LETTERS
https://doi.org/10.1038/s41564-018-0124-8

nature microbiology

Towards predicting the environmental metabolome from metagenomics with a mechanistic model

Daniel R. Garza[1], Marcel C. van Verk[2,3], Martijn A. Huynen[1] and Bas E. Dutilh[1,2]*

The environmental metabolome and metabolic potential of microorganisms are dominant and essential factors shaping microbial community composition. Recent advances in genome annotation and systems biology now allow us to semiautomatically reconstruct genome-scale metabolic models (GSMMs) of microorganisms based on their genome sequence[1]. Next, growth of these models in a defined metabolic environment can be predicted in silico, mechanistically linking the metabolic fluxes of individual microbial populations to the community dynamics. A major advantage of GSMMs is that no training data is needed, besides information about the metabolic capacity of individual genes (genome annotation) and knowledge of the available environmental metabolites that allow the microorganism to grow. However, the composition of the environment is often not fully determined and remains difficult to measure[2]. We hypothesized that the relative abundance of different bacterial species, as measured by metagenomics, can be combined with GSMMs of individual bacteria to reveal the metabolic status of a given biome. Using a newly developed algorithm involving over 1,500 GSMMs of human-associated bacteria, we inferred distinct metabolomes for four human body sites that are consistent with experimental data. Together, we link the metagenome to the metabolome in a mechanistic framework towards predictive microbiome modelling.

Microbial communities constantly adapt to exploit available resources[3]. As a result, the presence of specific microorganisms or distributions of microorganisms allow us to infer environmental features. For example, the altered metabolic conditions in the microenvironment of colorectal cancer tumours select for the outgrowth of specific species in the human colorectal cancer microbiome[4], allowing cancer detection[5]. Similarly, microorganisms can serve as biosensors for geochemical features such as solvent or uranium contamination[6]. These and many other empirical examples of significant associations between the environment and the microbiota[7,8] suggest that the composition and metabolic potential of microbial communities can be used to reconstruct the metabolic environment of a biome through a reverse engineering strategy.

Shotgun metagenomics rapidly inventories the composition and genomic content of microbiomes, but obtaining comparably high-resolution metabolomic measurements of microbial environments and linking them to microorganisms remains challenging[2]. It is often even more challenging to translate the presence/absence of genes or the taxonomic composition of a given environment into predictions about the metabolic status of the community[9]. Although several tools exist that allow researchers to translate shotgun metagenomes into functions or functional profiles of a given environment[10–13], current methods often require extensive training datasets and provide knowledge at the community-level, without mechanistically attributing differences to specific microorganisms or metabolites.

Capitalizing on advances in the functional annotation of genes, GSMMs attempt to mechanistically explain bacterial growth and metabolism in a defined environment without requiring training data[1]. GSMMs provide a minimally biased description of the metabolic processes that are encoded in a microbial genome by integrating database knowledge about protein functions into a reaction network that describes the metabolic potential of a genome. Constraint-based approaches such as flux balance analysis (FBA) allow the growth or biomass production of GSMMs to be formulated as an optimization problem and therewith estimated[14]. However, integrating multiple GSMMs into a microbial community model, and producing meaningful predictions about the metabolic status of the environment, is still an open research challenge[9]. Here, we addressed this challenge by developing an optimization framework to predict the metabolic environment that best explains the observed species abundance distribution profiles. The algorithm takes metagenomic species abundances and their GSMMs as input, and does not require any further training data.

The main premise of our approach, named MAMBO for Metabolomic Analysis of Metagenomes using fBa and Optimization, is that the abundance distribution of microbial genomes and their encoded metabolic potential reveal how microorganisms can exploit the metabolic resources that are available in the environment. We qualitatively infer these resources by searching for the metabolic environment that yields microbial growth that is best correlated with the relative abundances observed in the metagenome profiles. This computational approach is outlined in Fig. 1. First, we used reference genome sequences to generate GSMMs of the species encountered in a given metagenomic dataset. Also from the metagenomics, we extracted the relative abundances of these organisms in the sample. We can now ask the question, which metabolic environment would lead to relative growth rates of the GSMMs that best correlate with the observed relative abundances of the organisms? In this context, growth is defined as the fluxes in the biomass reactions of the GSMMs. We constrain the GSMMs of all the microorganisms found in a sample by providing them with the same metabolic environment (modelled as a limit to the import reactions of metabolites) and assuming the same cellular objective of growth.

