Manuscript Click here to access/download;Manuscript;manuscript.docx

Click here to view linked References

Infer disease-associated microbial biomarkers based on metagenomic and metatranscriptomic data

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

[1] School of Mathematics, Shandong University, Jinan, Shandong, 250200, China.

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

[#]To whom correspondence should be addressed.

Email

Zhaoqian Liu --- lzqian8023@163.com;

Qi Wang --- wangqi1994_sdu@163.com;

Dongjun Chung --- dongjun.chung@osumc.edu;

Qin Ma --- qin.ma@osumc.edu;

Jing Zhao --- jing.zhao2@osumc.edu;

Bingqiang Liu --- bingqiang@sdu.edu.cn.

Abstract

Unveiling disease-associated microbial biomarkers is crucial for disease diagnosis and therapy. However, the heterogeneity, high-dimensionality, and large amounts of microbial data bring tremendous challenges for fundamental characteristics discovery. We present IDAM, a novel method for disease-associated biomarker inference from metagenomic and metatranscriptomic data, without requiring prior metadata. It integrates gene context conservation (uber-operon) and regulatory mechanism (gene co-expression patterns) through a mathematical graph model. We applied IDAM to inflammatory bowel disease associated matched metagenomic and metatranscriptomic datasets, which showed superior performance in biomarker inference. IDAM is freely available at https://github.com/OSU-BMBL/IDAM.

Keywords

Metagenome; Metatranscriptome; Microbial biomarkers; Gene regulation; Gene context conservation; Unsupervised learning.

Background

Trillions of microbes colonize the human body and play critical roles in multiple fundamental physiological processes, such as immune system development and dietary energy harvest [1]. Moreover, massive evidence reveals that 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, worldwide projects, such as the Human Microbiome Project [3], the integrative Human Microbiome Project [4], and the Metagenomics of the Human Intestinal Tract project [5], have generated vast amounts of microbial data for diverse diseases. Thereafter, increasing efforts have been aimed at bringing these data to clinical insights

and deepening the understanding of crucial mechanisms responsible for disease progression and treatment. For example, recent studies have suggested that transferring several microbes from a donor into a sick recipient can promote the recipient's recovery and cause changes in the donor phenotype (e.g., increased adiposity) [6]. Furthermore, some pathogenic organisms have been found closely related to disease states and can be used as a potential marker 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, i.e., microbial biomarkers.

Till now, a great deal of research has been explored for identifying microbial biomarkers due to their critical roles in diseases. One of the classical paradigms is identifying taxa and functions that can statistically significantly differentiate two or more groups as biomarkers [9-11]. Additionally, given machine learning's success under multiple scenarios, there are several attempts to adopt them in microbial biomarker inference, such as random forest and deep feedforward networks [12, 13]. However, these approaches face several challenges.

First, most of them focus on 16S ribosome RNA (16S rRNA) and metagenomic (MG) data. 16S rRNA data suffer from a low taxonomic resolution and an absence of functional information [14]. The taxonomic analysis alone may induce spurious biomarkers since diverse microbial communities from different patients can perform remarkably similar functional capabilities [14]. While MG data can provide information at all taxonomic levels as well as potential functional information, researchers have found that: (i) A great proportion of reads cannot be successfully mapped to existing reference genomes during taxonomic identification, which leads to the potential loss of valuable information [15]. Intuitively, one may propose to use de novo assembled genomes as a complement. However, the difficulties in metagenome assembly and binning generally cause incorrect contigs, which substantially affect further taxonomic inferences [16]. (ii) The functional

analyses from metagenomics are not equivalent to the true activities within the microbiome. It has been observed that multiple metagenomically abundant genes in the human gut microbiome were significantly down-regulated at the transcriptional level [17].

Second, existing methods are designed based on the strong assumption that the data with sufficient sample size and accurate and detailed metadata information is available to design groups or train models. However, the current metadata of a considerable number of sequencing samples is incomplete, misleading, or not publicly available [18], which may lead to these methods being infeasible or causing bias in biomarker inference. Moreover, their intrinsic design in using known phenotype information makes them incapable of revealing new subtypes or stages of diseases. 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.

In this study, we propose a novel method to Infer Disease-Associated Microbial functional biomarkers (IDAM), which to a large extent surpasses these limitations discussed above. Specifically, the data we used was matched MG and metatranscriptomic (MT) data sequenced from microbial samples for assessing active functions within communities. To make full use of available data, we focused on a community-level functional analysis that goes beyond species level (by which 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. Furthermore, we considered identifying functional biomarkers by detecting genes with specific co-expression patterns across certain subsets of samples, without requiring prior metadata. Undeniably, co-expression does not necessarily imply functional relevance [19]. Moreover, one concern is that gene expression referred here is assessed by the reads from various species (Fig. 1a), which represents an ensemble expression of multiple homologous genes. Hence, we accept the fact that one cannot fully characterize the complex functional

activities based solely on the ensemble co-expression, by which the genes with no biological relevance may be clustered together (Fig. 1b). Fortunately, recent studies have suggested that functionally associated gene sets are generally highly conserved during evolution, i.e., uber-operon structures [20, 21]. This finding strongly motivated us to integrate gene context conservation to study cross-species functional-related mechanisms, thereby alleviating the risk of false positives or false negatives in biomarker inference. Specifically, we used uber-operon structures as the evolutionary footprints of cross-species functionally related genes. 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. Therefore, we integrated both gene regulation and gene context conservation to identify cross-species functional biomarkers in IDAM (Fig. 1c).

We applied IDAM to the published matched MG and MT datasets of inflammatory bowel diseases (IBD) (*n* = 813; including two subtypes, ulcerative colitis (abbreviated as UC) and Crohn's disease (abbreviated as CD)) [22]. It could identify 41 gene modules, corresponding to 41 biomarkers. We validated the reliability of these biomarkers based on known phenotypes and well-studied species/functions related to IBD reported previously. Results suggested superior performance of IDAM in key biomarker inference compared to popular tools, including both metadata-based (LEfSe [11]) and non-metadata-based approaches (ICA [23], QUBIC [24], ISA [25], and FABIA [26]). Furthermore, IDAM classified two IBD subtypes with high accuracy. Hence, we believe that IDAM can be a highly advantageous method for biomarker inference based on MG and MT data. It can potentially pave the way for understanding the role of the microbiome in human diseases and improving disease diagnosis and treatment in clinical practice.

Please put Fig. 1 here.

Fig. 1 The diagram of 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 metagenomic shotgun sequencing does not isolate microorganisms, the reads from different species (represented by black purple, black green, and black blue) were jumbled together. As a result, we cannot distinguish them easily at the species level but we can observe the expression of each gene only in a manner that reads from all species are merged together. Note that the microbial community transcript number (MT data) can be corrected using the underlying genomic copy number (MG data) in practice. Here, we simply showed the number of transcriptomic reads as expressions. The colors of circles (grey, blue, and red) indicate the expression level of genes within each sample. Here, the ground truth 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.** The co-expression modules. There were three co-expression modules, which are defined as the gene and sample subsets with coordinated expression patterns. Note that however they were not equivalent to the known functional relationship. **c.** The functional modules by integrating uber-operon structures and expression similarity. Each ellipse, including multiple genes, represents an uber-operon. Here we search 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 the problem formulation of this article and then provide an algorithmic overview of IDAM.

