the application of some ML model (an approach we used in this study as a
561 baseline of disease predictability). Indeed, intermediate integration methods have been gaining
562 increased attention in recent years, in concordance with the proliferation of multi-omic study
563 designs, due to their ability to uncover intricate multivariable interactions, which is desperately
564 required when dealing with complex biological systems[35]. Our motivation to use a sCCA-based
565 method was the assumption that gut microbiome-disease associations may be manifested in
566 microbiome multi-omic data as one or more "modules" of features from all omics that take part
567 in a specific mechanism effected by or promoting host disease. Furthermore, sCCA-based
568 frameworks have been previously applied to various microbiome multi-omic data with
569 insightful results[38,39,45]. To overcome challenges related to the high-dimensionality of the data
570 and consequent instability of results, however, we further expanded this analytical technique
571 with repeated iterations using data subsamples and consensus analysis, to identify *robust*
572 consensus modules based on multiple sCCA-based putative modules. This conservative
573 approach indeed resulted in markedly lower risk of false discovery, while maintaining similar
574 performance across other evaluation metrics, when compared to similar methods (see
575 Supplementary Note 1).

576 In essence, MintTea is a data-driven hypothesis generation tool that aims to discover
577 microbiome signatures reflected by multiple layers of information. The evaluation of such
578 signatures, however, is nontrivial, and remains an open challenge and major limitation of
579 multiple intermediate integration methods. We specifically recognize 3 main types of valuable
580 evaluations: robustness evaluation (to data perturbations or method parameters),
581 generalizability evaluation (the ability to replicate results in independent cohorts), and
582 assessment of the biological coherence or relevance of each finding. MintTea inherently
583 improves robustness by requiring that modules are consistent across multiple data perturbations
584 (e.g., subsampling), and indeed this addition yielded lower false discovery rates as previously

noted. We note that as multi-omic data sharing becomes more common and more standardized, mitigating the challenges of cross-study comparisons involving multiple data types, such validations will become essential. Lastly, we note that another major limitation of intermediate integration methods in general, and MintTea specifically, is the difficulty to systematically validate the biological meaningfulness of the identified modules, largely since systematic mapping of such multi-omic mechanisms is lacking (and in many cases also largely unknown).

We also note that clearly, any attempt to identify disease-associated modules based on metagenomic and other features can only be as accurate and reliable as the underlying features. Though there are several well-established tools for taxonomic and functional profiling from metagenomics data, they may each induce biases and inaccuracies[116]. Moreover, additional, possibly invaluable features can be recovered from metagenomics and metabolomics data, for example by using assembly-based methods to add unknown-genome abundances or strain-level information to taxonomic profiles[116], or adding unannotated metabolites from untargeted metabolomic profiles. Another limitation is the insufficient control for major confounders known to exert microbiome shifts (e.g., stool consistency, diet, and medication usage)[117]. This limitation could be potentially mitigated by careful sample stratification, inclusion of these factors as covariates in models, or even adding such data as an additional "view". Diet, for example, is likely a substantial mediator of many of the identified microbiome-metabolome modules, and it could therefore be especially interesting to include "diet"-view profiles in future analyses to elucidate potential diet-driven mechanisms. Finally, since microbiome omic data has unique characteristics, linear approaches such as CCA may be insufficient, and alternative algorithmic and statistical approaches should definitely be explored and developed.

Overall, this work serves as a proof of concept for the potential benefit of applying advanced intermediate integration methods to microbiome multi-omic data for generating multifaceted hypotheses concerning the role of the microbiome in human disease. We advocate for using such multi-omic intermediate integration models to capture complementary biological insights into the complex mechanisms underlying microbiome-disease associations.

References

1. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* doi:10.1038/s41422-020-0332-7

2. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).

617 3. Neish, A. Microbes in gastrointestinal health and disease. *Gasteroenterology* **30**, 2008 (2009).

618 4. Sharon, G. *et al.* Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms
619 in Mice. *Cell* **177**, 1600-1618.e17 (2019).

620 5. Mars, R. A. T. *et al.* Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable
621 Bowel Syndrome. *Cell* **182**, 1460-1473.e17 (2020).

622 6. Lavelle, A. & Sokol, H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease.
623 *Nat. Rev. Gastroenterol. Hepatol.* **17**, 223–237 (2020).

624 7. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*
625 **535**, 376–381 (2016).

626 8. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat.*
627 *Rev. Microbiol.* **12**, 661–672 (2014).

628 9. Turnbaugh, P. J. & Gordon, J. I. An Invitation to the Marriage of Metagenomics and Metabolomics. *Cell*
629 **134**, 708–713 (2008).

630 10. Zierer, J. *et al.* The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* **50**, 790–
631 795 (2018).

632 11. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the
633 gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).

634 12. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease.
635 *Nat. Microbiol.* **4**, 293–305 (2019).

636 13. Fromentin, S. *et al.* Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat.*
637 *Med.* **28**, 303–314 (2022).

638 14. Chong, J. & Xia, J. Computational Approaches for Integrative Analysis of the Metabolome and
639 Microbiome. *Metabolites* **7**, 62 (2017).

640 15. Jiang, D. *et al.* Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and
641 Opportunities. *Front. Genet.* **10**, 995 (2019).

642 16. Daliri, E. B. M., Ofosu, F. K., Chelliah, R., Lee, B. H. & Oh, D. H. Challenges and Perspective in
643 Integrated Multi-Omics in Gut Microbiota Studies. *Biomol. 2021, Vol. 11, Page 300* **11**, 300 (2021).

644 17. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–
645 60 (2012).

646 18. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.*
647 **10**, 766 (2014).

648 19. Montassier, E. *et al.* Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection.
649 *Genome Med.* **8**, 1–11 (2016).

650 20. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the
651 sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 1–10 (2016).

652 21. Sinha, R. *et al.* Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. *PLoS One* **11**,
653 e0152126 (2016).

654 22. Tang, Z.-Z. *et al.* Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals
655 Microbiome-Dependent Relationships Between Diet and Metabolites. *Front. Genet.* **10**, 454 (2019).

656 23. Eicher, T. *et al.* Metabolomics and multi-omics integration: A survey of computational methods and
657 resources. *Metabolites* **10**, (2020).

658 24. Pérez-Cobas, A. E. *et al.* Gut microbiota disturbance during antibiotic therapy: A multi-omic approach.
659 *Gut* **62**, 1591–1601 (2013).

660 25. Oh, T. G. *et al.* A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. *Cell Metab.* **32**, 878-
661 888.e6 (2020).

662 26. Beck, D. & Foster, J. A. Machine Learning Techniques Accurately Classify Microbial Communities by
663 Bacterial Vaginosis Characteristics. *PLoS One* **9**, e87830 (2014).

664 27. Gou, W. *et al.* Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features
665 Associated With Type 2 Diabetes. *Diabetes Care* **44**, 358–366 (2021).

666 28. Bar, N. *et al.* A reference map of potential determinants for the human serum metabolome. *Nature* **588**,
667 (2020).