[1]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands. [2]Theoretical Biology and Bioinformatics, Sience4Life, Utrecht University, Utrecht, The Netherlands. [3]Plant–Microbe Interactions, Science4Life, Utrecht University, Utrecht, The Netherlands. *e-mail: bedutilh@gmail.com

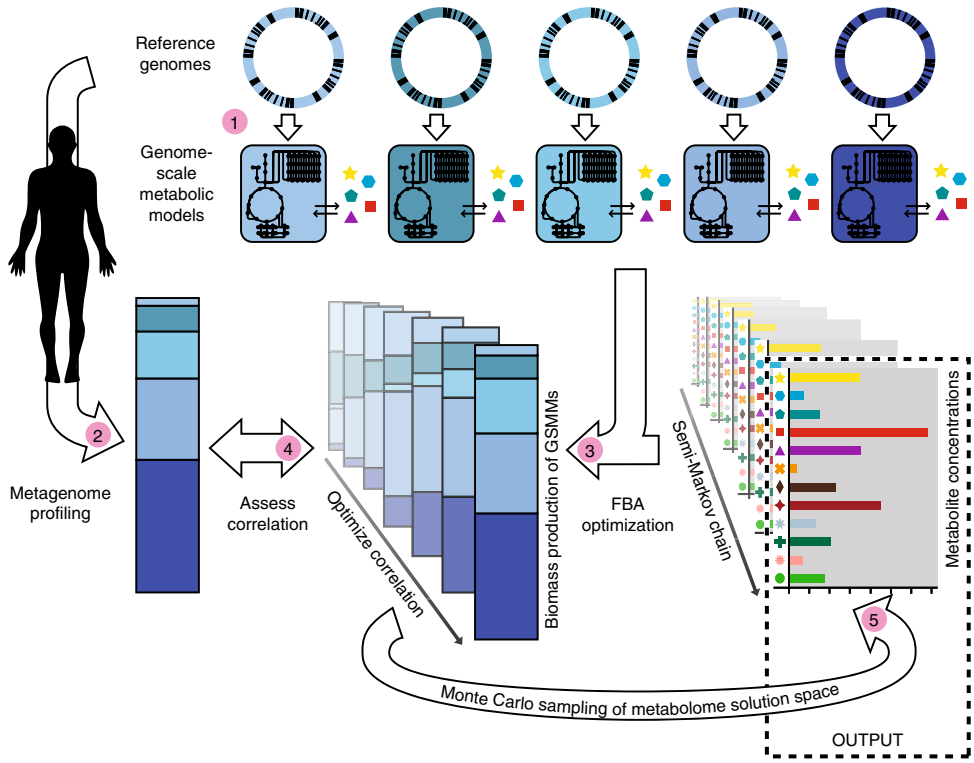

Fig. 1 | Overview of the MAMBO algorithm. Reference microbial genomes are used to reconstruct GSMMs (1); community abundance profiles are obtained through reference mapping (2); and biomass production of the GSMMs, obtained through FBA (3), is correlated with the metagenomic community abundance profile (4). This correlation is optimized by multiple iterations of a Monte Carlo-based sampling algorithm (5) (see main text and Methods section).

Finally, we use a semi-Markov chain to sample the highly dimensional metabolome space, optimizing for the metabolomic composition that leads to an optimal correlation of the GSMMs growth profile with the microbiome metagenomic abundance profile. Thus, an optimization run predicts the relative metabolite abundances in an environment, as shown for a typical run in Fig. 2. Importantly, our approach mechanistically links the environmental metabolome to the metabolic fluxes in genome-scale reconstructions of the metabolism of each individual microorganism, because the optimization is performed by FBA solutions on a per-genome basis.

