Perz *et al. BMC Bioinformatics* 2019, **20**(Suppl 2):96
https://doi.org/10.1186/s12859-019-2623-x

BMC Bioinformatics

RESEARCH Open Access

MNEMONIC: MetageNomic Experiment Mining to create an OTU Network of Inhabitant Correlations

Aleksandra I. Perz[1*], Cory B. Giles[1,4], Chase A. Brown[1,3], Hunter Porter[1,3], Xiavan Roopnarinesingh[1,2] and Jonathan D. Wren[1,2,3,4*]

From The 15th Annual MCBIOS Conference
Starkville, MS, USA. March 29 - 31 2018

Abstract

Background: The number of publicly available metagenomic experiments in various environments has been rapidly growing, empowering the potential to identify similar shifts in species abundance between different experiments. This could be a potentially powerful way to interpret new experiments, by identifying common themes and causes behind changes in species abundance.

Results: We propose a novel framework for comparing microbial shifts between conditions. Using data from one of the largest human metagenome projects to date, the American Gut Project (AGP), we obtain differential abundance vectors for microbes using experimental condition information provided with the AGP metadata, such as patient age, dietary habits, or health status. We show it can be used to identify similar and opposing shifts in microbial species, and infer putative interactions between microbes. Our results show that groups of shifts with similar effects on microbiome can be identified and that similar dietary interventions display similar microbial abundance shifts.

Conclusions: Without comparison to prior data, it is difficult for experimentalists to know if their observed changes in species abundance have been observed by others, both in their conditions and in others they would never consider comparable. Yet, this can be a very important contextual factor in interpreting the significance of a shift. We've proposed and tested an algorithmic solution to this problem, which also allows for comparing the metagenomic signature shifts between conditions in the existing body of data.

Keywords: Human microbiome, Differential abundance, Case-control shift, Meta-analysis

Introduction

Communities of microbial species have co-evolved within a number of microenvironments, both outside and inside of larger organisms. For example, humans do not produce all the enzymes necessary to metabolize the spectrum of nutrients they take in as food, and the gut environment, through some uncertain mechanism, permits microbes that metabolize nutrients for the host in exchange for non-essential products to effectively be permanent yet dynamic co-habitants. However, pathogenic microbes can also occupy niches in the human body and often come with unwanted consequences for the host. The link between pathogenic microbes and acute conditions (e.g., diarrhea) has long been known but, in part, current studies are exploring whether or not chronic conditions may be due to changes in microbial communities. By cataloging microbial composition in these environments, we hope to better understand what role they may play in an array of physiological processes and diseases.

Changes in relative microbial abundance, known as differential abundances (DA), are a common measure of microbial variability and a starting point for understanding

* Correspondence: jdwren@gmail.com; jonathan-wren@omrf.org
[1] Arthritis and Clinical Immunology Program, Division of Genomics and Data Sciences, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104-5005, USA
Full list of author information is available at the end of the article

how certain mutualistic or pathogenic species may contribute to vital functions or diseases. Thus, changes in the relative abundance of microbial species could either cause or correlate with certain diseases, either by removing symbiotic microbes or by introducing hostile/non-beneficial microbes. And although a metagenomic experiment can quantify the shift in species abundance, interpreting its potential relevance and significance relies in part upon putting the newly observed shift within the context of previously observed shifts.

Metagenomic studies can be motivated by several goals, including the discovery of novel microbial genes of interest [1, 2], validation of metabolic hypotheses [3–5], profiling of the relationship between microbial community composition and variation in environmental or geographic parameters [6, 7] and assessment and comparison of the global metabolic complement found in one or more habitats [8–12]. In particular, there has been a substantial increase in examining potential associations between microbial changes and either chronic or late-onset human disease [13].

However, in human gut microbiome studies it is often the case that the composition of the microflora varies greatly depending on variables that are not directly related to a studied disease or condition, such as geographical location or project. The public availability of metagenomic data provides a powerful opportunity to corroborate the significance of microbial changes by searching for similar changes, thus showing robustness. The conditions in which the changes observed could either help corroborate one's observations (e.g., if the experiment was a similar experiments) or could raise interesting questions (e.g, if the experiment was very different, yet yielded similar results).

Meta-analytic approaches are useful to identify statistically significant changes, but are likely limited when it comes to understanding the biological significance of microbial changes [14]. Nonetheless, identifying similar and opposing shifts in species abundance accelerates both one's confidence in the robustness of the results and biological interpretation of the changes.

The gut microbiome

Although the number of experiments that could be analyzed is growing rapidly (Figs. 1 and 2), because they lack standardized meta-data and annotations describing which sets belong to experimental groups and which to control groups, automatically determining this is still an open problem. However, the American Gut Project (AGP) [15] is one of the largest studies conducted to date and has a highly structured description of potential covariates, such as dietary preferences, so we chose the AGP subset for analysis.

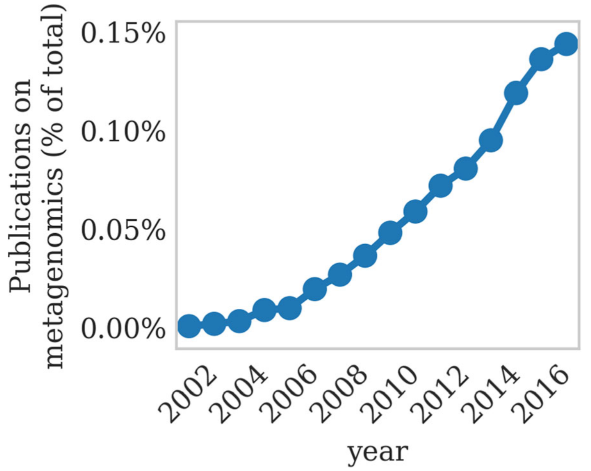

Fig. 1 The percentage of publications in PubMed containing the word "metagenomics" per year is growing rapidly relative to the overall growth of PubMed

The gut microbiome also has the advantage that it has been studied in a variety of contexts and has been shown to change in a range of diseases. Gastrointestinal tract and metabolic diseases have been studied particularly extensively, but the effect of gut microbes seems to be broader than that: it has been suggested that it might be implicated in autoimmune, mental, and cardiovascular diseases as well. Rheumatoid arthritis (RA) and inflammatory bowel disease (IBD) [16–18] diarrhea [19] as well as antibiotic treatment were shown to decrease the microbiota diversity. Further, obesity, metabolic syndrome, and type II diabetes [7, 11, 20–22], colitis [23, 24], colorectal cancer [25] have all been shown to be associated with changes in microbiome. There is some evidence that the microbiome might be implicated in autism spectrum disorders and cardiovascular disease,

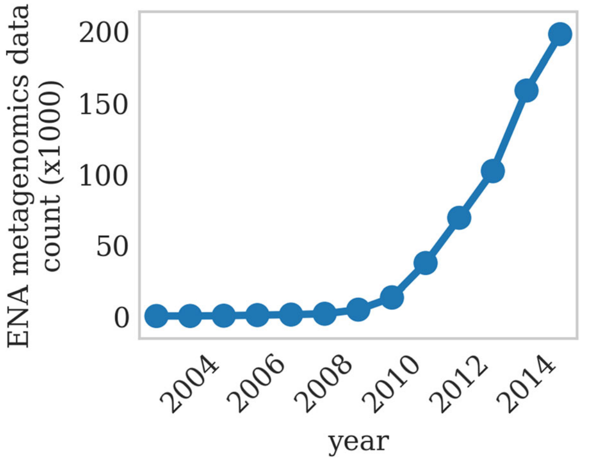

Fig. 2 Growth of metagenomics data in ENA

but the exact nature of this link has not been established ([26–28].

