Inference of disease-associated microbial biomarkers based on metagenomic and metatranscriptomic data

Zhaoqian Liu[1], Qi Wang[1], Anjun Ma[2], Dongjun Chung[2], Jing Zhao[2], Qin Ma[2,#], Bingqiang Liu[1,3,#]

[1] School of Mathematics, Shandong University, Jinan, Shandong, 250100, China.

[2] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210, USA.

[3]Shandong National Center for Applied Mathematics, Jinan, Shandong, 250100, China.

[#]To whom correspondence should be addressed: Dr. Qin Ma, qin.ma@osumc.edu

Abstract

The identification of disease-associated microbial biomarkers is crucial for disease diagnosis and therapy. However, the heterogeneity, high dimensionality, and large amounts of microbial data present tremendous challenges for the discovery of biomarkers. In this paper, we present IDAM, a novel method for disease-associated biomarker inference from metagenomic and metatranscriptomic data, which does not require prior metadata. This method integrates gene context conservation (uber-operon) and regulatory mechanisms (gene co-expression patterns), using a mathematical graph model. We demonstrated the excellent performance of IDAM in biomarker inference, through the datasets from inflammatory bowel disease, melanoma, type 1 diabetes mellitus, and irritable bowel syndrome. The software of IDAM is freely available at https://github.com/OSU-BMBL/IDAM.

Keywords

metagenome; metatranscriptome; microbial biomarkers; gene regulation; gene context conservation; unsupervised learning.

Introduction

Trillions of microbes colonize the human body and play critical roles in multiple fundamental physiological processes, and changes in microbial composition and functions are intimately interwoven with multiple diseases, ranging from obesity to cancer and autism [1,2]. With the rapid development of sequencing technologies, the Human

Microbiome Project [3], integrative Human Microbiome Project [4], and Metagenomics of the Human Intestinal Tract project [5] have generated large amounts of microbial data related to numerous diseases. Considerable effort has been aimed at using these data to produce clinical insights and deepening the understanding of the mechanisms responsible for disease progression and treatment [6]. Some pathogenic organisms have been found to be closely related to disease states, and can be used as a potential markers of disease development [7,8]. These findings hint at a promising avenue for clinical application—disease diagnosis and therapy—via the use of key microbial taxa or functions as microbial biomarkers.

One of the classical paradigms is the identification of taxa and functions that can statistically significantly distinguish between two or more groups as biomarkers [9-11]. Given the success of machine learning in multiple scenarios, there have been several attempts to use random forest and deep feedforward networks in the identification of microbial biomarkers [12,13]. However, these approaches face several challenges. *First*, most of them focus on 16S ribosome RNA and metagenomic (MG) data. 16S ribosome RNA data suffers from low taxonomic resolution and an absence of functional information [14], which may induce spurious biomarker identification since diverse microbial communities from different patients can have very similar functional capabilities [14]. While MG data can provide functional information at all taxonomic levels, researchers have found that a large proportion of reads cannot be successfully mapped to existing reference genomes during taxonomic identification, which leads to the potential loss of valuable information [15]. Meanwhile, the functional analyses produced by metagenomic analyses are not necessarily equivalent to the true activities within the microbiome. Multiple metagenomically abundant genes in the human gut microbiome are significantly down-regulated at the transcriptional level [16]. *Second*, existing methods for biomarker inference are designed based on the strong assumption that data with sufficiently large sample sizes and accurate metadata information are available to design groups or train models. However, the current metadata related to many sequencing samples is incomplete, misleading, or not publicly available [17], which may lead to these methods being infeasible or causing bias in biomarker inference. These challenges drive the need to develop an easy-to-use and effective method that keeps up with current data for disease-associated microbial biomarker inference.

We proposed a novel method to **I**nfer **D**isease-**A**ssociated **M**icrobial functional biomarkers (**IDAM**), which to a large extent overcomes these limitations. The data we used was matched MG and metatranscriptomic (MT) data sequenced from microbial samples for assessing active functions within communities. We focused on a community-level functional analysis that goes beyond species level, so that the reads do not need to be mapped to reference genomes, and provided a comprehensive overview of functions performed by either individuals or different microbial assemblages. We investigated the identification of biomarkers by the detection of gene modules with specific co-expression patterns across certain subsets of samples, without requiring prior metadata. Here, we accept that one cannot fully characterize complex functional activities based solely on ensemble co-expression, by which gene expression is assessed using reads from various species (**Fig. 1a**) and genes with no biological relevance may be clustered together (**Fig. 1b**). Fortunately, previous studies have suggested that functionally associated gene modules, such as uber-operon structures [18,19], are generally highly conserved during evolution. This finding motivated us to integrate gene context conservation to study cross-species functional-related mechanisms, thereby alleviating the bias in biomarker inference. We used uber-operon structures as the evolutionary footprints of cross-species, functionally related genes (**Fig. 1c**). We believe that genes are more likely to be functionally related if they have similar expression patterns and share the same uber-operon during evolution.

We applied IDAM to published matched MG and MT datasets of inflammatory bowel diseases (IBD) (*n* = 813; including ulcerative colitis (UC) and Crohn's disease (CD)) [20]. The results suggested that IDAM performed better in key biomarker inference than popular tools, including both supervised and unsupervised approaches. We further demonstrated the generalizability and good performance of IDAM through other publicly available datasets associated with melanoma [21], type 1 diabetes mellitus (T1DM) [22], and irritable bowel syndrome (IBS) [23]. Hence, we believe that IDAM can be a highly advantageous method for biomarker inference based on MG and MT data. This approach may be valuable for investigating the role of the microbiome in human diseases and improving disease diagnosis and treatment in clinical practice.

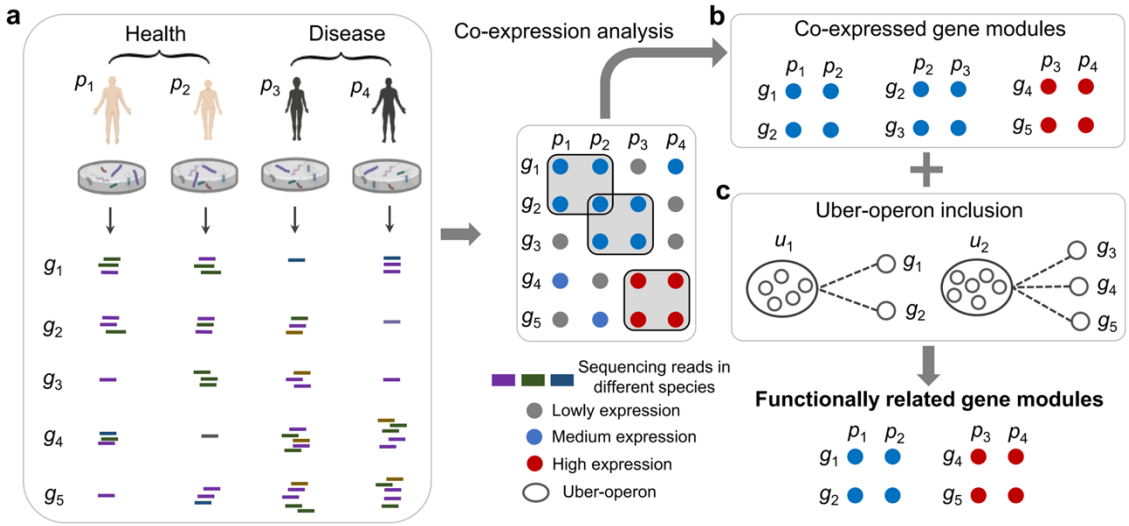


Fig. 1 Key ideas of IDAM. **a.** Sequencing data from four samples (p_1 and p_2 from healthy cohorts, and p_3 and p_4 from disease cohorts). Since MG shotgun sequencing does not isolate microorganisms, the reads from different species (represented by purple, green, and blue) were jumbled together. We therefore could not distinguish them easily at the species level, but we could observe the community-level expression of each gene based on the merged reads from all species. The microbial community transcript number (MT data) can be corrected using the underlying genomic copy number (MG data) in practice. Here, we simply show the number of transcriptomic reads as expressions. The colors of the circles (grey, blue, and red) indicate the expression level of the genes within each sample. It is assumed that g_1 and g_2 are functionally related in healthy samples, while g_4 and g_5 are functionally related in disease samples. **b.** Co-expression modules. There were three co-expression modules, which are defined as the gene and sample subsets with coordinated expression patterns. They were not equivalent to known functional relationship. **c.** Functional modules identified by integrating uber-operon structures and expression similarity. Each ellipse, including multiple genes, represents an uber-operon. We searched for modules consisting of genes with both co-expressed patterns and shared uber-operons, which resulted in two identified modules. Therefore, we inferred there were functional relationships between g_1 and g_2, as well as g_4 and g_5, which were consistent with the ground truth.

