Shaffer et al. BMC Bioinformatics (2019) 20:614
https://doi.org/10.1186/s12859-019-3176-8

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data

M. Shaffer[1], K. Thurimella[1], K. Quinn[2], K. Doenges[2], X. Zhang[2,3], S. Bokatzian[2], N. Reisdorph[2] and C. A. Lozupone[1*]

Abstract

Background: Untargeted metabolomics of host-associated samples has yielded insights into mechanisms by which microbes modulate health. However, data interpretation is challenged by the complexity of origins of the small molecules measured, which can come from the host, microbes that live within the host, or from other exposures such as diet or the environment.

Results: We address this challenge through development of AMON: Annotation of Metabolite Origins via Networks. AMON is an open-source bioinformatics application that can be used to annotate which compounds in the metabolome could have been produced by bacteria present or the host, to evaluate pathway enrichment of host verses microbial metabolites, and to visualize which compounds may have been produced by host versus microbial enzymes in KEGG pathway maps.

Conclusions: AMON empowers researchers to predict origins of metabolites via genomic information and to visualize potential host:microbe interplay. Additionally, the evaluation of enrichment of pathway metabolites of host versus microbial origin gives insight into the metabolic functionality that a microbial community adds to a host: microbe system. Through integrated analysis of microbiome and metabolome data, mechanistic relationships between microbial communities and host phenotypes can be better understood.

Keywords: Microbiome, Metabolome, Data-integration

Background

The host-associated microbiome can influence many aspects of human health and disease through its metabolic activity. Examples include host:microbe co-metabolism of dietary choline/carnitine to Trimethylamine N-oxide (TMAO) as a driver of heart disease [1], microbial production of branched chain amino acids as a contributor to insulin resistance [2], and microbial production of 12,13-DiHOME as a driver of $CD4^+$ T cell dysfunction associated with childhood atopy [3]. A key way of exploring which compounds might mediate relationships between microbial activity and host disease is untargeted metabolomics (e.g. mass spectrometry) of host materials such as stool, plasma, urine, or tissues. These analyses result in the detection and relative quantitation of hundreds to thousands of compounds, the sum of which is referred to as a "metabolome". Host-associated metabolomes represent a complex milieu of compounds that can have different origins, including the diet of the host organism and a variety of environmental exposures such as pollutants. In addition, the metabolome contains metabolic products of these compounds, i.e. metabolites, that can result from host and/or microbiome metabolism or co-metabolism [4].

One way to estimate which metabolites in host samples originate from host versus microbial metabolism is to use metabolic networks described in databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [5]. These networks capture the relationship between metabolites, the enzymes that produce them, and the genomes of organisms (both host and microbial) that contain genes encoding those enzymes. These networks thus provide a framework for relating the genes present

* Correspondence: catherine.lozupone@cuanschutz.edu
[1]Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
Full list of author information is available at the end of the article

in the host and colonizing bacteria, and the metabolites present in a sample. Several papers have explored use of metabolic networks to understand likely products of microbial metabolism [6–14]. Algorithms that consider the combined influence of microbial and host metabolism have also been explored [2, 8, 10–12, 15]. Although these studies together show great promise in this field, these methods often rely on comprehensive, validated metabolic models [6, 8, 13, 14], focus only on subsets of carefully measured metabolites [15], or focus on other aspects of community ecology such as predicting metabolic interactions [11], limiting their application to relating complex untargeted metagenomics and metabolomic datasets [16]. Furthermore, algorithms developed in this field often do not have a user interface allowing researchers to apply them to their own data [2, 15, 17]. One exception is the predicted relative metabolic turnover (PRMT) scoring metric [16, 18], and MIMOSA [6], an application that uses PRMT to relate metabolite levels and predicted microbial metabolic capabilities in untargeted metabolomes and metagenomes. However, MIMOSA does not currently evaluate contributions of host metabolism to metabolite levels.

Here we present a tool for annotation of metabolite origins via networks (AMON), which uses information in KEGG to predict whether measured metabolites are likely to originate from singular organisms or collections of organisms based on a list of the genes that they encode. As an example, AMON can be used to predict whether metabolites may originate from the host versus from host-associated microbiomes as assessed with 16S ribosomal RNA (rRNA) gene sequences or shotgun metagenomics. We demonstrate our tool by applying it to a dataset from a cohort of HIV positive individuals and controls in which the stool microbiome was assessed with 16S rRNA gene sequencing and the plasma metabolome was assessed with untargeted liquid chromatography mass spectrometry (LC/MS). We also illustrate how much information is lost when we only focus on compounds and genes of known identity/function, emphasizing the need for complimentary approaches to general metabolomic database searches for the identification of microbially produced compounds.

Methods

AMON implementation

AMON is an open source program implemented in python 3. It is available at https://github.com/lozuponelab/AMON as well as in the python package index. AMON takes as input lists of KO (KEGG Orthology) identifiers that are predicted to be present in different potential sources (e.g. the metagenome of a host-associated microbiome or the genome of host organism) and a list of KEGG compound IDs, such as from an annotated metabolome (Fig. 1). Microbiome KO lists can be generated from 16S rRNA data using PICRUSt [19] or Tax4Fun [20], or from a shotgun metagenome using annotation tools such as HUMAnN [21]. The KOs from any KEGG organism can be acquired using the extract_ko_genome_from_organism.py script supplied with AMON, which determines the KOs for a given organism from files retrieved using the freely available KEGG API (https://www.kegg.jp/kegg/rest/) or from a user-supplied KEGG FTP file for those with a KEGG subscription.

The goal of AMON is to determine the compounds that a set of KEGG KOs can potentially generate. First, the reactions associated with each KO and formulas describing substrates and products of each reaction are retrieved from the KEGG "reactions" file or the KEGG API. The products of all reactions are the putative set of compounds that the given KOs could produce. The KEGG reaction file does not directly define reversibility of reactions so AMON assumes that the primary direction of reactions is from the left to the right in the equations and therefore the compounds on the right side of the equation are the products. As an example, if the supplied set of KOs included K00929 (butyrate kinase), the following formula from the reaction performed by this enzyme (R01688) would be retrieved: C02527 (Butanoyl phosphate) = > C00246 (butyrate). Butyrate would then be added to the list of compounds that could be generated by this set of KOs.