Problem formulation

The problem of biomarker inference 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 RNA abundance (MT data) against DNA abundance (MG data) of g_i within sample j. Based on the matrix, the co-expression patterns present in gene and sample subsets can be detected to identify local low-rank submatrices [27]. Given the constraints of gene context conservation, we here integrate the submatrices with a graph model to maximize the number of genes connecting to a single uber-operon. Specifically, 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 identifying a set of local low-rank submatrices within A that maximizes the number of connected components between the corresponding gene sets and V_1 on G. Note that here we maximize the number of connected components between gene subsets and V_1 to minimize the heterogeneity of genes for each uber-operon, thereby ensuring the genes within each identified submatrix as functionally related as possible. Based on the identified submatrices within A (each of which contains a subset of genes and samples), the gene sets can be regarded as biomarkers of the shared phenotypes of the sample sets. However, this problem is theoretically intractable (NP-hard, Additional file 1: Section 1).

Hence, instead of trying to solve it directly, we develop 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). Firstly, the community-level expression within each sample is assessed by HUMAnN2 based on the matched MG and MT data [28]. This step provides a gene-sample expression matrix, each entry of which represents the expression of a gene within a sample. Secondly, IDAM combines expression similarity and gene distribution within uber-operons to measure the likelihood that each pair of genes is involved in the same function. This step provides a list based on combined assessment for module initialization (Fig. 2b). Thirdly, the gene pairs in the obtained list are used as seeds one by one, and each seed is expanded by iteratively adding other functionally related genes to generate modules until the new module starts to become smaller than the previously identified ones (Fig. 2c). Finally, the gene modules consisting of gene sets and corresponding sample sets will be considered as the final output of the algorithm. Here we will regard gene sets as functional biomarkers of the most common phenotype or state of corresponding samples. Details are available in the “Methods” section.

Please put Fig. 2 here.

Fig. 2 The workflow of IDAM. **a.** The flow chart. We consider matched MG and MT data of multiple samples as input, from which we obtain an expression matrix. By module initialization and expansion, the modules will generate as output. **b.** Details of module initialization. For each gene pair, we assess both expression similarity and context conservation for 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 context

conservation measurement. Expression similarity is assessed according to the number of samples under which the rows are identical, and we get a score S (Methods). Finally, U and S are combined via a function C (w is a tuning parameter), obtaining a score C_{34}. All gene pairs are sorted in a decreasing order based on combined scores. **c.** Details of module expansion. We use gene pairs in the list as initial modules and expand them gradually. Specifically, 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, respectively, 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 $g_3g_4g_5$ is of the highest score, we replace the previous with this one. Next, 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 $g_3g_4g_5g_2$ is of 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 well distinguished different phenotypes and enabled the discovery of disease subtypes

To assess the usefulness of biomarkers identified by IDAM, we assessed the phenotype classification performance using identified biomarkers. We applied IDAM to the publicly available dataset of patients with IBD and controls (referred to as Non-IBD) (Methods, all following analyses are based on these datasets). 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, IDAM found 41 gene modules with genes spanning from 2,340 to 41, totally covering 19,327 genes (Additional file 2: Table S1). The gene set contained in each module was regarded as a biomarker. That is, there

were 41 identified biomarkers. Ideally, samples with the same biomarker should have highly consistent phenotypes. We used Z-test to evaluate the phenotype consistency of samples in each module (Methods). Results showed that 32 biomarker-associated sample sets were consistently enriched for a particular phenotype (19 enriched for IBD and 13 for Non-IBD, Fig. 3a, Additional file 2: Table S2). On the other hand, we recognized 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 identified biomarkers from IDAM had reliable associations with phenotypes, by applying a propensity score-based matching for each biomarker to confirm its association with phenotypes. Specifically, we constructed paired samples by matching the propensity score estimated from other covariates, such as age, gender, and comorbidities (Methods). Wilcoxon signed-rank test based on the paired samples indicated that 82.93% (34 of 41) biomarkers were still statistically significantly related to a specific phenotype (with Benjamini-Hochberg-adjusted p-value < 0.05, Fig. 3b, Additional file 2: Table S3), demonstrating high reliability of identified biomarkers from IDAM.

Additionally, we observed statistically significant enrichment of the disease subtypes within each module. A noteworthy fact was that the identified biomarkers could recognize the subtypes of IBD, in which 11 correspond to UC and eight correspond to CD (Fig. 3c). We here gave an example of three modules to show the phenotype specificity (Fig. 3d), namely Module 1, Module 17, and Module 30, each of which is enriched for a particular phenotype. Limited by space, we showed the expression of the top 3% genes of each module (17/618, 19/689, and 14/521 for Modules 1, 17, and 30, respectively). We can see that IDAM detected the classical checker-board substructures from the original expression data. Specifically, Biomarker 1 is specific to Non-IBD while Biomarker 17 and

30 are specific to UC and CD, respectively, suggesting the capability of IDAM to uncover biomarkers of disease subtypes. We further carried out functional enrichment analysis for inferred biomarkers from IDAM (Methods). Results showed that these biomarkers were associated with 168 pathways (Additional file 2: Table S4), of which 27 were also identified as phenotype-differential pathways based on pathways abundance analysis of HUMAnN2 (Methods, Additional file 2: Table S4). For example, there was a significant enrichment of short-chain fatty acids metabolism (e.g., butyrate metabolism) in Non-IBD-associated biomarkers. 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. Additionally, the superpathway of lipopolysaccharide biosynthesis was enriched in CD-associated biomarkers while methylphosphonate degradation was enriched in UC-associated biomarkers. These findings are paralleling with previous observations [29-32], indicating that the biomarkers from IDAM were biologically informative in distinguishing phenotypes.

We compared IDAM with several popular biclustering-based and decomposition-based approaches, including ICA [23], QUBIC [24], ISA [25], and FABIA [26] (Methods). These methods have not been applied to microbial data for biomarker inference but intuitively they can reveal functional characteristics as biomarkers under different phenotypes without the need for prior metadata. 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 were with a consistent phenotype, only 61.76% (21/34), 33.33% (10/30), and 30.77% (4/13) for QUBIC, ICA, and FABIA, respectively (Fig. 3e). This revealed that the inferred biomarkers from IDAM can best cluster the samples with the same phenotypes together compared

with others, demonstrating the high effectiveness in distinguishing phenotypes. Indeed, 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 intrinsically used only gene expression. Detailed analyses can be found in the following subsection “The integration of gene context conservation substantially improves the performance of biomarker inference”. Finally, we compared the running time of these methods (Fig. 3f). IDAM spent 7 hours, QUBIC and FABIA took 16 and 22 hours, respectively, while ISA was terminated after 24 hours without results. ICA costs 8 minutes; however, the application power of ICA is limited by the relatively low accuracy (only one-third of the biomarkers can distinguish phenotypes).

Collectively, the biomarkers identified by IDAM performed best in distinguishing phenotypes and can provide hints for new disease subtypes or states in clinical application.

Please put Fig. 3 here.

