HAL open science

Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species

Arnaud Belcour, Clémence Frioux, Méziane Aite, Anthony Bretaudeau, Anne Siegel

To cite this version:

Arnaud Belcour, Clémence Frioux, Méziane Aite, Anthony Bretaudeau, Anne Siegel. Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species. 2019. hal-02395024v1

HAL Id: hal-02395024

https://inria.hal.science/hal-02395024v1

Preprint submitted on 5 Dec 2019 (v1), last revised 7 Jan 2021 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species

Arnaud Belcour*[1], Clémence Frioux*[1,2], Méziane Aite[1], Anthony Bretaudeau[1,3,4], Anne Siegel[1]

[1]Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France
[2]Quadram Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UA, UK
[3]INRA, UMR IGEPP, BioInformatics Platform for Agroecosystems Arthropods (BIPAA), 35000, Rennes, France
[4]Inria, IRISA, GenOuest Core Facility, 35000, Rennes, France

*: equal contributions

October 11, 2019

Abstract

Capturing the functional diversity of microbiotas entails identifying metabolic functions and species of interest within hundreds or thousands. Starting from genomes, a way to functionally analyse genetic information is to build metabolic networks. Yet, no method enables a functional screening of such a large number of metabolic networks nor the identification of critical species with respect to metabolic cooperation.

Metage2Metabo (M2M) addresses scalability issues raised by metagenomics datasets to identify keystone, essential and alternative symbionts in large microbiotas communities with respect to individual metabolism and collective metabolic complementarity. Genome-scale metabolic networks for the community can be either provided by the user or very effi-

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

ciently reconstructed from a large family of genomes thanks to a multi-processing solution to run the Pathway Tools software. The pipeline was applied to 1,520 genomes from the gut microbiota and 913 metagenome-assembled genomes of the rumen microbiota. Reconstruction of metabolic networks and subsequent metabolic analyses were performed in a reasonable time.

M2M identifies keystone, essential and alternative organisms by reducing the complexity of a large-scale microbiota into minimal communities with equivalent properties, suitable for further analyses.

Keywords: Metabolic networks — Microbiota — Metagenomics Community selection — Network expansion

Background

Understanding the interactions within microbiotas is crucial for ecological [53] and health [11] applications. With the improvements of metagenomics, and in particular the rise of methods to assemble individual genomes from metagenomes, unprecedented amounts of data are available to disentangle the functioning of microbiotas. This provides ways to tackle the potential role of microbes, whereas previous metataxonomics [34] analyses could only provide information on "who is there". Henceforth, the main challenge is to handle both the scale of metagenomics datasets, and the incompleteness of their data. Hundreds or thousands of genomes can be reconstructed from various environments [41, 16, 61, 52], either with the help of reference genomes or through metagenome-assembled genomes (MAGs).

These genomes are the starting point to a large set of analyses dedicated to gather the functions of the considered organisms, and possibly study them with regard to functions performed by a host. A first level of analyses is to annotate these genomes and characterise the families of molecular processes likely to happen in the species, which can rely on ontologies [4, 24]. Another possibility is to target the whole metabolism and build a genome-scale metabolic network

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

(GSMN) for each individual genome.

GSMNs gather all the expected metabolic reactions of an organism. Thiele et al [54] defined a precise protocol for building them, associating the use of automatic methods and a thorough curation of the model, based on expertise, literature, and mathematical analyses. This has been the basis for many implementations of GSMN reconstruction as all-in-one platforms [13, 27, 3], toolboxes [1, 57, 46], or individual tools for targeted refinements and analyses on GSMN [43, 55, 56]. Automatic reconstruction of GSMNs relies on the annotation of the genomes, as well as on the search for orthologues in known species. As expected based on the initial protocol for reconstruction described in [54], and despite the improvements of methods since then, automatic reconstructions of GSMN are likely to be incomplete. They produce GSMN drafts that need to be further refined and/or gap-filled, requiring human expertise that cannot be easily contemplated for a large number of genomes. [28]. Nevertheless, in the context of microbiotas and poorly-described organisms, it is hypothesised that some gaps in the metabolism can be explained by the dependency of organisms to each others [36]. It is therefore relevant to systematically analyse the complementarity between metabolic networks of microbiota species as a index for putative cooperation between them [58].