AMON produces a table indicating which compounds could be produced by each of the provided KO sets or both. For instance if one KO set is from the host and one from the microbiome, AMON will indicate whether compounds that were the products of the reactions that these compounds encoded originated from the microbiome KO set only, host KO set only, or both microbial and host KO sets. A file for input to KEGG mapper (https://www.genome.jp/kegg/mapper.html) is also produced, which can be used to overlay this information on KEGG pathway diagrams. AMON also generates information on pathway enrichment in the compounds produced by the user-supplied gene lists. Specifically, the pathway assignment of the set of metabolites predicted to be produced by each input KO list is tested for enrichment relative to the full set of all compounds in that pathway using the hypergeometric test. This calculation is performed for all KEGG pathways that had at least one metabolite predicted to be produced by the provided gene sets. Both raw and Benjamini-Hochberg FDR adjusted p-values are reported. AMON also produces a summary figure (Venn diagram) illustrating predicted metabolite origins. A set of example outputs are provided with the case study (Figs. 2b, 3 and Additional file 2: Table S2, Additional file 3: Table S3). We have found run times to typically be less than 1 min if KEGG files

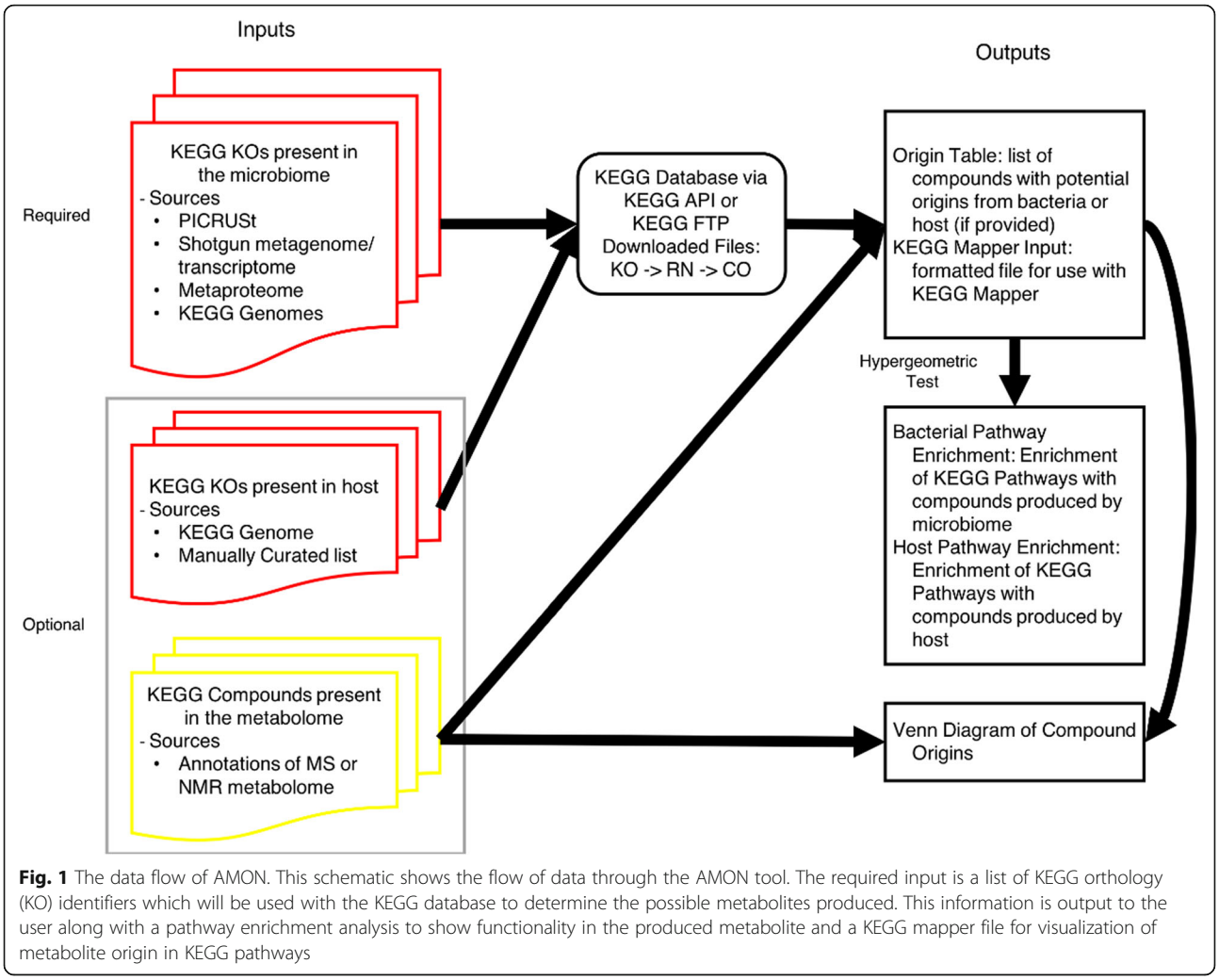

Fig. 1 The data flow of AMON. This schematic shows the flow of data through the AMON tool. The required input is a list of KEGG orthology (KO) identifiers which will be used with the KEGG database to determine the possible metabolites produced. This information is output to the user along with a pathway enrichment analysis to show functionality in the produced metabolite and a KEGG mapper file for visualization of metabolite origin in KEGG pathways

are provided. If KEGG files are not provided then run time is dependent on the length of the provided KO lists since the KEGG API limits the volume of data downloaded in a set period of time.

Case study

We illustrate the utility of AMON using a data set from the gut microbiome (16S rRNA) and blood metabolome (LC/MS) of HIV positive individuals and controls. The cohort and the fecal 16S rRNA data were previously described as part of a larger study of differences in the fecal microbiome in HIV positive and high risk populations [22]. These 16S rRNA data are paired with metabolome data as a part of a study described at ClinicalTrials.gov (Identifier: NCT02258685). Stool samples from 59 individuals, of which 37 were HIV positive and 22 were HIV negative, were collected at home in a commode specimen collector within 24 h of the clinic visit during which blood was drawn.

Generation of fecal 16S rRNA data

Stool samples were stored at − 20 °C during transit and at − 80 °C prior to DNA extraction with the MoBIO kit and preparation for barcoding sequencing using the Earth Microbiome Project protocol [23]. The 16S rRNA gene V4 region of stool microbes was sequenced using MiSeq (Illumina), denoised using DADA2 [24] and binned into 99% Operational Taxonomic Units (OTUs) using UCLUST [25] and the greengenes database (version 13_8) via QIIME 1.9.1 [26]. We used PICRUSt [19] to predict a metagenome and AMON to predict metabolites.

