AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

Check for updates

| Environmental Microbiology | Research Article

Refining microbial community metabolic models derived from metagenomics using reference-based taxonomic profiling

Marwan E. Majzoub,[1] Laurence D. W. Luu,[1] Craig Haifer,[2,3] Sudarshan Paramsothy,[4,5] Thomas J. Borody,[6] Rupert W. Leong,[4,5] Torsten Thomas,[7] Nadeem O. Kaakoush[1]

AUTHOR AFFILIATIONS See affiliation list on p. 12.

ABSTRACT Characterization of microbial community metabolic output is crucial to understanding their functions. Construction of genome-scale metabolic models from metagenome-assembled genomes (MAG) has enabled prediction of metabolite production by microbial communities, yet little is known about their accuracy. Here, we examined the performance of two approaches for metabolite prediction from metagenomes, one that is MAG-guided and another that is taxonomic reference-guided. We applied both on shotgun metagenomics data from human and environmental samples, and validated findings in the human samples using untargeted metabolomics. We found that in human samples, where taxonomic profiling is optimized and reference genomes are readily available, when number of input taxa was normalized, the reference-guided approach predicted more metabolites than the MAG-guided approach. The two approaches showed significant overlap but each identified metabolites not predicted in the other. Pathway enrichment analyses identified significant differences in inferences derived from data based on the approach, highlighting the need for caution in interpretation. In environmental samples, when the number of input taxa was normalized, the reference-guided approach predicted more metabolites than the MAG-guided approach for total metabolites in both sample types and non-redundant metabolites in seawater samples. Nonetheless, as was observed for the human samples, the approaches overlapped substantially but also predicted metabolites not observed in the other. Our findings report on utility of a complementary input to genome-scale metabolic model construction that is less computationally intensive forgoing MAG assembly and refinement, and that can be applied on shallow shotgun sequencing where MAGs cannot be generated.

IMPORTANCE Little is known about the accuracy of genome-scale metabolic models (GEMs) of microbial communities despite their influence on inferring community metabolic outputs and culture conditions. The performance of GEMs for metabolite prediction from metagenomes was assessed by applying two approaches on shotgun metagenomics data from human and environmental samples, and validating findings in the human samples using untargeted metabolomics. The performance of the approach was found to be dependent on sample type, but collectively, the reference-guided approach predicted more metabolites than the MAG-guided approach. Despite the differences, the predictions from the approaches overlapped substantially but each identified metabolites not predicted in the other. We found significant differences in biological inferences based on the approach, with some examples of uniquely enriched pathways in one group being invalidated when using the alternative approach, highlighting the need for caution in interpretation of GEMs.

KEYWORDS metabolic modelling, microbiome, human, environmental, metagenomics

Editor Samuel Chaffron, CNRS Delegation Bretagne et Pays de Loire, Nantes, France

Address correspondence to Nadeem O. Kaakoush, n.kaakoush@unsw.edu.au.

S.P. has served as a consultant for Finch Therapeutics and has received speaker fees from Ferring, Janssen, and Takeda. T.J.B. has a pecuniary interest in the Centre for Digestive Diseases, is a medical advisor to Finch Therapeutics, RedHill Bio, and Topelia Aust, and holds patents in FMT treatment. All other authors have no conflicts of interest to declare.

See the funding table on p. 13.

Received 30 May 2024
Accepted 10 July 2024
Published 13 August 2024

Microbial communities perform essential functions within their host, such as stimulating and developing the host immune system, providing protection against pathogens, maintaining metabolic homeostasis, and generating nutrients through pathways that do not exist in the host (1–3). The host-associated microbiome can perform these functions through the production of a range of metabolic products including short-chain fatty acids, secondary and tertiary bile acids, and by-products of tryptophan metabolism among many others (4–8). Moreover, microbial communities perform essential ecological functions within their environmental niches such as biogeochemical cycling (9, 10). Thus, there is a growing interest in deciphering the metabolic output of microbial communities.

Metabolic network modeling is a system biology approach that enables the inference of metabolic outputs from genomic information (11). The construction of genome-scale metabolic models (GEMs) has provided researchers with the ability to study strain- and environment-specific metabolism (12–14). Modeling can also be expanded from individual strains to a community level through several ways that are still subject to refinement (15). These *in silico* models provide a platform for researchers to predict metabolic interactions that occur within a microbial community, as well as metabolic activities under specific environmental conditions (16, 17). Thus, a GEM can allow for the prediction of pathways that may be important to the host or environment and optimal culture conditions for lesser known microorganisms (15).

When constructing a community-level GEM from metagenomes, one approach is to input metagenome-assembled genomes (MAGs); however, input of reference genomes based on taxonomic profiling has also been proposed (18). Advances in the sequencing technology has enabled deeper profiling of microbial communities through shotgun metagenomics and the capacity for production of more complete MAGs that subsequently improve the quality of GEMs. These technological advances have also led to a substantial increase in coverage of reference pangenomes for bacterial species, providing an opportunity to utilize these pangenomes in GEM construction. To date, little is known about the influence of the genomic input strategy on community-level metabolic models based on metagenomes.

Here, we constructed GEMs using shotgun metagenomic data from human stool samples as well as environmental samples (seawater and sediment) by using high- and medium-quality MAGs as input (MAG-guided approach) and, in parallel, constructed GEMs using an approach guided by species relative abundances calculated from taxonomic profiling (reference-guided approach). We validated inferred outputs from the human samples using longitudinal untargeted metabolomics data. We then compared the efficiency and differences in biological inferences of the two approaches when applied to microbiotas with different efficacies when used to treat patients with ulcerative colitis.

RESULTS

Selection of features from human metagenomic samples for modeling

The taxonomic composition within each human stool sample was profiled using MetaPhlAn4, and relative abundances of microbial species were determined (Fig. 1A and B). Mean relative abundances of species per individual (i.e., donor 1 or 2) were calculated based on all available samples from that donor. Species with mean relative abundances of greater than or equal to 5%, 2.5%, 1%, or 0.5% were selected for the reference-guided approach as four independent inputs per donor (Fig. 1A; Table 1). In addition, MAGs assembled from the same samples were classified as either high quality or medium quality and selected as two independent inputs per donor for the MAG-guided approach (Fig. 1A; Table 1). Thus, six predictions identifying the net metabolic output of the community were made for each donor microbiota using these reference-guided and MAG-guided approaches (Fig. 1A).