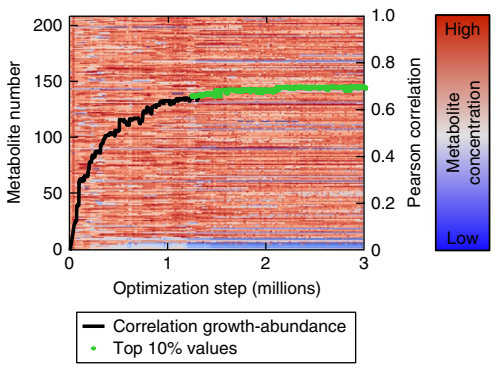

Fig. 2 | A typical optimization of a randomly chosen metagenome (buccal mucosa sample SRS058186). There are 209 metabolites in the combined GSMMs of the species in this sample (left *y* axis). The red/blue heatmap indicates the relative concentrations of these metabolites as they change throughout the optimization run. The black and green line indicates the Pearson correlation score between the GSMM biomass production rates and the metagenomic relative abundances (right *y* axis). The top 10% values are coloured green; metabolic profiles from these optimization steps are averaged, resulting in the final inferred values for each metabolite. This approach provides a robust representation of the target region in the multidimensional metabolome search space.

We tested our approach by inferring the metabolomic environment in four human body sites (oral, skin, stool and vaginal) using 175 metagenomic datasets[15] and GSMMs that we reconstructed for 1,562 detected bacteria[16]. The inferred metabolomes revealed four clusters corresponding to the four body site biomes (Fig. 3), as were previously observed for the body site-specific microbiomes[15]. This clustering was independent of the initialization of our algorithm in the high-dimensional metabolome search space. For example, searches based on oral metagenomes that were initiated with predicted skin metabolomes quickly converged to the oral metabolome cluster and the same pattern was observed for the other body sites (Supplementary Fig. 1a–d). Repeated metabolomes inferred from the same metagenome had an average Pearson correlation of 0.96 ± 0.02, showing high robustness and consistency of the algorithm.

To benchmark our algorithm on experimental data, we identified six annotated, quantified, high-throughput metabolomes from saliva, faeces and vagina[17–23] and correlated the metabolites measured in these studies with our predictions from 175 Human Microbiome Project (HMP) metagenomes. Figure 4 shows that the predicted and measured metabolomes for these body site biomes are consistent: metabolomes inferred from oral, stool and vaginal metagenomes correlated significantly better with those measured in saliva, faecal water and faecal incubator, and vagina, respectively, than with metabolomes from other body site biomes ($P = 3.5 \times 10^{-52}$, one-tailed unpaired *t*-test).

A recent screen of metabolites on human skin revealed the influence of skin care and hygiene products on the microbiome[24].

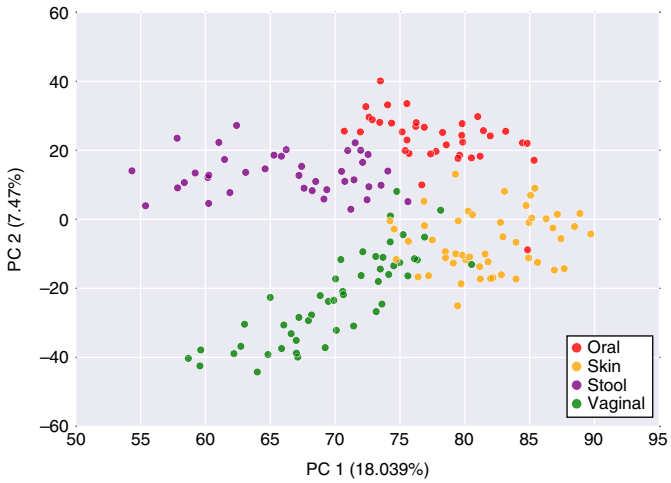

Fig. 3 | Principal component analysis of predicted metabolomes. Each dot represents one of the 175 Human Microbiome Project metagenomic samples, where the predicted relative concentrations of metabolites are plotted across the two principal components. Colours depict the body site from which the sample was obtained. PC, principal component.