Results

IDAM: a metagenomic and metatranscriptomic analysis framework for disease-associated microbial biomarker inference.

We first present a formulation of the problem addressed by this article, and then provide an algorithmic overview of IDAM.

Computational formulation for biomarker inference

The inference of biomarkers can be mathematically formulated as an optimization problem. For matched MG and MT data, a gene-sample expression matrix $A = (a_{ij})_{m\times n}$ can be constructed, which includes m genes ($g_1, g_2, \ldots, g_m$) and n samples ($p_1, p_2, \ldots, p_n$). a_{ij} ($1\leq i \leq m$, $1\leq j \leq n$) represents the expression of gene g_i within sample p_j, which is

calculated by normalizing the RNA abundance (MT data) against the DNA abundance (MG data) of g_i within sample p_j [24]. Based on the matrix, the co-expression patterns present in gene and sample subsets can be detected by identifying local low-rank submatrices [25]. Given the constraints of gene context conservation, we integrated the submatrices with a graph model to maximize the number of genes connecting to a single uber-operon. Specifically, we define a bipartite graph $G = (V_1, V_2, E)$, where each node $u \in V_1$ represents an uber-operon. Each node $v \in V_2$ represents *a* gene in A, and the edge $e \in E$ connecting a node $u \in V_1$ to a node $v \in V_2$ exists if the gene v belongs to uber-operon u (Methods). Using this graph and matrix A, the biomarker inference problem can be formulated as the identification of a set of local low-rank submatrices within A that maximizes the number of connected components between the corresponding gene modules and V_1 on G. We maximize the number of connected components between gene subsets and V_1, to minimize the heterogeneity of the gene subsets within the uber-operons, thereby ensuring that the genes within each identified submatrix are as functionally related as possible. Based on the identified submatrices within A, each of which contains a subset of genes and samples, the gene modules can be regarded as biomarkers of the shared phenotypes of the sample sets. However, this problem is theoretically intractable (NP-hard, **Supplementary Material Section 1**). Hence, instead of solving this problem in a globally optimized way, we developed a heuristic algorithm, IDAM, as an approximate optimal solution.

Overview of IDAM

The main goal of IDAM is to identify disease-associated microbial biomarkers (**Fig. 2a**). First, the community-level expression within each sample is assessed using HUMAnN2 based on matched MG and MT data [26]. This step provides a gene-sample expression matrix, each entry of which represents the expression of a gene within a sample. Second, IDAM combines the expression similarity and gene distribution within uber-operons to measure the likelihood that each pair of genes having the same function. This step provides a list based on the combined assessment for module initialization (**Fig. 2b**). Third, the gene pairs in the list are used as seeds, and each seed is expanded by iteratively adding other functionally related genes, to generate modules until the new module starts to become smaller than previously identified modules (**Fig. 2c**). Finally, the gene modules and corresponding sample sets are considered as the final output of the algorithm. We regard the gene modules as functional biomarkers of the most common

phenotype or state of the corresponding samples. Details are available in the **Methods** section.

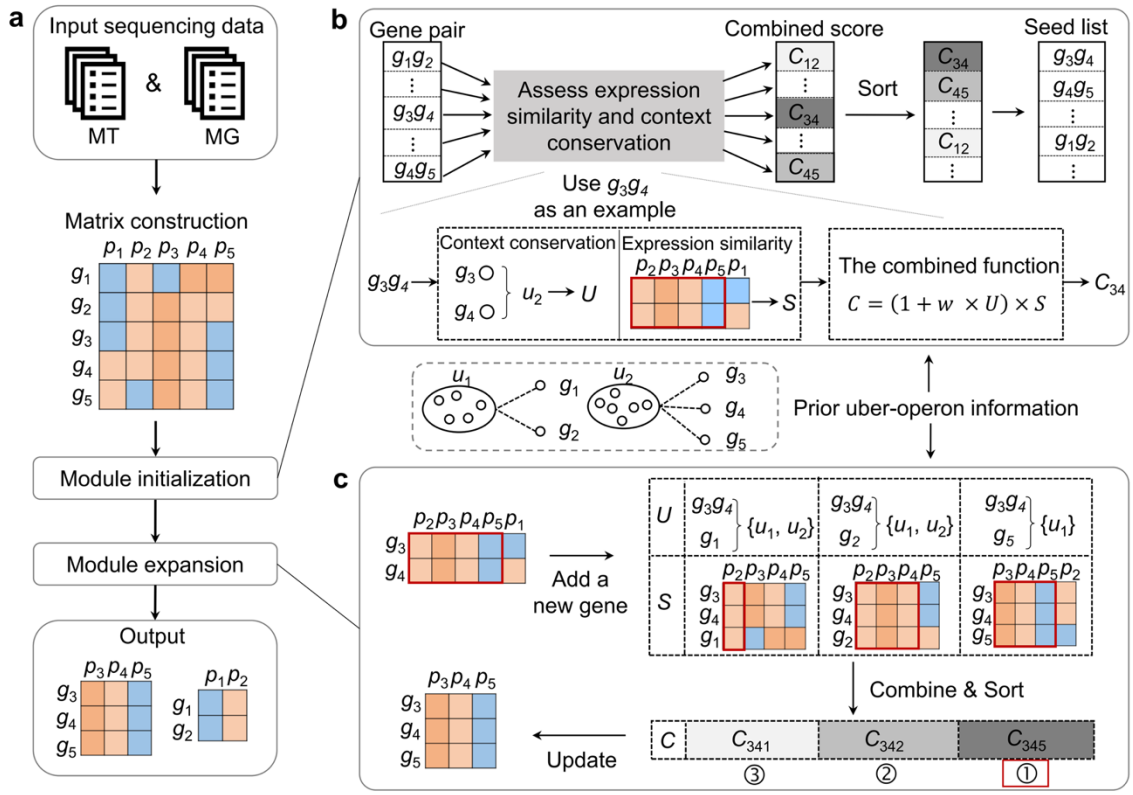


Fig. 2 Workflow of IDAM. **a.** The flow chart. We consider the matched MG and MT data of multiple samples as the input, from which we obtain an expression matrix. By module initialization and expansion, the modules will be generated as the output. **b.** Module initialization. For each gene pair, we assess both expression similarity and context conservation to produce a combined score. Let us use g_3g_4 as an example. Based on known uber-operon structures, the two genes belong to u_2. We assign a reward U for g_3g_4 as a context conservation measurement. The expression similarity is assessed according to the number of samples in which the rows are identical, and we obtain a score, S (Methods). Finally, U and S are combined via a function C (w is a tuning parameter), generating a score, C_{34}. All gene pairs are sorted in a decreasing order based on the combined scores. **c.** Module expansion. We use gene pairs in the list as the initial modules and expand them gradually. We iteratively add a new gene with the highest combined score into a module until the updated module does not get bigger. Let us use g_3g_4 as an example. First, we add three genes (g_1, g_2, and g_5) into g_3g_4, forming three new modules (denoted as $g_3g_4g_1$, $g_3g_4g_2$, and $g_3g_4g_5$). Then, combined score of each module is assessed (Methods). Since the module $g_3g_4g_5$ has the highest score, we replace the previous module with this one. Next, we iteratively add a new gene (g_1 and g_2) to the current module and implement the same processes. After comparing the combined scores, we find $g_3g_4g_5g_2$ to have the highest score, but the module is of size 4×2, which is smaller than $g_3g_4g_5$ (3×3). Therefore, we stop here and the module $g_3g_4g_5$ will be provided as the final output.