As metagenomics generates a large number of genomes for bacteria having the capability of exchanging compounds, their interactions deserve to be investigated at the metabolic scale. It is a challenging objective that entails turning hundreds or thousands of genomes into metabolic information, and analysing the latter according to their capabilities of individually or collectively produce metabolic compounds. Such a metabolic screening of organisms capabilities in microbiotas should enable the global comprehension of the functions expected in each member, based on its genetic content. Notice however that the objective of considering all the species together and study the complementarity of their respective metabolisms advocates for a very restrictive use of gap-filling procedures of individual metabolic networks. They indeed may artefactually assume that individual organisms each sustain their own growth in a restricted environ-

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

ment, whereas they rely on metabolic interactions to do so. As GSMNs are built individually, the purpose is instead to perform metabolic analyses on genomes that have been homogeneously prepared and take the most out of automatically built draft GSMNs, in order to highlight differences between them.

A variety of toolboxes have been proposed to study communities of organisms with GSMNs [30, 48]. Some studies are focused on pairwise interactions. Analyses on larger communities mainly rely on constraint-based modeling [8, 60, 29] at steady or unsteady states. However, these tools were applied to small size communities, usually no more than ten members, which suggests a computational bottleneck that needs to be faced for larger communities [30]. In addition, the GSMNs have to be of high-quality for accurate mathematical predictions. Therefore, as stated by [30], development of tools tailored to the analysis of large communities is needed. Being able to identify the main metabolic features of organisms beyond functional annotation is critical in microbiotas for the identification of important species, and further experiments or curation of their models.

Here we describe a software, metage2metabo (M2M), to analyse metabolic complementarity starting from individual annotated genomes in large microbiotas. M2M first characterises the added-value of organisms complementarity in terms of metabolic compounds production. Then it identifies communities and keystone species with respect to a family of metabolic compounds selected from the previous step. M2M uses the algorithm of network expansion [14] to capture the set of producible metabolites in a GSMN, therefore handling stoichiometry inaccuracy which is commonly faced with automatically reconstructed models. We therefore advocate for the identification of keystone species and the added-value of cooperation within the microbiota using M2M to evaluate the individual GSMNs and screen the collective metabolism.

To illustrate the metabolic screening potential of our method, we selected two large-scale sets of genomes for analysis: 913 cow rumen MAGs [52], and a set of 1,520 draft bacterial reference genomes from the gut microbiota [61]. We show that M2M can efficiently reconstruct metabolic networks for each genome,

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

identify potential metabolites produced by cooperating bacteria, and suggest minimal communities and keystone species associated to their production.

Implementation

Metage2Metabo (M2M) is a Python package. It can be used on a personal computer or on a cluster (using the Python package, Docker or Singularity) to benefit from its multi-processing functionality on large microbiotas datasets. A detailed documentation is available on metage2metabo.readthedocs.io.

M2M's main pipeline consists in three main steps performed sequentially: i) automatic reconstruction of metabolic networks for a large number of annotated genomes, ii) analysis of metabolic capabilities for each metabolic network and computation of the cooperation potential, i.e. the set of metabolites predicted to become producible through complementarity of synthetic pathways, and iii) identification of minimal communities and keystone species for a targeted set of compounds.

Figure 1 depicts the pipeline of M2M. The inputs for the whole workflow are a set of annotated genomes, and a list of nutrients representing a growth medium. However, each step can also be run individually. For instance, one can perform the metabolic network analysis by providing the GSMNs and the growth medium as inputs, or the community selection by providing the GSMNs, the medium and a family of metabolic compounds aimed to be produced by the community.