Plasma sample preparation

A modified liquid-liquid extraction protocol was used to extract hydrophobic and hydrophilic compounds from the plasma samples [27]. Briefly, 100 μL of plasma spiked with internal standards underwent a protein crash with 400 μL ice cold methanol. The supernatant was dried under nitrogen and methyl *tert*-butyl ether (MTBE) and

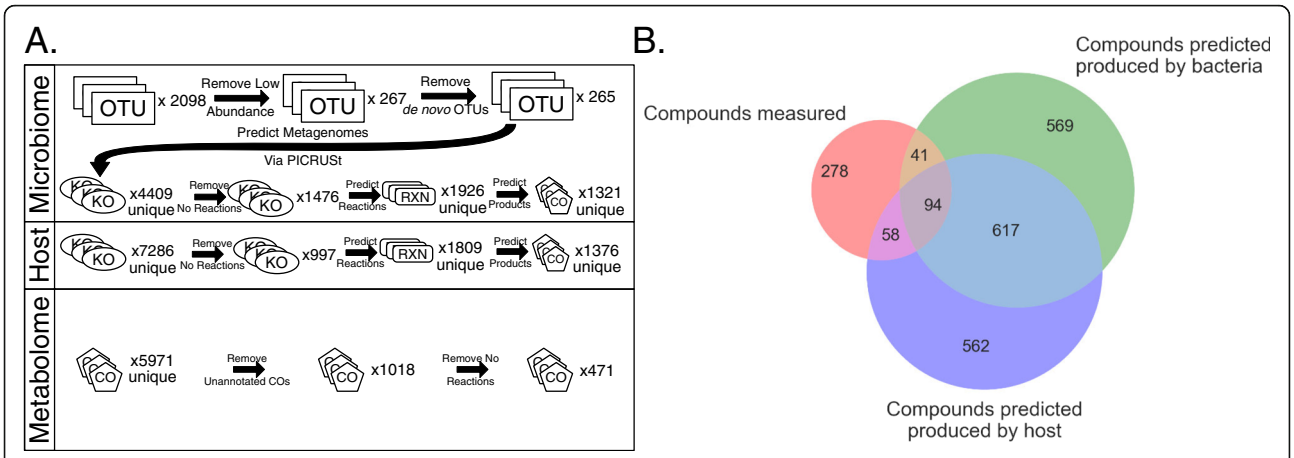

Fig. 2 The results of a case study running AMON with 16S rRNA sequencing data from stool and PICRUSt to predict the metagenome along with the KEGG human genome and an LC/MS untargeted metabolome. **a** A flow diagram showing how much data is lost between parts of analyses at all data levels. **b** A Venn diagram showing overlaps in compound sets. The red circle shows compounds detected with untargeted LC/MS with an annotated KEGG compound ID. The green and purple circles show compounds that the metabolic network tells us could have been produced by the bacteria present in the microbiome and the host respectively

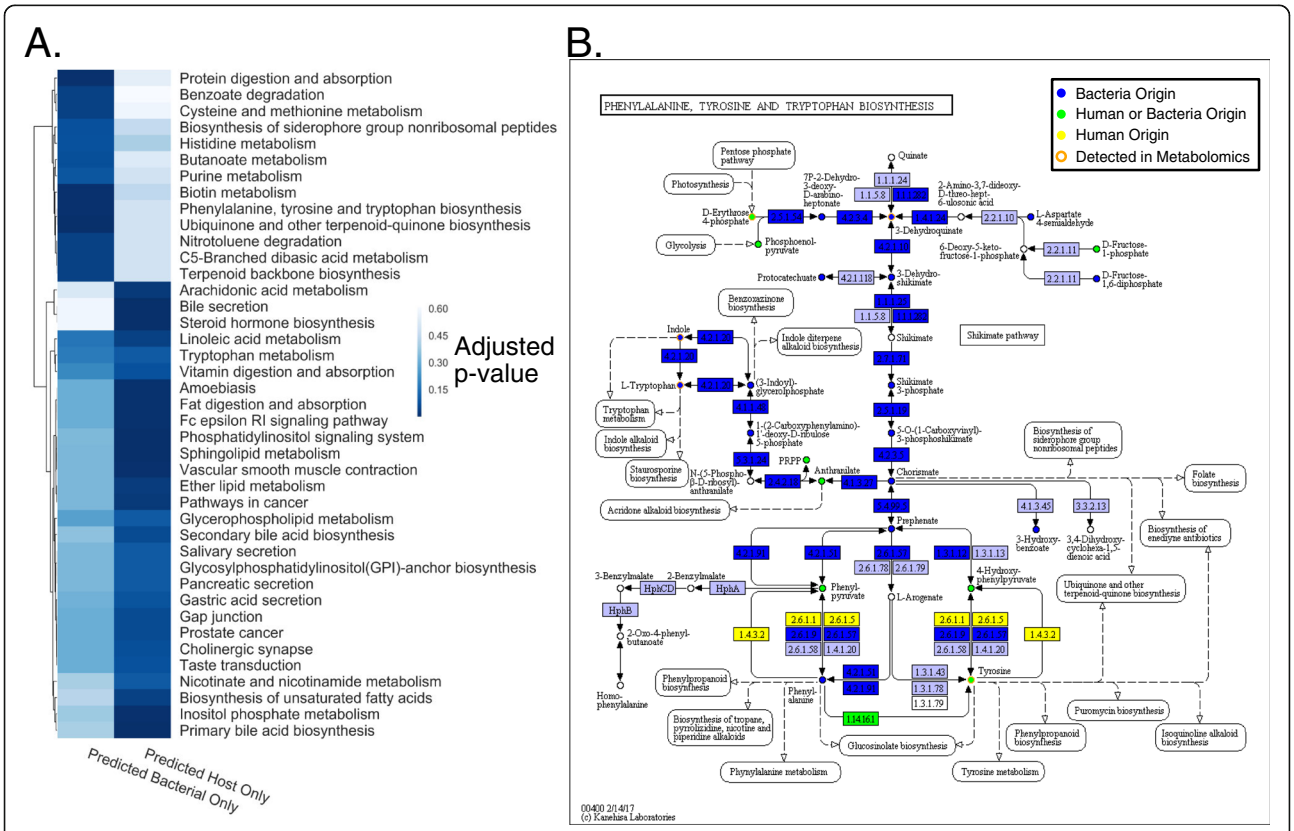

Fig. 3 Enrichment of pathways and a single enriched pathway colored with metabolite origin. **a** A heatmap showing the *p*-values associated with a pathway enrichment analysis with KEGG pathways. The first column is p-values for enrichment of KEGG pathways in compounds that were detected via untargeted LC/MS of plasma and we predict could be generated by members of the fecal microbiome. The second column is the same but for compounds that we predicted could have been generated by the human host. **b** This pathway map is colored by putative origin of the compound, which are circles, and presence of the reaction, which are rectangles. Dark blue is a compound or gene with a bacterial origin, yellow is a compound or gene with a human origin, orange outlined compounds are detected in the metabolomics. Circles or rectangles could be of human or bacterial origin

water were added to extract the hydrophobic and hydrophilic compounds, respectively. The upper hydrophobic layer was transferred to a new tube and the lower hydrophilic layer was re-extracted with MTBE. The upper hydrophobic layer was combined, dried under nitrogen and reconstituted in 200 μL of methanol. The hydrophilic layer was dried under nitrogen, underwent a second protein crash with water and ice-cold methanol (1:4 water-methanol). The supernatant was removed, dried by SpeedVac at 45 °C and reconstituted in 100 μL of 5% acetonitrile in water. Both fractions were stored at −80 °C until LCMS analysis.