The biomarkers inferred by IDAM distinguished different phenotypes and enabled the discovery of disease subtypes

To assess the usefulness of biomarkers from IDAM, we assessed its phenotype classification performance using identified biomarkers. We applied IDAM to a publicly available dataset of patients with IBD and controls (referred to as non-IBD). The following analyses are based on the dataset described in the Methods section unless otherwise specified. By assessing the community-level expression within each sample, we obtained a matrix consisting of the expression of 941,785 genes across 813 samples. Based on this matrix, IDAM found 41 gene modules with 2,340 to 41 genes, covering 19,327 genes (**Supplementary Table 1**). Each gene module was regarded as a biomarker, giving rise to 41 identified biomarkers. Ideally, samples with the same biomarker should have highly consistent phenotypes. We used Z-tests to evaluate the phenotypic consistency of samples in each module. The 32 biomarker-associated sample sets were consistently enriched for a particular phenotype, with 19 enriched for IBD and 13 for non-IBD (**Fig. 3a**, **Supplementary Table 2**). We recognize that microbial communities can be influenced by multiple factors, such as age, race, and diet. These confounding factors can lead to high inter-individual heterogeneity, make disease-associated microbial characteristics less clear, and increase the risk of false positives or negatives in biomarker inference. Hence, we further assessed whether the biomarkers identified using IDAM had reliable associations with phenotypes, by applying a propensity score-based matching for each biomarker to confirm its association with phenotypes. We constructed paired samples by matching the propensity score estimated from other covariates, such as age, gender, and comorbidities. The Wilcoxon signed-rank test based on the paired samples indicated that 82.93% (34 of 41) biomarkers were statistically significantly related to a specific phenotype, with a Benjamini-Hochberg-adjusted p-value < 0.05 (**Fig. 3b**, **Supplementary Table 3**), demonstrating the high reliability of the biomarkers identified using IDAM.

We observed a statistically significant enrichment of the disease phenotype within each module. The biomarkers identified could recognize the subtypes of IBD, with 11 corresponding to UC and eight corresponding to CD (**Fig. 3c**). The phenotype specificity is shown in **Fig. 3d**. Each of Module 1, Module 17, and Module 30 is enriched for a particular phenotype. Because of space limitations, we show the expression of the top 3% of genes in each module: 17/618, 19/689, and 14/521 for Modules 1, 17, and 30, respectively. IDAM detected the checker-board substructures from the original expression data. Biomarker 1, corresponding to Module 1, is specific to non-IBD, while

Biomarkers 17 and 30, corresponding to Modules 17 and 30, are specific to UC and CD, respectively, suggesting the capability of IDAM to uncover biomarkers of disease subtypes. These biomarkers were associated with 168 pathways (**Supplementary Table 4**), of which 27 were also identified as phenotype-differential pathways based on a pathway abundance analysis of HUMAnN2 (**Methods**, **Supplementary Table 4**). For example, there was a significant enrichment in short-chain fatty acid metabolism (e.g., butyrate metabolism) in non-IBD-associated biomarkers. The L-methionine biosynthesis, superpathway of L-citrulline metabolism, and superpathway of L-lysine, L-threonine, and L-methionine biosynthesis were enriched in non-IBD-associated biomarkers, while $preQ_0$ biosynthesis was enriched in IBD. The superpathway of lipopolysaccharide biosynthesis was enriched in CD-associated biomarkers, and methylphosphonate degradation was enriched in UC-associated biomarkers. These findings are in keeping with previous observations [27-30], indicating that the biomarkers from IDAM were valuable in distinguishing phenotypes.

We also compared IDAM with several decomposition-based and biclustering-based approaches, including ICA [31], QUBIC [32], ISA [33], and FABIA [34]. These methods have not been applied to microbial data for biomarker inference, but they may reveal the functional characteristics of biomarkers of different phenotypes without the need for prior metadata. The results suggested that there were 34, 30, and 13 biomarkers identified by QUBIC, ICA, and FABIA, respectively, while no biomarker was detected by ISA. We further evaluated the phenotype consistency of biomarker-associated sample sets. While 78.05% (32/41) sample sets from IDAM had a consistent phenotype, this number was only 61.76% (21/34), 33.33% (10/30), and 30.77% (4/13) for QUBIC, ICA, and FABIA, respectively (**Fig. 3e**). This finding indicated that the biomarkers inferred by IDAM can cluster samples with the same phenotypes more effectively than other tools. This finding was not surprising, since IDAM took both gene expression and gene context conservation into consideration, thereby identifying more biologically meaningful biomarkers, while others have used only gene expression. We also compared the running time of these methods (**Fig. 3f**). IDAM took seven hours, QUBIC and FABIA took 16 and 22 hours, respectively, while ISA was terminated after 24 hours without results. ICA took eight minutes; however, the application power of ICA was limited by the relatively low accuracy; only one third of the biomarkers could distinguish between phenotypes. Collectively, the biomarkers

identified by IDAM performed best in distinguishing the phenotypes and can provide leads for new disease subtypes or states in clinical applications.