Factors that have been shown to affect the microbiome composition of the human gut include diet, geographical location, culture, and genetics [11, 29–36]. Specifically, there seems to be a great difference between the Western and plant-based diet, to the point of distinguishing human gut 'enterotypes' based mostly on prevalence of taxa that are associated with these diets. Consumption of salt can increase the prevalence of specific genes in the gut [37].

Changes in the gut microbiome also occur during late pregnancy [38]. Route of delivery affects the initial gut microbiome [39, 40], and the samples from infant gut cluster more readily with samples from human vagina than adult gut [41]. The largest changes are seen within the first three years of life [31, 41–45].

Aging has also been shown to change the gut microbiota, specifically to increase the number of 'subdominant species' [42].

Thus, a number of connections between microbiome changes in the gut and human health have been established. Part of the growing scientific interest in metagenomics studies is the therapeutic potential for intervention. If we can establish and understand links between species presence and/or relative abundance and human disease, there are a number of safe and inexpensive ways the microbiome could be altered in an attempt to alter the course of the disease. Potential clinical applications include supplementation of beneficial bacteria to supplement or help regulate host metabolism, metabolize molecules that might be problematic within the human digestive tract, and identifying microbes that might serve as natural competitors to harmful species.

Metagenomics resources

Along with the rush to sequence microbial genomes within their environment, a number of bioinformatics resources were developed for storage, functional and taxonomic analysis, visualization, and retrieval of data. There are a number of repositories where users can store and browse metagenomic data: Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis CAMERA (no longer operating, but the data is still available through iMicrobe) [46], Integrated Microbial Genomes and Metagenomes IMG/M ([47] Metagenomics-Rapid Annotations using Subsystems Technology MG-RAST ([48]), and EBI Metagenomics ([49]). They offer storage of sequencing data, as well as processing and basic functional and taxonomic analyses. Raw sequencing reads may be stored in databases like SRA or ENA. SEED and KEGG databases are often used as the reference for the functional components of the metagenome.

Taxonomic abundance change (i.e., differential abundance between conditions), can be analyzed with MEGAN, Dendroscope3, LEfSe [50], ANCOM [[51]. as well as a number of dedicated R packages (Phyloseq, metagenomeSeq,, and MaAsLin [52], BhGLM [53], as well as R packages primarily used for RNA-Seq data and adapted to microbiome analysis (DESeq2, edgeR [54], limma-voom [55] and web applications (Metastats); QIIME and MEGAN aim to integrate many analysis steps into a pipeline, which may also include functional analysis. QIIME is a widely-used and rich suite of tools for command-line analysis and visualization of sequencing metagenomics data that also integrates other tools [56]. Building on top of some of the previously mentioned tools, Nephele offers a web interface and the ability to compare the uploaded data to a data from a selected body site from Human Microbiome Project [12, 57]). A specific type of analysis may also be performed with eudysbiome - an R package whose flag concept is the dichotomous character of host-microbe interaction, or MetaMIS, which simulates the interactions between microbes in a group of samples across time.

While many of these packages can aid investigators in identifying species that change between two conditions of interest, they are focused on many pairwise comparisons and lacking in the ability to compare entire vectors of changes. By focusing on comparing these vectors of differential abundance, DA shifts, our tool empowers researchers to both build confidence in their results by comparing to similar experiments and gain novel insight into microbiome changes by comparing to other shifts associated with perturbations and pathologies.

Mnemonic

We present a tool, MNEMONIC (MetageNomic Experiment Mining to create an OTU Network of Inhabitant Correlations), whose main goal is to calculate microbial shifts occurring in different conditions and then compare those shifts between the conditions, thereby assessing the similarity of conditions in terms of how the microbiome changes. It allows for the exploration of different shifts, as well as cross-referencing those changes with literature association data and microbial traits. One can also provide their own data and compare their shifts to the shifts that we observe in the AGP data. MNEMONIC is a python package and is publicly available for download at https://gitlab.com/wrenlab/mnemonic. It uses the publicly available data from the EBI Metagenomics portal by fetching microbial count data for the samples within the American Gut Project. To model the differential abundance between groups of samples it uses the R package edgeR. For other tasks, python was used.

One type of question that can be tackled with this approach is which taxonomic groups shift in the same or opposite direction as a response to a certain condition and whether there are other conditions that a similar change is seen in. Conceivably such shared shifts in microbiome may indicate that similar mechanisms are in play. For example a certain dietary habit might influence the host microbiome in a specific way, favoring certain species over other. A now highly-abundant species might influence the host's health state eliciting or reducing an immune response from the host, or affecting the host via a product of microbial metabolism. Diseases may have a common cause in terms of the microbiome composition and function that, if evaluated, may allow a researcher to form hypotheses on whether a certain condition may be improved with a treatment that is commonly used for another.

Another example of a hypothesis that can be addressed is evaluating which taxonomic groups shift with - or opposite to - each other regularly and would allow to make statements about the interactions between the microbes themselves.

To the best of our knowledge, no other software has been described that is capable of comparing and visualizing shifts between the differential abundance vectors between conditions.

In addition to this novel functionality, MNEMONIC is also capable of bringing publicly available information as a context for a new study. For example, if a researcher plans a study on how the microbiome changes in diabetic people as compared to the non-diabetic people, MNEMONIC can provide a summarization of the results for this comparison from the AGP data. This is very useful for validating, as well as contrasting, the results of a new study.

Methods and implementation

Metagenomics data

There are two major approaches to obtaining metagenome data. One is amplification of bacterial 16S rRNA hypervariable regions, followed by amplicon sequencing and assigning the sequences to specific operational taxonomic units (OTUs). This approach can detect bacteria only. The hypervariable region may not be specific to a species, therefore sometimes only allows annotation to a higher taxonomic level. Gene presence in the sample cannot be obtained directly from the DNA, and can only be estimated based on the OTU presence.

Another common approach is whole genome shotgun sequencing. WGS provides information about the full DNA sequence in the sample, therefore allowing for a more accurate, as well as more sensitive [58]. species annotation. It can also identify sequences form kingdoms other than bacteria. Is however more expensive and requires a high coverage [59] and the analysis is computationally heavier. WGS also allows to directly map DNA sequences to a gene reference database for functional profiling.

Data characterization

EBI's repository was chosen as the primary source of data for the project. As of today, it is a home to over 1500 public projects, with more than 90,000 samples from various environments, among which the most abundant are human digestive system, soil, rhizosphere, rumen of ruminal animals, water bodies, etc. The API for EBI Metagenomics portal is under development [https://github.com/ProteinsWebTeam/ebi-metagenomics]. The data was acquired from the EBI Metagenomics portal with the aid of the MGPortal data retrieval script [https://github.com/ProteinsWebTeam/ebi-metagenomics]. The sample metadata was downloaded from the American Gut repository [https://github.com/biocore/American-Gut]. The American Gut Project is the main source of data for MNEMONIC. From the project metadata, we extracted 139 variables that could be coerced into numeric data, assigned them to 5 major categories, as well as 273 VioScreen-derived variables, and used them for further modeling of abundance changes (Figs. 3 and 4).