Fig. 3 The performance of inferred biomarkers by IDAM in distinguishing phenotypes. **a.** The Z-scores of Non-IBD, UC, and CD. They were respectively shown as pink, green, and purple, and each column represents a biomarker. The yellow horizontal line means 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.** The 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 means adjusted *p*-value. * indicates adjusted p-value < 0.05. **c.** The number of biomarkers identified for each phenotype. A total of 19 biomarkers correspond to IBD (in which 11 to UC and 8 to CD) and 13 correspond to Non-

IBD. **d.** An example of gene modules associated with phenotypes, where Module 1, 17, and 30 are shown. The colors of each cell in the heatmap indicate the log2 value of the expression. **e.** The percent of biomarkers associated with samples enriched in certain phenotypes, identified by IDAM and competing algorithms. NA means no results from ISA. **f.** The running time of IDAM and competing algorithms.

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

We implemented species-level analysis for biomarkers from IDAM. Since differences in community-level functions ultimately imply species-level functional differences, we traced genes within each biomarker back to specific species (Methods). We found that these genes were from 111 species, which were dominated by *Bacteroidetes* and *Firmicutes*, followed by *Proteobacteria* at the phylum level (Fig. 4a). This observation was consistent with previous research about IBD-associated species [32]. We further classified the species according to biomarker-associated phenotypes (Methods) and obtained 37, 64, and 13 species associated with Non-IBD, UC, and CD, respectively (Additional file 2: Table S5). Three of them were overlapped between UC and CD, including *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 of IBD [33-36] (Additional file 2: Table S6). Among them, a total of 29 species were captured by IDAM (Fig. 4a), including 17 for Non-IBD, 6 for UC, and 9 for CD (shown in red in Additional file 2: Table S6). An interesting finding was that the three overlapped species from IDAM have been reported associated with both UC and CD [33-35], suggesting the rationality of IDAM results.

We compared IDAM with QUBIC [24], ICA [23], FABIA [26], and LEfSe [11] (Methods). Note that ICA, QUBIC, and FABIA are not dependent on metadata while LEfSe identifies differential abundant species as biomarkers based on prior phenotype information. Detailed results are shown in Fig. 4b. Not surprisingly, the average consistency with collected species across three phenotypes was highest for IDAM (41.52%), followed by QUBIC (19.82%), LEfSe (18.38%), ICA (17.62%), and FABIA (4.74%) (Methods, Fig. 4c). Collectively, the species-level analysis showed the inferred biomarkers from IDAM were biologically associated with phenotypes.

Please put Fig. 4 here.

Fig. 4 The species-level analysis of identified biomarkers. **a.** The phylogenetic tree of the species associated with biomarkers identified by IDAM. The colored dots indicated different species that have been validated by previous studies. **b.** The results of species-level analysis by different methods. The numbers in the header (34, 32, and 90) represent the number of collected species associated with the three phenotypes, respectively. 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 different methods, among which consistent with the collected species are represented as *X*, and *Z* indicates the consistency index of each method associated with phenotypes. **c.** The mean consistency across three phenotypes was used as the average consistency index for each method. The greater value means that identified biomarkers-associated species are more reliable.

The integration of gene context conservation substantially improved the performance of biomarker inference

We investigated the contribution of gene context conservation, given our underlying hypothesis that genes with similar expression patterns and high context conservation during evolution are more likely to be functionally related. To evaluate this, we sought to identify biomarkers without using gene context conservation in IDAM (referred to as IDAM without gene context conservation). Results showed that biomarkers identified by this version are significantly less capable of distinguishing different phenotypes than those from IDAM (Fig. 5a). Specifically, 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. In contrast, in the case of IDAM, 78.05% (32/41) sample subsets were of consistent phenotypes. The pathway analysis suggested that 67.65% (23/34) biomarkers from IDAM without gene context conservation were significantly enriched in MetaCyc functional categories, which is significantly lower than 82.93% (34/41) of IDAM. These strongly indicated that the integration of uber-operon structures contributed to identifying informative biomarkers.

We further explored underlying reasons for the superior performance of integrating gene context conservation. For this purpose, for each biomarker (from either IDAM or IDAM without gene context conservation), we evaluated the distribution of contained genes in the uber-operons. Overall, we found that the genes within biomarkers that cannot significantly enrich certain functions tended to be spread out among more uber-operons, compared with those within biomarkers that enrich. This finding supports our hypothesis that functionally related gene sets are of high context conservation. As an example, we further investigated Module 32 from IDAM without gene context conservation (abbreviated as ‘Module 32’ hereinafter, Fig. 5b). Since the 42 genes within the module were of coordinated expression within the 80 samples, they would be clustered together when we merely considered co-expression. However, these genes were not significantly functionally related and were from six uber-operons (Fig. 5b & 5c). They cannot reflect the

unique functional characteristics of a particular phenotype. Not surprisingly, the 80 samples within Module 32 were evenly scattered over different phenotypes (22 with Non-IBD, 22 with UC, and 36 with CD). When we took gene context conservation into consideration (i.e., IDAM), a new module occurred (e.g., Module 31 from IDAM, abbreviated as 'Module 31' hereinafter). It consists of 85 genes and 34 samples, and there was considerable overlap with Module 32 (Fig. 5b). This module was dominated by an uber-operon that closely associates with purine metabolism [21], and the significantly enriched functions were shown in Fig. 5c. Previous studies have shown that the imbalance between biosynthesis and degradation of purine metabolism could produce excessive uric acid in the gut, leading to the occurrence of IBD [37]. As expected, the 34 samples were enriched in phenotype CD. This demonstrated the strong power of gene context conservation in clustering genes with biological relevance together, which improves the ability of IDAM for biomarker inference.

Please put Fig. 5 here.

Fig. 5 The analysis of the contribution of gene context conservation. **a.** Comparison between IDAM and IDAM without gene context conservation. The red bars mean the results from IDAM, while the grey bars mean results from IDAM without gene context conservation. **b.** A specific case of identified modules. Each row represented a gene and each column represented a sample (totally 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 on the top represented different phenotypes of corresponding samples. Circle colors on the right represented the genes that belong to the same uber-operons. **c.** The pathways enriched for Module 32 and Module 31. The grey ones mean the pathways enriched for Module 32 while red means those enriched for Module 31. The dotted vertical lines mean the critical value with

a *p*-value of 0.05. The pathways located outside of the dotted lines correspond to significantly enriched pathways.

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 and highly heterogeneous microbial data. We developed a new methodology, IDAM, for matched MG and MT data to identify disease-associated biomarkers. IDAM is innovative in the sense that, unlike the previous microbial biomarker analysis methods, it integrates gene context conservation and regulation information by leveraging a multi-omic view of microbial communities for functional characteristics inference of diseases. Furthermore, IDAM does not require any prior knowledge about samples, which alleviates bias from misleading data noise and allows us to reveal novel disease subtypes or states. We tested IDAM for the purpose of IBD biomarker inference and found that it remarkably outperformed existing methods. We discovered 41 significant microbial biomarkers associated with IBD, where some of their functions are consistent with what has been reported previously. In addition, we 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 demonstrated that uber-operon structures played a crucial role in functional indication. Notably, although we use IBD to illustrate and evaluate IDAM, IDAM can contribute to the understanding of other diseases.