The whole pipeline is called with the command *m2m workflow*. We detail below the characteristics of the M2M through a description of its main three steps.

Large-scale metabolic network reconstruction

M2M enables a fast reconstruction of non-curated metabolic networks using Pathway Tools with the *m2m recon* command. It is the first multi-processing

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

solution available to run this software and is therefore highly suitable to get metabolic insights into the hundreds or thousands of genomes that can be retrieved from metagenomic experiments.

Let us recall that Pathway Tools [27] is a graphical user interface (GUI) based software suite for the generation of GSMNs, called Pathway/Genome Databases (PGDBs). These PGDBs can be obtained from annotated genomes using Pathway Tools's prediction component (PathoLogic) and curated afterwards. However, both Pathway-Tools GUI or command-line interface do not scale to the reconstruction of hundreds of GSMNs. With *m2m recon*, we propose an extension to Pathway Tools, that automatises the creation of these metabolic networks (in Systems Biology Markup Language (SBML) [22, 23] or PGDB the native Pathway Tools format) with PathoLogic, for large sets of genomes. Reconstructions are run in parallel, facilitating the scale-up of genome analysis.

To enable this reconstruction, we developed a Python wrapper named Mpwt (Multiprocessing Pathway Tools) that is included in M2M. This multiprocessing wrapper does not accelerate one PathoLogic run but it runs simultaneously multiple PathoLogic processes on different organisms. By default, Mpwt uses only one core but the user can allocate more to parallelise the runs. We recommend to give the number of available physical cores. Regarding memory requirements, they depend on the genome size but we advise to use at least 2 GB per core. The wrapper handles several types of genomic inputs (Genbank, Generic Feature Format (GFF) or PathoLogic format) and creates the input files needed by Pathway Tools. These files are then used by PathoLogic processes (one process by physical core) to create PGDBs. When all PGDBs are generated, Mpwt uses a lisp command of Pathway Tools to export the PGDB in an attribute-value flat files. These files are then used by the Padmet library [1] to generate SBML files suitable for exporting and sharing metabolic networks, for example for their use in other software for refinement.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Analysis of metabolites producibility and potential added-value of cooperation

Individual metabolic network analysis is a first step to compare the quality of the reconstructions, as well as the functional potential of all species in a given environment. As metabolic networks are automatically reconstructed from thousands of possibly incomplete genomes, constraint-based methods relying on flux balance analysis [40] are not suitable. We therefore chose to use the network expansion algorithm [14] to assess the metabolic potential of each species. The network expansion algorithm computes the scope of a metabolic network from a description of the growth medium (seeds), that is, the family of metabolic compounds which are reachable according to a boolean abstraction of the network dynamics assuming that cycles cannot be self-activated. This algorithm has been widely used to analyse and refine metabolic networks[35, 31, 10, 45, 43], including for microbiota analysis [9, 38, 39, 18]. As the scope ignores the stoichiometry of metabolites involved in reactions, it appears to be a good trade-off between the accuracy of metabolic predictions and the precision required for the input data. It is therefore adapted to the difficulties met when studying the metabolism of hundreds or thousands of non-model organisms.

The analysis of metabolism provided by M2M can be called with the *m2m iscope* command. It predicts the set of reachable metabolites, the scope, in a metabolic network, starting from a set of seeds. This individual metabolic capability is calculated for each GSMN. All the scopes are exported as a json file and a summary is provided to the user: the intersection (metabolites reachable by all GSMN) and the union of all scopes, as well as the average size of the scopes, the minimal size and the maximal size of all, to get a glance at the range of metabolic capabilities among the species. This step is based on features implemented in Python packages that are also available as stand-alone tools: Menetools and MiSCoTo [1, 18].

In addition to individual studies, all metabolic networks are studied as a whole, to get insights into the complementarity of their metabolic pathways.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Metabolic capabilities of the whole microbiota [18] can be computed using the network expansion algorithm with the command *m2m cscope*. This simulates the sharing of metabolic biosynthesis through a meta-organism composed of all GSMNs, and assesses the metabolic compounds that it can reach.