The ProTraits database annotates microbial species with variables that relate to their metabolism, phenotype, or ecosystem. The annotations were obtained by the means of text-mining of the scientific literature, and contain both pre-defined variables and novel variables inferred from the text. The predictions are augmented with comparative genomics, gene pattern similarities, codon usage, proteome composition and co-occurrence in the metagenomic data. The original dataset provides information on the probability that there is a link

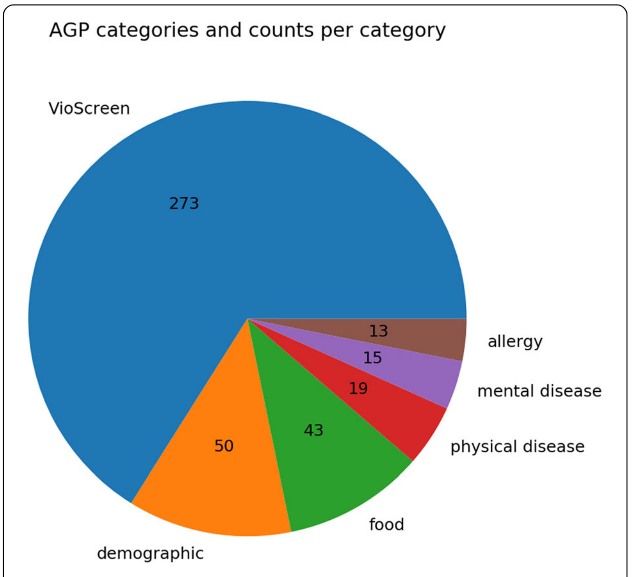

Fig. 3 AGP metadata categories

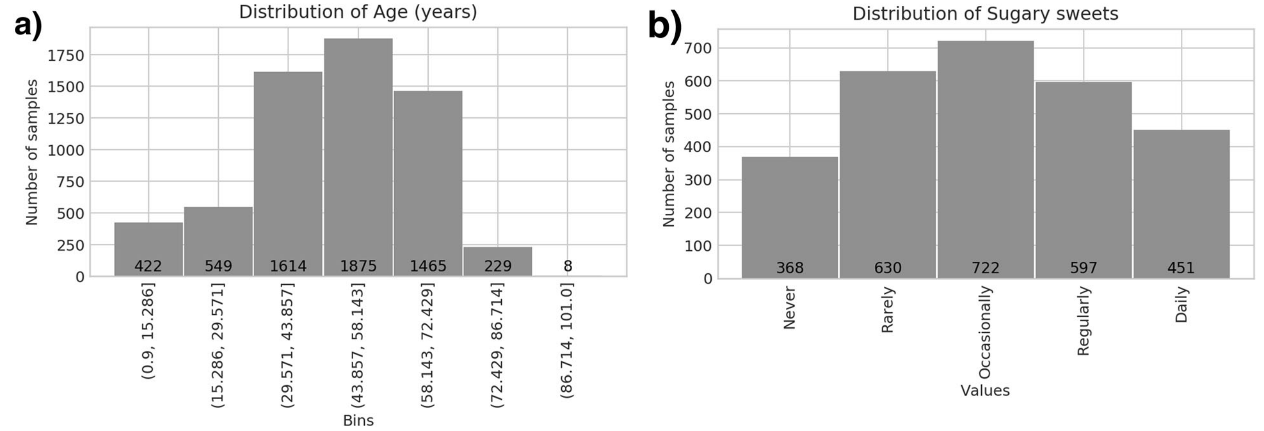

Fig. 4 Distributions of selected AGP metadata variables. Both age (**a**) and frequency of sugary sweet consumption (**b**) show a wide range of values enabling re-analysis. For variables involving frequencies of a given activity, such as dietary consumption, data was re-coded in terms of frequency per day (thus, a value of 1.0 in Fig. 4b indicates daily sugary sweet consumption)

between the taxon and the trait in question. It contains 3046 species that overlap with the AGP dataset, that can be annotated with 424 variables [60]. In this project, we download the binarized dataset with the cutoff at 0.9 from the ProTraits website [http://protraits.irb.hr].

EBI metagenomics database also provides the functional data, in the form of GO category annotation, paired with the corresponding OTU counts for sample. To evaluate the consistency of the OTU count data obtained from EBI, we compared the results obtained from the OTU matrix and GO matrix in terms of distances between samples that had both annotations. If the results are similar, the matrices obtained with either of the method should also be similar. A Mantel test was performed on affinity matrices calculated from both sources. A correlation of 0.25 on 10,439 samples from various environments was significant ($p < 0.001$), establishing that there is resemblance between the two similarity matrices. More importantly, the permutation determined that the similarity is non-random.

Another way to evaluate the data is to compare the similarities between taxonomic groups to a similarities derived from a permuted dataset, based on the OTU count data. Logically, the similarities within a higher-level taxonomic group (e.g. within a genus) should be higher than those of a sample of the same size of random taxa derived from the permuted table (species from mixed genera). We indeed do observe this behavior for all taxonomy level from kingdom to species (Fig. 5).

Differential abundance matrix

To determine differentially abundant taxa for each of the metadata variables in the AGP's metadata matrix, we first eliminated samples with fewer than 10,000 mapped 16S reads and taxa with fewer than 1 mapped read across all AGP samples. The primary reason for

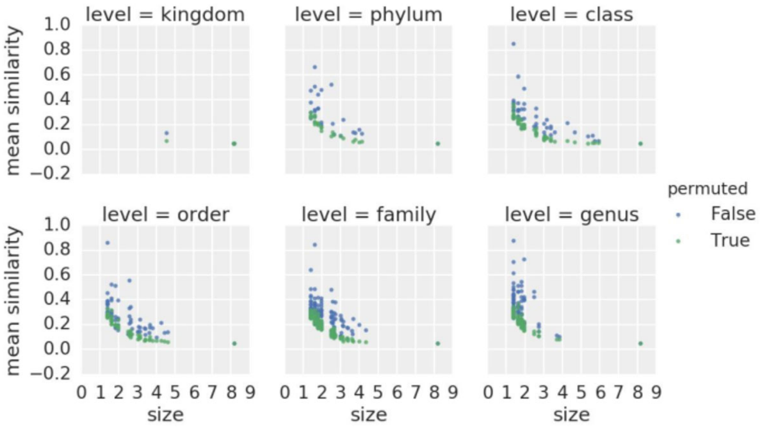

Fig. 5 Mean similarity within taxonomic levels. "size" is equal to log(number of species). The transformation of counts has been applied with relation to taxonomic units

excluding low-abundance taxa from further analysis was that we are interested in differential abundance *vectors* (i.e., the ordered set of fold-changes for differentially expressed taxa) for particular conditions, rather than individual taxa per se. As mean abundance of an OTU decreases, so too does the magnitude and variance of fold-changes for those OTUs, making it more difficult to stably compare conditions in terms of their overall shifts.

After these filtering steps, the differential abundance matrix was constructed as follows. For each metadata variable in the AGP:

1. Samples with missing values for the metadata variable were dropped.
2. A univariate negative binomial model was fit between the OTU count matrix and the metadata variable using the edgeR package [61]. The model is of the simple form: *Count ~ MetadataVariable*.
3. From the model, the magnitude and significance of each OTU's univariate association between abundance and that metadata variable were obtained using edgeR's generalized likelihood ratio test (glmLRT).

After repeating this procedure for each metadata variable, the results are concatenated into a single matrix. The result of this procedure can be conceptualized as a matrix of fold-changes for each OTU and experimental condition. This matrix of measures of differential abundance, or "shifts", can then be used to determine the similarity between each condition's abundance changes, or between a query "shift" and the database of AGP differential abundance vectors. Such matrices are fit for each level of the taxonomy (e.g., species, genus, phylum, etc) after collapsing the input abundance matrix to the appropriate level by summing the counts of all child taxa.