The species selected at different relative abundance thresholds for input (n = 6, 10, 20, and 34) into the reference-guided approach for donor 1 (Table 1; Table S1) were classified

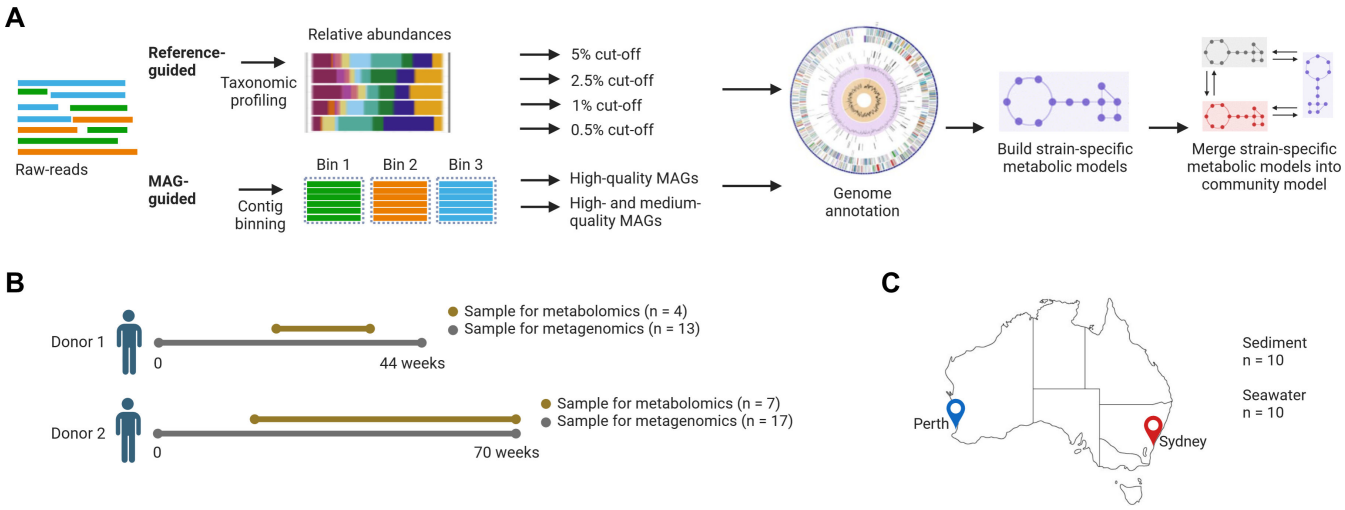

FIG 1 Metabolic modeling pipeline and study samples. (A) Framework for the community metabolic modeling using the reference-guided and MAG-guided approaches. (B) Sampling from the two healthy individuals for the metagenomics and metabolomics. (C) Sampling sites for the environmental metagenomic samples.

to a total of 23 genera that included several unnamed genus-level genomic bins that can be obtained from Pasolli et al. (19). Selected species for donor 2 (*n* = 2, 6, 24, and 37) included bacteria belonging to 24 genera. For the MAG-guided approach, a total of 56 high-quality MAGs and 86 medium- or high-quality MAGs as well as 67 high-quality MAGs and 84 medium- or high-quality MAGs were assembled from donor 1 and donor 2 samples, respectively (Tables S2 and S3). The analysis included MAGs from 70 and 65 genera for donors 1 and 2, respectively. The total number of uniquely classified taxa that were inputted into the GEMs was higher (>2-fold) in the MAG-guided approach than the reference-guided approach (i.e., high + medium vs 0.5% cutoff) (Table 1). Metabolite predictions from individual taxa inputted into the models, including the percentage of blocked reactions, are provided in Table S4.

TABLE 1 Microbial community metabolic modeling in human fecal samples for the reference-guided and MAG-guided approaches[a]

Donor	Parameter	Reference-guided				MAG-guided		
		5%	2.5%	1%	0.5%	High	High and medium	Normalized
Donor 1	Taxa	6	10	20	34	56	86	34
	Total predicted metabolites	5,568	9,149	17,812	28,583	44,928	67,924	28,172
	Total extracellular metabolites	131	145	184	200	191	196	185
	Unique metabolites	1,248	1,290	1,402	1,500	1,519	1,543	1,493
	Confirmed unique metabolites (total)	167	170	180	183	184	185	183
	Data loss (%)	31.41	30.62	30.03	30.13	30.22	30.33	30.21
	Confirmed unique metabolites (%)	19.51	18.99	18.35	17.46	17.36	17.21	17.56
Donor 2	Taxa	2	6	24	37	67	84	37
	Total predicted metabolites	1,910	5,570	21,318	33,661	55,818	69,840	31,902
	Total extracellular metabolites	102	148	176	219	223	223	221
	Unique metabolites	1,062	1,251	1,404	1,562	1,576	1,579	1,545
	Confirmed unique metabolites (total)	145	154	171	180	180	180	180
	Data loss (%)	27.50	31.73	30.77	30.15	30.71	30.72	30.36
	Confirmed unique metabolites (%)	18.83	18.03	17.59	16.50	16.48	16.45	16.73

[a]Summary includes the number of input bacterial species used for the metabolic predictions for each donor (taxa), the total number of predicted metabolites (duplicates not removed), total number of metabolites classified as extracellular (duplicates not removed), the number of non-duplicated total predicted metabolic products (unique metabolites), and the number of metabolites validated using untargeted metabolomics (confirmed unique metabolites). Data loss resulted from the predicted metabolic product not having an assigned Kyoto Encyclopedia of Genes and Genomes ID that can be matched to the untargeted metabolomics output and no standardized naming convention.

Outcomes of metabolic predictions in human metagenomic samples

A total of 5,568, 9,149, 17,812, and 28,583 metabolic compounds were predicted for donor 1, while a total of 1,910, 5,570, 21,318, and 33,661 were predicted for donor 2 using the 5%, 2.5%, 1%, and 0.5% relative abundance cutoffs, respectively (Table 1). These included a total of 1,248 and 1,062 unique metabolic compounds predicted for donor 1 and donor 2 samples, respectively, for the reference-guided approach when a relative abundance cutoff of 5% was implemented (Table 1). The total number of unique metabolic compounds further increased to 1,290, 1,402, and 1,500 for donor 1 and 1,251, 1,404, and 1,562 for donor 2 when the relative abundance cutoffs of 2.5%, 1%, and 0.5% were employed (Table 1). The number of novel predicted metabolites appeared to saturate with increasing number of input taxa (Fig. S1A), which is likely due to metabolic redundancy. In comparison, despite a higher number of input taxa and a higher number of total predicted metabolites (Table 1), 1,519 and 1,576 unique metabolic compounds were predicted in donor 1 and donor 2 samples, respectively, using the high-quality MAGs alone, while a total of 1,543 and 1,579 unique metabolic compounds were predicted in donor 1 and donor 2 samples, respectively, using both high- and medium-quality MAGs (Table 1). Thus, while the reference-guided approach predicted substantially lower total numbers of metabolites, it predicted a similar number of unique metabolites in both donors with less input taxa (i.e., 1,500 from 34 for donor 1 and 1,562 from 37 for donor 2 vs 1,543 from 86 for donor 1 and 1,579 from 84 for donor 2), possibly due to MAGs being incomplete. We tested this by normalizing the number of input taxa in the MAG-guided approach to the reference-guided approach (34 and 37 most complete high-quality MAGs from donor 1 and donor 2, respectively), showing a lower number of total and unique predicted metabolites (Table 1).

Next, we assessed if the two approaches were additive or complementary (Fig. 2A and B; Fig. S2). A substantial amount of overlap was observed across the two approaches (n = 1,475 of 1,543 and 1,530 of 1,579 for donor 1 and donor 2, respectively), with reference-guided approach predicting 25 and 32 unique metabolites for donor 1 and donor 2, respectively (Fig. 2A and B; Fig. S2). The MAG-guided approach also predicted unique metabolites not predicted by the reference-guided approach (68 and 49 for donor 1 and donor 2, respectively) (Fig. 2A and B; Fig. S2).