Liquid chromatography mass spectrometry

The hydrophobic fractions were analyzed using reverse phase chromatography on an Agilent Technologies (Santa Clara, CA) 1290 ultra-high precision liquid chromatography (UHPLC) system on an Agilent Zorbax Rapid Resolution HD SB-C18, 1.8um (2.1 × 100 mm) analytical column with an Agilent Zorbax SB-C18, 1.8 μm (2.1 × 5 mm) guard column. The hydrophilic fractions were analyzed using hydrophilic interaction liquid chromatography (HILIC) on a 1290 UHPLC system using a Phenomenex Kinetex HILIC, 2.6um (2.1 × 50 mm) analytical column with an Agilent Zorbax Eclipse Plus C8 5μm (2.1 × 12.5 mm) guard column. The hydrophobic and hydrophilic fractions were run on Agilent Technologies (Santa Clara, CA) 6520 and 6550 Quadrupole Time of Flight (QTOF) mass spectrometers, respectively. Both fractions were run in positive and negative electrospray ionization (ESI) modes, as previously described [28].

Mass spectrometry data processing

Compound data was extracted using Agilent Technologies (Santa Clara, CA) Mass Hunter Profinder Version B.08 (Profinder) software in combination with Agilent Technologies Mass Profiler Professional Version 14 (MPP) as described previously [28]. Specifically, a Profinder recursive workflow was used to extract compound data from all samples based on abundance profiles in m/z and retention time (RT) dimensions. The *aqueous positive* mode samples were extracted as follows: RT extraction range 0–14.7 min with noise peak height filter ≥2000 counts, ion species: +H, +Na, +K, +NH4 and charge state maximum of 2. Alignment tolerance for RT was 0% + 0.3 min with mass 20 ppm + 3 mDa. The 'Find by Molecule Feature' (MFE) parameters used were height ≥ 4500 counts and a score of 90. The 'Find by Ion' (FbI) parameters were height ≥ 3500 for EIC peak integration with post-processing filters using Abs height ≥ 3500 counts and score 50. The *aqueous negative* mode samples were extracted as follows: RT extraction range 0–14.7 min with noise peak height filter ≥1000 counts, ion species: -H, +Cl, +HCOO, +CH3COO and charge state maximum of 2. Alignment tolerance for RT was 0% + 0.3 min with mass 20 ppm + 3 mDa. The MFE parameters used were height ≥ 3000 counts and a score of 90. The FbI parameters were height ≥ 2500 for EIC peak integration with post-processing filters using Abs height ≥ 2500 counts and score 50. The *lipid positive* mode samples were extracted as follows: RT extraction range 0–10.4 min with noise peak height filter ≥500 counts, ion species: +H, +Na, +K, +NH4 and charge state maximum of 2. Alignment tolerance for RT was 0% + 0.25 min with mass 20 ppm + 2 mDa. The MFE parameters used were height ≥ 2000 counts and a score of 90. The FbI parameters were height ≥ 1500 for EIC peak integration with post-processing filters using Abs height ≥ 1500 counts and score 50. The *lipid negative* mode samples were extracted as follows: RT extraction range 0–10.4 min with noise peak height filter ≥300 counts, ion species: -H, +Cl, +HCOO, +CH3COO and charge state maximum of 2. Alignment tolerance for RT was 0% + 0.3 min with mass 20 ppm + 3 mDa. The MFE parameters used were height ≥ 4500 counts, and score 90. The FbI parameters were height ≥ 3500 for EIC peak integration with post-processing filters using Abs height ≥ 3500 counts and score 50. In all cases we required compounds had to be present in at least 2 sample files. Extracted data was imported into MPP and the KEGG database was used to putatively annotate plasma compounds based on exact mass, isotope ratios and isotopic distribution with a mass error cutoff of 10 ppm, whereby the predicted isotope distribution is compared to actual ion height and a score is generated. This corresponds to a Metabolomics Standards Initiative metabolite identification level 3 [29] and a Schymanski identification level 5 [31]. Although our approach in some cases output multiple KEGG compounds as possible "hits," we selected the compound with the highest score [29] such that each compound was assigned a single KEGG compound ID.

Results

We used AMON to relate the stool microbiome (as assessed with 16S rRNA gene sequencing) to the plasma metabolome (as assessed with untargeted LC/MS), in a cohort of HIV positive individuals and HIV-negative controls. The overall goal of our case study was to use AMON to determine the degree to which annotated compounds in the plasma metabolome of our study cohort may have been produced by bacteria present in fecal samples, the host, either (i.e. both are capable of production), or neither (i.e. neither the human or the fecal microbiome are predicted to be capable of producing the observed metabolite).

We used the 16S rRNA data and PICRUSt to predict the genome content of the OTUs detected in the fecal samples. PICRUSt drops OTUs from the analysis that do not have related reference sequences in the database and produces an estimate of the nearest sequenced taxon index (NSTI) which measures how close those sequences are to sequenced genomes (those more closely related to genomes have more power to make predictions regarding gene content). Since human gut bacteria are well represented in genome databases, only 0.7% of total reads of the detected sequences were dropped on account of not having a related reference sequence in the database. Furthermore, the average NSTI across samples was 0.08, indicating that most OTUs were highly related to an organism with a sequenced genome. We applied PICRUSt to the 16S rRNA dataset with only OTUs present in more than 11 of 59 samples (20%) included. The 267 remaining OTUs were predicted to contain 4409 unique KOs using PICRUSt. We used the KEGG list of KOs in the human genome to represent human gene content.