668 29. Pedersen, H. K. *et al.* A computational framework to integrate high-throughput '-omics' datasets for the
669 identification of potential mechanistic links. *Nat. Protoc.* **13**, 2781–2800 (2018).

670 30. Kappel, B. A. *et al.* Cross-omics analysis revealed gut microbiome-related metabolic pathways underlying
671 atherosclerosis development after antibiotics treatment. *Mol. Metab.* **36**, 100976 (2020).

672 31. McHardy, I. H. *et al.* Integrative analysis of the microbiome and metabolome of the human intestinal
673 mucosal surface reveals exquisite inter-relationships. *Microbiome* **1**, 17 (2013).

674 32. Wang, X. *et al.* Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and
675 rodents. *Gut* **69**, 2131–2142 (2020).

676 33. Cantoni, C. *et al.* Alterations of host-gut microbiome interactions in multiple sclerosis. *eBioMedicine* **76**,
677 103798 (2022).

678 34. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics
679 data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).

680 35. Wang, Y. *et al.* Multi-modal intermediate integrative methods in neuropsychiatric disorders: A review.
681 *Comput. Struct. Biotechnol. J.* **20**, 6149–6162 (2022).

682 36. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to
683 genomic data. *Stat. Appl. Genet. Mol. Biol.* **8**, (2009).

684 37. Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostatistics* **15**,
685 569–583 (2014).

686 38. Priya, S. *et al.* Identification of shared and disease-specific host gene–microbiome associations across

687 human diseases using multi-omic integration. *Nat. Microbiol.* **7**, 780–795 (2022).

688 39. Kostic, A. D. *et al.* The Dynamics of the Human Infant Gut Microbiome in Development and in
689 Progression toward Type 1 Diabetes. *Cell Host & Microbe* **17**, 260–273 (2015).

690 40. Nguyen, Q. P. *et al.* Associations between the gut microbiome and metabolome in early life. *BMC*
691 *Microbiol.* **21**, 1–19 (2021).

692 41. Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. & Li, H. Structure-constrained sparse canonical
693 correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**, 244–258 (2013).

694 42. Fukuyama, J. *et al.* Multidomain analyses of a longitudinal human microbiome intestinal cleanout
695 perturbation experiment. *PLoS Comput. Biol.* **13**, e1005706 (2017).

696 43. Komesu, Y. M. *et al.* Defining the relationship between vaginal and urinary microbiomes. *Am. J. Obstet.*
697 *Gynecol.* **222**, 154.e1-154.e10 (2020).

698 44. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics
699 assays. *Bioinformatics* **35**, 3055–3062 (2019).

700 45. O'Connor, J. B. *et al.* Network Analysis to Identify Multi-Omic Correlations in the Lower Airways of
701 Children With Cystic Fibrosis. *Front. Cell. Infect. Microbiol.* **12**, 1–11 (2022).

702 46. Galié, S. *et al.* Effects of Mediterranean Diet on plasma metabolites and their relationship with insulin
703 resistance and gut microbiota composition in a crossover randomized clinical trial. *Clin. Nutr.* **40**, 3798–
704 3806 (2021).

705 47. Rodríguez-Pérez, R., Fernández, L. & Marco, S. Overoptimism in cross-validation when using partial
706 least squares-discriminant analysis for omics data: a systematic study. *Anal. Bioanal. Chem.* **410**, 5981–
707 5992 (2018).

708 48. Lambert, Z. V & Durand, R. M. Some precautions in using canonical analysis. *J. Mark. Res.* **12**, 468–475
709 (1975).

710 49. Hubert, M., Rousseeuw, P. J. & Van Aelst, S. High-breakdown robust multivariate methods. *Stat. Sci.* **23**,
711 92–119 (2008).

712 50. Kiers, H. A. L. & Smilde, A. K. A comparison of various methods for multivariate regression with highly
713 collinear variables. *Stat. Methods Appl.* **16**, 193–228 (2007).

714 51. Shi, W. J. *et al.* Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics* **35**,
715 4336–4343 (2019).

716 52. Gross, S. M. & Tibshirani, R. Collaborative regression. *Biostatistics* **16**, 326–338 (2015).

717 53. Forslund, S. K. *et al.* Combinatorial, additive and dose-dependent drug–microbiome associations. *Nature*
718 **600**, 500–505 (2021).

719 54. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat.*
720 *Commun.* **6**, (2015).

721 55. Osman, M. A. *et al.* Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and

722 Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer. *Sci. Reports 2021 111*
723 **11**, 1–12 (2021).

724 56. Tsoi, H. *et al.* Peptostreptococcus anaerobius Induces Intracellular Cholesterol Biosynthesis in Colon
725 Cells to Induce Proliferation and Causes Dysplasia in Mice. *Gastroenterology* **152**, 1419-1433.e5 (2017).

726 57. Nakatsu, G. *et al.* Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* **6**,
727 30–32 (2015).

728 58. Nannini, G. *et al.* Fecal metabolomic profiles: A comparative study of patients with colorectal cancer vs
729 adenomatous polyps. *World J. Gastroenterol.* **27**, 6430–6441 (2021).

730 59. Coker, O. O. *et al.* Altered gut metabolites and microbiota interactions are implicated in colorectal
731 carcinogenesis and can be non-invasive diagnostic biomarkers. *Microbiome* **10**, 1–12 (2022).

732 60. Dai, Z.-L., Guoyao, W. & Wei-Yun, Z. Amino acid metabolism in intestinal bacteria: links between gut
733 ecology and host health. *Front. Biosci.* **16**, 1768–1786 (2011).

734 61. Avuthu, N. & Guda, C. Meta-Analysis of Altered Gut Microbiota Reveals Microbial and Metabolic
735 Biomarkers for Colorectal Cancer. *Microbiol. Spectr.* **10**, (2022).

736 62. Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **10**,
737 (2019).

738 63. Deng, K. *et al.* Comparison of fecal and blood metabolome reveals inconsistent associations of the gut
739 microbiota with cardiometabolic diseases. *Nat. Commun.* **14**, (2023).

740 64. Sone, H. *et al.* Physiological changes in circulating mannose levels in normal, glucose-intolerant, and
741 diabetic subjects. *Metabolism.* **52**, 1019–1027 (2003).

742 65. Newgard, C. B. *et al.* A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates
743 Obese and Lean Humans and Contributes to Insulin Resistance. *Cell Metab.* **9**, 311–326 (2009).

744 66. Maltais-Payette, I., Allam-Ndoul, B., Pérusse, L., Vohl, M. C. & Tchernof, A. Circulating glutamate level
745 as a potential biomarker for abdominal obesity and metabolic risk. *Nutr. Metab. Cardiovasc. Dis.* **29**,
746 1353–1360 (2019).

747 67. Utzschneider, K. M., Kratz, M., Damman, C. J. & Hullarg, M. Mechanisms linking the gut microbiome
748 and glucose metabolism. *J. Clin. Endocrinol. Metab.* **101**, 1445–1454 (2020).

749 68. Newgard, C. B. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell Metab.* **25**, 43–56
750 (2017).