Using MAMBO, we predicted metabolomes from 50 skin metagenomes, confirming the abundance of various cosmetic and hygiene ingredients (Supplementary Table 1). Moreover, we assessed the abundance of skin metabolites across twelve samples where both microbiomes and metabolomes had previously been measured[24], including triplicates from the hand and foot of a male and female volunteer. In these samples, differently from the metabolomic studies above, metabolites were measured by using an untargeted approach, which only allowed for a metabolite-by-metabolite comparison across samples. For the majority of metabolites the measured and predicted abundances correlated positively across these four skin sites ($P=0.0064$, one-tailed binomial test; Supplementary Fig. 2), showing that most metabolites can be distinguished between samples from the same biome.

We evaluated if the gene composition alone is sufficient to infer the environmental metabolome without the need to reconstruct GSMMs and use optimization. For this purpose, we adjusted an existing algorithm, named Predicted Relative Metabolic Turnover (PRMT) score[10,25], that predicts the metabolic turnover in one sample relative to another and applied it to the same reference mapping of 175 HMP samples as above[15,16], and evaluated how well this gene-based analysis predicted the relative abundances of metabolites. As shown in Supplementary Fig. 3, the resulting metabolite lists showed a lower correlation with the measured metabolomes than the GSMM-based MAMBO analysis. On closer inspection, this limited correlation of the gene-based predictions results from some metabolites being spuriously predicted at high abundance. For example, malate is consistently predicted to be abundant in stool because several stool bacteria contain multiple malate dehydrogenase genes. However, the malate concentration is low in faecal water[20] and faecal incubator[17] metabolomes, but high in the vaginal metabolome[18]. In contrast, by exploiting GSMMs rather than individual genes, MAMBO predicted that a high concentration of malate was not important to achieve the observed abundance profile of bacteria in most stool samples, while it was important for most vaginal samples, consistent with the experimentally measured metabolomes.

It should be noted that PRMT predicts the relative bacterial consumption or production of metabolites compared to an average, and is thus not directly aimed at predicting the net production or consumption of metabolites[10,25]. The main differences between the gene-based predictions and MAMBO are (1) PRMT does not offer an assessment of the metabolome on a sample-by-sample basis, but rather a comparison of a sample to an average metabolome in order to highlight the largest differences; (2) PRMT assumes that the relative abundance of a metabolite is proportional to the relative number of genes coding for an enzyme, an assumption we avoid by using FBA where enzyme fluxes are defined by optimizing for growth; and (3) the PRMT connectivity matrix is not compartmentalized by species, so PRMT does not require GSMMs.

The MAMBO metabolome prediction framework takes an important step towards predictive microbiome modelling. GSMMs have previously been applied to microbial communities[26,27] but those GSMMs depend on an explicit definition of the environment including the metabolites and their relative abundances, to allow the modelled microorganisms to grow. However, the environmental metabolome remains unclear for many measured microbiomes, either because the metabolomic experiments are lacking, or because they were measured in a slightly different system or at a different scale than observed by the microorganisms in the metagenome. Here, we bridge this gap by predicting the environmental metabolome directly from the metagenome. There is still some noise in these predictions; for example, several predicted metabolomes show high correlations to measured metabolomes from other biomes (Fig. 4). First, there are many factors that influence the microbial abundances besides the available metabolites, including bacteriophages[28] and human factors[29], to name a few. Taking these and other factors into account may improve the performance of a tool to infer the environmental metabolome, and further contribute to predictive microbiome modelling. From a technical perspective, the experimentally measured metabolomes used in our comparisons are derived from biofluids with a complex composition, and it remains challenging to link the mass-over-charge (m/z) peaks in