We provided these lists of gut microbiome and human KOs to AMON to produce a list of compounds generated from the gut microbiome and the human genome. We also provided AMON with a reaction file downloaded from KEGG January of 2015. Of the 4409 unique KOs that PICRUSt predicted to be present in the gut microbiome, only 1476 (33.5%) had an associated reaction in KEGG. Those without associated reactions may represent orthologous gene groups that do not perform metabolic reactions (such as transporters), or that are known to exist but for which the exact reaction is unknown, showing gaps in our knowledge (Fig. 2a). Using information in KEGG, AMON predicted these KOs to produce 1321 unique compounds via 1926 unique reactions. The human genome was predicted to produce 1376 metabolites via 1809 reactions.

Our metabolomics assays detected 5971 compounds, of which only 1018 (17%) could be putatively annotated with KEGG compound identifiers via a database search and based on match of measured m/z to KEGG compound mass within 10 ppm. Further, only 471 (6%) of the 5971 detected compounds were associated with a reaction in KEGG (Additional file 1: Table S1). Of these 471 annotated compounds in the plasma metabolome with associated KEGG reactions, 189 were predicted to be produced by enzymes in either human or stool bacterial genomes as follows: 40 compounds were exclusively produced by bacteria, 58 exclusively by the host, and 91 by either human or bacterial enzymes (Fig. 2b; Additional file 2: Table S2). There were a remaining 282 compounds that had KEGG compound IDs associated with at least one reaction but were not predicted to be from the human or the gut microbiome. These may be 1) from the environment, 2) produced by microbes in other body sites, 3) host or gut microbial products from unannotated genes, 4) artifacts derived from metabolite decompositions in the samples and/or are misannotations via the matching based on m/z alone.

We used AMON to assess enrichment of pathways in the detected human and bacterial metabolites using the hypergeometric test (Fig. 3a; Additional file 3: Table S3). The 40 compounds predicted to be produced by stool bacteria and not the host were enriched in xenobiotic degradation pathways, including nitrotoluene and atrazine degradation, and pathways for amino acids metabolism, including the phenylalanine, tyrosine and tryptophan biosynthesis pathway and the cysteine and methionine metabolism pathway. The metabolite origin data was visualized using KEGG mapper for the phenylalanine, tyrosine and tryptophan biosynthesis pathway (Fig. 3b). This tool helps to visualize the host-microbe co-metabolism and which genes are important for compounds that may have come from multiple sources. For instance, Fig. 3b allows us to see that indole is a compound found in our metabolome that could only have been produced by bacterial metabolism via the highlighted enzyme (K01695, tryptophan synthase). Also, tyrosine is a compound found in our metabolome that could have been synthesized by a variety of enzymes found only in bacteria, only in humans, or in both and so further exploration would be needed to understand origins of this compound. The 58 compounds which were detected and predicted to be produced by the human genome were enriched in pathways that include bile secretion, steroid hormone biosynthesis and gastric acid secretion.

Comparison of AMON with MIMOSA

The functionality of AMON is related to that of another tool called MIMOSA [6], in that MIMOSA also uses PICRUSt and KEGG to integrate microbiome (16S rRNA) and metabolome data. Unlike AMON, MIMOSA does not relate contributions of microbial versus host metabolism. However, MIMOSA determines quantitative relationships between the relative abundance of genes in a metagenome and the abundance of the particular compounds in a metabolome that their gene products produce/degrade. To compare the results of AMON and MIMOSA when applied to the same dataset, we analyzed our HIV case study with MIMOSA (Additional file 4: Table S4). We supplied MIMOSA with 1) a table of compound abundances measured in our HIV samples with untargeted LC/MS, 2) a gene abundance and gene contributions file generated using 16S rRNA data and PICRUSt and 3) a reaction_mapformula.lst file downloaded from KEGG in January 2015. Of 1018 compounds with KEGG

annotations, MIMOSA was able to successfully analyze the potential microbe contributions for 57 different compounds, and of these 10 (17.5%) had significant correlations to metabolic potential scores and were thus considered "well-predicted". In contrast, AMON predicted 135 compounds in the plasma metabolome to have derived either exclusively from the microbiome ($n = 40$) or from the microbiome or host ($n = 91$). Metabolites that AMON predicted to be of exclusive microbial (but not host) origin that MIMOSA was unable to analyze included important microbially-produced signaling molecules such as indole [32, 33], butyrate [34], D-alanine [35], and known microbial metabolites of dietary components such as 4-hydroxybenzoic acid [36] and diacetyl [37].

Of the 57 metabolites analyzed by MIMOSA, only 22 were predicted to be of bacterial origin by AMON. Some compounds analyzed by MIMOSA that were not predicted by AMON to be of microbial origin were substrates and not products in microbial reactions. This reflects the different goals of the programs to predict metabolite origins (AMON) versus metabolite turnover that may be influenced by production or degradation (MIMOSA). Three compounds that AMON determined that the host and the microbiome could produce were well-predicted by MIMOSA. These included biliverdin (C00500) and cell membrane components phosphatidylethanolamine (C00350) and 1-Acyl-sn-glycero-3-phosphocholine (C04230).

Discussion

Taken together, these analyses show that AMON can be used to predict the putative origin of compounds detected in a complex metabolome. Our case study shows the specific application of predicting origins of plasma compounds as being from the fecal microbiome versus the host. However, this tool can be used to compare any number of different sources – e.g. from the microbiomes of different body sites or compounds that may come directly from plants consumed in the diet. Also, the outputs of AMON can be used in conjunction with lists of metabolites that were determined to significantly differ with disease state or correlate with other host phenotypes to predict origins of metabolites of interest.

AMON uses the latest updates of KEGG while not requiring the user to purchase a KEGG license, by using either user supplied files for those with a license or the KEGG API which is freely available. However, we do note that the KEGG API option is comparatively slow and limits the maximum dataset size (due to limits of the KEGG API). AMON is built to be flexible to the methods used to obtain the list of KOs present in each source sample and compounds present in a metabolome. Although our example uses PICRUSt to predict compounds of bacterial origin using 16S rRNA sequence data, AMON requires a list of KEGG Orthology identifiers as input and so could also be used with shotgun sequencing data. This can allow for a more thorough interrogation of host microbiomes that account for strain level variation in genome content and opens its application to environments with less understood genomes.