Validation of metabolic predictions in human metagenomic samples with untargeted metabolomics

To validate metabolite predictions from the community models from the reference-guided and MAG-guided approaches, they were compared against untargeted metabolomics data generated longitudinally in each donor. A total of 167 and 145 metabolic products were validated in donor 1 and donor 2 samples, respectively, for the reference-guided approach when a relative abundance cutoff of 5% was implemented (Table 1). The total number of validated non-redundant metabolic products further increased to 170, 180, and 183 for donor 1 and 154, 171, and 180 for donor 2 when the relative abundance cutoffs of 2.5%, 1%, and 0.5% were employed (Table 1). A total of 184 and 180 metabolic products were validated in donor 1 and donor 2 samples, respectively, using the high-quality MAGs alone, while a total of 185 and 180 metabolic products were identified in donor 1 and donor 2 samples, respectively, using both high- and medium-quality MAGs (Table 1). Overall, a similar number of non-redundant metabolites were validated using the reference-guided approach and the MAG-guided approach (Table 1). Data loss, resulting from a lack of Kyoto Encyclopedia of Genes and Genomes (KEGG) IDs for a portion of predicted metabolites and non-standardized naming conventions, was similar in both the reference-guided data and MAG-guided data (Table 1). Both approaches had a similar susceptibility to error given that similar percentages of total unique metabolites were validated across both donors (Table 1). A further important point is that we did not calculate accuracy, specificity, and sensitivity due to our ability to only validate non-redundant metabolites rather than the total predicted

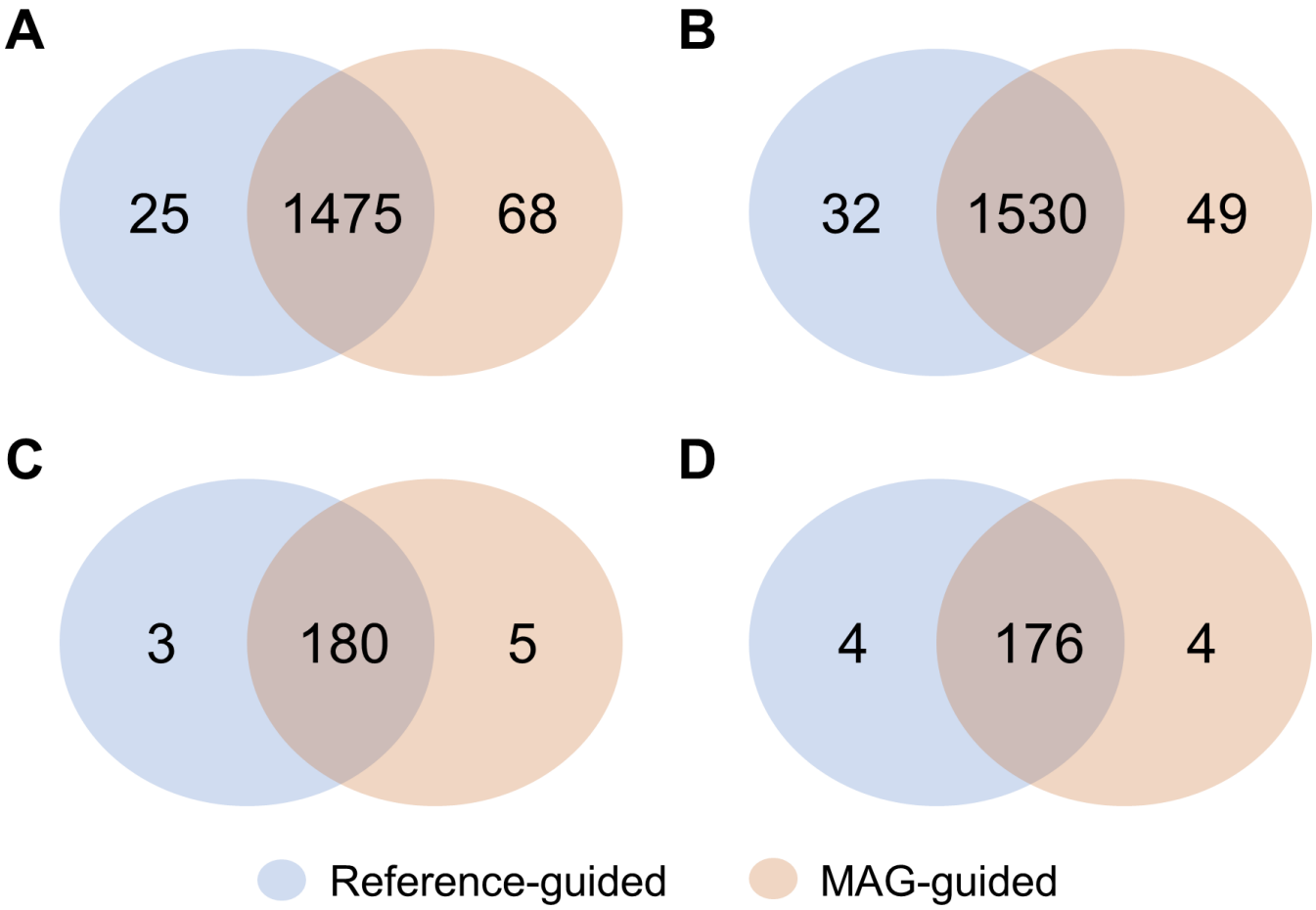

FIG 2 Number of unique metabolites across the reference-guided (0.5% cutoff) and MAG-guided (high- and medium-quality) inputs for human fecal samples. (A) Predicted metabolites for donor 1. (B) Predicted metabolites for donor 2. (C) Confirmed metabolites for donor 1. (D) Confirmed metabolites for donor 2.

metabolic output of the community, which we believe confounds the calculation of these parameters.

We then assessed if the approaches were complementary or additive for the confirmed metabolites (Fig. 2C and D). For donor 1, five unique metabolites in the MAG-guided approach were not predicted by the reference-guided approach. In contrast, three validated metabolites were predicted by the reference-guided approach but not the MAG-guided approach. Similarly, in donor 2, the MAG-guided approach predicted four metabolites unique to it while the reference-guided approach also predicted four metabolites unique to it, highlighting that there is substantial overlap across the approaches.

Changes to biological inferences based on the selected approach

To establish differences in biological inference gained by the selected approach, we performed pathway enrichment analysis across the different predicted unique metabolites lists. We first assessed within-donor variation by determining the significantly enriched pathways ($q < 0.05$) for metabolite outputs from each approach (Table S5). We identified 141 and 147 pathways to be common across the approaches for donor 1 and donor 2, respectively, whereas a total of 10 and 4 were unique to one approach (Table S5). We then compared the significantly enriched pathway outputs from each of the donor 1 and donor 2 microbiotas (Table S6). The microbiota from these donors have been shown to have variable efficacies (100% vs 36.4% for $n = 15$ patients, Fisher's $P = 0.026$) in the context of treating patients with ulcerative colitis (20), making the

identification of true unique metabolic pathways of therapeutic relevance. When using the reference-guided approach, five and seven pathways were uniquely enriched in donor 1 and donor 2, respectively. In contrast, only one and five pathways were uniquely enriched in donor 1 and donor 2 when using the MAG-guided approach. Notably, lysine degradation (SMP00037) was identified to be uniquely enriched in donor 1 in the reference-guided approach, whereas the KEGG version of lysine degradation (map00310) was uniquely enriched in donor 2 using the MAG-guided approach. This highlighted the variation in biological inferences depending on the approach employed.