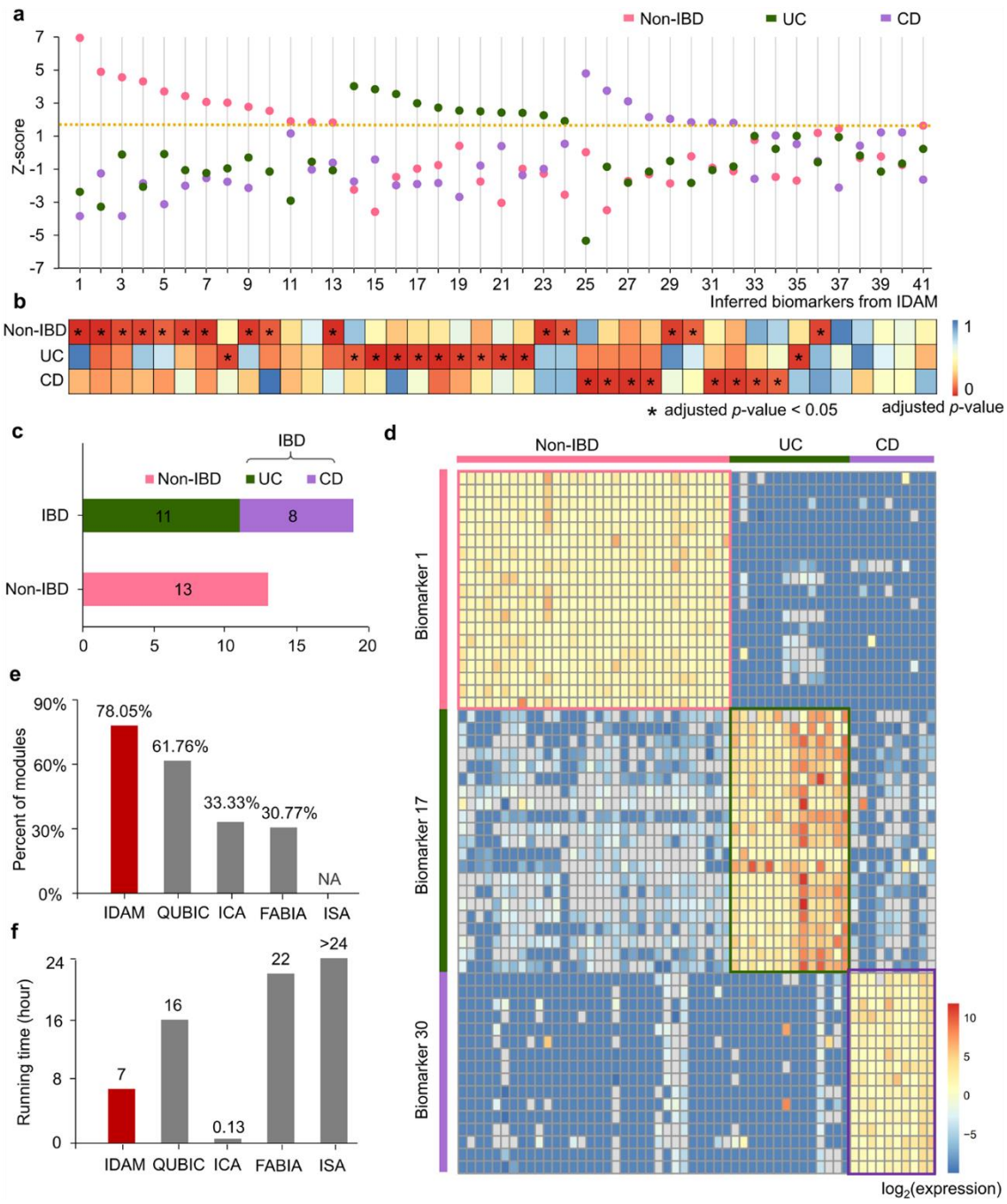


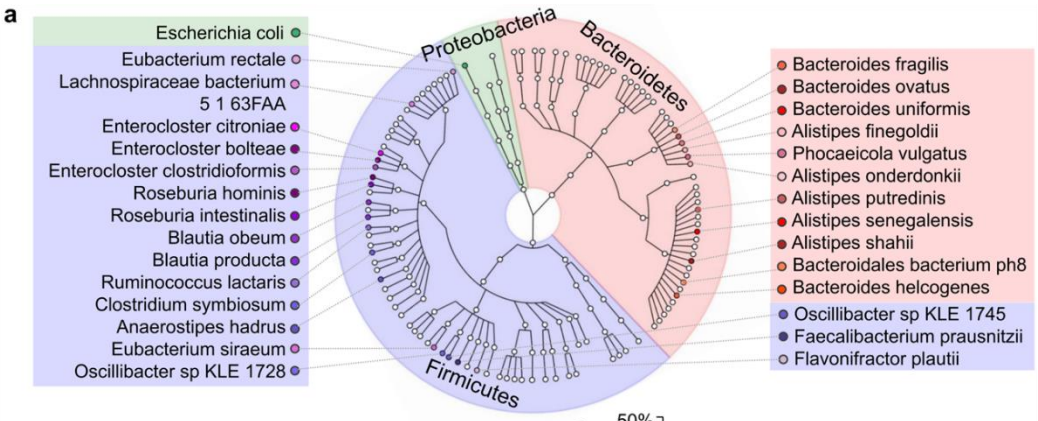
Fig. 3 Performance of biomarkers inferred by IDAM in distinguishing phenotypes. **a.** Z-scores of non-IBD, UC, and CD, shown as pink, green, and purple, respectively. Each column represents a biomarker. The yellow horizontal line indicates the critical value of 1.645. The color of circles above this line means that the corresponding phenotype exhibited strong enrichment within the biomarker-associated sample sets. **b.** Adjusted *p*-values of each biomarker with respect to different phenotypes. *p*-values were obtained using the Wilcoxon signed-rank test based on

propensity score matching, followed by multiple testing adjustments using the Benjamini-Hochberg procedure. Each row represents a biomarker, and the color indicates the adjusted *p*-value. * indicates an adjusted p-value of < 0.05. **c.** Number of biomarkers identified for each phenotype. A total of 19 biomarkers corresponded to IBD (11 UC and 8 CD) and 13 corresponded to non-IBD. **d.** An example of the gene modules associated with phenotypes, in which Modules 1, 17, and 30 are shown. The color of each cell in the heatmap indicate the log2 value of the expression. **e.** Percent of biomarkers associated with samples enriched in certain phenotypes, identified by IDAM and competing algorithms. NA means no results from ISA. **f.** Running time of IDAM and competing algorithms.

Species-level analysis suggested that the biomarkers inferred by IDAM were biologically associated with phenotypes

Since differences in community-level functions ultimately imply species-level functional differences, we traced genes within each biomarker back to individual species. These genes were from 111 species, which were dominated by *Bacteroidetes* and *Firmicutes*, followed by *Proteobacteria* at the phylum level (**Fig. 4a**). These observations were consistent with previous research into IBD-associated species [30]. We further classified the species according to biomarker-associated phenotypes and obtained 37, 64, and 13 species associated with non-IBD, UC, and CD, respectively (**Supplementary Table 5**). Three of these species overlapped between UC and CD: *Flavonifractor plautii*, *Bacteroides vulgatus*, and *Bacteroides fragilis*. We collected IBD-related species (34 for non-IBD, 32 for UC, and 90 for CD) from Peryton, a database of microbe-disease associations, and previous taxonomic studies into IBD [35-38] (**Supplementary Table 6**). Among them, a total of 32 species were captured by IDAM (**Fig. 4a**), including 17 for non-IBD, six for UC, and nine for CD (shown in red in **Supplementary Table 6**). The three overlapped species from IDAM have been reported to be associated with both UC and CD [35-37], suggesting that the IDAM results are reliable.

We compared IDAM with QUBIC [32], ICA [31], FABIA [34], and LEfSe [11]. ICA, QUBIC, and FABIA are not dependent on metadata, while LEfSe identifies differentially abundant species as biomarkers based on prior phenotype information. Detailed results are shown in **Fig. 4b**. The average consistency with collected species across the three phenotypes was significantly higher for IDAM (41.52%), than for QUBIC (19.82%), LEfSe (18.38%), ICA (17.62%), or FABIA (4.74%) (**Methods**, **Fig. 4c**). The species-level analysis showed that the biomarkers inferred by IDAM were biologically associated with phenotypes.



b

Method	Non-IBD (34)	UC (32)	CD (90)
IDAM	**17/37 (45.95%)**	**6/64 (9.38%)**	**9/13 (69.23%)**
QUBIC	12/36 (33.33%)	2/35 (5.71%)	10/49 (20.41%)
ICA	18/95 (18.95%)	6/40 (15.00%)	7/37 (18.92%)
FABIA	0	7/151 (4.64%)	11/115 (9.57%)
LEfSe	11/37 (29.73%)	2/37 (5.41%)	4/20 (20%)

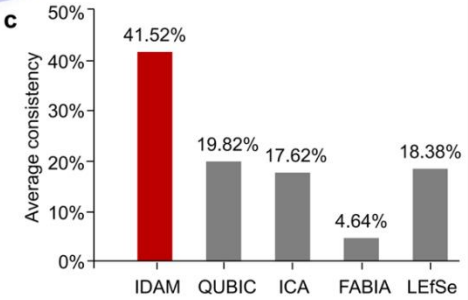


Fig. 4 Species-level analysis of identified biomarkers. **a.** Phylogenetic tree of the species associated with biomarkers identified by IDAM. Colored dots indicate different species that have been validated by previous studies. **b.** Results of species-level analysis by different methods. The numbers in the header (34, 32, and 90) represent the number of species associated with the three phenotypes. Each cell in the table is represented in the form of *X*/*Y* (*Z*), where *Y* represents the number of species associated with a particular phenotype based on the different methods, among which those consistent with the collected species are represented as *X*, and *Z* indicates the consistency index of each method associated with the phenotypes. **c.** Mean consistency across three phenotypes was used as the average consistency index for each method. A higher value indicates that identified biomarker-associated species are more reliable.

Integration of gene context conservation substantially improved the performance of biomarker inference

Our hypothesis is that genes with similar expression patterns and high context conservation during evolution are likely to be functionally related. To evaluate the validity of this hypothesis, we sought to identify biomarkers without using gene context conservation in IDAM (referred to as IDAM without gene context conservation). The biomarkers identified by this version were significantly less capable of distinguishing different phenotypes than those from IDAM (**Fig. 5a**). IDAM without gene context conservation detected 34 gene modules from the expression matrix, of which only 61.76% (21/34) sample subsets were of consistent phenotypes (the percentage of IDAM is 78.05%). Pathway analysis suggested that 67.65% (23/34) biomarkers from IDAM without gene context conservation were significantly enriched in MetaCyc functional categories, a figure which is significantly lower than the 82.93% (34/41) of IDAM. These

findings indicated that the integration of uber-operon structures contributed to the identification of informative biomarkers.

We explored the underlying reasons for the superior performance of integrating gene context conservation. We observed the distribution of the genes in the uber-operons for each biomarker from both IDAM and IDAM without gene context conservation. Overall, the genes within biomarkers that cannot significantly enrich certain functions tended to be spread out among more uber-operons. This finding supports our hypothesis that functionally related gene modules are of high context conservation. As an example, we investigated Module 32 from IDAM without gene context conservation (abbreviated as 'Module 32' hereinafter, **Fig. 5b**). Since the 42 genes within the module had coordinated expression within the 80 samples, they would be clustered together if we merely considered co-expression. However, these genes were not significantly functionally related, and were from six uber-operons (**Fig. 5b & c**). They cannot reflect the unique functional characteristics of a particular phenotype. The 80 samples within Module 32 were evenly scattered over different phenotypes: 22 in non-IBD, 22 in UC, and 36 in CD. When we took gene context conservation into consideration, a new module occurred (Module 31 from IDAM, abbreviated as 'Module 31' hereinafter). This module consists of 85 genes and 34 samples, and there was considerable overlap with Module 32 (**Fig. 5b**). The module was dominated by an uber-operon that is closely associated with purine metabolism [19], and the significantly enriched functions are shown in **Fig. 5c**. Previous studies have shown that the imbalance between the biosynthesis and degradation of purines can produce excessive uric acid in the gut, leading to the occurrence of IBD [39]. The 34 samples were enriched in phenotype CD. This finding demonstrated the value of gene context conservation in the clustering of genes with biological relevance, which improves the ability of IDAM for biomarker inference.