Based on the individual and community metabolic potentials, the global expected added-value (*m2m addedvalue*) of cooperation in a microbiota consists in the set of metabolites that can only be produced if several organisms share their metabolic biosynthesis. We advise the user to look at the list of producible compounds proposed by M2M, as it can be that some metabolites are false positive that do not necessitate cooperation for production, but were nonetheless selected due to missing annotations in the initial genomes. The *m2m addedvalue* command computes the identification of these compounds by comparing the results of individual and community scopes. It creates a target SBML file with these newly producible metabolites, for a possible use of these compounds in a community selection step.

Identification of keystone species

This step, called with the *m2m mincom* command provides the reduction of the initial large-scale community through the identification of minimal size communities to ensure a metabolic objective is met. The latter is the reachability of metabolic compounds (targets) starting from nutrients (seeds) under the network expansion algorithm. The targets can be the added-value of cooperation as presented above, or subcategories of them as defined by the user. This step relies on a functionality of MiSCoTo [18] to propose a minimal community of organisms suitable for the objective. *M2m mincom* assumes that all metabolic transports have equal costs as transport reactions are not well identified in automatically reconstructed GSMNs. Further analysis introducing different costs based on additional knowledge on transport reactions can be performed by using the MiSCoTo package [18].

Many equivalent minimal communities are expected to exist but their enu-

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

meration can be computationally fastidious, as well as their analysis. An originality of M2M is to efficiently sample the space of solutions without the need for a full enumeration, thanks to the underlying logic programming solving. The intersection of solutions i.e. species occurring in every minimal communities or *essential symbionts* are computed. We describe as *keystone species* the organisms occurring in at least one minimal community (union of solutions). Keystone species therefore contain the essential and *alternative symbionts*, the latter occuring in some minimal communities but not all of them. These terms were inspired by the terminology used in flux variability analysis [40] for the description of reactions in all optimal flux distributions.

Results

M2M was applied to two metagenomic experiments: a set of 1,520 bacterial high-quality draft reference genomes from the gut microbiota presented in [61], and 913 MAGs from the cow rumen published in [52]. The gut dataset consists in genomes of culturable bacteria, isolated from a large number of fecal samples and assembled into 338 species-level clusters. They cover major phyla: 796 Firmicutes, 447 Bacteroidetes, 235 Actinobacteria, 36 Proteobacteria and 6 Fusobacteria. We show that M2M is applicable to both types of datasets. We analyse in the next subsections the reconstructed GSMNs, the computation of the potential cooperation added-value, and keystone species. Finally we provide further analyses on the gut dataset with a deeper study of minimal communities composition.

The whole M2M workflow (from GSMN reconstruction to keystone species computation) took 155 minutes for the gut microbiota dataset. This reasonable time observed for computation is confirmed with the cow rumen dataset which ran in 81 minutes. In both cases, they were run on a cluster with 72 CPUs and 144 Go of memory.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

GSMN reconstruction

The genomes from the cow rumen were not annotated, a required feature for metabolic network reconstruction with M2M. We therefore annotated them using Prokka (v. 1.13.4) [47] as a preliminary step.

Results for the GSMNs reconstructions of both datasets and the analysis of individual metabolic potentials are presented in Table 1. The universe of metabolic reactions included in the reconstructions is of size 3,932 for the gut, and 4,418 for the rumen. Likewise, the universe of metabolites is of size 4,001 for the gut dataset, 4,466 for the rumen. The gut metabolic networks contained in average 1,144 ($\pm$ 255) reactions and 1,366 ($\pm$ 262) metabolites. 74.6% of the reactions were associated to genes, the remaining being spontaneous reactions or reactions added by the PathoLogic algorithm (they can be removed in M2M using the *–noorphan* option). GSMNs of the rumen dataset consisted in average of 1,155 ($\pm$ 199) reactions and 1,422 ($\pm$ 212) metabolites. 73.8% of the reactions were associated to genes. Supplementary Figure 1 displays the distributions of the numbers of reactions, pathways, metabolites and genes for both datasets. Altogether, these distributions are very similar for both datasets although the initial number of genes in the whole genomes varies a lot (Supp Fig. 1 g), a difference that is expected between MAGs and reference genomes. Interestingly, the average number of reactions per GSMN is slightly higher for the MAGs of the rumen than for the reference genomes of the gut. However, the smallest GSMN size is observed in the rumen (340 reactions vs 617 for the smallest GSMN of the gut). The similarity in the characteristics displayed by both datasets suggests a level of quality of the rumen MAGs close to the one of the gut reference genomes regarding the genes associated to metabolism.