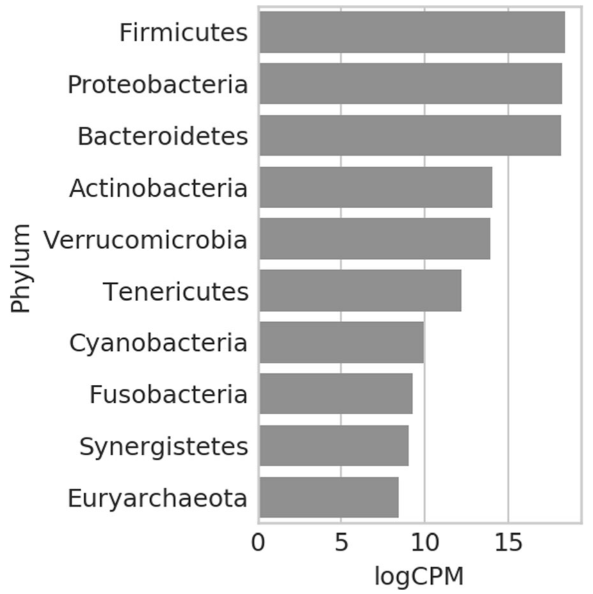

Fig. 6 Most abundant phyla in AGP dataset. Globally, within the AGP data, we observe that the most abundant phyla are Firmicutes, Proteobacteria, and Bacteroidetes. This is consistent with previously reported results

Results

The MNEMONIC package allows users to query and generate analyses of AGP and EBI metagenomic abundance data and differential abundance vectors, as well as compare user-provided data or shifts to public datasets. Figure 6 presents the most abundant phyla in the AGP dataset, among which are Firmicutes, Proteobacteria, and Bacteroidetes. This is consistent with the previously reported findings. Figure 7 shows a clustermap of diet-related terms between food metadata variables in the AGP and 20 differentially abundant microbial taxa within this term set. Microbial abundance associated with food terms results in clusters reflective of dietary considerations. The effect of refined sugar-containing diets on the microbiome has been well-characterized, even after controlled for obesity [62]. Similarly, we see foods with related nutritional content like sugar-related food terms including sugary sweets, frozen dessert, and sugar sweetened drinks cluster closely. Recent studies have shown microbial shifts associated with meat-containing diets and vegetarian diets, reflecting the clustering shown in Fig. 7 [63]. While meat terms like red meat, poultry, and seafood are closely related terms in the dendrogram, their closest related terms like eggs, vegetable and grain suggest food processing, and food terms indicating minimal processing, have a notable effect on microbial abundance. This effect also extends to other foods with limited processing also typically considered part of a healthy diet - vegetable, whole grain, olive oil, and home-cooked meals are closely clustered terms in the dendrogram. This type of change in microbial abundance has been previously shown in dysbiosis resulting from diets containing processed food additives like dietary emulsifiers [64].

Also of note is that probiotics and lactose have similar DA shifts and are grouped somewhat closely, possibly indicating that lactose-annotated samples were capable of lactose metabolism. This effect on microbial abundance is then observed in samples annotated with probiotic use, which is often used to confer lactose metabolizing bacteria in the case of lactose intolerance. This type of alleviation of lactose intolerance has been previously shown in studies like Vonk et al. using probiotic yogurt [65].

Nitrogen-restricted diets are also known to promote healthy aging in mice, possibly via improving microbial community structure [66], and a variety of diseases are

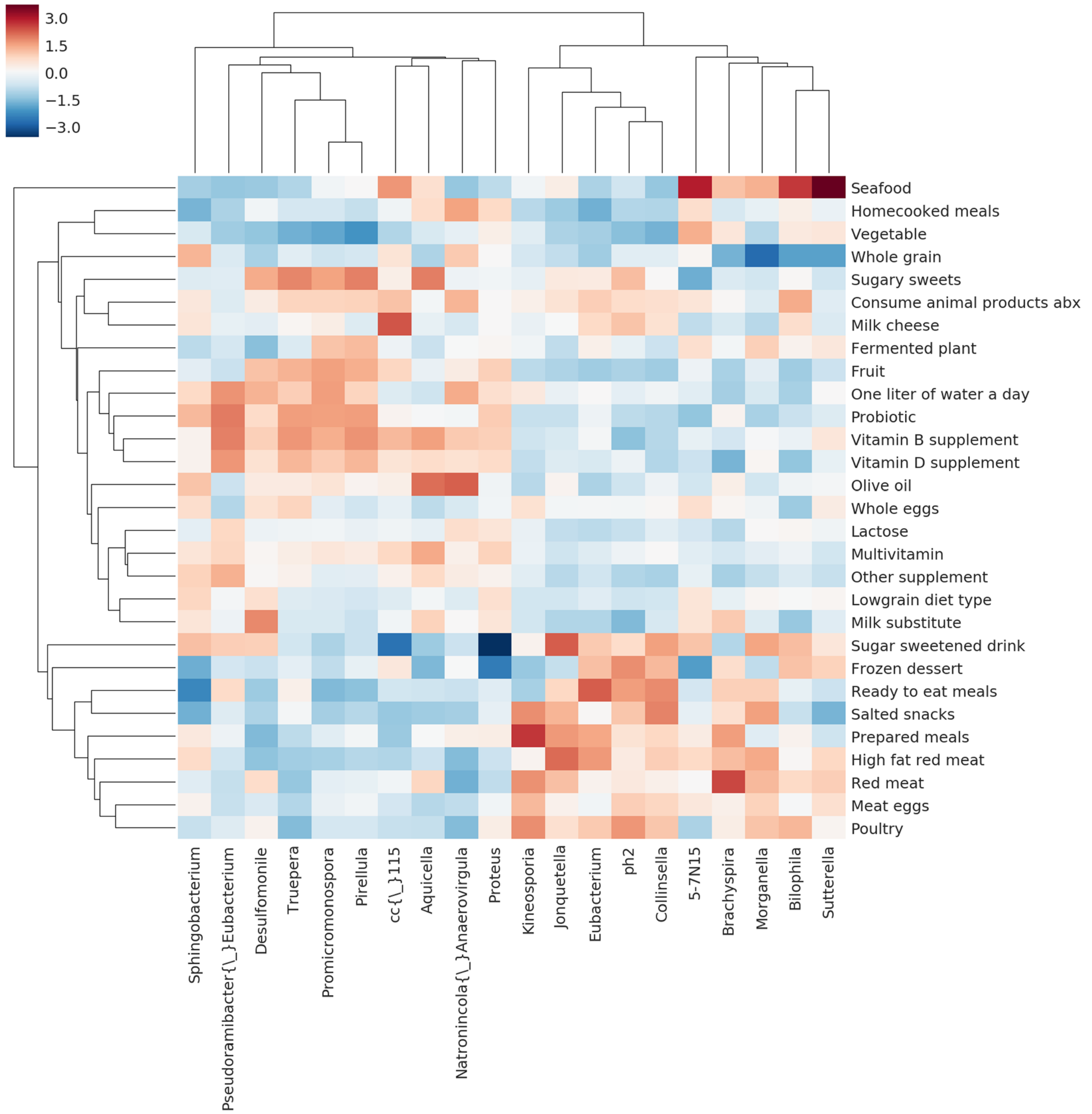

Fig. 7 AGP food differential abundance vectors cluster dietary terms by metabolic content. Each value in the heatmap represents the log-fold-change of each species-term combination, normalized across rows by Z-score. Feature selection on genera was performed by identifying the 20 genera most capable of dividing terms into 2 clusters using the ANOVA F-statistic. The resulting clustering of diet-related terms broadly divides dietary differential abundance vectors into protein-rich, simple carbohydrate, and complex carbohydrate clusters

correlated in severity or occurrence with microbial diversity; either directly, such as colon cancer [67] and systemic lupus erythematosus [68], or inversely, such as inflammatory bowel disease [69] and rheumatoid arthritis [70]. To assess whether dietary effects were associated with changes in microbial diversity in the AGP, we correlated the ACE metric [71] for microbial diversity with the frequency of consumption of various dietary items. Figure 8 shows that higher consumption of high-protein items is broadly associated with decreased microbial diversity, whereas consumption of probiotics, fruits and vegetables, and milk products is associated with increased diversity. This is consistent with previous findings showing that a high-protein diet decreases microbial diversity compared with a balanced protein/carbohydrate diet [72], fruit, legume, and vegetable-rich “agrarian” diets increase gut microbial diversity [73].