Admittedly, IDAM is still not free of limitations. First, IDAM 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 attribute to unknown species (labeled as unclassified species in HUMAnN2). Therefore, cautious inference is needed in interpretation of these biomarkers since we cannot confidently say whether they are actually false biomarkers, or instead simply correspond to species that have not been well studied. However, we believe that this issue will be alleviated as more microbes are discovered and explored. Given this, we plan to develop a three-dimensional analysis approach in the future, which takes both microbial species and genes (pathways) of multiple samples into consideration for biomarker inference. In this way, the specific modules with diseases consisting of function (genes/pathways) and species will be detected, simultaneously providing functional and taxonomic insights into diseases. Second, IDAM requires relatively large-scale data. Since IDAM uses expression similarity to assess the relevance of gene regulation, a larger sample size will result in more reliable measurement from data. In practice, the sample pool can be an approach that addresses this issue and guarantee the practicability of IDAM. Finally, the provided biomarkers of IDAM only provided associations between microbiomes and diseases. Therefore, rigorous experiments are required to establish causality between the identified biomarkers and the disease of interest, and establish clinical prevalence and utility.

Conclusions

IDAM provides a highly effective approach to identify disease-associated microbial biomarkers based on both gene regulation and context conservation during evolution, based on matched MG and MT data. The identified biomarkers of diseases can be useful in understanding disease pathogenesis and providing guidance for future development of disease diagnosis and treatment.

Methods

Datasets

We downloaded all matched quality-controlled MG and MT data, as well as the corresponding HUMAnN2 results and metadata of samples at the IBDMDB website in March 2020 (https://ibdmdb.org) [22]. In total, there are 198 samples from Non-IBD subjects and 615 samples from patients with IBD (consisting of 381 UC samples and 234 CD samples). Uber-operon structures were from the paper of Che *et al.* [21].

IDAM: a framework of 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 idea 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 (fasta, fasta.gz, fastq, or fastq.gz). The datasets of each sample are firstly processed using HUMAnN2 for gene relative expression within communities [28]. In this study, we directly used the HUMAnN2 results of the datasets from IBDMDB. 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 (totally *m* genes indicating by UniRef90 gene family [38] and forming a gene set *X*), and each column represents a sample (totally *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 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) [38], we align the gene family sequences with genes in uber-operons to determine which uber-operons these gene families belong to. Specifically, the sequences of genes in all uber-operons are collected as a custom database. For each gene family g_i within *X* ($1 \leq i \leq m$), we extract the corresponding sequence from the Uniprot database [38] and align it with the custom database using Blast [39]. We define a gene-to-uber-operon mapping function $\varphi: D \rightarrow H$, where *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 ($1 \leq i \leq m$) 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 single gene g_i. If multiple sequences from different uber-operons satisfy the *E*-value less than 0.001, we will choose the uber-operon corresponding to the smallest *E*-value. By contrast, if no sequence can satisfy the *E*-value less than 0.001, we set $\varphi(G_i) = \{\emptyset\}$. In this way, we construct a map between uber-operon and a gene.

We further define the relationship between uber-operons and gene sets 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 single gene g_i (*i* = 1, 2, …, b). For two gene sets *I* and *J*, we set a reward *U*(*I*, *J*) as gene context conservation measurement based on relative positions of their contained 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 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 the vector composed of expression values of each gene within all samples. Note that here we need to measure expression similarity not only between two genes, but also across multiple genes, while expanding modules in Step 3. This cannot be achieved by existing similarity measurements (e.g., Pearson correlation coefficient) since they are generally used for paired data points. To address this, the continuous expression value of each gene is first discretized to an integer representation by qualitative representation [24] (Additional file 1: Section 2). Based on this, 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 The integration of gene context conservation and expression similarity

A combined score $C(G_i, G_j)$ integrating gene context conservation and 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 (Additional file

1: Section 3 and Additional file 2: Table S7). Finally, the gene pairs are sorted in a descending order for the combined scores, generating a seed list *L.*

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

Since the gene set *X* is a pan-genome of all microorganisms from multiple samples, it consists of a large number of genes. To improve computational efficiency, we partition the rows within the matrix equally into *t* subsets [40, 41], where *t* is determined by the stochastic model [40, 41]. 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 computation burden.

A submatrix is regarded as biologically relevant if it does not occur randomly in 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 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. Note that, the submatrices we target are of rank one mathematically since the corresponding genes are with co-expressed patterns. Suppose a submatrix with rank one is of size *k* × *s*, its probability of random occurrence in $A_{m\times n}$ is calculated as follows [40, 41]:

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

We here think the submatrix is 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 mentioned above. By this, the smallest row size of biologically relevant submatrices can

be determined. Based on the above, users can set the number of subsets ($t = k$ -1). In this paper, the default value of t is set to 10. That is, the matrix will be divided into 10 subsets, which can greatly reduce the computation burden and keep biologically relevant submatrices detectable to a large extent.

3.1 Initialization

We start with the first gene pair in L that satisfies i) at least one gene of these two has not been included in previous modules, or ii) these two genes are respectively included within different previous modules, and there is no overlap between the gene sets of the two modules. We aggregate samples under which the integers of the two genes are identical (represented as set P), forming current module $M = \{Q, P\}$, where Q represents the gene set consisting of the paired genes. Then, the gene pair is 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 ($1 \leq z \leq m$)) and Q, namely $C(G_z, Q)$, is similar to that of two genes g_i and g_j. Specifically, 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 expression similarity 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 will set $M = M'$ and repeat Step 3.2. Otherwise, go to Step 3.3.

3.3 Filter and output

If the proportion of overlapped genes between *Q* and the gene sets of previously identified modules is less than a threshold *f*, module *M* will be provided as the final output, and it is ignored otherwise. In this paper, the default *f* is 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-test was used to assess whether samples with the same biomarker were of highly consistent phenotypes. Specifically, we extracted the phenotypes of all samples from metadata as background and the corresponding phenotypes of samples in each gene module as test sets. For each phenotype (Non-IBD, UC, and CD) and each test set, we counted the phenotype occurrence in the test set and background, respectively. We evaluated the significance of each phenotype using a right-tailed Z-test with the null hypothesis that the observed frequency in the test set was less than or equal to the occurrence probability within the background. The threshold of significant level 0.05 was used with a critical Z-value of 1.645. If Z-score was greater than 1.645 (p-value > 0.05), the null hypothesis would be rejected, meaning the phenotype significantly occurred within the test set. In this case, we would consider the samples within the gene module was 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 firstly identified underlying confounding variables that affected its occurrence. Here, we incorporated diet, age, race, health conditions of other diseases, as well as the other biomarkers and calculated a

propensity score for each sample (using 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 test was used to measure the significance of each biomarker that associates with different phenotypes, and the *p*-value was adjusted by the Benjamini-Hochberg procedure, with an adjusted *p*-value < 0.05 as significantly associated.

Functional analysis of biomarkers

We performed an analysis of the enriched functions of each biomarker. All UniRef90 identifiers were firstly mapped to MetaCyc reactions [28, 42]. The reactions were further associated with MetaCyc pathways [28, 42], 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 associate with UniRef90 gene families and pathways in MetaCyc, m_i is the number of reactions within the i th biomarkers, n_j is the total number of reactions within the j th pathway, and T_{ij} is the observed number of reactions of pathway j occurred in the 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.

Additionally, we used a widely used tool, LEfSe [11], to infer differentially abundance pathways among three phenotypes. This was performed using pathway abundance information of each sample from HUMAnN2.