The pathway enrichment of compounds predicted to be unique to the gut microbiome and the host provide a level of validation for AMON results. The pathways enriched with compounds predicted to only be from microbes are consistent with known roles for gut bacteria in degrading various xenobiotics [38–42] and for influencing amino acid [43, 44] and vitamin metabolism [45]. Likewise, the pathways enriched with compounds predicted to be human only include host processes such as taste transduction and bile secretion. Further, since the microbial community measured was from the human gut and the metabolome from plasma, these results suggest that these may represent microbial metabolites that have translocated from the gut into systemic circulation, although validation of the identity of these compounds with authentic standards would be needed to confirm these results. Several studies that have shown a strong influence of the gut microbiome on the plasma metabolome (reviewed in [4]) and the gut microbiome has been linked with many diseases that occur outside of the gut. Examples include interactions between the gut and brain via microbially derived compounds such as serotonin [44], and branched chain amino acids from the gut microbiome as a contributor to insulin resistance [2].

The most similar tool to AMON is MIMOSA [6]. While AMON's goal is to predict whether a compound could have been produced by community of bacteria versus the host, MIMOSA is a relatively quantitative tool that produces information on which particular microbes may influence which particular microbial metabolite levels, and considers both productive and consumptive relationships in these calculations. Unlike AMON, MIMOSA does not incorporate knowledge of host metabolism.

AMON designated many more compounds in the plasma metabolome of being of potential microbial origin compared to MIMOSA when run on the same dataset, and these included important microbially-produced signaling molecules such as indole [32, 33]. One potential reason for this may be more strict criteria needed for forming a metabolic potential score in MIMOSA, as they note in their paper that roughly 50% of metabolites in each data set could not be scored [6]. However, another source of this difference may be the KEGG source file used to define reactions. AMON uses the "reaction" file provided by KEGG which details all

reactions in the KEGG database and MIMOSA uses the "reaction_mapformula.lst" file, which also gives pathway specific information for each reaction (although MIMOSA does not currently use this additional information). We chose to use the "reaction" file of KEGG because it contains information for more reactions than the reaction_mapformula.lst file (e.g. 11,196 versus 7420 for files downloaded on June 9, 2019). The PRMT algorithm used by MIMOSA also makes many assumptions to perform a quantitative analysis that AMON does not, including that that relative abundance of genes for a unique enzyme function reflects levels of expressed functional proteins and reaction rates. Although the PRMT algorithm generally and MIMOSA specifically have been shown to provide strong correlations between microbiome functionality and metabolites and biological insights [6, 17], these weaknesses indicate that the broader information of microbe produced metabolites that is not reliant on this quantitative information that AMON produces is also valuable.

However, for compounds that were evaluated by both MIMOSA and AMON, using the two tools together provided interesting and complimentary insights. In particular, 3 compounds that AMON determined that both the host and the microbiome could produce were well-predicted by MIMOSA, supporting that gut microbe metabolism is an important driver of levels of these compounds despite the ability of the host to produce them. One of these is biliverdin, which is produced by macrophages during heme catabolism but also produced by heme oxygenenases encoded by a variety of bacteria that utilize heme as a source of iron [46]. The other two were lipids that are common components of bacterial cell membranes, supporting that cellular components of bacteria shape the plasma metabolome.

Our analysis also highlights limitations of these approaches that use functional databases such as KEGG due to issues with annotation of both metabolites and the enzymes that may produce them. Overall, it is striking that of 5971 compounds in the LC/MS data, only 471 could be linked to enzymatic reactions in KEGG. For example the human genome is known to contain approximately 20,000 genes [47]; however, there are only 7286 KOs annotated in KEGG. These KOs only predict the creation of 1376 unique compounds while the Human Metabolome Database 4.0 contains 114,100 [48]. Part of this discrepancy is because multiple species of lipids are, generally, reduced to a single compound in KEGG. For example, while KEGG includes a single phosphatidylcholine (PC) lipid molecule in the glycerophospholipid pathway, in fact, there are over 1000 species of PCs. It is also important to note that metabolite annotations are based on peak masses and isotope ratios, which can often represent multiple compounds and/or in-source fragments; our confidence in the identity of these compounds is only moderate. As with any metabolomics dataset, we caution the user to limit their biological conclusions when level 3 annotations are used in downstream applications such as AMON. As it is not feasible to verify compound identities using authentic standards or MS/MS for hundreds of compounds, AMON provides a valuable tool for prioritizing compounds for additional analysis, including identification using authentic standards, by providing information on their potential origins.

The limitations are more stark for complex microbial communities, where there are fewer genes of known function. Because of these gaps in our knowledge of metabolite production, efforts to identify microbially produced metabolites that affect disease should also use methods that are agnostic to these knowledge-bases. These include techniques such as 1) identifying highly correlated microbes and metabolites to identify potential productive/consumptive relationships that can be further validated 2) molecular networking approaches which take advantage of tandem mass spectroscopy data to annotate compounds based on similarity to known compounds with related tandem mass spectrometry (MS/MS) profiles [49] or 3) coupling LC/MS runs with data from germ-free versus colonized animals [1, 50, 51] or antibiotic versus non-antibiotic treated humans [52, 53]. Because AMON takes only KO identifiers and can pull database information from the KEGG API or user provided KEGG files, our tool will become increasingly useful with improvements from KEGG as well as other parts of the annotation process. In addition, AMON can also accept metabolomics datasets with Level 1 identifications; i.e. where the identity of the compounds has been verified with authentic standards.

Although our application is specifically designed to work with the KEGG database, similar logic could be used for other databases such as MetaCyc [54]. Our tool also does not apply methods such as gap-filling [7, 55] and metabolic modeling [12, 57] in its estimates. The goal is not to produce precise measurements of the contributions of the microbiome and host to the abundance of a metabolite. Rather, AMON is designed to annotate metabolomics results to give the user an understanding of whether specific metabolites could have been produced directly by the host or microbial communities. If a metabolite is identified by AMON as being of microbial origin and is associated with a phenotype, this result should motivate the researcher to perform follow up studies. These can include confirming the identity of the metabolite, via methods such as tandem mass spectrometry, and performing experiments to confirm the ability of microbes of interest to produce the metabolite.

AMON also does not account for co-metabolism between the host and microbes. An example of this is the production of TMAO from dietary choline. Our tool would list TMAO as a host compound and its precursor trimethylamine (TMA) as a microbiome derived compound but would not indicate that TMAO could overall not be produced from dietary substrates unless a microbiome was present. Further inspection of metabolic networks, which is enabled by AMON's functionality in producing outputs for visualization in KEGG mapper may be needed to decipher these co-metabolism relationships. Previously described methods for constructing possible biotransformation pathways, while discriminating between microbiota and host reactions [15] could also be incorporated into AMON in the future.