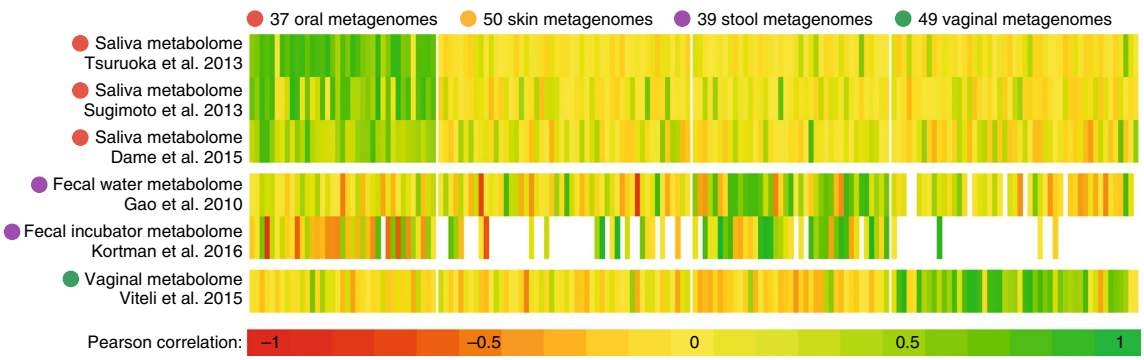

Fig. 4 | Pearson correlations between 175 metabolomes from four body sites predicted by MAMBO and six experimentally measured metabolomes from literature. Correlations are only shown if >5 metabolites match between the predicted and measured metabolomes. See Supplementary Table 3 for details.

mass spectrometry spectra to specific metabolites[30]. For example, we could only identify the metabolites in untargeted metabolomics experiments[24] by mapping the m/z values to published standards (see Methods). Structurally different metabolites often have identical chemical composition and m/z values, so identification of metabolites based on m/z alone is inherently inaccurate. Finally, the metabolomic and metagenomic datasets that we exploited were measured and published independently in samples that may differ in unknown ways; for example, a faecal incubator versus fresh stool. Nevertheless, we could predict metabolomes for different body site biomes that significantly correlate with the experimental data (P value $=3.5\times10^{-52}$, one-tailed unpaired t-test), and find positive correlations for most metabolites across paired samples from the same body site.

Without requiring training data, MAMBO implicitly exploits the fact that microorganisms in an ecosystem constantly compete for resources, leading to a relative abundance distribution that reflects their ability to exploit these resources. Metagenome-guided modelling enables a deeper understanding of microbiomes by linking the environmental metabolome to the metabolic network of individual microbial populations. By explicitly modelling the fluxes of individual GSMMs that are matched with the species composition of the system in a probabilistic fashion, our approach provides a starting point for mechanistic models of microbial ecology, including the potential for systems with more complex cross-feeding networks[9].

Methods

Datasets. From the US Department of Energy Systems Biology Knowledgebase (KBase, http://www.kbase.us) and the HMP (http://www.hmpdacc.org) we downloaded the human microbiome reference genomes[15], as well as taxon-abundance profiles for 175 metagenomes, respectively, including 37 oral, 50 skin, 39 stool and 49 vaginal metagenomes (listed in Supplementary Table 2). The abundance profiles were previously generated according to the HMP standard operating procedure[16] (http://www.hmpdacc.org/doc/ReadMapping_SOP.pdf), where 57.6% of the sequenced reads were aligned to the reference database across all HMP metagenomes (see section 4.5 of the SOP document). Additionally, we included 372 GSMMs from a recent study[26] of genomes that were not in the HMP list. Metabolites were matched to the SEED database using the conversion table provided by the authors, and the genome sequences were obtained from the NCBI nucleotide database.

Experimentally measured metabolomic profiles were obtained from the Human Metabolomics Database[19], including one from faecal water[17] and three from saliva[21–23]. One faecal incubator[17] and one vaginal[18] metabolome were obtained from recent literature.

We used data from a recent study of the metabolites on human skin[24] to compare predicted and measured metabolites across four different skin sites where both the microbiome and metabolome were measured. We obtained 16 S amplicon datasets and raw capillary gas chromatography mass spectrometry spectra for 12 skin samples, including triplicates from two body sites of two individuals[24]. We used the Burrows Wheeler Aligner[31] to map the 16 S reads to the genomes in our database (average 74.5% of reads mapped). We created a database containing all 214 metabolites exported by the GSMMs for which the retention time and mass-to-charge ratio (m/z) were annotated in the Human Metabolomics Database[19]. We then used these parameters to annotate the capillary gas chromatography mass spectrometry peaks from the skin study. We used MZ-Mine[32] to identify features and align and deconvolute the raw spectra, generating a normalized peak list consisting of retention time versus m/z. Thus, 21 peaks could be unambiguously mapped to GSMM metabolites across all twelve samples. Next, we compared metabolite abundances across, rather than within, samples, since the raw spectra of different metabolites in an untargeted metabolomics study are not comparable within a sample[33]. Thus, the area under the peaks was used as an indicator of metabolite abundance across samples, and used in our analysis (Supplementary Fig. 2).