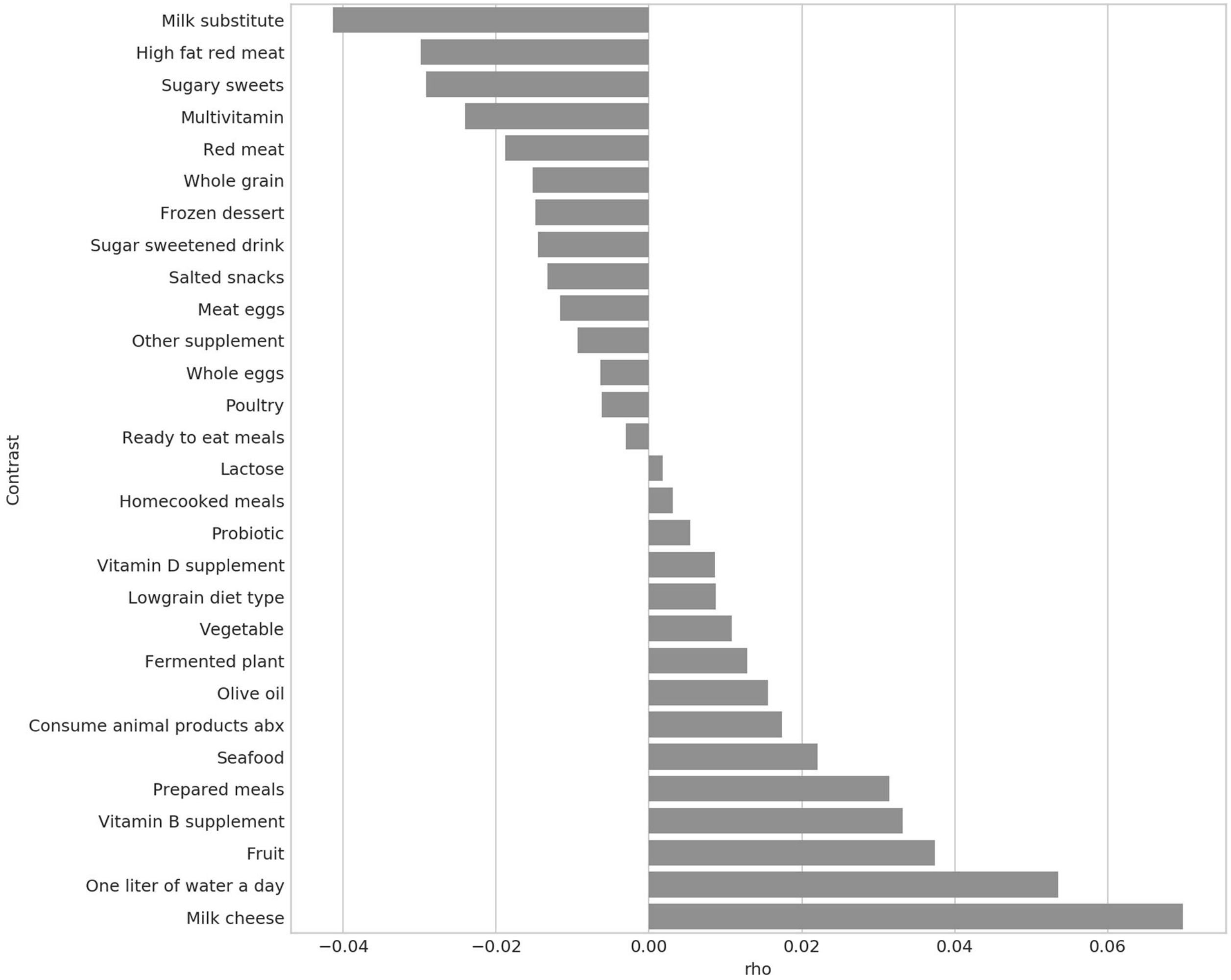

Fig. 8 Protein-rich foods are associated with decreased microbial diversity in the AGP. For each AGP sample, the Abundance-covered Coverage Estimator (ACE) metric of microbial alpha diversity was computed using the scikit-bio Python package (https://github.com/biocore/scikit-bio), and the resulting vector compared to each dietary metadata variable using Spearman's correlation. A high or low rho indicates a respective increase or decrease in microbial diversity is associated with increased frequency of consumption of this type of food within the AGP

In order to further assess the role of metabolic changes during microbiome-associated diseases, we obtained a matrix of predicted metabolic roles and other traits of microbial species from the ProTraits database [60] and correlated those predicted probabilities with the vector of $\log_2$ fold changes for AGP differential abundance vectors. We found that dietary fruit consumption was associated with a significant increase in OTUs annotated with fructose metabolism, whereas red meat consumption was significantly associated with increases in microbes annotated with metabolism of amino acids including aspartate and glutamate, as well as alkaline phosphatase and lipase C14 metabolic activity (Additional file 1).

We then applied this analysis to diabetes mellitus (DM; Fig. 9) and inflammatory bowel disease (IBD; Fig. 10), and found that whereas both were associated with significant increases in glycogen-metabolizing microbes, DM was also marked by an increase in metabolism of the monosaccharide ribose and disaccharide cellobiose, as well as the sorbitol, which is often used as a sugar substitute, and a decrease in maltose metabolism. Decreasing maltose metabolism has been reported in literature to improve blood glucose after sugar challenge and reduction in maltase activity is a desired effect in anti-diabetic therapies such as trigonelline [74, 75]IBD was associated with increases in starch and beta-galactosidase activity. Broadly, then, both disorders are marked by a shift of the microbiome towards carbohydrate metabolism.

There is increasing amount of epidemiological evidence that the microbiome might be involved in the development of late-onset autism. Antimicrobial therapy seems to precede the symptoms [76] and a subsequent vancomycin therapy can alleviate the symptoms short-term [77]. It has been proposed that some gut microbes may produce