Conclusions

When researchers are seeking to integrate microbiome and metabolome data, identifying the origin of metabolites measured is an obvious route. AMON facilitates the annotation of metabolomics data by tagging compounds with their potential origin, either as bacteria or host. This allows researchers to develop hypotheses about the metabolic involvement of microbes in disease.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3176-8.

Additional file 1: Table S1. Table of Metabolites with KEGG Annotations from LC/MS of Human Plasma

Additional file 2: Table S2. Example Output of AMON Metabolite Origin Table

Additional file 3: Table S3. Example Output of AMON Pathway Enrichment Table

Additional file 4: Table S4. MIMOSA Results Table from Case Study Data

Abbreviations

AMON: Analysis of Metabolite Origins Using Networks; HILIC: Hydrophilic interaction liquid chromatography; KEGG: Kyoto Encyclopedia of Genes and Genomes; KO: KEGG Orthology; LC/MS: Liquid Chromatography / Mass Spectrometry; MPP: Mass Profiler Professional; MS/MS: Tandem mass spectrometry; MTBE: Methyl tert-butyl ether; OTU: Operational Taxonomic Unit; PRMT: Predicted Relative Metabolic Turnover; QTOF: Quadrupole Time of Flight; rRNA: ribosomal RNA; RT: Retention Time; TMA: Trimethylamine; TMAO: Trimethylamine N-oxide; UHPLC: Ultra-high precision liquid chromatography

Acknowledgements

The authors would like to thank Elmar Pruesse for discussions on design of the software and Nancy Moreno Huizar for feedback on software usage. We also thank Dr. Brent Palmer, Dr. Tom Campbell, Suzanne Fiorillo and Christine Griesmer for providing samples for the HIV case study.

Authors' contributions

M.S. designed and wrote the software, analyzed the example data and wrote the manuscript. K.T. analyzed example data, contributed to the software, and to the manuscript. K.Q., K.D., X.Z. and S.B. extracted compounds, ran LC/MS metabolomics, and processed metabolomics data. N.R. supervised the metabolomics work and interpreted results. C.L. conceived of the study, interpreted results, and helped in the writing of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by National Institutes of Health R01 DK104047 for the collection of the microbiome data and the associated metabolomics supplement by the National Institutes of Health Common fund and 4 T15 LM009451–10 for the collection of the metabolomics data. High performance computing was supported by a cluster at the University of Colorado Boulder funded by National Institutes of Health 1S10OD012300. Mike Shaffer was supported by a National Institutes of Health Bioinformatics Research training award 5 T15 LM009451–12. The metabolomics workbench used to make the metabolomics data available is supported by NIH grant U2C-DK119886.

Availability of data and materials

Microbiome data is available in the European Nucleotide Archive repository **PRJEB28485** (https://www.ebi.ac.uk/ena/data/view/PRJEB28485). The metabolomics data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, https://www.metabolomicsworkbench.org, where it has been assigned Project ID **(ST001268)**. The data can be accessed directly via its Project DOI:**(**https://doi.org/10.21228/M8F108**)**. The genes predicted to be present in this data set and the compounds detected in the metabolomics data are listed in the AMON repository (https://github.com/lozuponelab/AMON/tree/master/data).

Ethics approval and consent to participate

All study participants were recruited from University of Colorado Hospital with a protocol that was approved by the Colorado Multiple Institutional Review Board (CoMIRB 14–1595). Informed consent was obtained with written consent forms from all study participants.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA. [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, 80045CO, Aurora, USA. [3]Present address: BioElectron Technology Corporation, Mountain View, CA 94043, USA.

Received: 14 December 2018 Accepted: 28 October 2019
Published online: 28 November 2019

References

1. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. 2011;472:57–63. https://doi.org/10.1038/nature09922.
2. Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BAH, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. Nature. 2016;535:376–81. https://doi.org/10.1038/nature18646.
3. Fujimura KE, Sitarik AR, Havstad S, Lin DL, Levan S, Fadrosh D, et al. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. Nat Med. 2016;22:1187–91. https://doi.org/10.1038/nm.4176.
4. Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. Microbiome and metabolome data integration provides insight into health and disease. Transl Res. 2017;189:51–64. https://doi.org/10.1016/J.TRSL.2017.07.001.
5. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092.
6. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, et al. Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and