751 69. Yoshida, N. *et al.* Bacteroides spp. promotes branched-chain amino acid catabolism in brown fat and
752 inhibits obesity. *iScience* **24**, 103342 (2021).

753 70. Chang, C.-H., Lin, C.-H. & Lane, H.-Y. D-glutamate and gut microbiota in Alzheimer's disease. *Int. J.*
754 *Mol. Sci.* **21**, 1–17 (2020).

755 71. Barra, N. G. *et al.* Micronutrients impact the gut microbiota and blood glucose. *J. Endocrinol.* **250**, R1–
756 R21 (2021).

757 72. Xu, M. *et al.* Bacteroides vulgatus Ameliorates Lipid Metabolic Disorders and Modulates Gut Microbial
758 Composition in Hyperlipidemic Rats. *Microbiol. Spectr.* **11**, (2023).

759 73. Mills, R. H. *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome link Bacteroides vulgatus
760 proteases with disease severity. *Nat. Microbiol.* **7**, 262–276 (2022).

761 74. Yoshida, N. *et al.* Bacteroides vulgatus and Bacteroides dorei Reduce Gut Microbial Lipopolysaccharide
762 Production and Inhibit Atherosclerosis. *Circulation* **138**, 2486–2498 (2018).

763 75. You, H. J. *et al.* Bacteroides vulgatus SNUG 40005 Restores Akkermansia Depletion by Metabolite
764 Modulation. *Gastroenterology* **164**, 103–116 (2023).

765 76. Qiao, S. *et al.* Gut Parabacteroides merdae protects against cardiovascular damage by enhancing
766 branched-chain amino acid catabolism. *Nat. Metab.* **4**, 1271–1286 (2022).

767 77. Devkota, S. & Chang, E. B. Interactions between diet, bile acid metabolism, gut microbiota, and
768 inflammatory bowel diseases. *Dig. Dis.* **33**, 351–356 (2015).

769 78. Natividad, J. M. *et al.* Bilophila wadsworthia aggravates high fat diet induced metabolic dysfunctions in
770 mice. *Nat. Commun.* **9**, 1–15 (2018).

771 79. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563
772 (2014).

773 80. Chen, H. & Jiang, W. Application of high-throughput sequencing in understanding human oral
774 microbiome related with health and disease. *Front. Microbiol.* **5**, 508 (2014).

775 81. Wei, X. *et al.* Abnormal gut microbiota metabolism specific for liver cirrhosis. *Front. Microbiol.* **9**, 1–4
776 (2018).

777 82. Davis, B. C. & Bajaj, J. S. The Human Gut Microbiome in Liver Diseases. *Semin. Liver Dis.* **37**, 128–140
778 (2017).

779 83. Zhong, X. *et al.* Streptococcus, the Predominant Bacterium to Predict the Severity of Liver Injury in
780 Alcoholic Liver Disease. *Front. Cell. Infect. Microbiol.* **11**, 1–10 (2021).

781 84. Rigsbee, L. *et al.* Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable
782 bowel syndrome. *Am. J. Gastroenterol.* **107**, 1740–1751 (2012).

783 85. Santoru, M. L. *et al.* Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian
784 cohort of IBD patients. *Sci. Rep.* **7**, 1–14 (2017).

785 86. Qi, Y. feng *et al.* Intestinal Microbiota Is Altered in Patients with Gastric Cancer from Shanxi Province,
786 China. *Dig. Dis. Sci.* **64**, 1193–1203 (2019).

787 87. Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowel disease and
788 obesity. *Nat. Commun.* **11**, 1–12 (2020).

789 88. Chalmers, N. I., Palmer, R. J., Cisar, J. O. & Kolenbrander, P. E. Characterization of a Streptococcus sp.-
790 Veillonella sp. community micromanipulated from dental plaque. *J. Bacteriol.* **190**, 8145–8154 (2008).

791 89. Lai, Y., Masatoshi, H., Ma, Y., Guo, Y. & Zhang, B. Role of Vitamin K in Intestinal Health. *Front.*

792 *Immunol.* **12**, 1–19 (2022).

793 90. Habu, D. *et al.* Role of Vitamin K 2 in the Development of Hepatocellular Carcinoma in Women. *Jama*
794 **292**, 1–4 (2004).

795 91. Ramotar, K., Conly, J. M. & Louie, T. J. Production of menaquinones by intestinal anaerobes. *J. Infect.*
796 *Dis.* **150**, 213–218 (1984).

797 92. Karl, J. P. *et al.* Fecal menaquinone profiles of overweight adults are associated with gut microbiota
798 composition during a gut microbiota-targeted dietary intervention. *Am. J. Clin. Nutr.* **102**, 84–93 (2015).

799 93. Bentley, R. & Meganathan, R. Biosynthesis of vitamin K (menaquinone) in bacteria. *Microbiol. Rev.* **46**,
800 241–280 (1982).

801 94. Zhou, P., Li, X. & Qi, F. Identification and characterization of a haem biosynthesis locus in Veillonella.
802 *Microbiol. (United Kingdom)* **162**, 1735–1743 (2016).

803 95. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–
804 1024 (2017).

805 96. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).

806 97. Rubel, M. A. *et al.* Lifestyle and the presence of helminths is associated with gut microbiome composition
807 in Cameroonians. *Genome Biol.* **21**, 1–32 (2020).

808 98. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**,
809 59–65 (2010).

810 99. López-Almela, I. *et al.* Bacteroides uniformis combined with fiber amplifies metabolic and immune
811 benefits in obese mice. *Gut Microbes* **13**, 1–20 (2021).

812 100. Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**,
813 732–739 (2022).

814 101. Das, P., Babaei, P. & Nielsen, J. Metagenomic analysis of microbe-mediated vitamin metabolism in the
815 human gut microbiome. *BMC Genomics* **20**, 208 (2019).

816 102. Uebanso, T., Shimohata, T., Mawatari, K. & Takahashi, A. Functional Roles of B-Vitamins in the Gut
817 and Gut Microbiome. *Mol. Nutr. Food Res.* **64**, 1–10 (2020).

818 103. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial
819 communities with biobakery 3. *Elife* **10**, (2021).

820 104. Caspi, R. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids*
821 *Res.* **34**, 511–516 (2006).

822 105. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat.*
823 *Methods* **10**, 1196–1199 (2013).

824 106. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection
825 and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).

826 107. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal*

Complex Syst. 1–9 (2006).

108. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **36**, 1–13 (2010).

109. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, (2015).

110. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

111. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).

112. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Comput. Biol.* **12**, e1004977 (2016).

113. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).

114. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 1–27 (2021).

115. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).

116. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

117. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

118. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).

119. Sun, L., Ji, S., Yu, S. & Ye, J. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. in *IJCAI International Joint Conference on Artificial Intelligence* 1230–1235 (2009).

120. Tenenhaus, A. & Tenenhaus, M. Regularized Generalized Canonical Correlation Analysis. *Psychom. 2011 762* **76**, 257–284 (2011).