To determine if biological inferences change further when combining outputs from the two approaches, we merged the predicted metabolite lists from the reference- and MAG-guided approaches, removed duplicate metabolites, and then compared pathway enrichment within and between the two donors. This resulted in 148 and 149 pathways to be significantly ($q < 0.05$) enriched in donor 1 and donor 2, respectively (Table S7). For donor 1, the 148 pathways corresponded to 131 common across approaches, 3 specific to the reference-guided approach, 14 specific to the MAG-guided approach. For donor 2, the 149 pathways corresponded to 144 common across approaches, two specific to the reference-guided approach, two specific to the MAG-guided approach, and one novel pathway, highlighting that within-donor inferences can change. For the comparison between donors using the combined data, two and three enriched pathways were identified to be unique to each donor at $q < 0.05$, with the only result found to be consistent across the reference-only, MAG-only, and combined approaches being enrichment of propanoate metabolism in donor 2 (Table 2; Table S7). Furthermore, biosynthesis of terpenoids and steroids (map01062), initially shown to be uniquely enriched in donor 1 using both approaches, was identified in donor 2 when using the combined approach (Tables S6 and S7).

Outcomes of metabolic predictions in environmental samples

To determine if the differences in metabolite predictions between the two types of approaches were observed beyond human samples, we applied them to environmental samples originating from seawater or sediment (Fig. 1C). In total, 32 species were found to have a mean relative abundance ≥0.5% for the sediment samples (Table 3; Table S8) and these included bacteria belonging to 23 genera. Forty-four species were identified at the same cutoff for the seawater samples (Table 3; Table S8), and these included bacteria belonging to 41 genera. A total of 30,051 and 20,519 predicted metabolites and 1,610 and 1,597 unique metabolic compounds were predicted in sediment and seawater samples, respectively, using a mean relative abundance cutoff of 0.5% (Table 3). Of note, draft metabolic models could not be built for two species in the sediment samples and a further 22 species for the seawater samples using our method due to a lack of publicly available reference genomes or representative MAGs for those species. This issue was apparent in the environmental samples as not all MAGs used for classification by MetaPhlAn4 were publicly available at the time of analysis, unlike those commonly found in human samples (19). Similar to the human samples, the number of novel predicted metabolites appeared to saturate with increasing number of input taxa (Fig. S1B).

A total of 51 high-quality MAGs and 131 high- or medium-quality MAGs and 75 high-quality MAGs and 158 high- or medium-quality MAGs were assembled from sediment and seawater samples, respectively (Tables S9 and S10). In total, the analysis

TABLE 2 Metabolic pathways significantly enriched only in donor 1 or donor 2 using the combined non-redundant metabolic prediction outputs from both approaches

Donor	Pathway ID	Pathway name	Source	*q*-value
Donor 1	SMP00037	Lysine degradation	HMDB	0.031
	map00562	Inositol phosphate metabolism	KEGG	0.046
Donor 2	SMP00016	Propanoate metabolism	HMDB	0.0071
	SMP00020	Arginine and proline metabolism	HMDB	0.021
	SMP00450	Phytanic acid peroxisomal oxidation	HMDB	0.034

TABLE 3 Microbial community metabolic modeling in environmental samples for the reference-guided and MAG-guided approaches[a]

Sample	Parameter	Reference-guided				MAG-guided		
		5%	2.5%	1%	0.5%	High	High and medium	Normalized
Sediment	Taxa	2 (3)	8 (9)	19 (20)	30 (32)	51	131	30
	Total predicted metabolites	2,192	8,010	20,147	30,051	46,185	120,790	27,813
	Total extracellular metabolites	102	162	193	198	200	211	193
	Unique metabolites	1,301	1,463	1,560	1,610	1,646	1,716	1,621
Seawater	Taxa	2 (3)	6 (7)	13 (23)	21 (44)	75	158	21
	Total predicted metabolites	1,835	5,930	13,391	20,519	68,130	136,462	18,331
	Total extracellular metabolites	93	167	185	195	214	216	181
	Unique metabolites	1,098	1,433	1,569	1,597	1,687	1,728	1,498

[a]Summary includes the number of input bacterial species used for the metabolic predictions for each sample type (taxa), the total number of predicted metabolites (duplicates not removed), total number of metabolites classified as extracellular (duplicates not removed), and the number of non-duplicated total predicted metabolic products (unique metabolites). Numbers in brackets represent actual species detected at the relative abundance cutoff as opposed to true number of species used as input due to absence of genomic information. Normalized refers to metabolite predictions when the total number of input MAGs was matched to number of input taxa at the 0.5% cutoff.

included MAGs from 84 genera for the sediment samples and 86 genera for the seawater samples. In contrast to the human samples, a higher number of total and unique metabolites were observed using the MAG-guided approach compared to the reference-guided approach for the environmental samples. A total of 120,790 and 136,462 predicted metabolic compounds and 1,716 and 1,728 unique metabolic compounds were predicted in the sediment and seawater communities, respectively (Table 3). Metabolite predictions from individual taxa inputted into the models, including percentage of blocked reactions, are provided in Table S11.

To assess if this may be due to the >4- and 7-fold input in species-level taxa for the MAG-guided approach, we normalized the number of input MAGs to the number of input species in the 0.5% cutoff (sediment: 30; seawater: 21), selecting the most complete high-quality MAGs. For seawater samples, the total number (20,519 vs 18,331 metabolites) and number of unique predicted metabolites (1,597 vs 1,498 metabolites) were higher for the reference-guided approach than the MAG-guided approach (Table 3). While the total number of predicted metabolites was higher for the reference-guided approach in the sediment samples (30,051 vs 27,813 metabolites), the number of unique metabolites from the MAG-guided approach was higher (1,621 metabolites) than the reference-guided approach (1,610 metabolites) (Table 3).

We then assessed if the approaches were complementary or additive. For the sediment samples, there was substantial overlap across the approaches (1,576 metabolites), but each approach also identified unique metabolites not detected in the other (reference-guided: 34; MAG-guided: 140 metabolites) (Fig. 3; Fig. S3). Similarly, for the seawater samples, 1,577 metabolites overlapped across the approaches, with a further 20 unique to the reference-guided approach and 151 metabolites unique to the MAG-guided approach (Fig. 3; Fig. S3).