For both experiments, we designed a set of seeds metabolites representing a nutritional environment that is required for the metabolic analyses. It consists in components of a classical diet for the gut microbiota (93 metabolites), and basic nutrients (26 metabolites including inorganic compounds, carbon dioxide, glucose and cellobiose) for the rumen (Supp Tables 1 and 2). The scope rep-

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

resents the metabolic potential (reachable metabolites) starting from available seeds (nutrients) according to the network expansion algorithm [14], that is a simulation of a boolean abstraction of the GSMN dynamics. The average size of the individual scopes is relatively small compared to the universe of metabolites for both datasets, which is highly dependent to the chosen seeds and the potential gaps in the GSMNs. The union of all individual scopes is of size 828 and 368 for the gut and the rumen respectively (21 % of the universe of compounds for the gut, 8.2 % for the rumen). Supplementary Figure 1 (h, i, j and k) displays the distributions of the scopes for both datasets.

Altogether, these results suggest that despite a good size of GSMN reconstructions, the individual metabolic potential of each species is relatively small. This can be an impact of the choice of seeds nutrients, but it can also be due to missing annotations or gaps in the networks. In the latter case, it is likely that metabolic cooperation between the species fills the gaps of biosynthesis pathways and enables the putative producibility of more metabolites, which can be calculated with M2M.

Added value of metabolic cooperation in the datasets

The metabolism of a given bacterium can be completed by the metabolism of others by filling gaps that exist in the first one, and therefore enabling the activation of more reactions than what can be expected when considered in isolation. By taking into account the complementarity between GSMNs in each dataset, it is possible to capture the benefit of metabolic cooperation over the producibility of metabolic compounds. Running *m2m cscope* evidenced that 296 and 156 new metabolites are potentially producible by the rumen GSMNs and the gut GSMNs respectively if cooperation between their members is allowed. This increases up to 15% and 25% the proportion of reachable metabolites in the whole universe of metabolites in the rumen and in the gut respectively. These are the metabolites that could not be reached by any GSMN when considered individually and therefore require shared metabolic capabilities to be produced.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

We analysed the composition of the 156 newly producible metabolites for the gut dataset using the ontology provided for metabolic compounds in the MetaCyc database [26]. We divided the metabolites into 6 categories: amino acids and derivatives (5 metabolites), aromatic compounds (11), carboxy acids (14), coenzyme A (CoA) derivatives (10), lipids (28), sugar derivatives (58) (Supp. Table 3). The remaining 30 compounds were highly heterogeneous, we therefore restrained our subsequent analyses to subcategories of homogeneous targets. By default, the M2M pipeline sets all newly producible metabolites as targets to perform community reduction in the following steps. Yet, it is possible to change them by selecting a subset of these metabolites, for example the subcategories of targets as we will present in the next paragraphs.

Community reduction

After the prediction of the cooperation added-value, M2M performs a community selection step based on a metabolic objective i.e. a list of metabolic compounds. It computes the size of a minimal community to ensure their producibility and their associated expected keystones species based on the GSMNs contents.