Comparing IDAM with other methods

Biomarker inference of IDAM was compared with the other four top performing unsupervised tools for grouping genes into functional modules [43], including biclustering-based (QUBIC [24], ISA [25], and FABIA [26]) and decomposition-based (fastICA, an improved method of ICA [23]). All tools were tested and compared based on the filtered gene-sample matrix with 941,785 genes and 813 samples. We run these tools on the Pitzer cluster of the Ohio Supercomputer Center with memory usage set to 300GB [44] with default parameters. Specifically, QUBIC was implemented using the published source code, while the other three were implemented by the R packages, isa2 0.3.5, fabia 2.36.0, and fastICA 1.2.2, respectively.

Species-level analysis

For biomarkers from IDAM, ICA [23], QUBIC [24], and FABIA [26], we extracted species that contribute to the abundance of genes within identified biomarkers, respectively, based on HUMAnN2 output [28]. By this, community-level gene expression (abundance) was decomposed into species-level data. The phylogenetic tree was generated using the software GraPhlAn [45]. Since the biomarkers have been associated with phenotypes by Z-test mentioned above, we assigned the species of each biomarker to the corresponding

phenotype of the biomarker. For each phenotype, we counted the total number of species, as well as the number that matched with the collected species. Then, the total number of species in the phenotype dividing by the number of matched species was calculated as the consistency index.

Besides, the tool, LEfSe [11], was used here for comparison. It is based on species profiling from MetaPhlAn2 to identify statistically different species among the three phenotypes [46, 47]. The significantly differential species of each phenotype were aligned to collected ones, by which we determined the number of matched species and calculated the consistency of each phenotype.

Abbreviations

16S ribosome RNA: 16S rRNA;

Metagenomic: MG;

Metatranscriptomic: MT;

Inflammatory bowel disease: IBD;

Ulcerative colitis: UC;

Crohn's disease: CD.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets supporting the conclusions of this article are available in the IBDMDB database (https://ibdmdb.org).

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

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

Acknowledgements

The authors would like to thank Anjun Ma for his assistance in the design of Fig. 1 and Fig. 2.

References

1. Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science. 2021; 371: eabc4552.
2. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature medicine. 2018; 24:392-400.
3. Consortium THMP. A framework for human microbiome research. Nature. 2012; 486:215-221.
4. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, Buck GA, Snyder MP, Strauss JF, Weinstock GM, et al. The Integrative Human Microbiome Project. Nature. 2019; 569:641-648.
5. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59-65.
6. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006; 444:1027-1031.
7. Dubinsky M, Braun J. Diagnostic and Prognostic Microbial Biomarkers in Inflammatory Bowel Diseases. Gastroenterology. 2015; 149:1265-1274.
8. Lecuit M, Lortholary O. Immunoproliferative small intestinal disease associated with Campylobacter jejuni. Med Mal Infect. 2005; 35:S56-58.
9. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013; 10:1200-1202.
10. Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. Network analyses in microbiome based on high-throughput multi-omics data. Briefings in Bioinformatics. 2020; 22:1639-1655.
11. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biology. 2011; 12:R60.
12. Zhu Q, Jiang X, Zhu Q, Pan M, He T. Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. Frontiers in Genetics. 2019; 10:1182.
13. Wu Y, Jiao N, Zhu R, Zhang Y, Wu D, Wang A-J, Fang S, Tao L, Li Y, Cheng S, et al. Identification of microbial markers across populations in early detection of colorectal cancer. Nature Communications. 2021; 12:3063.
14. Snipen L, Angell I-L, Rognes T, Rudi K. Reduced metagenome sequencing for strain-resolution taxonomic profiles. Microbiome. 2021; 9:79.
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. 2019; 20:154.
16. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology. 2021; 39:105-114.
17. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, et al. Relating the metatranscriptome and metagenome of the human gut. Proceedings of the National Academy of Sciences. 2014; 111:E2329-E2338.
18. Kasmanas JC, Bartholomäus A, Corrêa FB, Tal T, Jehmlich N, Herberth G, von Bergen M, Stadler PF, Carvalho André Carlos Ponce de Leon Ferreira d, Nunes da Rocha U. HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Research. 2020; 49:D743-D750.

19. Yang J, Qin Y, Zhang T, Wang F, Peng L, Zhu L, Yuan D, Gao P, Zhuang J, Zhang Z, et al. Identification of human age-associated gene co-expressions in functional modules using liquid association. Oncotarget. 2017; 9:1063-1074.
20. Lathe III WC, Snel B, Bork P. Gene context conservation of a higher order than operons. Trends in biochemical sciences. 2000; 25:474-479.
21. Che D, Li G, Mao F, Wu H, Xu Y. Detecting uber-operons in prokaryotic genomes. Nucleic acids research. 2006; 34:2418-2427.
22. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019; 569:655-662.
23. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Networks. 2000; 13:411-430.
24. Li G, Ma Q, Tang H, Paterson AH, Xu Y. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Research. 2009; 37:e101-e101.
25. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys. 2003; 67:031902.
26. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Van Sanden S, Lin D, Talloen W, et al. FABIA: factor analysis for bicluster acquisition. Bioinformatics. 2010; 26:1520-1527.
27. Xie J, Ma A, Zhang Y, Liu B, Cao S, Wang C, Xu J, Zhang C, Ma Q. QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. Bioinformatics. 2020; 36:1143-1149.
28. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods. 2018; 15:962-968.
29. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology. 2012; 13:R79.
30. Imhann F, Vich Vila A, Bonder MJ, Fu J, Gevers D, Visschedijk MC, Spekhorst LM, Alberts R, Franke L, van Dullemen HM, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut. 2018; 67:108-119.
31. Klaassen MAY, Imhann F, Collij V, Fu J, Wijmenga C, Zhernakova A, Dijkstra G, Festen EAM, Gacesa R, Vich Vila A, Weersma RK. Anti-inflammatory Gut Microbial Pathways Are Decreased During Crohn’s Disease Exacerbations. Journal of Crohn's and Colitis. 2019; 13:1439-1449.
32. Schirmer M, Garner A, Vlamakis H, Xavier RJ. Microbial genes and pathways in inflammatory bowel disease. Nature reviews Microbiology. 2019; 17:497-511.
33. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature microbiology. 2019; 4:293-305.
34. Kuballa A, Geraci M, Johnston M, Sorrentino D. The Gut Microbial Profile of Preclinical Crohn's Disease Is Similar to That of Healthy Controls. Inflamm Bowel Dis. 2020; 26:1682-1690.
35. Vich Vila A, Imhann F, Collij V, Jankipersadsing SA, Gurry T, Mujagic Z, Kurilshikov A, Bonder MJ, Jiang X, Tigchelaar EF, et al. Gut microbiota

composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci Transl Med. 2018; 10:eaap8914.
36. Skoufos G, Kardaras FS, Alexiou A, Kavakiotis I, Lambropoulou A, Kotsira V, Tastsoglou S, Hatzigeorgiou Artemis G. Peryton: a manual collection of experimentally supported microbe-disease associations. Nucleic Acids Research. 2020; 49:D1328-D1333.
37. Wu J, Wei Z, Cheng P, Qian C, Xu F, Yang Y, Wang A, Chen W, Sun Z, Lu Y. Rhein modulates host purine metabolism in intestine through gut microbiota and ameliorates experimental colitis. Theranostics. 2020; 10:10665-10679.
38. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23:1282-1288.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403-410.
40. Wang Z, Li G, Robinson RW, Huang X. UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data. Sci Rep. 2016; 6:23466.
41. Ben-Dor A, Chor B, Karp R, Yakhini Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. J Comput Biol. 2003; 10:373-384.
42. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 2020; 48:D445-d453.
43. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nature Communications. 2018; 9:1090.
44. Center OS. Ohio Supercomputer Center. 1987.
45. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015; 3:e1029.
46. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012; 9:811-814.
47. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015; 12:902-903.