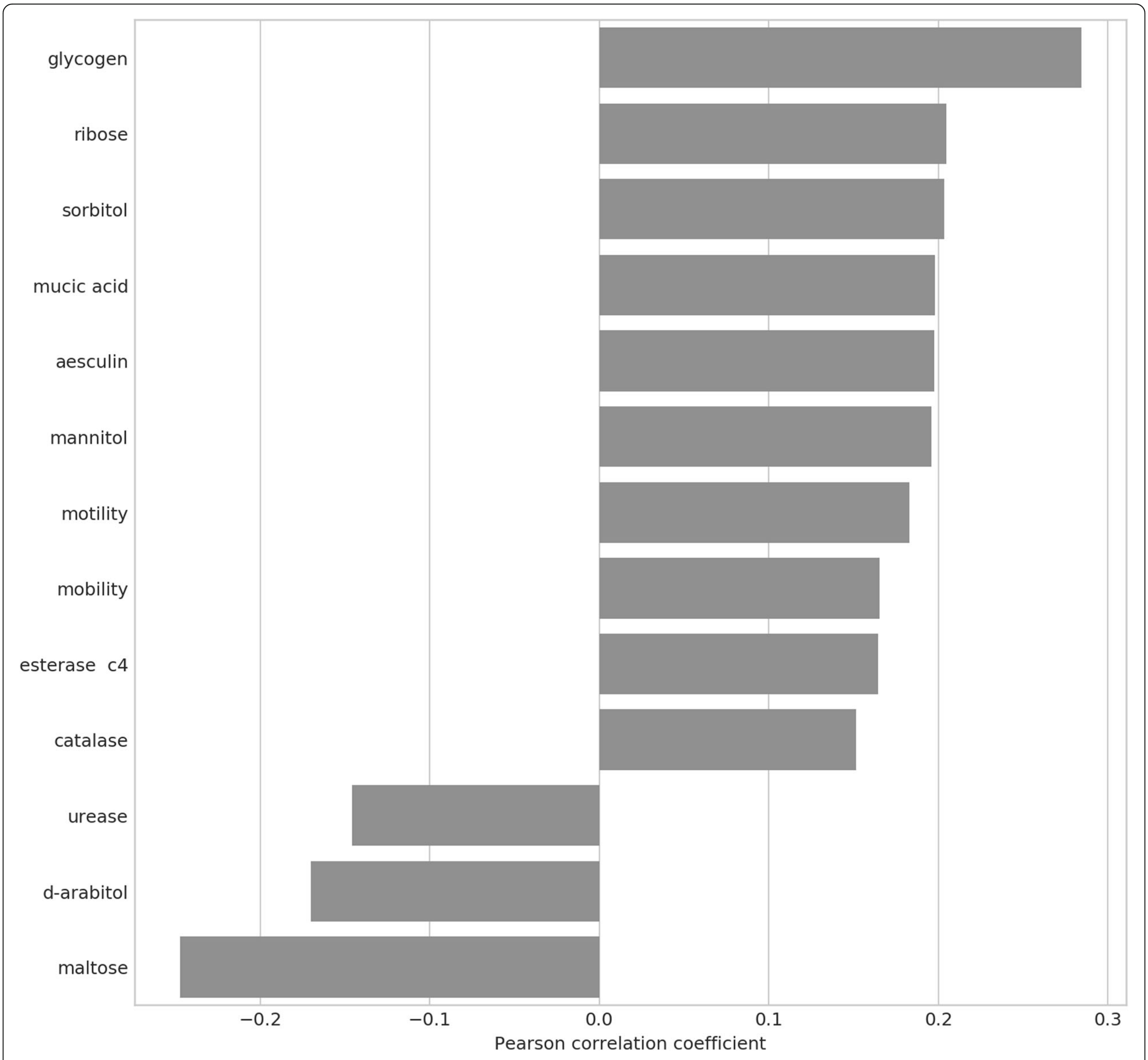

Fig. 9 ProTraits metabolic traits associated with diabetes. The differential abundance vector of log2FCs per species in diabetes was correlated with the vector of probabilities for association of each OTU with that metabolic property as provided by the ProTraits classifier. Only significant ($p < 0.05$) associations are shown. A positive correlation coefficient thus indicates an increase in abundance of OTUs predicted or known to metabolize the given metabolite

neurotoxins that would make the autism symptoms worse [78]. Many autism patients also exhibit GI tract problems during the onset of the disease which often persisting [79, 80]. In AGP samples, we observe an increase in some species associated with poor water quality (*Bacillus flexus, Kocuria palustris*), and food poisoning (*Campylobacter uroelyticus*, *Clostridium perfringens*) (Fig. 11) [81]. *Campylobacter ureolyticus* has been shown to be increased in Crohn's disease and other GI symptoms [82]. Autistic children have been reported to have elevated levels of ammonia in stool [83], a compound which *K. palustris* can degrade well [patent no CN103103141-A]. *Kocuria* species have been reported to be contributing to brain abscess and meningitis, as well as a cause of urinary tract infections [84, 85]. Some of the bacteria that showed up are seemingly unrelated to the condition, like *Xylophilus ampelinus* (a plant pathogen). Most of the bacteria decreased in ASD are non-pathological, environmental species: *Pseudoxanthomonas mexicana*, *Pseudomonas citronellolis, Blastomonas natatoria*, with the exception of *Staphylococcus haemolyticus*, which is a known hospital pathogen [86].

Discussion

The widespread availability of public metagenomic data enables new data to be interpreted within the context of

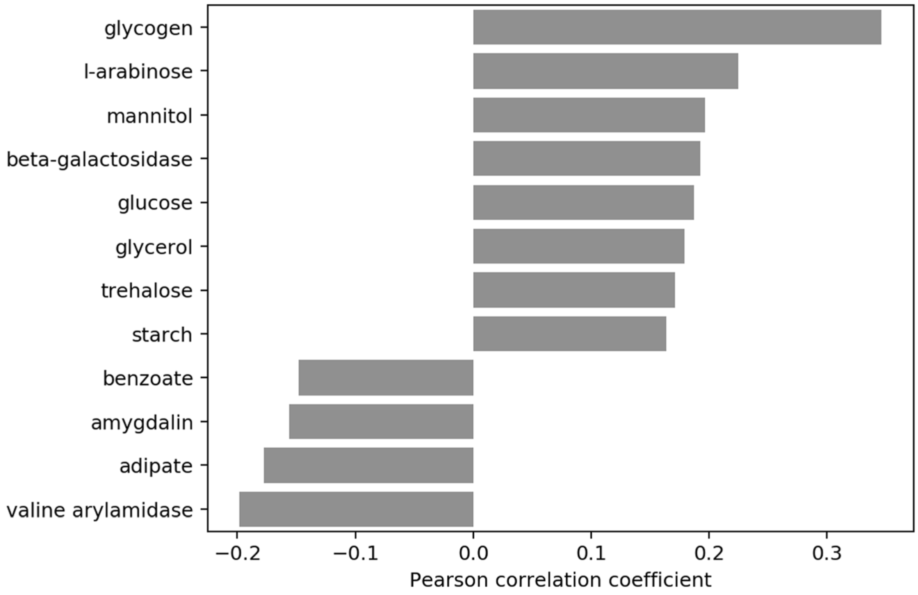

Fig. 10 ProTraits metabolic traits associated with IBD. The differential abundance vector of log2FCs per species in IBD was correlated with the vector of probabilities for association of each OTU with that metabolic property as provided by the ProTraits classifier. Only significant ($p < 0.05$) associations are shown. A positive correlation coefficient thus indicates an increase in abundance of OTUs predicted or known to metabolize the given metabolite

prior experiments. As more data is collected, more and more accurate statements can be formulated about the interactions within the microbiome, as well as between the microbes and the environment. As of today however, despite the fact that there are many samples available publicly spanning different environments, the data may prove somewhat selective. Some conditions are represented by a single project, some lack a 'control', or a suitable comparison; some are just lacking the sample annotation needed to draw any conclusions.

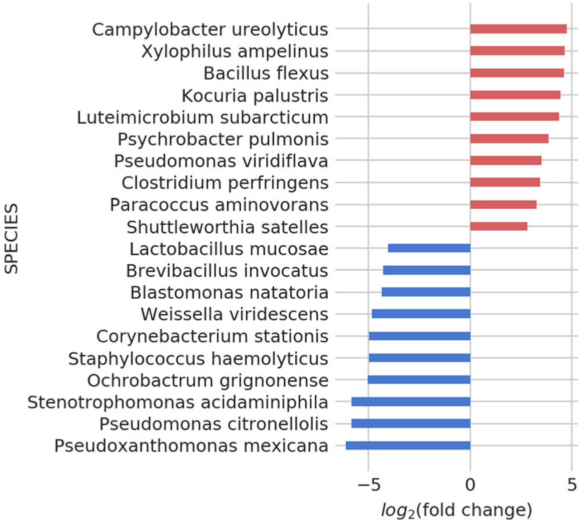

Fig. 11 Changes in microbial species in autism spectrum disorder patients

Moreover, there are a plethora of confounding factors which cannot always be controlled for. For example, human subjects find it difficult to stick to a strict diet, which introduces higher variability in the dietary input, which in turn changes the microbiome in a more convoluted way. We also have highly variable - individually and between each other - behavioral patterns. We differ genetically and with respect to background, which have also been shown to affect the gut microbiome. Finally, most of the variables - notably diet-related - are self-reported, which makes the data prone to a bias introduced by many observers, as well as simple forgetfulness.

All of the factors mentioned above limit the applicability of the method we propose. That said, the American Gut Project is a source of highly standardized information about the samples, and there is a decent number of samples in it. We also expect the trend in data accumulation to stay increasing. In a few years, there will be more data available, and this will widen the bottleneck we're faced with currently.