Data and code availability

Data used in this study was retrieved from either the curatedMetagenomicData package[95] (available at: https://waldronlab.io/curatedMetagenomicData/), supplementary tables and deposited data from 4 specific studies[11,12,32,53] (exact source of each table and accession numbers are listed in Supplementary Table S1). The R code used for this analysis, including

the MintTea pipeline, is available on GitHub (https://github.com/borenstein-lab/multi_view_integration_analysis).

Acknowledgements

We would like to thank the Borenstein lab members for helpful feedback and insightful discussions during the process. We thank all authors of the multiple microbiome studies whose data was included in our analysis, for generating invaluable data and making them publicly available. We are also grateful to Professor Uri Gophna for his important inputs, support, and assistance. Lastly, we also thank the Tel Aviv University Computer Science IT team for their technical support during computational work. This work was supported in part by National Institutes of Health grant U19AG057377, and Israel Science Foundation grant 2435/19 to EB. EM and IS are supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary notes

Supplementary note 1: Conceptual and quantitative comparison between MintTea and other related methods

As described in the main text, MintTea's objective is to identify cohesive disease-associated multi-omic microbiome modules, and as such aims to simultaneously capture both associations between omics and associations with the host disease state. Though many methods that focus on either of these two aims exist, only a few methods tackle both objectives simultaneously, each taking a unique approach to achieve these two tasks, with various, and sometime subtle, differences. In the table below, we outlined the capabilities and key objectives of several different approaches for identifying associations between the microbiome and a given disease and for pinpointing disease-associated signatures in the microbiome using multi-omic data. Importantly, in this comparison, we focused on the *conceptual* differences between these approaches, rather than on their specific performances as captured by some evaluation metric. We specifically indicated for each method whether it is explicitly designed for multi-omic data, whether it directly attempts to optimize classification ability, and whether it is easily interpretable in terms of pinpointing the most important features. We further noted whether each method also directly aims to optimize associations between the different omics (or between features from different omics). Assuming that informative features are identified by the method, we further indicated whether these features are partitioned or clustered into modules to generate distinct disease signatures, and also whether these modules are forced to include all omics (even at the cost of reduced association) or not. Lastly, given that a major limitation of many statistical and machine learning (ML) methods is their sensitivity to noisy and high-dimensional data (as is the case for microbiome data), which in turn impacts reproducibility of the obtained results, we indicated whether each method inherently incorporates mechanisms to reduce likely spurious results.

To span the space of potential approaches, we included in the comparison both ML approaches for multi-omic data, CCA extensions discussed herein (see Supplementary note 2), as well as a reasonable heuristic approach that can be used to achieve a similar goal, based on first filtering out non-informative features and then clustering the remaining features based on pairwise associations (using WGCNA for example, a highly popular method for clustering highly correlated features).

Method	Description	Specific for multi-omic data (≥ 2 omics)	Optimizes classification accuracy	Pinpoints informative features (interpretable)	Optimizes inter-omic associations	Partitions informative features into modules	Supports modules with a subset of omics[1]	Inherently controls for high-dimensional and noisy data
ML: no integration	A ML model is trained for each omic independently	✗	✓✗ [2]	✓✗ [3]	✗	✗	NA	✓
ML: early integration	A ML model is trained on a concatenation of all omics	✓	✓	✓✗ [3]	✗	✗	✓	✓
ML: late integration	A ML model is trained for each omic independently, with a final ensemble model for final classification	✓	✓	✓✗ [3]	✗	✗	✓	✓ [4]
Feature selection + WGCNA[118]	Features are filtered based on disease-association, and WGCNA is applied to find correlated modules	✗	✓✗ [5]	✓	✓✗	✓	✓	✓✗
sCCA / sGCCA / sMCCA[36,37]	A sparse version of classic CCA, and extensions of sCCA for ≥ 2 omics	✓	✗	✓	✓	✓	✗	✓✗ [6]
DIABLO[44]	An extension of sGCCA to a supervised framework. See Ref[44]	✓	✓	✓	✓	✓	✗	✓✗ [6]
MintTea	Our framework	✓	✓	✓	✓	✓	✓	✓

ML: Machine learning; CCA: Canonical correlation analysis; sCCA: Sparse CCA; sGCCA/sMCCA: sparse generalized/multiple CCA; WGCNA: Weighted correlation network analysis.
[1] As opposed to requiring each module to contain features from all omics.
[2] Does not leverage the availability of multiple omics, hence optimization is only within each omic.
[3] Depends on the selected ML algorithm and the selected explainability methods.
[4] Depends on the selected ML algorithm (e.g., whether bootstrapping/bagging mechanisms are involved).
[5] Features are filtered by their univariate association with disease. As a consequence, some of the discarded features may have actually contributed to classification accuracy when combined with others.
[6] The user must design/perform post-hoc analysis to determine the stability of components.

As evident from this comparison, only a few of these approaches directly target, as does MintTea, both aims described above. It is also important to note that while it may be possible, in principle, to first apply some ML method that optimizes predictability and then, as a second step, try to cluster informative features into modules (as in the "Feature selection + WGCNA" entry in the table above), such a two-step approach may yield different or partial modules, failing, for example, to identify features that are not sufficiently informative by themselves but that can be incorporated into informative modules to enhance predictability (as discussed in the main text).

A few other properties further differentiate MintTea from other approaches, including specifically its ability to generate modules that do not incorporate all omics (which may be crucial in future studies when the number of omics further increases), and its inherent subsampling and consensus detection procedure, which allows MintTea to better handle the extremely high-dimensionality and noise associated with microbiome data and ultimately to

918 provide more conservative results. To quantify the impact of this subsampling procedure on
919 the obtained modules' quality, robustness, and false discovery rates, we next compared
920 MintTea's modules to those produced by the two CCA-based method listed in the above table:
921 sMCCA (as implemented by the 'MultiCCA' function in the 'PMA' R package[36], version
922 1.2.1) and DIABLO[44] (Methods). For the comparisons, we used both the liver cirrhosis dataset,
923 as it was one of the largest datasets at hand (237 samples), and the CD (Franzosa, 2019) dataset.
924 To allow reasonable comparisons, we also tuned the penalty parameters of each method so that
925 the average module size (or number of non-zero loadings in MultiCCA/DIABLO) is
926 approximately 20 features (from all omics combined), with DIABLO and MultiCCA further
927 set to output 3 modules only.

928 We first compared the modules in terms of their ability to classify disease and the strength of
929 between-omic associations in unseen data. To this end, we ran each model on 70% of the
930 samples, serving as a training set, and then computed the resulting modules on the remaining
931 left-out 30% test set. For DIABLO and MultiCCA, the test set modules were computed using
932 the trained loadings (feature weights), and for MintTea we took the 1st PC of all features from
933 all omics (as done in the main text). Notably, DIABLO and MultiCCA result in one predicted
934 variable *per omic* (3 omics in this case) that are supposedly highly correlated, and so evaluation
935 is carried out on each such vector separately and then averaged. The entire procedure was
936 repeated 100 times to ensure the results consistency and significance. We found that modules
937 from all 3 methods achieved similar disease classification performance (with a notable, yet not
938 statistically significant advantage for MintTea), with an AUC of 0.7 (standard deviation
939 $\pm$0.14), 0.71 ($\pm$0.13) and 0.74 ($\pm$0.11) for DIABLO, MultiCCA and MintTea, respectively, in
940 the cirrhosis datasets (Supplementary Figure S6A). In the CD datasets, AUC's were 0.71
941 ($\pm$0.14), 0.71 ($\pm$0.11), and 0.73 ($\pm$0.16) (Supplementary Figure S6F). Similarly, modules were
942 comparable in their average correlation between features from different omics (Supplementary
943 Figure S6B, S6G).