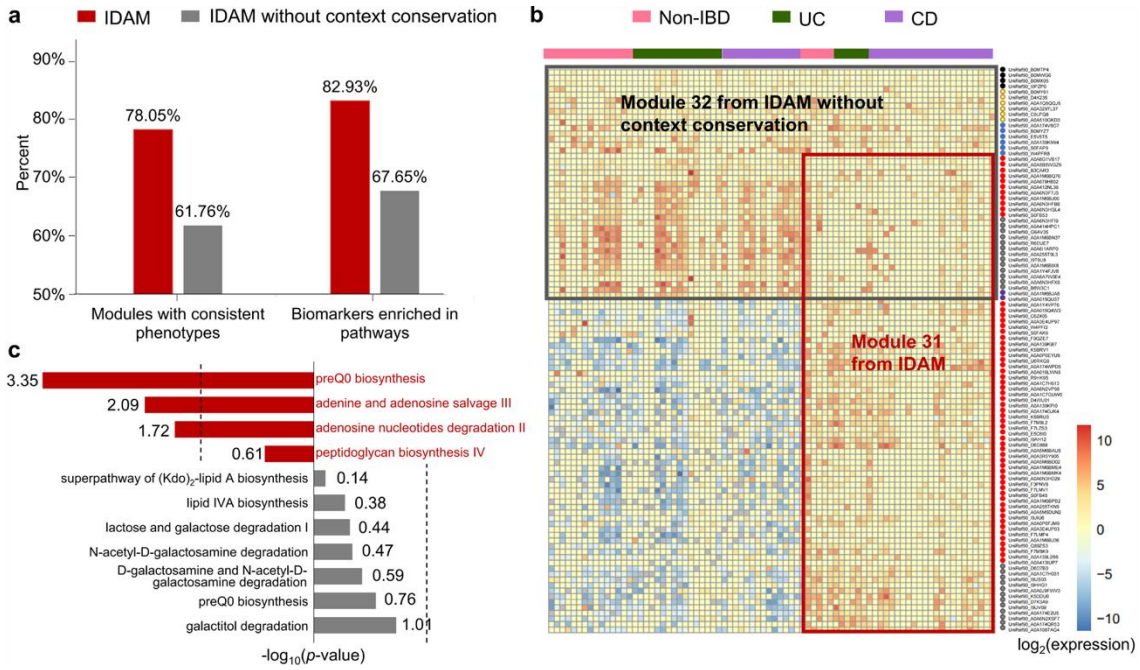


Fig. 5 Analysis of the contribution of gene context conservation. **a.** Comparison between IDAM and IDAM without gene context conservation. The red bars indicate the results from IDAM, while the grey bars indicate the results from IDAM without gene context conservation. **b.** A specific case of identified modules. Each row represents a gene, and each column represents a sample (a total of 101 genes and 80 samples). The grey rectangle highlights Module 32, identified by IDAM without gene context conservation, and the red highlights Module 31, identified by IDAM. The colored bar at the top represents the different phenotypes of the corresponding samples. The colored circles on the right represent genes that belong to the same uber-operons. **c.** Pathways enriched for Module 32 and Module 31. The pathways in grey indicate the pathways enriched for Module 32, while red indicates those enriched for Module 31. The dotted vertical lines indicate the critical value, with a *p*-value of 0.05. The pathways located outside the dotted lines correspond to significantly enriched pathways.

IDAM showed good generalizability in biomarker inference and provided insights into melanoma-related progression-free survival after immunotherapy

To demonstrate the power of IDAM, we applied it to an integrated dataset containing 60 samples from four different cohorts: cohorts with melanoma, T1DM, without T1DM (non-T1DM), and IBS (**Supplementary Material Section 2 & Supplementary Table 7**) [21-23]. The data processing included quality control and the measurement of the community-level expression of each sample (**Supplementary Material Section 2**). A matrix with the expression of 470,305 genes across 60 samples was produced. Based on this matrix, IDAM inferred 55 biomarkers covering 13,145 genes.

Analogous to the analyses of the IBD-associated dataset, we first evaluated the phenotype consistency of each biomarker-associated sample set using Z-tests and found

that 32 of 55 (58.15%) were consistently enriched for a particular phenotype: seven enriched for T1DM, 18 for non-T1DM, and seven for melanoma. Comparison with other methods, QUBIC, ICA, FABIA, and ISA, showed that the biomarkers inferred by IDAM could distinguish between the phenotypes most effectively (**Fig. 6a**, **Supplementary Table 8**). We then found that 20.43% of the species from IDAM were consistent with the collected disease-specific species, and the average consistency across the four phenotypes was the highest of all the methods (**Fig. 6b, Supplementary Table 9**). Pathway enrichment for the genes within each biomarker showed that IDAM could detect eight phenotype-specific pathways identified by HUMAnN2 (**Supplementary Table 10**) and many pathways consistent with previous findings [40-42]. For example, the biosynthesis of low-carbon-number saturated lipid classes, including myristic, stearic, and palmitic acid, was likely to be higher in those with T1DM than in those without T1DM [41]. Aromatic amino acids, phenylalanine, and tyrosine were higher in T1DM patients [40], which tends to increase the risk of T1DM. Collectively, these results demonstrated the high reliability of the biomarkers identified by IDAM.

We also noted that progression-free survival (PFS) after immunotherapy of melanoma individuals was publicly available. We explored the association between inferred microbial biomarkers and PFS. Four out of seven biomarker-related groups, with or without the biomarker, differed significantly in their PFS, with a log-rank p-value < 0.05 (**Fig. 6c & Supplementary Material Section 3**). We assessed the progression risk of patients using Cox proportional hazards models, and found that the patients with Biomarkers 2, 4, and 5 had a significantly lower risk of progression than those without these biomarkers, as assessed during follow-up (**Fig. 6d, Supplementary Table 11**). We inferred that the three biomarkers were associated with longer PFS. In contrast, biomarker 1 was associated with shorter PFS, and individuals with this biomarker had a high risk of melanoma progression (Hazard ratio [95% CI] = 9.1195 [3.987,20.615], p-value = 0.016; **Fig. 6d**). We collected all species and pathways that have been identified to be associated with shorter or longer PFS from the paper by Peters *et al*. [21] (**Fig. 6e**). Most of these (6/11 species and 5/8 pathways) can be found in the biomarkers identified by IDAM (shown as red in **Fig. 6e**). These findings support the contention that the gut microbiome has an impact on the efficiency of immunotherapy on melanoma [21], and provide mechanistic insights into means for improving the efficiency of immunotherapy by manipulating the microbiome.

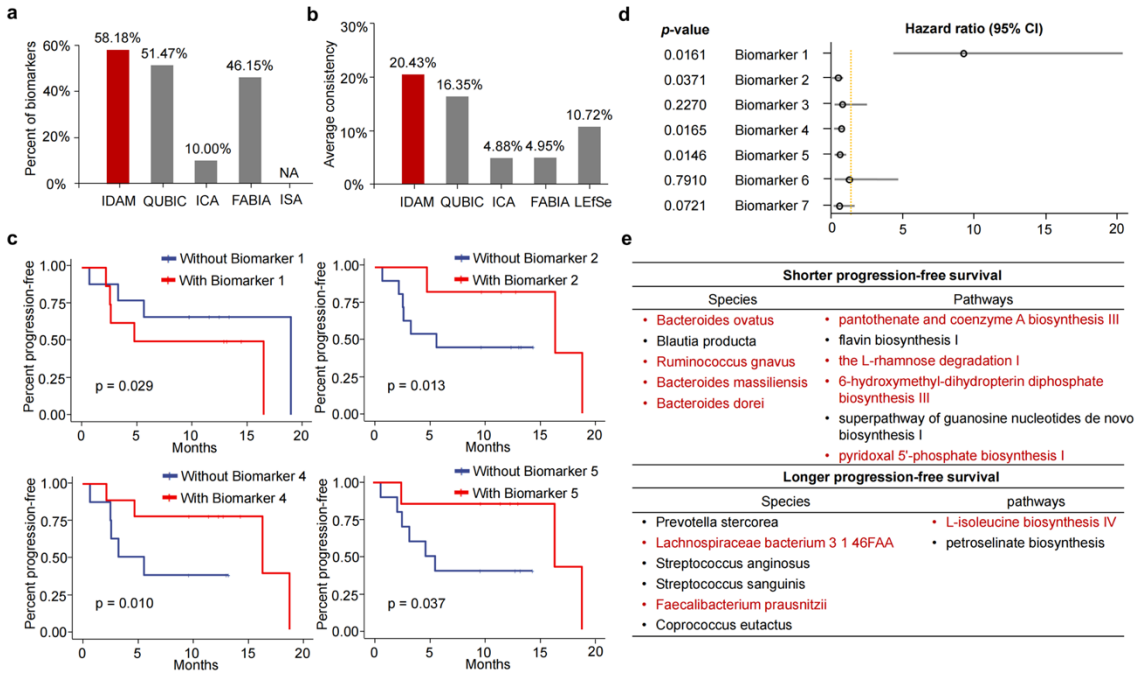


Fig. 6 Results of IDAM on the integrated dataset. **a.** Comparison of phenotype consistency among IDAM, QUBIC, ICA, FABIA, and ISA. The red bars indicate the percent of biomarker-associated sample sets with consistent phenotypes from IDAM, while the grey bars indicate those from other methods. **b.** Comparison of the average consistency of collected species across the four phenotypes. **c.** Kaplan-Meier curves of biomarker-associated groups along with corresponding log-rank test *p*-values. **d.** Results from Cox proportional hazard models. The yellow line indicates a hazard ratio equal to 1. CI: confidence interval. **e.** Species and pathways related to PFS. The red highlights represent those that can be found by biomarkers from IDAM.