Metabolic Variation mSystems 2016;1:e00013–e00015. doi:https://doi.org/10.1128/mSystems.00013-15.
7. Thiele I, Vlassis N, Fleming RMT. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014;30:2529–31. https://doi.org/10.1093/bioinformatics/btu321.
8. Sung J, Kim S, Cabatbat JJT, Jang S, Jin YS, Jung GY, et al. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nat Commun. 2017;8.
9. Sridharan GV, Choi K, Klemashevich C, Wu C, Prabakaran D, Bin PL, et al. Prediction and quantification of bioactive microbiota metabolites in the mouse gut. Nat Commun. 2014;5:5492. https://doi.org/10.1038/ncomms6492.
10. Heinken A, Thiele I. Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. Gut Microbes. 2015;6:120–30. https://doi.org/10.1080/19490976.2015.1023494.
11. Mendes-Soares H, Mundy M, Soares LM, Chia N. MMinte: an application for predicting metabolic interactions among the microbial species in a community. BMC Bioinformatics. 2016;17:343. https://doi.org/10.1186/s12859-016-1230-3.
12. Mendes-Soares H, Chia N. Community metabolic modeling approaches to understanding the gut microbiome: bridging biochemistry and ecology. Free Radic Biol Med. 2017;105:102–9. https://doi.org/10.1016/J.FREERADBIOMED.2016.12.017.
13. Chiu H-C, Levy R, Borenstein E. Emergent biosynthetic capacity in simple microbial communities. PLoS Comput Biol. 2014;10:e1003695. https://doi.org/10.1371/journal.pcbi.1003695.
14. Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, et al. Quantifying diet-induced metabolic changes of the human gut microbiome. Cell Metab. 2015;22.
15. Sridharan GV, Choi K, Klemashevich C, Wu C, Prabakaran D, Bin PL, et al. Prediction and quantification of bioactive microbiota metabolites in the mouse gut. Nat Commun. 2014;5:5492. https://doi.org/10.1038/ncomms6492.
16. Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, Henry CS, et al. Predicted relative Metabolomic turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. Microb Inform Exp. 2011;1:4. https://doi.org/10.1186/2042-5783-1-4.
17. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome. 2013;1:17. https://doi.org/10.1186/2049-2618-1-17.
18. Larsen PE, Dai Y. Metabolome of human gut microbiome is predictive of host dysbiosis. Gigascience. 2015;4:42. https://doi.org/10.1186/s13742-015-0084-3.
19. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotech. 2013;31:814–21 https://doi.org/10.1038/nbt.2676.
20. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data: fig. 1. Bioinformatics. 2015;31:2882–4. https://doi.org/10.1093/bioinformatics/btv287.
21. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012;8:e1002358. https://doi.org/10.1371/journal.pcbi.1002358.
22. Armstrong AJ, Shaffer M, Nusbacher NM, Griesmer C, Fiorillo S, Schneider JM, et al. An exploration of Prevotella-rich microbiomes in HIV and men who have sex with men. Microbiome. 2018:424291. https://doi.org/10.1101/424291.
23. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551:457. https://doi.org/10.1038/nature24621.
24. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.
25. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1. https://doi.org/10.1093/bioinformatics/btq461.
26. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6. https://doi.org/10.1038/nmeth.f.303.
27. Yang Y, Cruickshank C, Armstrong M, Mahaffey S, Reisdorph R, Reisdorph N. New sample preparation approach for mass spectrometry-based profiling of plasma results in improved coverage of metabolome. J Chromatogr A. 2013;1300:217–26. https://doi.org/10.1016/j.chroma.2013.04.030.
28. Heischmann S, Quinn K, Cruickshank-Quinn C, Liang L-P, Reisdorph R, Reisdorph N, et al. Exploratory metabolomics profiling in the Kainic acid rat model reveals depletion of 25-Hydroxyvitamin D3 during Epileptogenesis. Sci Rep. 2016;6:31424. https://doi.org/10.1038/srep31424.
29. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). Metabolomics. 2007;3:211–21. https://doi.org/10.1007/s11306-007-0082-2.
30. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol. 2014;48:2097–8.
31. Kim CH. Immune regulation by microbiome metabolites. Immunology. 2018;154:220–9. https://doi.org/10.1111/imm.12930.
32. Jaglin M, Rhimi M, Philippe C, Pons N, Bruneau A, Goustard B, et al. Indole, a signaling molecule produced by the gut microbiota Negatively Impacts Emotional Behaviors in Rats. Front Neurosci. 2018;12:216. https://doi.org/10.3389/fnins.2018.00216.
33. Louis P, Flint HJ. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS Microbiol Lett. 2009;294:1–8. https://doi.org/10.1111/j.1574-6968.2009.01514.x.
34. Radkov AD, Moe LA. Bacterial synthesis of d-amino acids. Appl Microbiol Biotechnol. 2014;98:5363–74. https://doi.org/10.1007/s00253-014-5726-3.
35. Gonthier M-P, Cheynier V, Donovan JL, Manach C, Morand C, Mila I, et al. Microbial aromatic acid metabolites formed in the gut account for a major fraction of the polyphenols excreted in urine of rats fed red wine polyphenols. J Nutr. 2003;133:461–7. https://doi.org/10.1093/jn/133.2.461.
36. Bartowsky EJ, Henschke PA. The 'buttery' attribute of wine—diacetyl—desirability, spoilage and beyond. Int J Food Microbiol. 2004;96:235–52. https://doi.org/10.1016/J.IJFOODMICRO.2004.05.013.
37. Maurice CFF, Haiser HJJ, Turnbaugh PJJ. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell. 2013;152:39–50. https://doi.org/10.1016/j.cell.2012.10.052.
38. Lu K, Mahbub R, Fox JG. Xenobiotics: interaction with the intestinal microflora. ILAR J. 2015;56:218–27. https://doi.org/10.1093/ilar/ilv018.
39. Das A, Srinivasan M, Ghosh TS, Mande SS, Alastrue C, Dore J. Xenobiotic metabolism and gut microbiomes. PLoS One. 2016;11:e0163099. https://doi.org/10.1371/journal.pone.0163099.
40. Saad R, Rizkallah MR, Aziz RK. Gut Pharmacomicrobiomics: the tip of an iceberg of complex interactions between drugs and gut-associated microbes doi:https://doi.org/10.1186/1757-4749-4-16.
41. Clayton TA, Baker D, Lindon JC, Everett JR, Nicholson JK. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. Proc Natl Acad Sci U S A. 2009;106:14728–33. https://doi.org/10.1073/pnas.0904489106.
42. Neis E, Dejong C, Rensen S. The role of microbial amino acid metabolism in host metabolism. Nutrients. 2015;7:2930–46. https://doi.org/10.3390/nu7042930.
43. O'Mahony SM, Clarke G, Borre YE, Dinan TG, Cryan JF. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. Behav Brain Res. 2015;277:32–48. https://doi.org/10.1016/j.bbr.2014.07.027.
44. Streit WR, Entcheva P. Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. Appl Microbiol Biotechnol. 2003;61:21–31. https://doi.org/10.1007/s00253-002-1186-2.
45. Wilks A, Ikeda-Saito M. Heme utilization by pathogenic Bacteria: not all pathways Lead to Biliverdin. Acc Chem Res. 2014;47:2291–8. https://doi.org/10.1021/ar500028n.
46. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. Hum Mol Genet. 2014;23:5866–78. https://doi.org/10.1093/hmg/ddu309.
47. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46:D608–17. https://doi.org/10.1093/nar/gkx1089.
48. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci. 2012;109:E1743–52. https://doi.org/10.1073/pnas.1203689109.
49. Rothhammer V, Mascanfroni ID, Bunse L, Takenaka MC, Kenison JE, Mayo L, et al. Type I interferons and microbial metabolites of tryptophan

modulate astrocyte activity and central nervous system inflammation via the aryl hydrocarbon receptor. Nat Med. 2016;22:586. https://doi.org/10.1038/nm.4106.
50. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013;155:1451–63. https://doi.org/10.1016/j.cell.2013.11.024.
51. Tang WHW, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, et al. Intestinal microbial metabolism of Phosphatidylcholine and cardiovascular risk. N Engl J Med. 2013;368:1575–84. https://doi.org/10.1056/NEJMoa1109400.
52. Antunes LCM, Han J, Ferreira RBR, Lolić P, Borchers CH, Finlay BB. Effect of antibiotic treatment on the intestinal metabolome. Antimicrob Agents Chemother. 2011;55:1494–503. https://doi.org/10.1128/AAC.01664-10.
53. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2014;42:D459–71. https://doi.org/10.1093/nar/gkt1103.
54. Orth JD, Palsson BØ. Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng. 2010;107:403–12. https://doi.org/10.1002/bit.22844.
55. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28:245–8. https://doi.org/10.1038/nbt.1614.

Publisher's Note