Metabolic modelling. We used the ModelSEED pipeline[1] to generate GSMMs for the 1,562 HMP reference genomes that were present in at least one of the metagenomic datasets (available at https://github.com/danielriosgarza/MAMBO). Briefly, genomic annotations were used to identify the biochemical reactions in a species' metabolic network. The molecular stoichiometry of these reactions was expressed in a matrix that transforms reaction rates to the time-derivative of metabolite concentrations. The nullspace of this matrix contains the equilibria solutions for reaction rates. Parsimonious gap filling was applied by adding the minimal possible set of reactions to the model that are essential for a model to grow; that is, to yield a flux through the biomass reactions[1]. Gap-filled reactions were probably missed during sequencing, assembly or genome annotation. We excluded dead-end exchange reactions from the models that remained unresolved after gap filling or had no influence on the objective function. FBA simulations were performed in a Python 2.7 environment, using the COBRApy package for constraint-based modelling[34] and Gurobi 5.6.3 (http://www.gurobi.com) or GLPK 4.35 (http://www.gnu.org/software/glpk) as linear programming solvers. To reflect the constant competition between microorganisms, we used growth as the objective function in the FBA[14].

Metabolomic Analysis of Metagenomes using fBa and Optimization algorithm. For the MAMBO algorithm, explained in the main text and depicted in Fig. 1, we constrained the GSMMs of all the microorganisms found in a metagenome by providing the same metabolic environment (modelled as an upper bound to the import reactions of metabolites) and assuming the same cellular objective of growth. We used semi-Markov chain sampling embedded on a Metropolis-Hastings algorithm[35] to identify the metabolomic composition that optimally correlates with the abundance profile of the microbial genomes observed in the metagenome.

The input of the algorithm consists of (1) a list of microorganisms and their relative abundances and (2) a database of GSMMs generated from the genomes of these microorganisms. Thus, the approach depends on the availability of high-quality draft reference genome sequences, as are available for the microorganisms found in the human microbiome and increasingly also for other environments. Typically, one GSMM will have 35–80 exchange reactions representing the metabolic compounds that the organism can utilize. Depending on the complexity of the microbiome, the GSMMs of all the microorganisms in a community together will be able to utilize >200 different metabolites. These combined exchange reactions represent the metabolites whose environmental concentrations are inferred by MAMBO.

At the core of the approach is an optimization algorithm that searches the > 200-dimensional metabolome search space for a composition of the metabolomic environment that, when applied simultaneously to the GSMMs of all coexisting microorganisms using FBA, yields a relative biomass production profile b that correlates with the abundance profile m of the microorganisms in the metagenome. The metabolic compound concentrations are modelled in the FBA as an upper bound to the influx reaction. We use Monte Carlo optimization following a semi-Markov chain to search the highly dimensional solution space. After random initialization (or initialization with a decoy metabolome as in Supplementary Fig. 1a–d), a new candidate environment e' is generated from the current environment e, by slightly altering the concentration of one metabolite following a uniform distribution. The maximum biomass production rates of all GSMMs are then evaluated for the candidate environment, and the change is accepted if the Pearson correlation of the metagenomic abundances with the growth rates in the candidate environment, $\rho(m,b_{e'})$, is higher than for the current environment, $\rho(m,b_e)$, or with a uniform probability $\rho(m,b_{e'})/\rho(m,b_e)$ otherwise. Every 150 search steps, the algorithm evaluates the past outcomes and chooses the environment that yielded the highest correlation[36]. Samples were first subjected to 100,000 search steps, and 100,000 steps were subsequently added until a high Pearson correlation (ρ 0.6) with the target metagenomic abundance profile was achieved. Finally, the 10% time points with the highest Pearson correlation scores between the biomass profile and the metagenomic abundance profile were averaged, yielding a robust predicted metabolome (Fig. 2). Note that the correlation that is optimized using the semi-Markov chain is the correlation $\rho(m,b_e)$ between the metagenomic species profile m and the biomass production rates b_e, while the correlations that are shown in our results (for example, in Fig. 4 and Supplementary Table 3) are correlations between the predicted metabolome e and experimentally measured metabolic concentrations.