944 Second, we compared each method's stability when trained on different data subsets. To this
945 end, we ran each method on two randomly selected data subsets of the same size and compared
946 the resulting modules, assuming that more similar modules imply higher robustness of the
947 method. The two sets of modules were compared both in terms of individual features included
948 in the modules (Supplementary Figure S6C, S6H) and in terms of pairs of features that co-
949 occurred in the same module in both sets (Supplementary Figure S6D, S6I). In the latter
950 comparison, we used a z-score to quantify how likely were *pairs* of features co-occurring in

951 the same module in one output to also co-occur in the same module in the other output at
952 random. The data subsets used ranged from 50% of the samples to 100%, with stability
953 estimations increasing as subset size increased, as expected. Overall, this analysis suggests that
954 MintTea is slightly more stable than the other methods in terms of stability of individual
955 features selected, while in terms of stability of feature pairs, MintTea is significantly superior
956 in one dataset, while DIABLO is superior in the other dataset, suggesting that different methods
957 may be favored for different datasets.

958 Importantly, one main caveat of CCA-based methods is that, by design, they always output a
959 set of transformations, even when no correlation structure in the data is expected to exist. When
960 using these methods, users are thus required to carefully evaluate the obtained modules
961 themselves, and be aware of potentially high false-positive rates especially when the number
962 of features greatly exceeds the number of samples (again, an unfortunately common setting in
963 microbiome studies). To demonstrate this point, we assessed the false discovery rates of all
964 methods by applying them to a shuffled version of the data where each omic was shuffled
965 independently (row-wise) so that any existing correlations between omics (and between omics
966 and disease) were removed. We kept only DIABLO and MultiCCA transformations that
967 resulted in significant between-omic correlations (FDR-corrected P value < 0.01), and only the
968 features in each module that had a non-redundant loading (> 0.001). Supplementary Figure
969 S6E and S6I illustrate the results obtained over 100 random samples of 70% of the data.
970 Evidently, even when strictly filtering the output of DIABLO/MultiCCA, their false discovery
971 rates were still significantly higher than that of MintTea, demonstrating that MintTea is
972 markedly more conservative and reliable.

973 In summary, we argue that MintTea's dual objective, both identifying associations with disease
974 and uncovering relationships between different omics, distinguishes it from several other
975 supervised and unsupervised methods, and that even when compared to closely related CCA-
976 based methods, MintTea represents a successful balance between multiple properties and
977 metrics while maintaining substantially lower false discovery rates. Nonetheless, given the
978 substantial variability between datasets' characteristics, and the particular complexity of multi-
979 omic data, it is important to acknowledge that no single method can be universally optimal,
980 and it is strongly advisable to explore and apply a range of different methods for comprehensive
981 analysis.

Supplementary note 2: Explanation and formulation of sCCA and its expansions

The intermediate integration approach implemented in this study is based on sparse canonical correlation analysis (sCCA) and specifically inspired by a recently introduced framework named DIABLO[44]. The modifications of sCCA which we implemented in the MintTea framework are described in the Methods section. Further explanations and formulations of sCCA and its relevant expansions are provided below.

Sparse canonical correlation analysis (sCCA)

Briefly, sCCA is a method used to identify sparse linear combinations of two sets of features describing the same set of samples, so that the resulting latent variables are maximally correlated with each other. Formally, we denote 2 standardized feature tables ('views') as $X^{(1)} \in \mathbb{R}^{n \times P_1}$ and $X^{(2)} \in \mathbb{R}^{n \times P_2}$, P_1 and P_2 being the number of features in each view respectively, and n denoting the number of samples. Sparsity is enforced using a lasso (ℓ_1) penalty function on the loading vectors, essentially making most of the loading values equal to zero, with a hyper-parameter λ determining the degree of sparsity. It is particularly helpful in cases where the number of features substantially exceeds the number of samples. sCCA then solves the following optimization problem iteratively per component (note, that the term component here refers to one set of omic-transformations) $h \in 1, \dots, H$, H being the total number of requested components (a user-defined hyper-parameter):

$$\underset{a_h^{(1)}, a_h^{(2)}}{\operatorname{argmax}} \ \operatorname{cov}\left(X_h^{(1)} \cdot a_h^{(1)},\ X_h^{(2)} \cdot a_h^{(2)}\right)$$
$$s.t. \left\|a_h^{(1)}\right\|_2 = 1\,,\ \left\|a_h^{(2)}\right\|_2 = 1\,,\ \left\|a_h^{(1)}\right\|_1 \le \lambda^{(1)},\ \left\|a_h^{(2)}\right\|_1 \le \lambda^{(2)}$$

In the above formula, $a_h^{(q)}$ is the loading vector for view q and component h, and $X_h^{(q)}$ is the residual matrix of the regression between $X_{h-1}^{(q)}$ *and* $a_{h-1}^{(q)}$. The use of residual matrices guarantees orthogonality of components within each block. We note that the use of *covariance* in the objective function, instead of *correlation* as in the original CCA, is clarified in Tenenhaus et al.[37] and is closely related to the PLS method[119].

Sparse generalized canonical correlation analysis (sGCCA) and DIABLO

To expand sCCA to $Q \ge 2$ views, i.e. generalizing sCCA, the objective function shown above is set to maximize the *sum* of correlations between all pairs of latent variables. Tenenhaus and

Tenenhaus (2011)[120] further introduced the notion of a design matrix, C, to weigh each pair of views according to the expected level of association between them, based on prior knowledge. Formally, a design matrix $C = \{c_{ij}\}$ is defined by the user, where $c_{ij} \in [0,1]\ \forall i,j \in 1,\dots,Q$, so that $c_{ij} = 1$ if view i and view j are expected to be closely related, and 0 if they are expected to be unrelated. Intermediate design values, between 0 and 1, are also supported. The resulting optimization problem for sGCCA therefore becomes:

$$\underset{a_h^{(1)},\dots,a_h^{(Q)}}{\operatorname{argmax}} \sum_{\substack{i,j=1 \\ i \neq j}}^{Q} c_{ij} \cdot \left(\operatorname{cov}\left(X_h^{(i)} \cdot a_h^{(i)},\ X_h^{(j)} \cdot a_h^{(j)}\right)\right)$$
$$s.t. \left\|a_h^{(q)}\right\|_2 = 1\ \ and\ \left\|a_h^{(q)}\right\|_1 \leq \lambda^{(q)}\ \ \forall q \in 1,\dots,Q$$

The optimization problem is solved by an efficient iterative algorithm described in detail in Tenenhaus et al.[37]. DIABLO extends sGCCA to classification problems by adding an additional "dummy" view, namely X^{Q+1}, with one or more indicator variables representing the label of each sample (here: host disease state). By doing so, the identified "components" reflect both inter-view relationships as well as associations with the disease. Furthermore, DIABLO substitutes the lasso's λ parameter with a "*keepX*" parameter indicating the exact number of variables to select from each view, for ease of use[44].