Supplementary Information

Additional file 1: 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. Sections 2 and 3 describe how to do discretization and select parameter *w*.

Additional file 2: Supplementary results

This file includes supplementary results in the Result part and parameter selection of w in the Method part. The first is an overview of Additional file 2. Table S1 shows the size of identified gene modules from IDAM. Table S2 shows the Z-score and *p*-value of each module for three phenotypes. Table S3 shows the adjusted *p*-value from Wilcoxon signed-rank test. Table S4 shows pathways enriched for identified biomarkers from IDAM. Table S5 shows the species associated with different phenotypes from IDAM. Table S6 shows the species associated with different phenotypes collected from previous studies. Table S7 shows the performance of IDAM with different parameter *w*, which is attached to Methods and Additional file 1: Section 3.

Figure 1

Click here to access/download;Figure;figure_1.tif

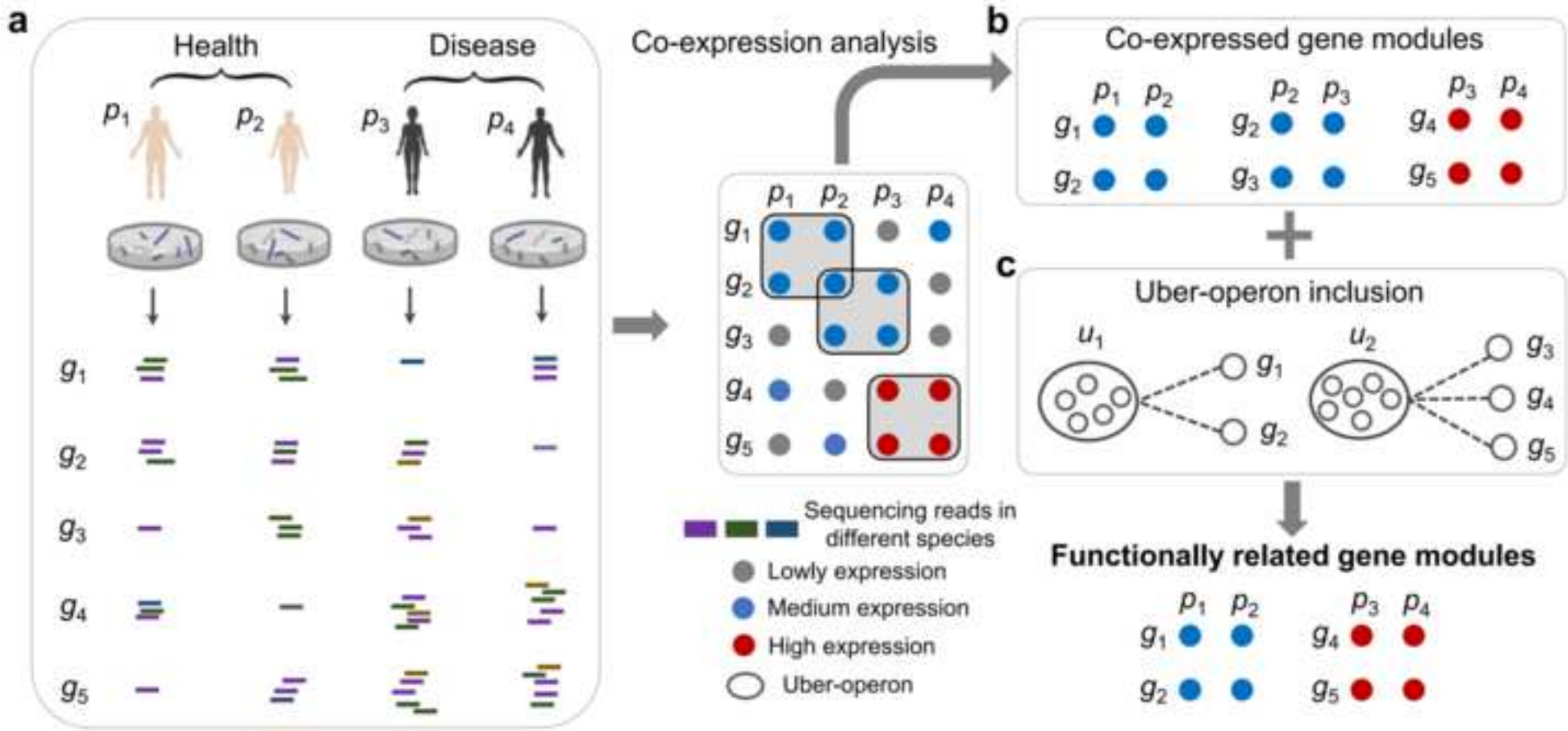


Figure 2

Click here to access/download;Figure;figure_2.tif

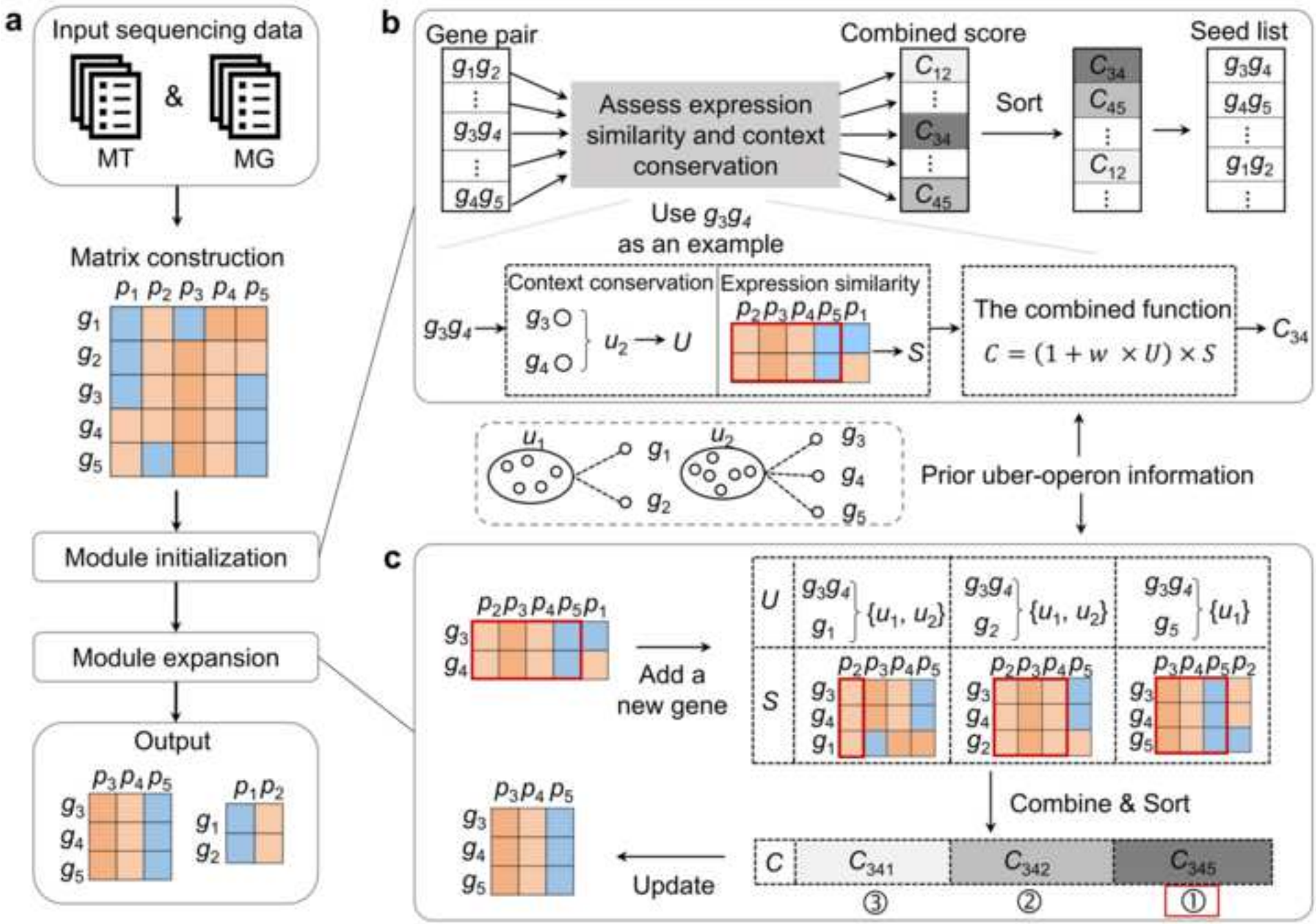


Figure 3

Click here to access/download;Figure;Figure_3.tif

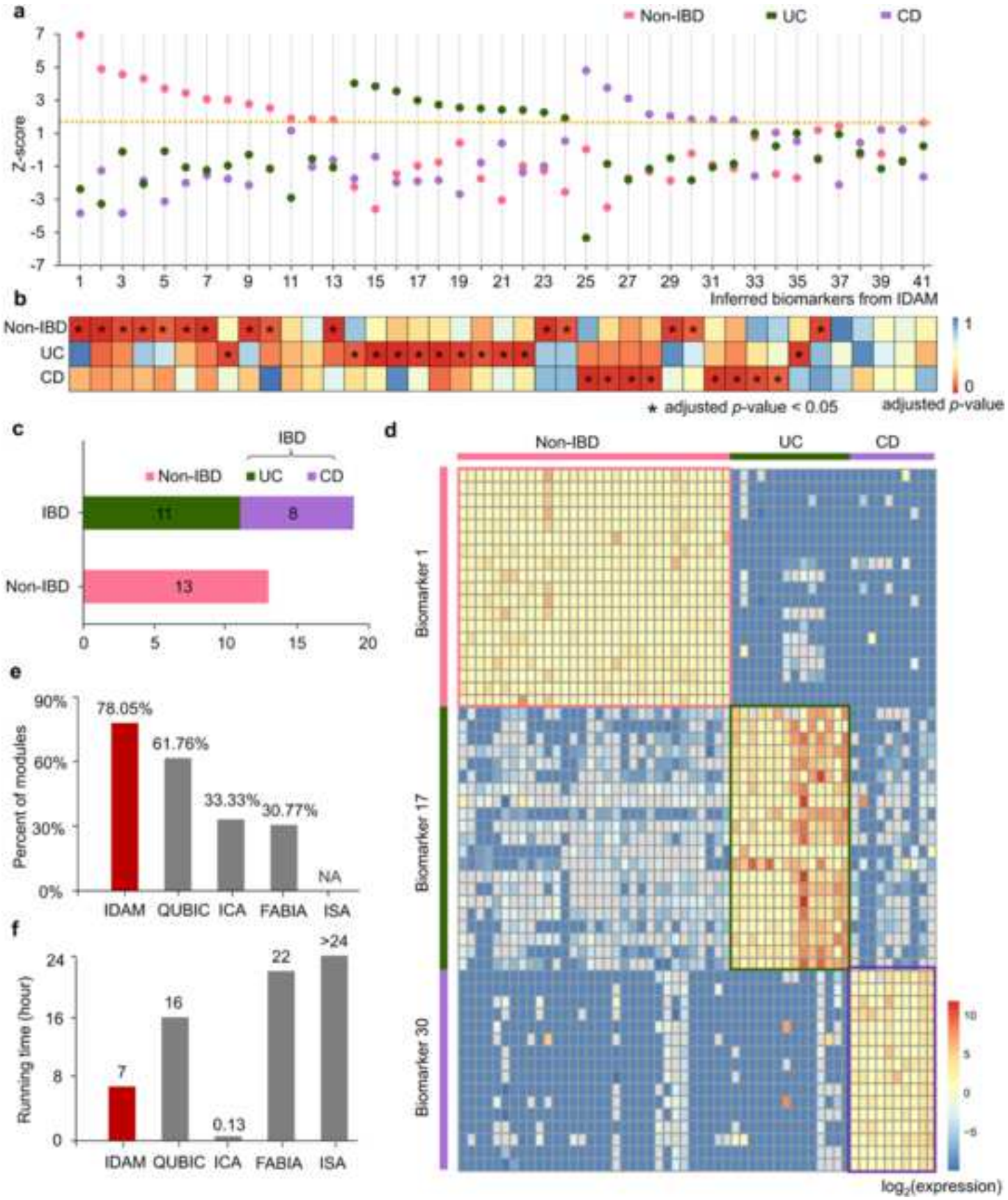
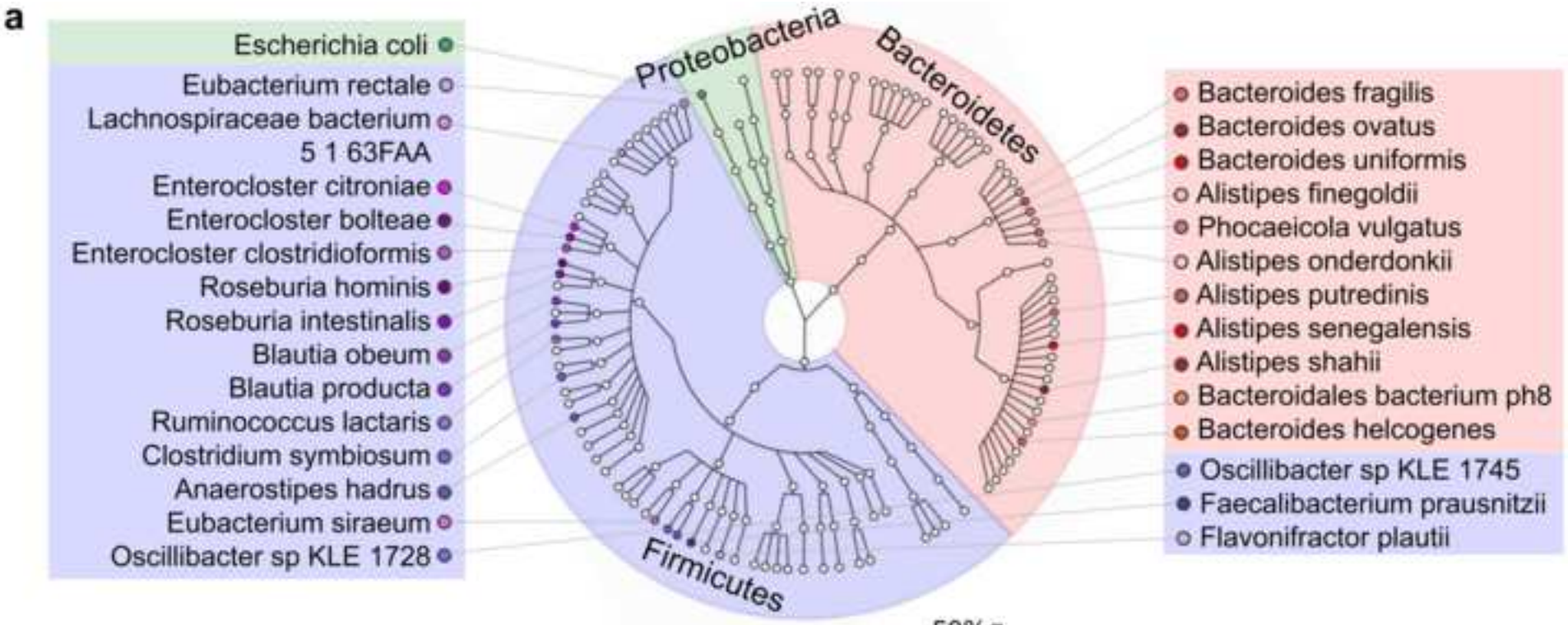