Effect of gapfilling on predictions of metabolic output in human and environmental samples

Next, we examined if gapfilling of the GEMs alters the conclusions from the human fecal and the environmental samples. Gapfilling increased the total number of predicted metabolites and the number of unique metabolites across both the reference-based and MAG-based approaches in both donors (Tables S4, S11 .and S12). When comparing the fold increase following gapfilling between the reference-based approach (0.5% cutoff) and the MAG-based approach (normalized), gapfilling had a similar effect on the reference-based and MAG-based approaches (donor 1: 1.05-fold vs 1.03-fold; donor 2: 1.04-fold vs 1.04-fold) and number of unique (donor 1: 1.04-fold vs 1.02-fold; donor 2: 1.02-fold vs 1.02-fold) predicted metabolites. The total number of unique predicted metabolites remained higher in the reference-based approach than the MAG-based

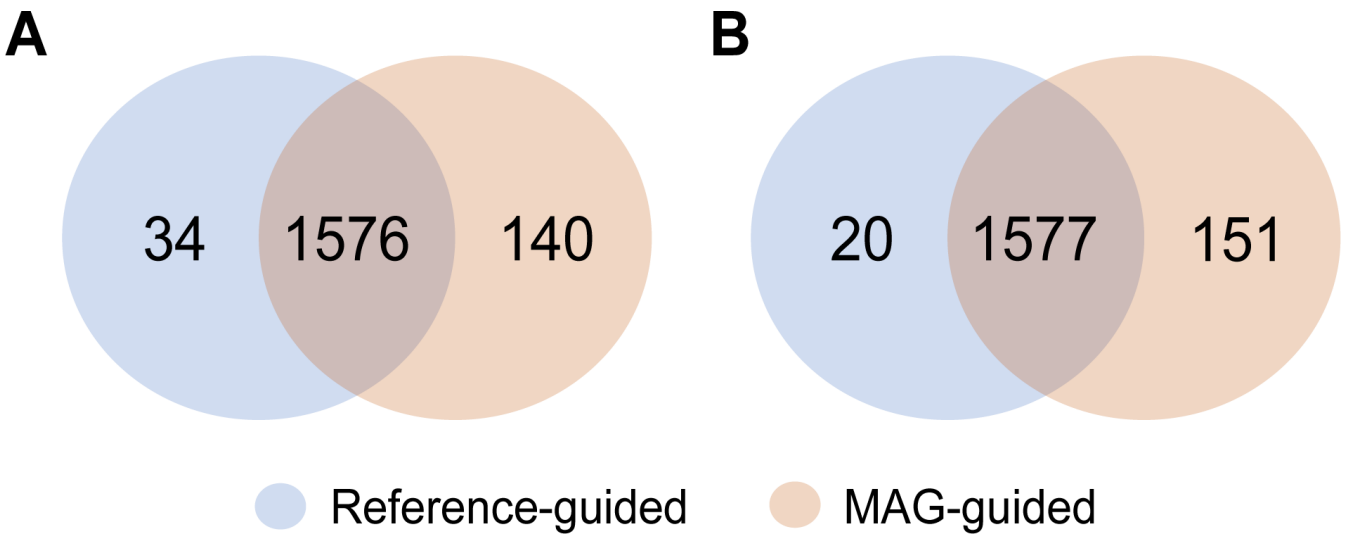

FIG 3 Number of unique predicted metabolic compounds across the reference-guided (0.5% cutoff) and MAG-guided (high- and medium-quality) inputs for environmental samples. (A) Sediment. (B) Seawater.

approach in both donors (donor 1: 1,558 vs 1,536 metabolites; donor 2: 1,599 vs 1,574 metabolites; Table S12).

In the environmental samples where the reference-based approach predicted more metabolites than the MAG-based approach prior to gapfilling (when numbers of input taxa were normalized), the conclusions did not change following gapfilling, with an increase in predicted metabolites across all analyses (Table S12). Gapfilling had a similar effect (by fold-change of metabolites) on the reference-based models (0.5% cutoff) as the MAG-based models (normalized) for total number of predicted metabolites (sediment: 1.07-fold vs 1.05-fold; seawater: 1.02-fold vs 1.05-fold) and unique metabolites (sediment: 1.03-fold vs 1.03-fold; seawater: 1.02-fold vs 1.03-fold) (Table S12).

DISCUSSION

Elucidating the metabolites produced by microbial communities is key to understanding their impact on their ecological niches and understanding their overall contribution to host metabolism and environmental processes (21–23). Although untargeted profiling of the metabolome allows for the detection of thousands of metabolites from a given sample, output is dependent on extraction protocols and the type of detection technique used, and can include metabolites not produced by the community (e.g., host). An alternative practice has been to generate draft reconstructions and model metabolic outputs by the microbial communities from metagenomic sequencing data. This is commonly performed by using MAGs binned from the data as input. However, this strategy remains computationally intensive, and it requires deep shotgun sequencing for efficient MAG binning. There has been an increase in the number of genomes for bacterial species that are available in public databases, and this provides an opportunity to leverage these genomes for construction of GEMs. Here, we systematically compared drafting GEMs using reference-guided and MAG-guided approaches for both human and environmental samples and tested our findings using untargeted metabolomics. We showed that while the reference-guided approach initially predicted less total number of metabolites than the MAG-guided approach for the human samples, it predicted more than the MAG-guided approach when the number of input taxa was normalized. The validation findings suggested that both approaches had a similar level of error. The two approaches appeared to be complementary, but despite this substantial overlap, each predicted the production of unique metabolites. Our data also suggested that biological inferences can change based on the approach used, and thus, there may be utility in

integrating both approaches or validating one approach with the other. In environmental samples, similar results were observed where the MAG-guided approach predicted production of more metabolites which was due to the higher input of genome bins into this approach relative to the reference-guided approach. Despite this, the contribution of metabolic uniqueness of the input taxa cannot be discounted as an influencer of non-redundant metabolite predictions in a community GEM given the predictions of unique metabolites in sediment samples. In this sample type, the two approaches were also found to be complementary showing substantial overlap but each identified unique metabolites.

There are several strengths and limitations to the reference-guided approach that differentiate it from the MAG-guided approach. One strength includes the lack of requirement for MAG binning and refinement which is a computationally intensive process and one that remains the subject of research, specifically into accuracy of the assembled and refined genome bins (15). Through the use of genomes from pure isolates (i.e., no contamination during assembly), it can be speculated that this approach would provide more precise tracking of metabolite origin to specific taxa. The lack of need for MAGs also allows for the construction of GEMs from shallow shotgun sequencing data sets which are becoming more popular and more readily available as sequencing costs decrease. In contrast, one key limitation of the reference-guided approach was seen in the environmental samples where the reliance on reference genomes that were not readily available posed a problem. With the increase in publicly available genomes and MAGs from different species, this limitation should potentially become less of an issue; however, it is important to note that we could not obtain neither a reference genome nor a high- or medium-quality MAG for taxa that were detected at >5% relative abundance in sediment and seawater, indicating that additional manual curation of environmental MAGs is required due to known difficulties in assembly of some highly abundant organisms (24, 25). Another limitation of the reference-guided approach worth mentioning is the lack of specificity of its input to the strains within the sequenced samples, and thus, bacterial strain variations would be unaccounted for, leading to possible false predictions. However, integration of our GEM predictions with the untargeted metabolomics showed a similar percentage of validated metabolites relative to total predicted in the reference-guided approach. Lastly, when we attempted to lower the relative abundance cutoff to 0.1%, several taxa were found to be unclassified at the species level (data not shown), which would restrict their inclusion into the predictions. A key limitation of both approaches worth highlighting is the limited predictive potential due to lack of curation against experimental data; however, it is plausible to assume that curation would be more readily possible for reference genome-based draft reconstructions.