M2M proposes a community composition for the objective. Yet given the redundancy of functions in microbiotas [37], more than one minimal community solution is expected to exist. There might be thousands of them, and it can can be computationally difficult to enumerate due to the high combinatorics of the problem, especially for large sets of targets. Thanks to the logic programming solving assets of M2M, keystone species can be calculated without the need for all solutions to be enumerated, which is highly efficient computationally.

Applied to our datasets, *m2m mincom* led to a minimal community of size 25 to produce the 156 targets of the gut microbiota (Table 2). 11 members are essential symbionts, i.e. found in every minimal community, and the total number of keystone species is 205. Therefore, all minimal communities are composed of the same 11 members and a set of 14 others picked among 194

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

alternative symbionts. For the rumen dataset, the size of a minimal communities was 44, with an intersection of 20 essential symbionts and 107 alternative ones. The main pipeline of M2M stops after the computation of these sets of GSMNs, predicted to be relevant regarding the metabolic objective provided as input.

Keystone species in the gut dataset

We ran the community reduction step (*m2m mincom* command) with the 6 groups of targets that we isolated from the cooperation added-value for further analyses. The keystone species were computed for each group, and we studied their compositions in terms of phyla. The number of keystone species varies between 59 and 227, which is a strong reduction compared to the initial number of 1,520 GSMN used for the analysis.

Analysis of minimal communities in the gut

We enumerated all minimal communities for each individual group of targets using features of *m2m_analysis*. The number of optimal solutions (size of the enumeration) is large, reaching more than 7 million equivalent minimal communities for the sugar-derivated targets (Table 3). We observe however that the size of the minimal community is quite small for each targets group (between 4 and 11).

The sizes of the enumerations make them difficult to analyse. Yet, our analyses suggest that the large number of optimal communities comes from numerous possibilities of combinatorial choices among a rather small family of bacteria. More precisely, the identification of keystone, essential and alternative groups of species to capture the diversity of the minimal communities yielded to the results depicted in Table 3. It gathers the composition of the three groups for the whole set of targets and the targets categories (see Supp. Tables 5 to 10 for the contents of the groups). In particular, essential symbionts are of high importance in minimal communities as they are found in each solution. More generally, compositions vary across the targets categories: a high proportion of

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

keystone species for the production of lipids targets are Bacteroidetes whereas Firmicutes are highly represented for aminoacids and derivatives production.

Importantly, the keystones species for the categorical targets are not subsets of the ones for the whole set of targets. Large and heterogeneous sets of targets will lead to the selection of keystone species with more diversified biosynthesis pathways, likely to unblock the producibility of more compounds. On the contrary, for smaller and more homogeneous sets of targets, it is likely that a larger number of organisms with equivalent biosynthesis capabilities will be selected. Indeed, the number of keystone species is not necessarily smaller for small groups of targets.

Evidencing organisms with equivalent roles in microbial communities

In order to visualise the association of GSMNs in individual solutions, we created a graph whose nodes are the keystone species, and whose edges represent the association between two members if they co-occur in at least one of the enumerated communities. We created such graphs for each of the targets sets (lipids, sugar derivatives, aromatic compounds, aminoacids derivatives, carboxyacids and coA derivatives). Yet, they were very dense (185 nodes, 6888 edges for the lipids, 142 nodes and 6602 edges for the sugar derivatives), which is expected given the large number of optimal communities and the relatively small number of keystone species.

We compressed these graphs into power graphs to capture the mechanisms underlying the combinatorics of associations within minimal communities. Power graphs enable a lossless compression of re-occurring motifs within a graph: cliques, bicliques and star patterns [44]. We generated them using Power-GrASP [6] and visualised them with Cytoscape (version 2.8.3) [49] and the CyOog plugin (version 2.8.2) developed by [44]. As compressed graphs have a better readability, the power graphs generated from the association graphs of the enumerations enable to pinpoint metabolic equivalency between members of the keystone species.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure 2 presents the compressed graphs for each set of targets. Graph nodes are the keystone species, coloured by their phylum. A version of the figures with nodes identification is available in Supplementary Figures 2 to 7 (see. Supp. Table 4 for a mapping between identifiers and taxonomy). Nodes are included into power nodes, connected by power edges, depicting the equivalency between species with respect to the enumerated solutions. Symbionts belonging to a power node play the same role in the construction of the minimal communities. Essential symbionts are conveniently represented in the power graphs, either into power nodes with loops (Fig 2 a, e) or individual nodes connected to power nodes (Fig 2 a, c, d, f).