Being still a relatively young field, metagenomics suffers from the lack of standardized methods for data analysis, as is the case with e.g. gene expression analysis. The EBI analysis pipeline is an example of a proposed standardized solution for metagenomics data analysis up to converting the raw data to counts of different taxa or genes. Using such a pipeline allows us to minimize the bias introduced by technical variability. The choice of methods for downstream analysis is less obvious. There have been efforts undertaken to compare different computational tools to each other and evaluate the performance of any individual method for differential abundance calculation, clustering, etc. The modeling strategy we used in MNEMONIC seems to be performing well, but may not be optimal. One caveat of using gene expression analysis tools for metagenomics data is that, because of the sequencing technology limitations, the latter is compositional and does not represent absolute counts, which in turn implies that data is not independent [87, 88]. This characteristic will cause many standard statistical tools to yield false positive results. The data is also more sparse than the RNA-seq counts, which further complicates modeling [89]. However, despite difficulties associated with the analysis of metagenomics data, studies consistently show that some signal can be achieved even using simple statistical methods, and is strong enough to overcome the technical bias [41, 90].

In addition to these issues, without a gold standard, we cannot quantitatively evaluate the performance of the approach and are relegated mostly to "sanity checks" (i.e., replicating findings well-established by others). Lastly, when studying metagenomics in the context of human gut, it is not clear, in most cases, whether the observed or potential changes in microbiome are the cause of a pathology, the effect, or just accidental. It could also be the case that they contribute both to the cause and the effect, with potential complex interactions and feedback loops, and the interplay of host condition and microbiota composition converges, over time, to a recognizable disease state. For example,a patient with rheumatoid arthritis might go grocery shopping less frequently because of the pain it causes them to walk to the grocery store, resulting in a less diverse or otherwise changed diet and, in turn, changes in microbiome. And perhaps the microbiome changes, in turn, exacerbate their health issues, resulting in a cycle of less frequent shopping and worsened condition.

One should be also wary when interpreting results from a metagenomics experiment in a certain context. As an example, dysbiosis is widely defined as a generic disturbance of the "correct" microbiome [91] and seems to be associated with a variety of diseases. Interestingly, different diseases seem to have a similar pattern of change with respect to the 'healthy' microbiome [90]. However, it's been suggested that the host-microbiome interactions may be much more complex and that the term is not only overly general, but also misleading [92]. Even with so many factors influencing the microbial composition, it may be possible to define a 'healthy' human gut microbiome or a 'core' gut microbiome, but necessarily with tolerance to the relatively high group or individual variation.

Conclusion

We present an approach to help interpret new metagenomics experiments, specifically as an algorithmic means of searching for similar "shifts" that have been reported within public metagenomic repositories in terms of differential microbial abundance between conditions. The question itself is important to establish whether a newly observed change in the microbiome has been seen before and, if so, under what conditions. This is important in interpreting new results, as a merely descriptive report of changes in microbial fractions does not answer the question about what that change might mean in terms of its potential relevance to the host. If our observations of microbial abundance shifts are similar to others, then the nature of their experiments informs us as to how to interpret ours. For example, examining the experiments that led to similar microbiome changes might yield a unifying theme such as changes in dietary composition (salt, protein, fat, sugar, etc), immune activation/repression, or response to stress. In turn, if we are studying diseases, then similar shifts might lend themselves to testable hypotheses regarding causality. Alternatively, if no previous experiments are highly similar, then knowing this enables us to claim our observations are novel.

The main limitation of this report is that we cannot yet automatically (algorithmically) detect from the meta-data alone which samples within an experiment are control and which ones experimental. In some cases, there may be only controls (e.g., a survey of microbial abudance), or there may be multiple comparisons that involve either different experimental perturbations or multiple time-points for one perturbation. The AGP enabled us to bypass this limitation for now by focusing on one that came with well-annotated structure. In the future, for the potential for this approach to be fully realized, the reporting of experimental vs control conditions needs to be easy to recognize algorithmically. One solution would be to require increased structure to the meta-data reporting in new microbial experiments, but another would also be to increase our ability to algorithmically extract the necessary values from a free-form description either in the meta-data or publication itself.

Additional file

Additional file 1: Association of fruit and meat consumption with metabolic variables. Dietary fruit consumption is associated with a significant increase in OTUs annotated with fructose metabolism, whereas red meat consumption was significantly associated with increases in microbes annotated with metabolism of amino acids, alkaline phosphatase, and lipase C14 metabolic activity. (PNG 70 kb)

Abbreviations

AGP: American Gut Project; DA: Differential abundance; EBI: European Bioinformatics Institute

Acknowledgments

Not applicable.

Funding
We would like to acknowledge NIH grants NIH grants 5P20GM103636 and 1P30AG050911 and a pilot project grant from The Presbyterian Health Foundation (to JDW) for supporting this work.

Availability of data and materials
The software and the datasets and the information on how to use it is available in the GitLab repository, https://gitlab.com/wrenlab/mnemonic.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 20 Supplement 2, 2019: Proceedings of the 15th Annual MCBIOS Conference. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-2

Authors' contributions
AP designed and prototyped the package. CG made the final version of the package. CB has helped with the analysis of the AGP metadata. XR contributed to the analysis of the results. HP helped to clarify the concepts. XR contributed to the analysis of the results. All authors participated in the process of writing and editing the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

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

Author details
[1]Arthritis and Clinical Immunology Program, Division of Genomics and Data Sciences, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104-5005, USA. [2]Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [3]Oklahoma Center for Neuroscience, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [4]Department of Geriatric Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA.

Published: 14 March 2019

References

1. Beja O, et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science. 2000;289(5486):1902–6.
2. Yooseph S, et al. The sorcerer II Global Ocean sampling expedition: expanding the universe of protein families. PLoS Biol. 2007;5(3):e16.
3. Garcia Martin H, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol. 2006; 24(10):1263–9.
4. Hallam SJ, et al. Reverse methanogenesis: testing the hypothesis with environmental genomics. Science. 2004;305(5689):1457–62.
5. Mou X, et al. Bacterial carbon processing by generalist species in the coastal ocean. Nature. 2008;451(7179):708–11.
6. Dinsdale EA, et al. Microbial ecology of four coral atolls in the northern Line Islands. PLoS One. 2008;3(2):e1584.
7. Ley RE, et al. Microbial ecology: human gut microbes associated with obesity. Nature. 2006;444(7122):1022–3.
8. Brulc JM, et al. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc Natl Acad Sci U S A. 2009;106(6):1948–53.
9. Gerdes S, et al. Essential genes on metabolic maps. Curr Opin Biotechnol. 2006;17(5):448–56.
10. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6(11):805–14.
11. Turnbaugh PJ, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4.
12. Henschel A, Anwar MZ, Manohar V. Comprehensive meta-analysis of ontology annotated 16S rRNA profiles identifies Beta diversity clusters of environmental bacterial communities. PLoS Comput Biol. 2015;11(10): e1004468.
13. Buford TW. (dis)trust your gut: the gut microbiome in age-related inflammation, health, and disease. Microbiome. 2017;5(1):80.
14. Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. Bioinformatics. 2010;26(6):715–21.
15. Gut A. American Gut Project. Available from: http://americangut.org/. Accessed 24 Oct 2018.
16. Jakobsson HE, et al. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. PLoS One. 2010;5(3):e9836.
17. Dethlefsen, L. and D.A. Relman, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc Natl Acad Sci U S A, 2011. 108 Suppl 1: p. 4554–4561.
18. Willing, BP., et al., A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology, 2010. 139(6): p. 1844–1854 e1.
19. Schubert AM, et al. Microbiome data distinguish patients with Clostridium difficile infection and non-C. Difficile-associated diarrhea from healthy controls. mBio. 2014;5(3):e01021–14.
20. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65.
21. Graessler J, et al. Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters. Pharmacogenomics J. 2013;13(6):514–22.
22. Vrieze A, et al. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. Gastroenterology. 2012;143(4):913–6 e7.
23. Chang YY, Ouyang Q. Expression and significance of mucosal beta-defensin-2, TNFalpha and IL-1beta in ulcerative colitis. Zhonghua Nei Ke Za Zhi. 2008; 47(1):11–4.
24. Khoruts A, et al. Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. J Clin Gastroenterol. 2010;44(5):354–60.
25. Wang T, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J. 2012;6(2):320–9.
26. Kang D-W, et al. Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. PLoS One. 2013;8(7):e68322.
27. Rosenfeld CS. Microbiome disturbances and autism Spectrum disorders. Drug Metab Dispos. 2015;43(10):1557–71.
28. Wang Y, et al. Dynamic gut microbiome across life history of the malaria mosquito Anopheles gambiae in Kenya. PLoS One. 2011;6(9):e24767.
29. De Filippo C, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A. 2010;107(33):14691–6.
30. Hehemann JH, et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature. 2010;464(7290):908–12.
31. Yatsunenko T, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222–7.
32. Zupancic ML, et al. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. PLoS One. 2012;7(8):e43052.
33. Muegge BD, et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science. 2011;332(6032): 970–4.
34. Wu GD, et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science. 2011;334(6052):105–8.
35. Hansen, E.E., et al., Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. Proc Natl Acad Sci U S A, 2011. 108 Suppl 1: p. 4599–4606.
36. Dicksved J, et al. Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J. 2008;2(7):716–27.
37. Wilck N, et al. Salt-responsive gut commensal modulates TH17 axis and disease. Nature. 2017;551(7682):585–9.
38. Koren O, et al. A guide to Enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol. 2013;9(1):e1002863.
39. Huurre A, et al. Mode of delivery - effects on gut microbiota and humoral immunity. Neonatology. 2008;93(4):236–40.
40. Dominguez-Bello MG, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A. 2010;107(26):11971–5.
41. Lozupone CA, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23(10):1704–14.