Discussion

Although many studies have focused on the microbial biomarker inference of diseases, it is still a fundamental challenge to explore critical characteristics in large-scale, highly heterogeneous microbial data. We developed a new methodology, IDAM, for matched MG and MT data to identify disease-associated biomarkers. The above results indicate that IDAM is a highly effective approach to the identification of disease-associated microbial biomarkers based on both gene co-expression and context conservation during evolution. The biomarkers of diseases identified can be useful in understanding disease pathogenesis and providing guidance for the future development of disease diagnosis and treatment.

IDAM is innovative in that, unlike previous microbial biomarker analysis methods, it integrates gene context conservation and gene expression information by leveraging a

multi-omic view of microbial communities for the inference of diseases using functional characteristics. IDAM does not require any prior knowledge about the samples, eliminating bias from misleading data noise and allowing us to identify novel disease subtypes and states. We illustrated and evaluated IDAM using IBD, melanoma, T1DM, and IBS associated datasets, and found that it outperformed existing methods. We discovered microbial biomarkers significantly associated with specific diseases, some of whose functions are consistent with what has been reported previously. We also evaluated the gain of using gene context conservation information in IDAM and found that ignoring this information can lead to significantly decreased biological relevance between genes within a module. This research demonstrated that uber-operon structures play a crucial role in functional indication.

IDAM is still not free of limitations. It identifies biomarkers at the community level, which needs further taxonomic analysis to improve their utility as clinical biomarkers. This limitation can be somewhat addressed by tracing the community-level functions back to the taxonomic level using HUMAnN2. However, there are some biomarkers that are attributed to unknown species, labeled as unclassified species in HUMAnN2. Therefore, caution is needed in the interpretation of these biomarkers, since we cannot confidently say whether they are false biomarkers, or instead simply correspond to species that have not been well studied. Overall, we believe that this issue will be alleviated as more microbes are discovered and explored. We plan to develop a three-dimensional analysis approach in the future [43], which takes both microbial species and genes and pathways of multiple samples into consideration for biomarker inference. In this way, modules involved in diseases consisting of genes and pathways with specific functions in specific species will be detected, providing functional and taxonomic insights into diseases. IDAM requires relatively large amounts of data. Since IDAM uses expression similarity to assess the relevance of gene regulation, a large sample size might be needed to guarantee reliable measurement from data. In practice, new paired MG and MT data from users can be integrated with the datasets used in this article to increase the sample pool, thereby improving the practicability of IDAM. Finally, the biomarkers identified by IDAM only provided associations between microbiomes and diseases. Rigorous biological experiments are required to establish the causality between the biomarkers identified and the disease of interest, and to establish their clinical prevalence and utility.

Methods

Datasets

We downloaded all matched quality-controlled MG and MT data associated with IBD, as well as the corresponding HUMAnN2 results and metadata of samples at the IBDMDB website in March 2020 (https://ibdmdb.org) [20]. There were 198 samples from non-IBD subjects and 615 samples from patients with IBD, consisting of 381 UC samples and 234 CD samples. We also downloaded matched MG and MT data in .fasta format from the NCBI Sequence Read Archive with the accession number of SRP197281 for melanoma-associated data [21], SRP072561 for T1DM- and non-T1DM-associated data [22], and ERP001739 for IBS-associated data [23]. Uber-operon structures were obtained from the paper by Che *et al.* [19].

IDAM: a framework for identifying disease-associated microbial biomarkers based on metagenomic and metatranscriptomic data

IDAM is a heuristic algorithm for disease-associated microbial biomarker inference. It is based on a greedy search to gather genes with similar expression patterns and high context conservation as functional biomarkers. IDAM consists of three steps.

Step 1*: *Extracting community-level gene expression of samples

The input data is quality-controlled (quality- and length-filtered and screened for residual host DNA) matched MG and MT sequencing datasets in fasta, fasta.gz, fastq, or fastq.gz format. The datasets of each sample are first processed using HUMAnN2 for gene relative expression within communities [26]. In this study, we directly used the HUMAnN2 results of the datasets from IBDMDB for the IBD-associated datasets. Then, we extract the community-level gene expression of each sample to construct a gene-sample expression matrix $A_{m\times n}$, in which each row represents a gene (total m genes indicating by UniRef90 gene family [44] and forming a gene set X), and each column represents a sample (total n samples).

Step 2*: *Measuring gene context conservation and expression similarity for each gene pair

We assess context conservation and expression similarity for each gene pair and sort them as a list for module initialization.

2.1 Gene context conservation

Gene context conservation is assessed based on the gene distribution in uber-operons. Since the genes within X referred here are indicated as gene families from UniRef90 (a non-redundant protein sequence database) [44], we align the gene family sequences with genes in uber-operons to determine to which uber-operons these gene families belong. The sequences of genes in all uber-operons are collected into a custom database. For each gene family, g_i, within X (1≤ $i \leq m$), we extract the corresponding sequence from the UniProt database [44] and align it with the custom database using BLAST [45]. We define a gene-to-uber-operon mapping function $\varphi: D \rightarrow H$, in which D consists of all subsets of X, Y is the set of all uber-operons, and H consists of all subsets of Y. We assume that a gene family g_i (1≤ $i \leq m$) belongs to an uber-operon u if the *E*-value of its alignment with one of the uber-operon's sequences is less than 0.001. This is denoted by $\varphi(G_i) = \{u\}$, where G_i is the set consisting of a single gene, g_i. If multiple sequences from different uber-operons have an *E*-value less than 0.001, we choose the uber-operon corresponding to the smallest *E*-value. In contrast, if no sequence has an *E*-value less than 0.001, we set $\varphi(G_i) = \{\emptyset\}$. In this way, we construct a map between an uber-operon and a gene.

We further define the relationship between uber-operons and gene modules with multiple genes. Suppose a gene set Q consists of b genes $\{g_1, g_2, \cdots, g_b\}$ (1 $< b \leq m$). We assume $\varphi(Q) = \varphi(\{g_1\} \cup \{g_2\} \cup \cdots \cup \{g_b\}) = \varphi(G_1 \cup G_2 \cup \cdots \cup G_b) = \varphi(G_1) \cup \varphi(G_2) \cup \cdots \cup \varphi(G_b)$, where G_i represents the set consisting of a single gene g_i ($i = 1, 2, \cdots, b$). For two gene modules, I and J, we set a reward $U(I,J)$ as a gene context conservation measurement, based on the relative positions of their genes in uber-operons:

$$U(I,J) = \begin{cases} 1, & \varphi(I) \subseteq \varphi(J) \; and \; \varphi(I) \neq \{\emptyset\}, \\ -1, & \varphi(I) = \{\emptyset\} \; and \; \varphi(J) \neq \{\emptyset\}, \\ -1, & \varphi(I) \neq \{\emptyset\} \; and \; \varphi(J) = \{\emptyset\}, \\ -2, & \varphi(I) \neq \varphi(J) \neq \{\emptyset\}, \\ 0, & otherwise. \end{cases}$$

Based on this, we calculate $U(G_i, G_j)$ as the gene context conservation of each pair of genes g_i and g_j (1 $\leq i \leq j \leq m$).

2.2 Expression similarity of genes

Expression similarity is assessed based on a vector composed of the expression values of each gene within all samples. We need to measure the expression similarity not only

between two genes, but also across multiple genes, while expanding the modules in Step 3. This cannot be achieved using existing similarity measurements, such as Pearson correlation coefficients, since they are generally used for paired data points. To address this issue, the continuous expression value of each gene is first discretized to an integer representation using qualitative representation [32] (**Supplementary Material Section 4**). Based on this representation, we measure the expression similarity $S(G_i, G_j)$ of gene pair (g_i, g_j) as the number of samples under which the corresponding integers along the rows of the two genes are identical (or identical integers but with opposite signs).