In conclusion, our work compares alternative approaches to GEM construction from metagenomes, showing utility in genomic input guided by reference-based taxonomy, which can complement current MAG-guided methods for deep shotgun metagenomics data. We demonstrate that the choice of approach alters biological inferences, emphasizing the need for caution when relying on model predictions. Notably, this reference-guided approach could also be applied to shallow sequencing data where MAGs cannot be generated.

MATERIALS AND METHODS

Human stool samples and generation of MAGs

The collection of stool samples from healthy individuals was part of a study on fecal microbiota transplantation in the treatment of ulcerative colitis (20), of which the healthy individuals were donors of the study (donors 1 and 2).

A total of 30 fecal samples were collected from two healthy individuals over a period of 44 and 70 weeks (Fig. 1B). DNA was extracted from the lyophilized material using

the QIAamp PowerFecal DNA Kit (Qiagen, Chadstone, Victoria, Australia) and sequenced on the Illumina NovaSeq 6000 (S4 2 × 150 bp) using the Illumina DNA prep kits (Illumina, Melbourne, Victoria, Australia) as previously reported (26). Raw sequencing reads belonging to the same sample were concatenated into a single set of forward and reverse fastq files and were then quality trimmed using the Read_qc module to trim adaptors and remove human contamination. Reads were assembled using MetaBAT v.2 (27), MaxBin v.2 (28) and CONCOCT within MetaWRAP v.1.3.2 with default parameters. MAGs were generated with default parameters and subsequently refined using MetaWRAP v.1.3.2 (29). The completeness and contamination of MAGs were determined with CheckM (30). MAGs were defined as "high-quality" when they were >90% complete with less than 5% contamination, or "medium-quality" when they had a completeness of ≥50% and less than 10% contamination (31). MAGs that were "high-quality" or "medium-quality" were then dereplicated at 99% average nucleotide identity using dRep v.2.3.2 to remove duplicate MAGs (32). Taxonomy was assigned to each MAG based on the Genome Taxonomy Database (GTDB) r207 (33) with GTDB-Tk v.1.5.1 (34).

Environmental samples and generation of MAGs

Sediment (n = 10) and seawater (n = 10) samples were taken from coastal locations in Sydney and Perth, Australia (Fig. 1C). DNA was extracted from seawater and sediment samples as per standard operating procedures (https://github.com/AusMicrobiome/scientific_manual) and sequenced on a NovaSeq 6000 sequencer (2 × 150 bp run) using the Illumina DNA prep kit at the Ramaciotti Centre for Genomics (UNSW Sydney, Australia). Raw sequencing reads belonging to the same sample were concatenated into a single set of forward and reverse fastq files. Reads were then quality trimmed with Trimmomatic v.0.38 (35) with parameters specified as "HEADCROP:10 SLIDINGWINDOW:4:30." Seawater samples were assembled with metaSPAdes v.3.15.0 (36) with k-mer options 21,41,61,81,101,121,127, while sediment samples were assembled using MEGAHIT v.1.2.2b (37) with k-mer options specified as –k-min 21 –k-max 141 –k-step 20. Scaffolds smaller than 2,000 bp and 2,500 bp were removed for sediment and seawater samples, respectively. The coverage of the remaining scaffolds was determined by mapping of the quality-filtered reads using bowtie v.2.3.5.1 (38). After converting and sorting the mapping format using samtools (39), the coverage was determined with the jgi_summarize_bam_contig_depths script (27). MAGs were generated using MetaBAT v.2.12.1 (27) and MaxBin v.2.2.3 (28) with default parameters, and subsequently refined using MetaWRAP v.1.3 (29). The completeness and contamination of MAGs were determined with CheckM, and MAGs were defined as "high-quality" or "medium-quality" as above. MAGs that were "high-quality" or "medium-quality" were then dereplicated and taxonomically assigned as above.

Generation of species-level count tables for reference-guided approach

Quality-filtered reads from above were analyzed using MetaPhlAn4 to profile the composition of the microbial communities and generate tables with relative abundances of microbial species (40). The mean relative abundances of the microbial species within each sample group (i.e., donor 1, donor 2, seawater, and sediment) were calculated. Microbial species were grouped according to the cutoffs of ≥5%, ≥2.5%, ≥1%, and ≥0.5% mean relative abundance for this approach.

Construction of the genome-scale metabolic models

Genomes/MAGs were annotated in KBase with default parameters using the Annotate Genome/Assembly with RASTtk v.1.073 App prior to building the draft metabolic models for each organism. This included a similarity e-value cutoff of 1e−06. GEMs were built using MS2–Build Prokaryotic Metabolic Models with OMEGGA App implemented in KBase based on the ModelSEED Pipeline for individual genomes or MAGs. Models for genomes and MAGs classified as *Synechococcus* could not be built with the MS2–Build

Prokaryotic Metabolic Models with OMEGGA App and therefore were built using the Build Metabolic Model App (41, 42). Based on this method, biomass components of species are included within reactions if their genome contains the appropriate subsystems and annotations (42). All reactions associated with enzymes encoded in the annotated genome are included in the models, with spontaneous reactions also added (42). Draft models from individual organisms were then merged into a joint model of a community of multiple organisms using the app "Merge two or more metabolic models into a compartmentalized community model." In this joint multi-species model, compounds and reactions in each species are placed in uniquely labeled compartments, with compounds transported out of any member placed into a shared extracellular environment that can be accessed by any member possessing a transport reaction that can import an available extracellular compound (41, 42).

For the MAG-guided models, high-quality MAGs and high- and medium-quality MAGs were used to construct the GEM. For the reference-guided models, the relative abundance of bacteria was calculated, and input taxa were selected based on different cutoffs (i.e., ≥5%, ≥2.5%, ≥1%, and ≥0.5%). Genomic information from the selected bacteria were then employed to construct the GEMs. Representative genomes for the reference-guided approach were downloaded from the KBase public database and, if not present there, were downloaded directly from NCBI. A total of 15, 27, 24, and 20 taxa had only one genome available for donor 1, donor 2, sediment, and seawater samples, respectively. Where more genomes were available, the genome with the highest number of features was chosen. This strategy was implemented to maximize metabolic output from the selected reference genomes. However, comparison of predictions from genomes with the highest number of features with those with the lowest number of features in the same bacterial species showed that this increased the total number of predictions in 15 of the 19 species in human samples where multiple genomes were available (Table S13). In contrast, four outputs decreased (Table S13). For environmental samples, this strategy resulted in increased, similar, or decreased predictions for three, one, and three species where multiple genomes were available (Table S13). In certain cases where a public genome was not available (some taxa classified by MetaPhlAn4 to species-level genome bins), the corresponding MAGs present within the data were uploaded and merged into the community model. Not all species that were selected for the reference-guided approach in the environmental samples could be included as neither reference genomes nor MAGs were publicly available, and this was attributed to difficulties in assembly of highly abundant organisms.