Figure 4

Click here to access/download;Figure;figure_4.tif



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%)

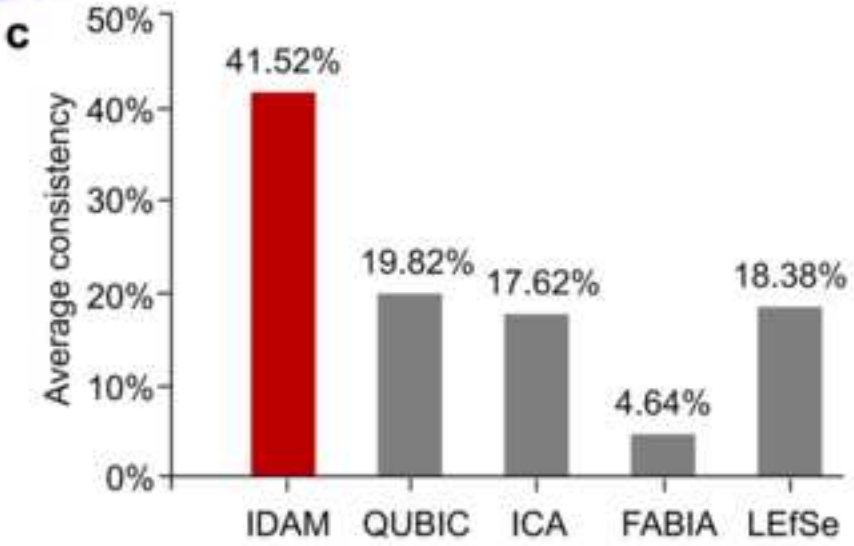
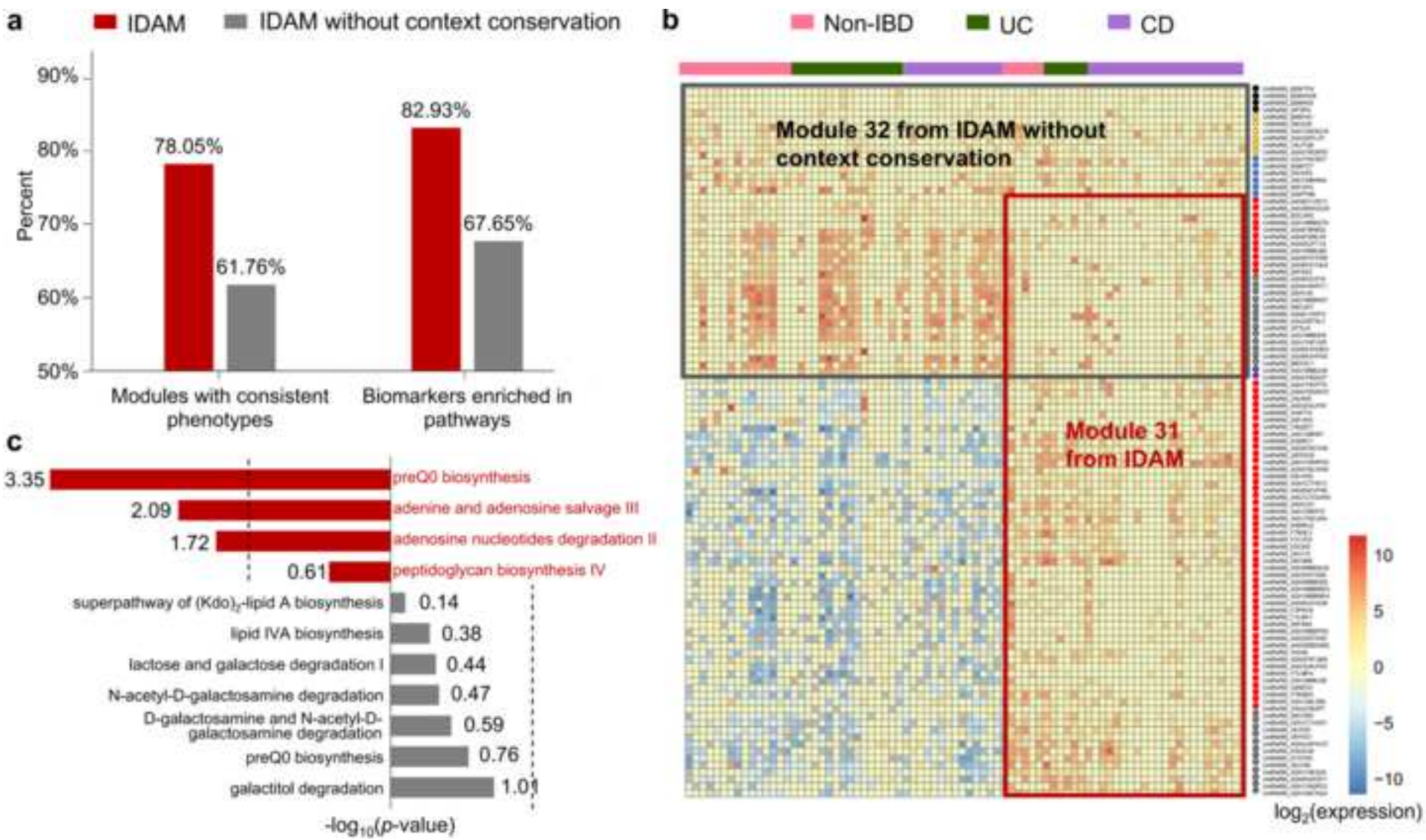


Figure 5

Click here to access/download;Figure;figure_5.tif



Click here to access/download
Supplementary Material
Additional file 1.docx

Supplementary table

Click here to access/download
Supplementary Material
Additional file 2.xlsx