2.3 Integration of gene context conservation and expression similarity

A combined score, $C(G_i, G_j)$, integrating gene context conservation and the expression similarity of g_i and g_j is obtained via the following function:

$$C(G_i, G_j) = \left(1 + w \times U(G_i, G_j)\right) \times S(G_i, G_j),$$

where w is a tuning parameter that can be set by users. w ranges between 0 and 1, and a larger value of w indicates that gene context conservation has a stronger impact on estimating functional relationships. Here we used 0.1 as the default value (**Supplementary material Section 5 and Supplementary Table 12**). Finally, the gene pairs are sorted in descending order for the combined scores, generating a seed list, L.

Step 3*: *Generating modules consisting of gene modules and corresponding sample sets

Since the gene set X is a pan-genome of all microorganisms from multiple samples, it includes a large number of genes. To improve the computational efficiency, we partition the rows within the matrix equally into t subsets [46,47], where t is determined using a stochastic model [46,47]. Then, we use gene pairs within each subset instead of all gene pairs in X, forming the seed list L, which can greatly decrease the number of seeds and reduce the computational burden.

A submatrix is regarded as biologically relevant if it does not occur randomly in a matrix $A_{m \times n}$. Since the output submatrices from our algorithm are biologically relevant and generated by a gene pair (see Step 3.1), we need to make sure that at least one gene pair within each meaningful submatrix can exist as a seed, to keep all biologically relevant submatrices detectable. This means we can at most divide the rows of matrix *A*

into k -1 subsets (*i.e.*, $t = k - 1$), where k represents the number of genes of the biologically relevant submatrix with the fewest genes (referred to with the smallest row size). We statistically assess the smallest row size of the submatrix that is unlikely to occur by chance. The submatrices we target are of rank one mathematically, since the corresponding genes have co-expressed patterns. Suppose a submatrix with rank one is of size $k \times s$, its probability of random occurrence in $A_{m\times n}$ is calculated as follows [46,47]:

$$P_{ks} = n \times (n-1) \times \cdots \times (n-s+1) \times \binom{m}{k} \times \left(\frac{2}{s!}\right)^k \times (1 - \frac{2}{s!})^{m-k}$$

The submatrix is considered to be biologically relevant if P_{ks} < 0.05. When users have prior knowledge of the largest column size, s, we can calculate k based on the formula above. The smallest row size of the biologically relevant submatrices can be determined. Users can then set the number of subsets ($t = k - 1$). In this work, the default value of t was set to 10. That is, the matrix was divided into 10 subsets, which greatly reduced the computational burden and kept biologically relevant submatrices detectable to a large extent.

3.1 Initialization

We start with the first gene pair in L that satisfies one of two conditions: i) at least one gene of these two has not been included in previous modules, or ii) these two genes are included within different previous modules, and there is no overlap between the gene components of the two modules. We aggregate samples in which the integers of the two genes are identical (represented as set P), forming a current module $M = \{Q, P\}$, where Q represents the gene set consisting of the paired genes. The gene pair is then removed from L.

3.2 Expansion

We expand M by adding a new gene (if any) that has the highest combined score with M, giving rise to the updated module $M' = \{Q', P'\}$. The calculation of the combined score between a gene (represented as g_z (1≤ $z \leq m$)) and Q, namely $C(G_z, Q)$, is similar to that of two genes g_i and g_j. We first assess the context conservation between g_z and the gene set Q of module M, i.e., $U(G_z, Q)$. Then, we measure the similarity of expression of g_z and Q, $S(G_z, Q)$, as the number of samples within P under which the corresponding integers along the rows are identical (or identical integers but with opposite signs). Based on these, a combined score $C(G_z, Q)$ can be obtained by

replacing $U(G_i, G_j)$ and $S(G_i, G_j)$ in the calculation of $C(G_i, G_j)$ with $U(G_z, Q)$ and $S(G_z, Q)$, respectively. If $min\{|Q'|, |P'|\} > min\{|Q|, |P|\}$, we set $M = M'$ and repeat Step 3.2. Otherwise, go to Step 3.3.

3.3 Filter and output

If the proportion of genes overlapping between Q and the component genes of previously identified modules is less than a threshold, f, module M will be provided as the final output, and is ignored otherwise. In this study, the default f was set to 0.1 for more separate modules. Go to Step 3.1 until the list L is empty.

Evaluating phenotype consistency of samples within one module

Z-tests are used to assess whether samples with the same biomarker are of highly consistent phenotypes. Specifically, we extract the phenotypes of all samples from the metadata as background, and the corresponding phenotypes of samples in each gene module as test sets. For each phenotype and each test set, we count the phenotype occurrence in the test set and background. We evaluate the significance of each phenotype using a right-tailed Z-test with the null hypothesis that the observed frequency in the test set is less than or equal to the occurrence probability within the background. The threshold of significant level 0.05 is used with a critical Z-value of 1.645. If the Z-score is greater than 1.645 (*p*-value < 0.05), the null hypothesis is rejected, meaning the phenotype significantly occurred within the test set. In this case, we consider the samples within the gene module to be of a highly consistent phenotype.

The propensity score matching method

Our aim was to assess whether the association between identified biomarkers and disease phenotypes was reliable. For each biomarker, we first identified the underlying confounding variables that affected its occurrence. We included diet, age, race, and health status, as well as the other biomarkers, and calculated a propensity score for each sample using the R package MatchIt 4.2.0. Then, the paired samples were constructed by matching with the optimal score by setting the parameter method = “optimal”. Finally, Wilcoxon signed-rank tests were used to measure the significance of each biomarker associated with a different phenotype, and the *p*-value was adjusted using the

Benjamini-Hochberg procedure, in which an adjusted *p*-value < 0.05 was considered to be significant.

Functional analysis of biomarkers

We performed an analysis of the enriched functions of each biomarker. All UniRef90 identifiers were firstly mapped to MetaCyc reactions [26,48]. The reactions were further associated with MetaCyc pathways [26,48], which enabled a transitive association between UniRef90 gene families and MetaCyc pathways. We defined a pathway that was satisfied when the reactions within the pathway significantly occurred via the hypergeometric test,

$$P(t_{ij}) = \frac{\binom{n_j}{t}\binom{N-n_j}{m_i-t}}{\binom{N}{m_i}}$$

$$p-value = \sum_{t_{ij}>T_{ij}}^{n_j} P(t_{ij})$$

where N is the total number of reactions that were associated with the UniRef90 gene families and pathways in MetaCyc, m_i is the number of reactions within the ith biomarkers, n_j is the total number of reactions within the jth pathway, and T_{ij} is the number of observed reactions of pathway j occurring in biomarker i. This yielded a *p*-value for each biomarker and pathway. We selected the pathways with a *p*-value less than 0.05 as the significantly enriched pathways of each biomarker.

We used a widely used tool, LEfSe [11], to infer differentially abundant pathways among phenotypes. We used pathway abundance information for each sample from HUMAnN2.

Comparison of IDAM with other methods

The performance of biomarker inference of IDAM was compared with the other four top-performing unsupervised tools for grouping genes into functional modules [49], including biclustering-based (QUBIC [32], ISA [33], and FABIA [34]) and decomposition-based (fastICA, an improvement of ICA [31]) methods. All tools were tested and compared based on the gene-sample matrix constructed. We ran these tools on the Pitzer cluster of the Ohio Supercomputer Center with memory usage set to 300GB [50] with default parameters. QUBIC was implemented using the published source code, while the other three

algorithms were implemented using the R packages, isa2 0.3.5, fabia 2.36.0, and fastICA 1.2.2, respectively.

Species-level analysis

For the biomarkers from IDAM, ICA [31], QUBIC [32], and FABIA [34], we extracted species that contribute to the abundance of genes within the identified biomarkers, based on the HUMAnN2 output [26]. Community-level gene expression (abundance) was decomposed into species-level data. A phylogenetic tree was generated using the software GraPhlAn [51]. Since the biomarkers have been associated with phenotypes using Z-tests, we assigned the species of each biomarker to the corresponding phenotype of the biomarker. We collected disease-associated species from the databases Peryton and MicrophenoDB [52], and previous taxonomic studies. For each phenotype, we counted the total number of species, as well as the number that matched with the collected species. Then, the consistency index was calculated as the total number of species in the phenotype dividing by the number of matched species.

LEfSe [11], was used here for comparison. It uses species profiling from MetaPhlAn2 to identify statistically different species among phenotypes [53,54]. The significantly different species of each phenotype were aligned to collected species, from which we determined the number of matched species and calculated the consistency of each phenotype.