To report the outputs of the model, values for "total predicted metabolites," "total extracellular metabolites," and "unique metabolites" were provided. Total predicted metabolites refer to all metabolites present in a system (all compartments, duplicates not removed). Total extracellular metabolites refer to metabolites present in a system classified as extracellular (duplicates not removed). Unique metabolites refer to the total predicted metabolites (all compartments) filtered for redundancy, in an effort to reflect metabolic diversity.

To assess its effect on metabolite predictions, all individual models for the reference-guided and MAG-guided approaches were gapfilled using the app MS2–"Improved Gapfill Metabolic Models with OMEGGA" and then merged into a compartmentalized community model. Flux balance analysis (FBA) was applied on gapfilled metabolic models for both the reference-guided and MAG-guided approaches using the app Run Flux Balance Analysis. The Run FBA method uses flux variability analysis (43) to classify if the reactions in the models are unable to carry flux (i.e., blocked).

Untargeted metabolomics from human stool samples

Untargeted metabolomics analysis on stool samples from the healthy individuals was performed using Metabolon's Precision Metabolomics liquid chromatography-mass spectrometry global metabolomics platform and was previously reported (26). These were included to validate the output inferences from the GEMs. The samples analyzed

were collected from the individuals on weeks 24, 26, 34, and 40 for donor 1 and weeks 20, 22, 31, 33, 41, 42, and 70 for donor 2 (Fig. 1B), and hence, the samples overlap longitudinally and extensively with the metagenomics data, overcoming biological variability associated with cross-sectional sampling. Raw metabolite abundance data prior to imputation was converted to presence/absence of metabolite per donor. Experimentally validated metabolites were reported as "confirmed unique metabolites."

For donor 1, a total of 1,348 metabolites were identified by untargeted metabolomics, of which 489 had KEGG IDs. An additional 34 KEGG IDs corresponded to the same metabolites (replicate IDs) and were included in the searches, making 523 the total number of KEGG IDs for donor 1. For donor 2, a total of 1,190 metabolites were identified by untargeted metabolomics, of which 453 had KEGG IDs. An additional 35 KEGG IDs corresponded to the same metabolites (replicate IDs) and were included in the searches, making 488 the total number of KEGG IDs for donor 1. It is important to note that fecal metabolomics is not restricted to bacterial metabolites and will include human metabolites and those from other members of the microbiota.

Pathway enrichment analysis of predicted metabolites

Following GEM construction, metabolite lists were imported into MBROLE 2.0 to perform pathway enrichment analyses (44) using pathways from the Small Molecule Pathway Database (human), KEGG, and UniPathway. MBROLE annotates the metabolites in the test and background lists with their respective pathways within the selected databases, then performs an over-representation analysis using cumulative hypergeometric distribution, after which the *P*-values are corrected for false discovery rate using the Benjamini and Hochberg method (45).

ACKNOWLEDGMENTS

The authors would like to acknowledge donors who took part in the clinical trial and the hospital staff that assisted with procedures. Figure 1 was generated using BioRender.

This study was supported by the Crohn's & Colitis Foundation of America (Litwin Award; 988415) and National Health and Medical Research Council of Australia (Ideas grant 2011047). The collection, DNA extraction, and sequencing of the environmental samples were supported by grants from BioPlatforms Australia (BPA) and the Integrated Marine Observing System (IMOS). S.P. is supported by an NHMRC Investigator Grant. N.O.K. is supported by a UNSW Scientia fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conceptualization: M.E.M., N.O.K. Data curation: M.E.M., L.D.W.L., N.O.K. Formal analysis: M.E.M., N.O.K., L.D.W.L. Funding acquisition: N.O.K., T.T. Investigation: M.E.M., N.O.K., T.T., L.D.W.L., C.H., S.P., T.J.B., R.W.L. Supervision: N.O.K., T.T., R.W.L., S.P. Visualization: M.E.M., N.O.K. Writing – original draft: M.E.M., N.O.K. Writing – review & editing: T.T., L.D.W.L., C.H., S.P., T.J.B., R.W.L.

AUTHOR AFFILIATIONS

[1]School of Biomedical Sciences, Faculty of Medicine and Health, UNSW Sydney, Sydney, New South Wales, Australia
[2]School of Clinical Medicine, Faculty of Medicine and Health, UNSW Sydney, Sydney, New South Wales, Australia
[3]Department of Gastroenterology, St. Vincent's Hospital, Sydney, New South Wales, Australia
[4]Concord Clinical School, University of Sydney, Sydney, New South Wales, Australia
[5]Department of Gastroenterology, Concord Repatriation General Hospital, Sydney, New South Wales, Australia
[6]Centre for Digestive Diseases, Sydney, New South Wales, Australia

[7]Centre for Marine Science and Innovation, School of Biological, Earth and Environmental Sciences, Faculty of Science, UNSW Sydney, Sydney, New South Wales, Australia

AUTHOR ORCIDs

Marwan E. Majzoub http://orcid.org/0000-0001-6461-8211
Craig Haifer http://orcid.org/0000-0002-3675-6550
Thomas J. Borody http://orcid.org/0000-0002-0519-4698
Nadeem O. Kaakoush http://orcid.org/0000-0003-4017-1077

FUNDING

Funder	Grant(s)	Author(s)
Crohn's and Colitis Foundation (CCF)	988415	Nadeem O. Kaakoush
DHAC | National Health and Medical Research Council (NHMRC)	2011047	Nadeem O. Kaakoush
Bioplatforms Australia (Bioplatforms)		Torsten Thomas
Integrated Marine Observing System (IMOS)		Torsten Thomas
DHAC | National Health and Medical Research Council (NHMRC)	Emerging leader grant	Sudarshan Paramsothy
University of New South Wales (UNSW)	Scientia fellowship	Nadeem O. Kaakoush

AUTHOR CONTRIBUTIONS

Marwan E. Majzoub, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft | Laurence D. W. Luu, Data curation, Formal analysis, Investigation, Writing – review and editing | Craig Haifer, Investigation, Writing – review and editing | Sudarshan Paramsothy, Investigation, Supervision, Writing – review and editing | Thomas J. Borody, Investigation, Writing – review and editing | Rupert W. Leong, Investigation, Supervision, Writing – review and editing | Torsten Thomas, Funding acquisition, Investigation, Supervision, Writing – review and editing | Nadeem O. Kaakoush, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Supervision, Visualization, Writing – original draft

DATA AVAILABILITY

The human raw metagenomics reads are available from the European Nucleotide Archive (ENA) under the accession PRJEB50699. The environmental raw metagenomic reads are available through the BioPlatforms Australia data portal (https://data.bioplatforms.com/). Table S14 includes the list of individual accession links. The KBase narratives can be accessed through the html links found in Table S15.