Overall, given Q view tables and a phenotype of interest, DIABLO's sGCCA variant returns one or more "components" which are sparse linear combinations of the original views, capturing a multi-omic set of features that are both closely related with one another and jointly associated with the phenotype. Importantly, there's a tradeoff between the two (a "correlation-discrimination tradeoff"), meaning that, at the extreme, some components may be strongly associated with the disease but have low inter-view correlations, and vice versa. Though DIABLO offers additional functionalities, mainly label prediction based on distances in the latent space, we utilized only its core sGCCA computation, as described above.

Supplementary figures

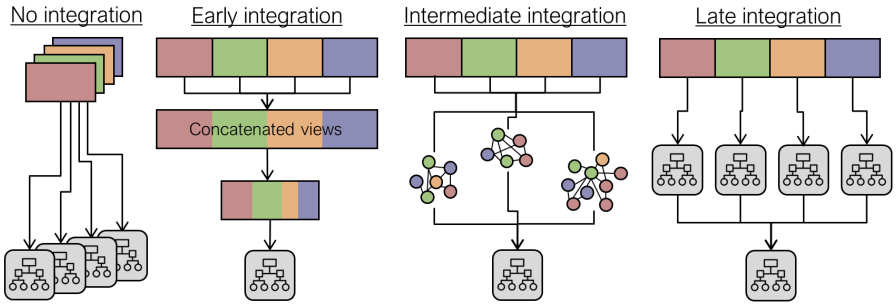


Figure S1: An illustration of different multi-view integration approaches. Different colors represent different omics.

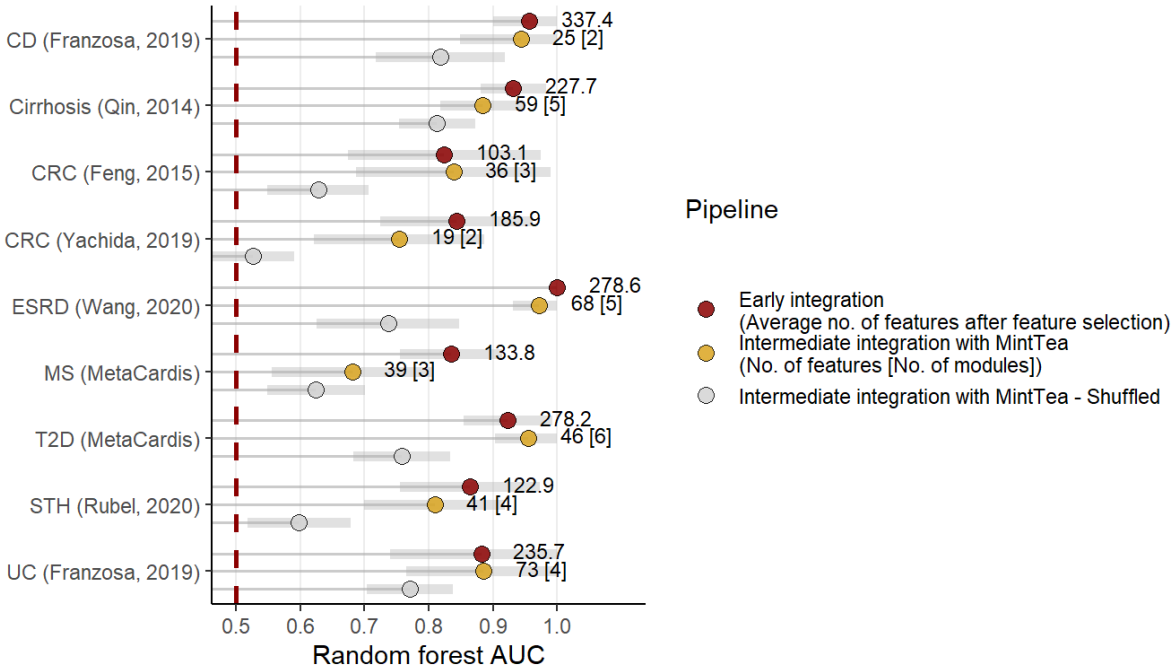
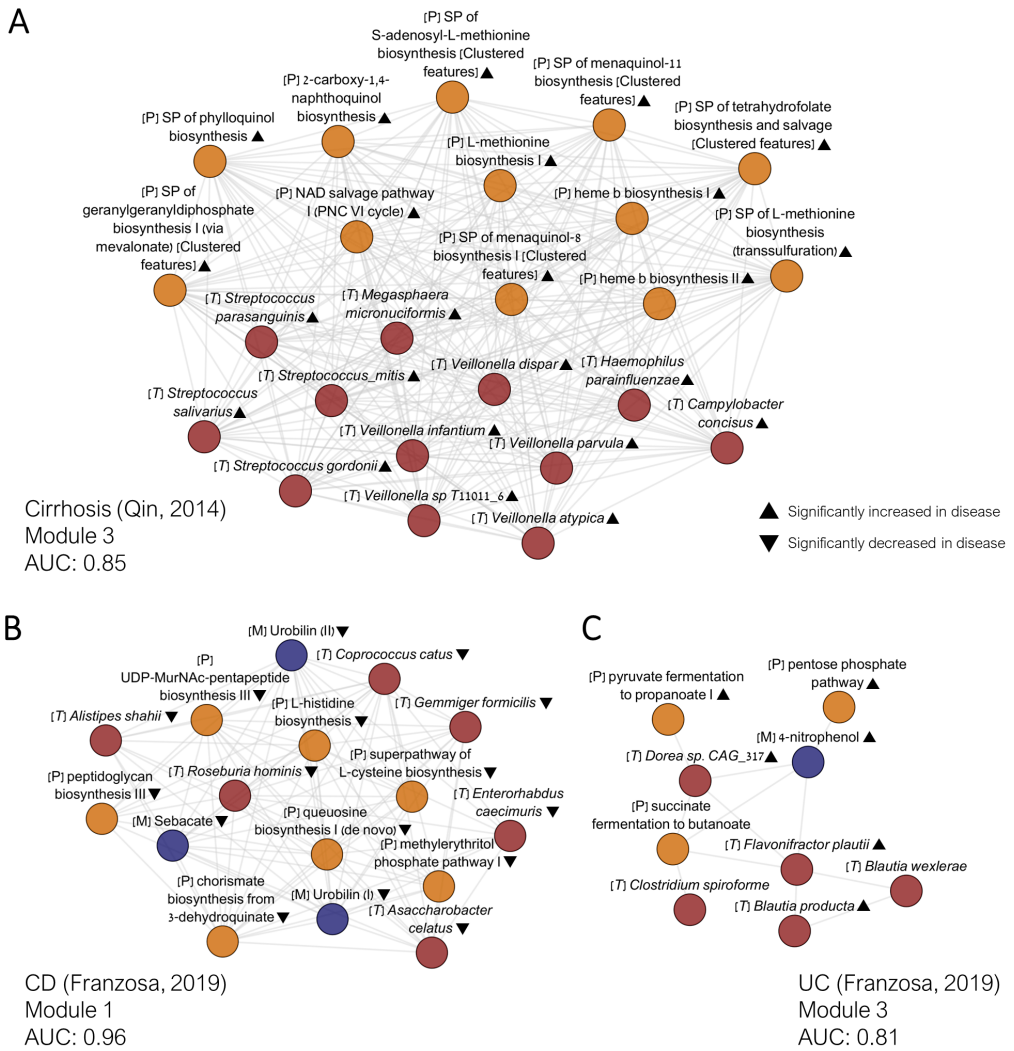