Survival analysis

We assessed the associations of the identified biomarkers of melanoma with PFS. The Kaplan-Meier curves of each group were plotted using the “ggsurvplot” function in the survival R package [55]. The Cox proportional hazards model was used to explore specific relationships between the biomarker and PFS, adjusting for age, antibiotic use in the last 6 months, sex, BMI, and disease stage. This was achieved using the “coxph” function in the survival and survminer R packages [56].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The IBD-associated dataset supporting the conclusions of this article is available in the IBDMDB database (https://ibdmdb.org). The melanoma, T1DM, and IBS associated datasets supporting the conclusions of this article are available on NCBI Sequence Read Archive, with the accession number of SRP197281, SRP072561, and ERP001739, respectively.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by National Key R&D Program of China (2020YFA0712400), National Nature Science Foundation of China (NSFC, 61772313 and 11931008), and Interdisciplinary Science Innovation Group Project of Shandong University (2019).

Authors' contributions

Q.M. and B.L. conceived the basic idea. Z.L. carried out the computational analysis and data interpretation. L.Z., Q.W., and A.M. designed and drew the figures. D.C. and J.Z. helped polish the manuscript and all the authors wrote the manuscript. All authors read and approved the final manuscript.

References

1 Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nature medicine* 24, 392-400 (2018).

2 Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Science* 371, eabc4552 (2021).

3 Consortium, T. H. M. P. A framework for human microbiome research. *Nature* 486, 215-221 (2012).

4 Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* 569, 641-648 (2019).

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

6 Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031 (2006).

7 Dubinsky, M. & Braun, J. Diagnostic and Prognostic Microbial Biomarkers in Inflammatory Bowel Diseases. *Gastroenterology* 149, 1265-1274.e3 (2015).

8 Lecuit, M. & Lortholary, O. Immunoproliferative small intestinal disease associated with Campylobacter jejuni. *Medecine et maladies infectieuses* 35, S56-58 (2005).

9 Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods* 10, 1200-1202 (2013).

10 Liu, Z. *et al.* Network analyses in microbiome based on high-throughput multi-omics data. *Briefings in Bioinformatics* 22, 1639-1655 (2020).

11 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biology* 12, R60 (2011).

12 Zhu, Q., Jiang, X., Zhu, Q., Pan, M. & He, T. Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. *Frontiers in Genetics* 10 (2019).

13 Wu, Y. *et al.* Identification of microbial markers across populations in early detection of colorectal cancer. *Nature Communications* 12, 3063 (2021).

14 Snipen, L., Angell, I.-L., Rognes, T. & Rudi, K. Reduced metagenome sequencing for strain-resolution taxonomic profiles. *Microbiome* 9, 79 (2021).

15 Zhu, Z., Ren, J., Michail, S. & Sun, F. MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. *Genome Biology* 20, 154 (2019).

16 Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences* 111, E2329-E2338 (2014).

17 Kasmanas, J. C. *et al.* HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research* 49, D743-D750 (2020).

18 Lathe III, W. C., Snel, B. & Bork, P. Gene context conservation of a higher order than operons. *Trends in biochemical sciences* 25, 474-479 (2000).

19 Che, D., Li, G., Mao, F., Wu, H. & Xu, Y. Detecting uber-operons in prokaryotic genomes. *Nucleic acids research* 34, 2418-2427 (2006).

20 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655-662 (2019).

21 Peters, B. A. *et al.* Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome medicine* 11, 61 (2019).

22 Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology* 2, 16180 (2016).

23 Durbán, A. *et al.* Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome. *FEMS microbiology ecology* 86, 581-589 (2013).

24 Niu, S.-Y. *et al.* Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Briefings in Bioinformatics* 19, 1415-1429 (2017).

25 Xie, J. *et al.* QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics* 36, 1143-1149 (2020).

26 Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* 15, 962-968 (2018).

27 Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology* 13, R79 (2012).

28 Imhann, F. *et al.* Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* 67, 108-119 (2018).

29 Klaassen, M. A. Y. *et al.* Anti-inflammatory Gut Microbial Pathways Are Decreased During Crohn's Disease Exacerbations. *Journal of Crohn's and Colitis* 13, 1439-1449 (2019).

30 Schirmer, M., Garner, A., Vlamakis, H. & Xavier, R. J. Microbial genes and pathways in inflammatory bowel disease. *Nat Rev Microbiol* 17, 497-511 (2019).

31 Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* 13, 411-430 (2000).

32 Li, G., Ma, Q., Tang, H., Paterson, A. H. & Xu, Y. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research* 37, e101-e101 (2009).

33 Bergmann, S., Ihmels, J. & Barkai, N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review. E, Statistical, nonlinear, and soft matter physics* 67, 031902 (2003).

34 Hochreiter, S. *et al.* FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 26, 1520-1527 (2010).

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

36 Kuballa, A., Geraci, M., Johnston, M. & Sorrentino, D. The Gut Microbial Profile of Preclinical Crohn's Disease Is Similar to That of Healthy Controls. *Inflammatory bowel diseases* 26, 1682-1690 (2020).

37 Vich Vila, A. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* 10, eaap8914 (2018).

38 Skoufos, G. *et al.* Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Research* 49, D1328-D1333 (2020).

39 Wu, J. *et al.* Rhein modulates host purine metabolism in intestine through gut microbiota and ameliorates experimental colitis. *Theranostics* 10, 10665-10679 (2020).

40 Koo, B. K. *et al.* Identification of novel autoantibodies in type 1 diabetic patients using a high-density protein microarray. *Diabetes* 63, 3022-3032 (2014).

41 Wang-Sattler, R. *et al.* Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular systems biology* 8, 615 (2012).

42 Mahendran, Y. *et al.* Glycerol and fatty acids in serum predict the development of hyperglycemia and type 2 diabetes in Finnish men. *Diabetes care* 36, 3732-3738 (2013).

43 Luo, Y., Ma, Q., Zhang, C. & Zhang, A. R. Provable Second-order Riemannian Gauss-Newton Method for Low-rank Tensor Estimation. *2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (2022).

44 Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, 1282-1288 (2007).

45 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* 215, 403-410 (1990).

46 Wang, Z., Li, G., Robinson, R. W. & Huang, X. UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data. *Scientific reports* 6, 23466 (2016).

47 Ben-Dor, A., Chor, B., Karp, R. & Yakhini, Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology : a journal of computational molecular cell biology* 10, 373-384 (2003).

48 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic acids research* 48, D445-D453 (2020).

49 Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications* 9, 1090 (2018).

50 Ohio Supercomputer Center. Pitzer Supercomputer. http://osc.edu/ark:/19495/hpc56htp (2018).

51 Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* 3, e1029 (2015).

52 Yao, G. *et al.* MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes. *Genomics, Proteomics & Bioinformatics* 18, 760-772 (2021).

53 Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9, 811-814 (2012).

54 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12, 902-903 (2015).

55 Therneau, T. M. & Lumley, T. Package ‘survival’. *R Top Doc* 128, 28-33 (2015).

56 Andersen, P. K. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100-1120 (1982).

Supplementary Information

Supplementary Material

This supplementary material includes details in the Problem formulation and Methods part. Section 1 illustrates why the problem we formulated is NP-hard.

Section 2 shows how to construct the integrated datasets and process the raw sequencing data.

Section 3 shows the Kaplan-Meier curves of biomarker-associated groups with non-significantly different PFS.

Sections 4 and 5 describe how to do discretization and select parameter *w*.

Supplementary Tables

This file includes supplementary results in the Result part and parameter selection of *w* in the Method part with an overview of Supplementary tables.

Table 1 shows the size of identified gene modules from IDAM.

Table 2 shows the Z-score and *p*-value of each module for three phenotypes.

Table 3 shows the adjusted *p*-value from Wilcoxon signed-rank test.

Table 4 shows pathways enriched for identified biomarkers from IDAM.

Table 5 shows the species associated with different phenotypes from IDAM.

Table 6 shows the species associated with different phenotypes collected from previous studies.

Table 7 shows detailed information of the integrated dataset.

Table 8 shows the number of identified modules and significantly phenotype-specific modules on the integrated dataset.

Table 9 shows the results of species-level analysis on the integrated dataset by different methods.

Table 10 shows the pathways enriched in biomarkers from IDAM on the integrated dataset.

Table S11 shows The results of Cox proportional hazards models.

Table 12 shows the performance of IDAM with different parameter *w*, which is attached to Methods and Supplementary Material Section 5.