Comparison with individual gene-based metabolome prediction. For gene-based comparison we generated PRMT[10,25] scores for the same 175 HMP samples that were used to benchmark the MAMBO algorithm. For this purpose, we first derived a matrix containing the number of genes per genome coding for a given enzyme reaction. This matrix was used to transform the vector of relative bacterial abundances per environment into a vector of normalized enzyme counts, which expresses the relative importance of enzymes given a metagenomic species abundance profile. Second, we built a table mapping all the enzymes found in the previous step to metabolites, which was expressed as a large connectivity matrix with metabolites as rows and enzymes as columns, and was normalized by rows. This matrix was used to transform the vector of normalized enzyme counts into a vector of predicted scores per metabolites. The predicted scores were quantile-normalized and compared to the average scores across all samples to produce the sample-by-sample PRMT scores, which are expressed as fold changes of metabolite importance in a given sample relative to the average importance across all samples.

Statistics. Statistical analyses were performed on a Python 2.7 environment, using the "stat" statistical package of Scipy 0.15.1. Principal coordinate analyses were performed using the scikit-learn package.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The Cython/Python implementation of MAMBO nr. 1 is available at https://github.com/danielriosgarza/MAMBO.

Data availability. This study strongly depended on recycled data generated by others, as referenced in the appropriate sections above. Moreover, we generated 1,562 GSMMs of human-associated bacteria that can be obtained from https://github.com/danielriosgarza/MAMBO. MAMBO-predicted metabolomic profiles of 37 oral, 50 skin, 39 stool and 49 vaginal metagenomes are listed in Supplementary Table 2, as well as six experimentally measured metabolomic profiles. Correlations between the measured and predicted metabolomes are listed in Supplementary Table 3.

Received: 21 November 2017; Accepted: 8 February 2018;
Published online: 12 March 2018

References

1. Henry, C. S. et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–982 (2010).
2. Marcobal, A. et al. A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *ISME J.* **7**, 1933–1943 (2013).
3. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**, aac9323 (2015).
4. Tjalsma, H., Boleij, A., Marchesi, J. R. & Dutilh, B. E. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nat. Rev. Microbiol.* **10**, 575–582 (2012).
5. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
6. Smith, M. B. Natural bacterial communities serve as quantitative geochemical biosensors. *mBio* **6**, e00326-15 (2015).
7. Merrifield, C. A. et al. Neonatal environment exerts a sustained influence on the development of the intestinal microbiota and metabolic phenotype. *ISME J.* **10**, 145–157 (2016).
8. Adams, R. I., Bateman, A. C., Bik, H. M. & Meadow, J. F. Microbiota of the indoor environment: a meta-analysis. *Microbiome* **3**, 49 (2015).
9. Garza, D. R. & Dutilh, B. E. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cell. Mol. Life Sci.* **72**, 4287–4308 (2015).
10. Larsen, P. E. et al. Predicted Relative Metabolomic Turnover aPRMTa: determining metabolic turnover from a coastal marine metagenomic dataset. *Microb. Inform. Exp.* **1**, 4 (2011).
11. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl Acad. Sci. USA* **110**, 12804–12809 (2013).
12. Hanson, N. W. et al. Metabolic pathways for the whole community. *BMC Genom.* **15**, 619 (2014).
13. Silva, G. G. Z., Green, K. T., Dutilh, B. E. & Edwards, R. A. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* **32**, 354–361 (2016).
14. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
15. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
16. Martin, J. et al. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS ONE* **7**, e36427 (2012).
17. Kortman, G. A. M. et al. Microbial metabolism shifts towards an adverse profile with supplementary iron in the TIM-2 in vitro model of the human colon. *Microb. Physiol. Metab.* **6**, 1481 (2015).
18. Vitali, B. Vaginal microbiome and metabolome highlight specific signatures of bacterial vaginosis. *Eur. J. Clin. Microbiol.* **34**, 2367–2376 (2015).
19. Wishart, D. S. et al. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41**, D801–D807 (2013).
20. Gao, X., Pujos-Guillot, E. & Sébédio, J.-L. Development of a quantitative metabolomic approach to study clinical human fecal water metabolome based on trimethylsilylation derivatization and GC/MS analysis. *Anal. Chem.* **82**, 6447–6456 (2010).
21. Tsuruoka, M. et al. Capillary electrophoresis-mass spectrometry-based metabolome analysis of serum and saliva from neurodegenerative dementia patients. *Electrophoresis* **34**, 2865–2872 (2013).
22. Sugimoto, M. et al. Physiological and environmental parameters associated with mass spectrometry-based salivary metabolomic profiles. *Metabolomics* **9**, 454–463 (2013).
23. Dame, Z. T. et al. The human saliva metabolome. *Metabolomics* **11**, 1864–1883 (2015).
24. Bouslimani, A. et al. Molecular cartography of the human skin surface in 3D. *Proc. Natl Acad. Sci. USA* **112**, 2120–2129 (2015).
25. Larsen, P. E. & Dai, Y. Metabolome of human gut microbiome is predictive of host dysbiosis. *GigaScience* **4**, 42 (2015).
26. Magnúsdóttir, S. et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2017).
27. Reed, J. L. in *The Chemistry of Microbiomes* (National Academies of Sciences, Engineering and Medicine) Chap. 12 (National Academies Press, Washington, DC, 2017); https://doi.org/10.17226/24751.
28. Rodriguez-Valera, F. et al. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).
29. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
30. Wishart, D. S. Advances in metabolite identification. *Bioanalysis* **3**, 1769–1782 (2011).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **11**, 395 (2010).
33. Zhou, J. & Yin, Y. Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. *Analyst* **141**, 6362–6373 (2016).
34. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7**, 74 (2013).
35. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
36. Barbu, V. S. & Limnios, N. *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications* 1st edn, Vol 191 (Springer-Verlag, New York, 2008).