42. Biagi E, et al. Gut microbiota and extreme longevity. Curr Biol. 2016;26(11): 1480–1485s.
43. Palmer C, et al. Development of the human infant intestinal microbiota. PLoS Biol. 2007;5(7):e177.
44. Koenig JE, et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A. 2011;108(Suppl 1):4578–85.
45. O'Sullivan O, et al. Correlation of rRNA gene amplicon pyrosequencing and bacterial culture for microbial compositional analysis of faecal samples from elderly Irish subjects. J Appl Microbiol. 2011;111(2):467–73.
46. Sun S, et al. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res. 2011; 39(Database):D546–51.
47. Markowitz VM, et al. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 2012;40(D1):D123–9.
48. Wilke A, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016;44(D1):D590–4.
49. Mitchell A, et al. EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. 2016;44(D1):D595–603.
50. Segata N, Huttenhower C. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. PLoS One. 2011;6(9):e24704.
51. Mandal S, et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663.
52. Morgan XC, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):R79.
53. Zhang X, et al. Negative binomial mixed models for analyzing microbiome count data. BMC Bioinformatics. 2017;18(1):4.
54. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
55. Law CW, et al. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
56. Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
57. Weber N, et al. Nephele: a cloud platform for simplified, standardized, and reproducible microbiome data analysis. Bioinformatics.
58. Ranjan R, et al. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016;469(4):967–77.
59. Kuczynski J, et al. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet. 2011;13(1):47–58.
60. Brbic M, et al. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 2016;44(21):10074–90.
61. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97.
62. Noble EE, et al. Early-life sugar consumption affects the rat microbiome independently of obesity. J Nutr. 2017;147(1):20–8.
63. Singh RK, et al. Influence of diet on the gut microbiome and implications for human health. J Transl Med. 2017;15(1):73.
64. Chassaing B, et al. Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome. Nature. 2015;519(7541):92–6.
65. He T, et al. Effects of yogurt and bifidobacteria supplementation on the colonic microbiota in lactose-intolerant subjects. J Appl Microbiol. 2008; 104(2):595–604.
66. Brown JM. Eating to boost gut microbial diversity. Sci Transl Med. 2016; 8(369):369ec198.
67. Hibberd AA, et al. Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. BMJ Open Gastroenterol. 2017;4(1):e000145.
68. Luo XM, et al. Gut Microbiota in Human Systemic Lupus Erythematosus and a Mouse Model of Lupus. Appl Environ Microbiol. 2018;84(4):e02288–17.
69. Gong, D., et al., Involvement of reduced microbial diversity in inflammatory bowel disease. Gastroenterol Res Pract, 2016. 2016: p. 6951091.
70. Picchianti-Diamanti A, Rosado MM, D'Amelio R. Infectious agents and inflammation: the role of microbiota in autoimmune arthritis. Front Microbiol. 2017;8:2696.
71. Chao A, et al. Estimating the number of shared species in two communities. Stat Sin. 2000:227–46.
72. Kim E, Kim DB, Park JY. Changes of mouse Gut microbiota diversity and composition by modulating dietary protein and carbohydrate contents: a pilot study. Prev Nutr Food Sci. 2016;21(1):57–61.
73. Simpson HL, Campbell BJ. Review article: dietary fibre-microbiota interactions. Aliment Pharmacol Ther. 2015;42(2):158–79.
74. Hamden K, et al. Experimental diabetes treated with trigonelline: effect on key enzymes related to diabetes and hypertension, beta-cell and liver function. Mol Cell Biochem. 2013;381(1–2):85–94.
75. Zhou J, Chan L, Zhou S. Trigonelline: a plant alkaloid with therapeutic potential for diabetes and central nervous system disease. Curr Med Chem. 2012;19(21):3523–31.
76. Finegold, S.M., et al., Gastrointestinal microflora studies in late-onset autism. Clin Infect Dis, 2002. 35(Supplement_1): p. S6-S16.
77. Sandler RH, et al. Short-term benefit from oral vancomycin treatment of regressive-onset autism. J Child Neurol. 2000;15(7):429–35.
78. Bolte ER. Autism and Clostridium tetani. Med Hypotheses. 1998;51(2):133–44.
79. Buie, T., et al., Evaluation, diagnosis, and treatment of gastrointestinal disorders in individuals with ASDs: a consensus report. Pediatrics, 2010. 125(Supplement 1): p. S1-S18.
80. Mayer EA, Padua D, Tillisch K. Altered brain-gut axis in autism: comorbidity or causative mechanisms? BioEssays. 2014;36(10):933–9.
81. O'Donovan D, et al. Campylobacter ureolyticus. Virulence. 2014;5(4):498–506.
82. Burgos-Portugal JA, et al. Pathogenic potential of campylobacter ureolyticus. Infect Immun. 2012;80(2):883–90.
83. Wang L, et al. Elevated fecal short chain fatty acid and ammonia concentrations in children with autism spectrum disorder. Dig Dis Sci. 2012; 57(8):2096–102.
84. Tsai CY, et al. Kocuria varians infection associated with brain abscess: a case report. BMC Infect Dis. 2010;10:102.
85. Kandi V, et al. Emerging bacterial infection: identification and clinical significance of Kocuria species. Cureus. 2016;8(8):e731.
86. Barros EM, et al. Staphylococcus haemolyticus as an important hospital pathogen and carrier of methicillin resistance genes. J Clin Microbiol. 2012; 50(1):166–8.
87. Lovell D, et al. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol. 2015;11(3):e1004075.
88. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016;62(8): 692–703.
89. Mohri M, Roark B. Structural zeros versus sampling zeros. Portland, OR, USA: Oregon Health & Science University; 2005.
90. Duvallet C. Meta-analysis generates and prioritizes hypotheses for translational microbiome research. Microb Biotechnol. 2018;11(2):273–6.
91. Beer, K., The Gut microbiome in type 2 diabetes. Clinical Reviews, 2018.
92. Duvallet C, et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8(1):1784.