Figure S2: **Overall predictive power of multi-omic modules**. For each of the nine datasets, a random forest was trained on the first principal component of each consensus module in that dataset. Cross-validated AUC's of these models are shown in yellow, and compared to standard random forest models on raw features (in red, "early integration" approach) and models trained on random modules (grey). Numbers next to red dots represent the average number of features used for the model (averaged over cross-validation iterations). Numbers next to yellow dots represent the total number of raw features included in modules, and in square brackets the respective number of modules in that dataset.



1045

1046 **Figure S3: Additional disease-associated multi-omic modules**. Colors and format are as in

1047 Figure 2D-E.

1048

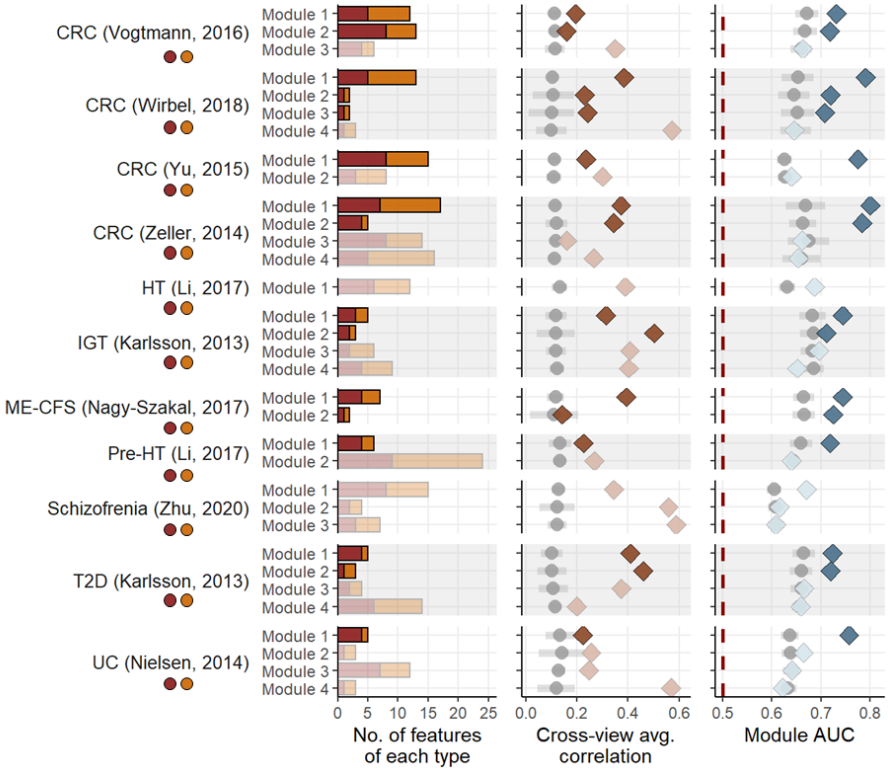


Figure S4: **Overview of MintTea modules identified in 11 additional shotgun datasets from the curatedMetagneomicData resource.** See legend of Figure 2A-C.