Acknowledgements

We thank M. Kooyman (SURFsara) for help implementing MAMBO on the Netherlands Life Science Grid, C.R. Berkers (Utrecht University) for insights regarding the annotation of untargeted metabolome datasets and the CMBI Comics Group for fruitful discussions. D.R.G. is supported by the Science Without Borders program of CNPQ/BRASIL. B.E.D. is supported by Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004.

Author contributions

D.R.G. created the algorithm and performed the experiments. All authors devised the study and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-018-0124-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.E.D.

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

nature research

Corresponding author(s): Bas E. Dutilh

☒ Initial submission ☐ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

For metagenome datasets, we used publicly available data from the Human Microbiome Project, which is to our knowledge the greatest sampling effort of human associated metagenomes with data available to the public. We tested the algorithm on 175 independent samples, selected randomly from this source, across four different body-sites (stated on the supplementary methods). The results also indicated that a sufficient sample size was chosen in order to separate the predicted metabolomes per body-site and to show the overall correlation with the few metabolome measurements that were publicly available for the body-sites that we tested. The approach is computationally intensive, which limited our choices of sample-size.

2. Data exclusions

Describe any data exclusions.

For the comparison of predicted metabolomes with experimentally determined concentration of metabolites, we excluded correlations between samples that had an overlap of 5 or less metabolites (This is stated on the relevant Figure legends).

3. Replication

Describe whether the experimental findings were reliably reproduced.

We reproduced the method for predicting microbial community metabolomes on four different body-sites and compared our predictions with different metabolome datasets that were available. Our main findings were reproduced on all four comparisons. Moreover, within each body site we present analyses of between 37-50 metagenomes, indicating high reproducibility of the method.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

The primary sources of data (metagenomes and metabolomes) that were used in our study were obtained from public published datasets from which we maintained the original labels.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

n/a

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The computer functions that are central for the Bottom-up ecology algorithm are publicly available at https://github.com/danielriosgarza/Bottom_Up_Ecology_Functions.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	n/a

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	n/a

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	n/a
b. Describe the method of cell line authentication used.	n/a
c. Report whether the cell lines were tested for mycoplasma contamination.	n/a
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	n/a

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	n/a

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	n/a