ETHICS APPROVAL

The study was approved by St. Vincent's Hospital Sydney's Human Research Ethics Committee (HREC/18/SVH/219). Written informed consent was obtained from all participants before participating in the study.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental Figures (mSystems00746-24-s0001.docx). Figures S1 to S3.
Supplemental Tables (mSystems00746-24-s0002.xlsx). Tables S1 to S15.

REFERENCES

1. Zhang Z, Tang H, Chen P, Xie H, Tao Y. 2019. Demystifying the manipulation of host immunity, metabolism, and extraintestinal tumors by the gut microbiome. Signal Transduct Target Ther 4:41. https://doi.org/10.1038/s41392-019-0074-5
2. Blumberg R, Powrie F. 2012. Microbiota, disease, and back to health: a metastable journey. Sci Transl Med 4:137rv7. https://doi.org/10.1126/scitranslmed.3004184
3. Levy M, Thaiss CA, Elinav E. 2016. Metabolites: messengers between the microbiota and the immune system. Genes Dev 30:1589–1597. https://doi.org/10.1101/gad.284091.116
4. Rauf A, Khalil AA, Rahman U-U-, Khalid A, Naz S, Shariati MA, Rebezov M, Urtecho EZ, de Albuquerque RDDG, Anwar S, Alamri A, Saini RK, Rengasamy KRR. 2022. Recent advances in the therapeutic application of short-chain fatty acids (SCFAs): an updated review. Crit Rev Food Sci Nutr 62:6034–6054. https://doi.org/10.1080/10408398.2021.1895064
5. Sun M, Wu W, Liu Z, Cong Y. 2017. Microbiota metabolite short chain fatty acids, GPCR, and inflammatory bowel diseases. J Gastroenterol 52:1–8. https://doi.org/10.1007/s00535-016-1242-9
6. Wahlström A, Sayin SI, Marschall HU, Bäckhed F. 2016. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. Cell Metab 24:41–50. https://doi.org/10.1016/j.cmet.2016.05.005
7. Li X, Zhang B, Hu Y, Zhao Y. 2021. New insights into gut-bacteria-derived indole and its derivatives in intestinal and liver diseases. Front Pharmacol 12:769501. https://doi.org/10.3389/fphar.2021.769501
8. Zhang B, Jiang M, Zhao J, Song Y, Du W, Shi J. 2022. The mechanism underlying the influence of indole-3-propionic acid: a relevance to metabolic disorders. Front Endocrinol (Lausanne) 13:841703. https://doi.org/10.3389/fendo.2022.841703
9. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621
10. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. Science 359:320–325. https://doi.org/10.1126/science.aap9516
11. Kitano H. 2002. Computational systems biology. Nature 420:206–210. https://doi.org/10.1038/nature01254
12. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jäger C, Baginska J, Wilmes P, Fleming RMT, Thiele I. 2017. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol 35:81–89. https://doi.org/10.1038/nbt.3703
13. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. 2019. Current status and applications of genome-scale metabolic models. Genome Biol 20:121. https://doi.org/10.1186/s13059-019-1730-3
14. Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, de Wouters T, Juste C, Rizkalla S, Chilloux J, Hoyles L, Nicholson JK, Dore J, Dumas ME, Clement K, Bäckhed F, Nielsen J, MICRO-Obes Consortium. 2015. Quantifying diet-induced metabolic changes of the human gut microbiome. Cell Metab 22:320–331. https://doi.org/10.1016/j.cmet.2015.07.001
15. Norsigian CJ, Fang X, Seif Y, Monk JM, Palsson BO. 2020. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. Nat Protoc 15:1–14. https://doi.org/10.1038/s41596-019-0254-3
16. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BO. 2001. Metabolic modeling of microbial strains *in silico*. Trends Biochem Sci 26:179–186. https://doi.org/10.1016/s0968-0004(00)01754-0
17. Bordbar A, Monk JM, King ZA, Palsson BO. 2014. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 15:107–120. https://doi.org/10.1038/nrg3643
18. Li P, Roos S, Luo H, Ji B, Nielsen J. 2023. Metabolic engineering of human gut microbiome: recent developments and future perspectives. Metab Eng 79:1–13. https://doi.org/10.1016/j.ymben.2023.06.006
19. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176:649–662. https://doi.org/10.1016/j.cell.2019.01.001
20. Haifer C, Paramsothy S, Kaakoush NO, Saikal A, Ghaly S, Yang T, Luu LDW, Borody TJ, Leong RW. 2022. Lyophilised oral faecal microbiota transplantation for ulcerative colitis (LOTUS): a randomised, double-blind, placebo-controlled trial. Lancet Gastroenterol Hepatol 7:141–151. https://doi.org/10.1016/S2468-1253(21)00400-3
21. Jia W, Li H, Zhao L, Nicholson JK. 2008. Gut microbiota: a potential new territory for drug targeting. Nat Rev Drug Discov 7:123–129. https://doi.org/10.1038/nrd2505
22. Patel S, Ahmed S. 2015. Emerging field of metabolomics: big promise for cancer biomarker identification and drug discovery. J Pharm Biomed Anal 107:63–74. https://doi.org/10.1016/j.jpba.2014.12.020
23. Wishart DS. 2016. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov 15:473–484. https://doi.org/10.1038/nrd.2016.32
24. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. 2021. The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. Appl Environ Microbiol 87:e02593-20. https://doi.org/10.1128/AEM.02593-20
25. Ramos-Barbero MD, Martin-Cuadrado A-B, Viver T, Santos F, Martinez-Garcia M, Antón J. 2019. Recovering microbial genomes from metagenomes in hypersaline environments: the Good, the Bad and the Ugly. Syst Appl Microbiol 42:30–40. https://doi.org/10.1016/j.syapm.2018.11.001
26. Haifer C, Luu LDW, Paramsothy S, Borody TJ, Leong RW, Kaakoush NO. 2022. Microbial determinants of effective donors in faecal microbiota transplantation for UC. Gut 72:90–100. https://doi.org/10.1136/gutjnl-2022-327742
27. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165. https://doi.org/10.7717/peerj.1165
28. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:26. https://doi.org/10.1186/2049-2618-2-26
29. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6:158. https://doi.org/10.1186/s40168-018-0541-1
30. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114
31. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893
32. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126
33. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229
34. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848
35. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
36. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116
37. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via

succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033
38. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
40. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol 41:1633–1644. https://doi.org/10.1038/s41587-023-01688-w
41. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, et al. 2018. KBase: the United States department of energy systems biology knowledgebase. Nat Biotechnol 36:566–569. https://doi.org/10.1038/nbt.4163
42. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28:977–982. https://doi.org/10.1038/nbt.1672
43. Mahadevan R, Schilling CH. 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276. https://doi.org/10.1016/j.ymben.2003.09.002
44. López-Ibáñez J, Pazos F, Chagoyen M. 2016. MBROLE 2.0-functional enrichment of chemical compounds. Nucleic Acids Res 44:W201–W204. https://doi.org/10.1093/nar/gkw253
45. Chagoyen M, Pazos F. 2011. MBRole: enrichment analysis of metabolomic data. Bioinformatics 27:730–731. https://doi.org/10.1093/bioinformatics/btr001