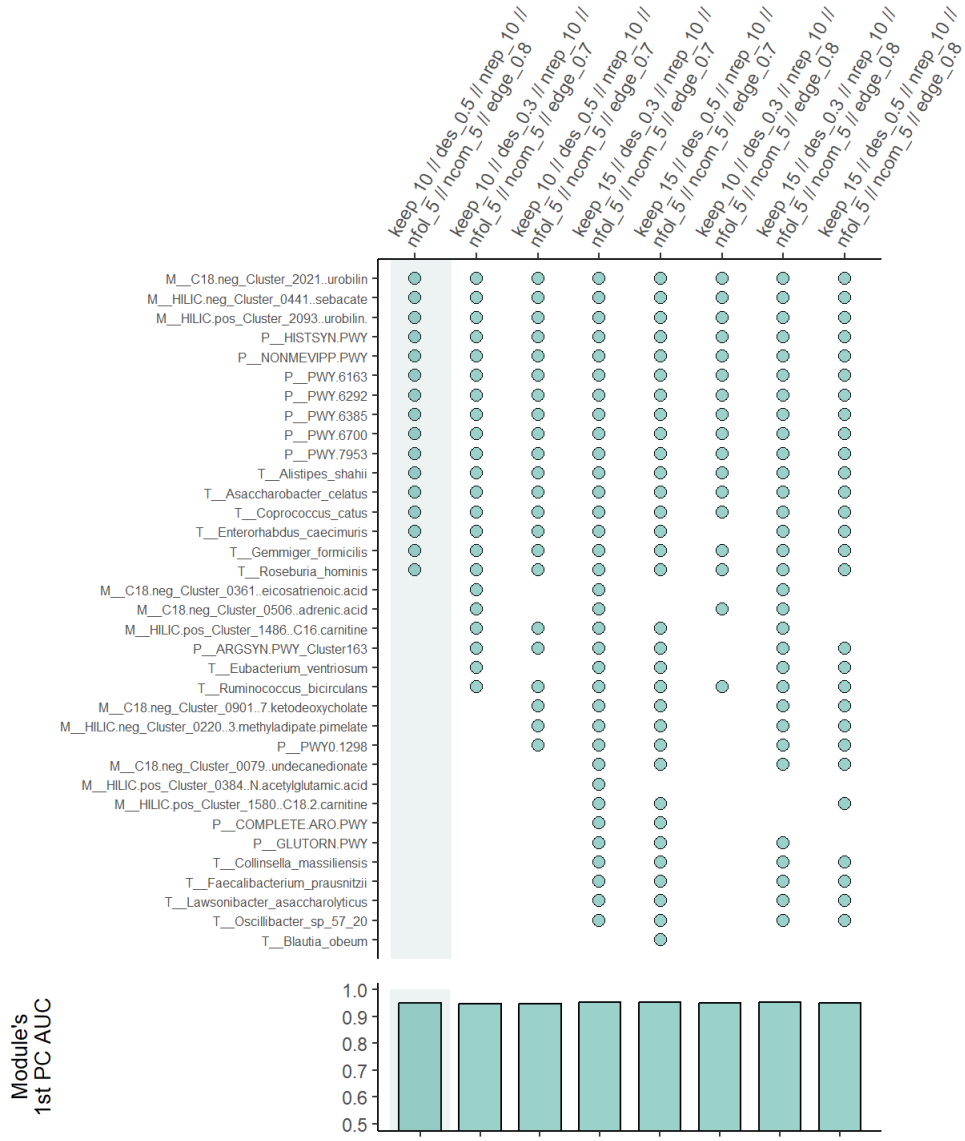


Figure S5: Example of sensitivity analysis for MintTea's pipeline parameters. The x axis lists various combinations of parameters required by our intermediate integration pipeline. A description of parameters is provided in Methods. The leftmost pipeline setting is the one used in the main results section of this study. Upper panel: Variations of module 1 from the CD (Franzosa, 2019) dataset (see also Supplementary Figure 3B), as outputted by each pipeline setting. The Y axis lists features from all omics, and points indicate that a feature was included in a multi-omic module that resulted from using that specific pipeline setting (on the x axis). Only modules overlapping with the original module (highlighted) are plotted, to illustrate how a module is generally robust to different pipeline settings. Bottom panel: Single module AUC's for each module appearing in the upper panel.

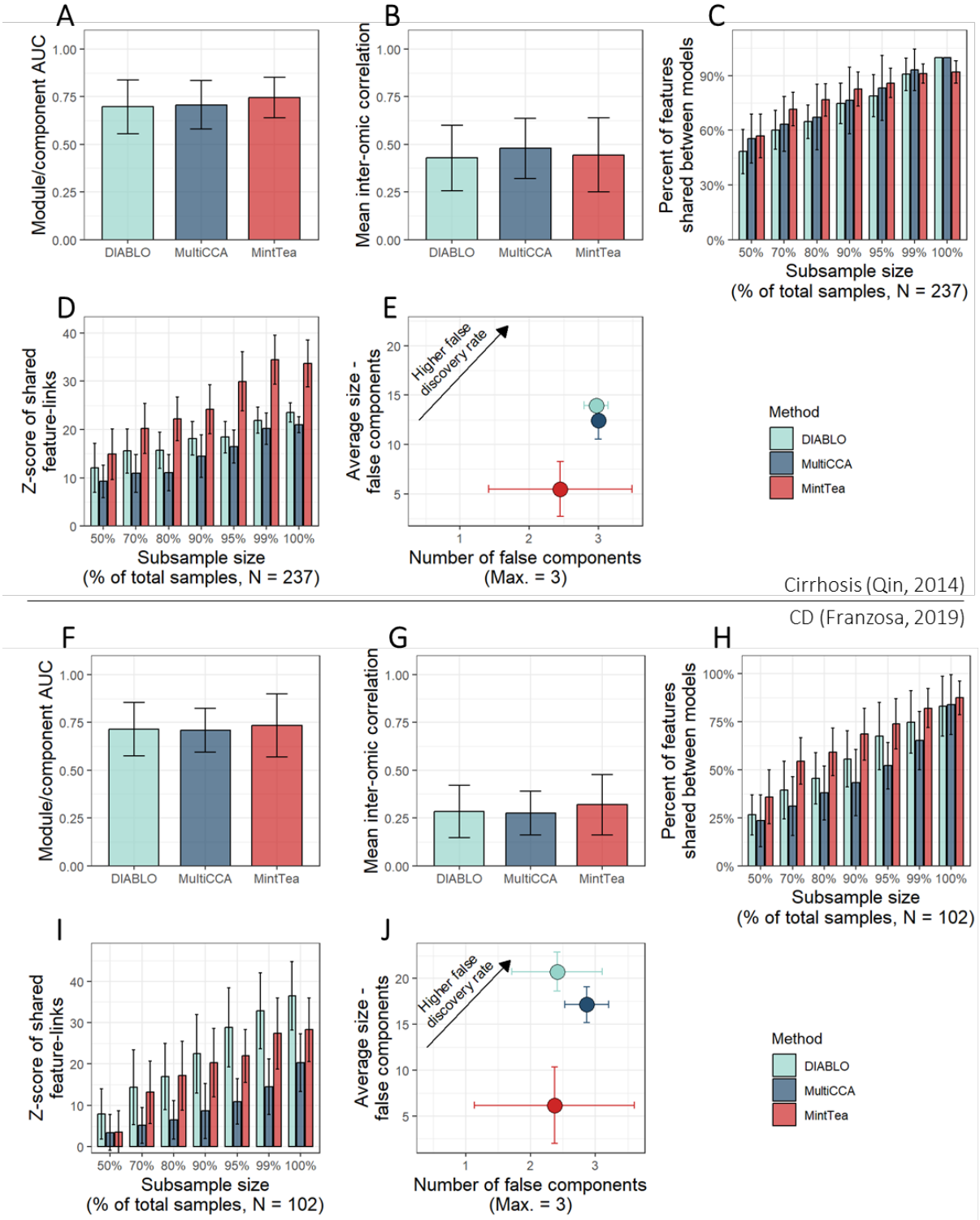


Figure S6: **MintTea comparison to DIABLO and MultiCCA**. Each panel represents a different comparison of MintTea to two other CCA-based methods: DIABLO[44] ('mixOmics' R package[106], version 6.18.1) and sparse multiple CCA (implemented by the MultiCCA function from the 'PMA' R package[36], version 1.2.1). The comparisons in panels A-E, and F-J, were made using the cirrhosis (Qin, 2014), and CD (Franzosa, 2019) datasets, respectively. (A,F) AUC of individual modules (or latent variables, in the case of DIABLO/MultiCCA)

1071 outputted by each method. (B,G) Correlations between features from different omics, within
1072 each module (or within each set of features assigned non-zero coefficients in
1073 DIABLO/MultiCCA). (C,D,H,I) Method stability assessment. For two random data subsets
1074 (ranging from 50% of the samples to 100%, noted on the x axis), modules (or sGCCA
1075 components) were identified and compared. The more the two sets of components were similar,
1076 both in terms of individual features included in the components (C,H) and in terms of pairs of
1077 features that co-occurred in a component in both sets (D,I), we conclude that the method is
1078 more stable. See Supplementary Note 1 for details on how we computed these estimates. (E,J)
1079 False discovery rate analysis. Here, all 3 methods were applied to data that was shuffled across
1080 samples within each omic. Higher number of identified modules as well as higher numbers of
1081 features within modules indicate higher false discovery rates.

1082 All results are averaged over 100 runs with random data subsets and error bars represent
1083 standard deviations. Penalty parameters were adjusted in each method so that the average
1084 module size (non-zero loadings in MultiCCA/DIABLO) is approximately 20 features (from all
1085 omics combined).

Supplementary tables

Table S1: Datasets included in the analysis

Table S2: Multi-view modules - Overview

Table S3: Multi-view modules - Detailed features

Table S4: Correlations between features within each module

Table S5: Random forest model performance results using early integration

Table S6: Random forest top feature contributors, early-integration models

Table S7: Overlapping modules across datasets

Table S8: Feature clusters