We observe that power nodes often contain GSMNs from the same phylum, indicating that an interesting function is shared by the taxonomic group or possibly the species-level clusters of genomes [61], for the producibility of the targets. The power graph gives information on the assembly of the GSMNs in these solutions. The power graphs for the groups of targets presented here are particularly informative to capture the composition of communities. Figure 2 a has additional comments to ease the reading. Each minimal community is composed of one Bacteroidetes from power node (PN) 1, one Actinobacteria from PN 2, the Firmicute member 3, one Proteobacteria from PN 4 and finally the two Firmicutes and the Proteobacteria from PN 5. For all the targets groups of this study, the numerous enumerations can be summarised with a boolean formula derivated from the graph compressions. For instance for the lipids of Figure 2 a, the community composition is the following $(\vee PN1) \wedge (\vee PN2) \wedge (PN3) \wedge (\vee PN4) \wedge (\wedge PN5)$. This visualisation of community compositions thus enables a better understanding of the associations of organisms into the proposed communities, and the relevance of keystone species to identify members with functions of interest for the producibility of targets.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Discussion

In this paper we present a new software for the functional analysis of metagenomic datasets at the metabolic level. M2M reconstructs metabolic networks using the Pathway Tools software and an efficient multi-processing wrapping. These GSMNs are then analysed individually and collectively to compute the potential added-value of metabolic cooperation and its associated minimal communities. Thanks to a graph-based modeling of producibility and powerful logic programming solving approaches, the whole space of minimal communities solutions can be parsed to retrieve all interesting bacteria, that we call keystone species, with respect to the metabolic objective. M2M can therefore suggest species for further analyses such as targeted curation of metabolic networks and deeper analysis of the genomes, that cannot be contemplated in an automatic way for the hundreds or thousands genomes of a microbiota.

The functionality of metagenomic sequences can be analysed at higher levels by directly computing functional profiles from reads [17, 51, 50, 42]. However, metabolic network reconstruction provides a more thorough study of the species by gathering the information about the reactions and pathways (complete or incomplete) they catalyse. Thanks to the improvements of the software suites for GSMN reconstruction, the automatically-built drafts provide a detailed view of the metabolic capabilities of species even without manual curation, although the latter is still needed for further analyses and accurate quantitative simulations of the metabolism.

Here we chose the Pathway Tools suite and its PathoLogic software [27] for GSMN reconstruction. It is based on the MetaCyc database [7] that covers a large taxonomic range of organisms and groups reactions into pathways that are convenient for a higher-level analysis of metabolism. Pathway Tools is distributed with a GUI that is very user-friendly for manual curation and analysis of a small number of GSMNs, but is limiting when aiming at reconstructing a large number of them. The multi-process wrapping we propose in this paper is entirely command-line based and performs metabolic network reconstruction

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

for a large number of species in a transparent way for the user. The resulting GSMNs can also be opened and refined with Pathway Tools GUI.

The assets of Pathway Tools are the decomposition of GSMN into pathways and the possibility to only use PathoLogic without further automatic gap-filling refinements to the model. The latter can be non-indicated since the species under study are not expected to be self sufficient for growth in a microbiota environment. Therefore we advocate for the use of GSMN drafts that can be likely to miss metabolic reactions, but have a lower risk of false positive reactions. The latter would otherwise possibly be added by automatic curation and gap-filling to sustain individual growth of the models. False positive reactions may lead to missed interactions between species when considering metabolic cooperation. GSMN of interest after M2M analysis can be further analysed and refined if needed. Nevertheless, there are several other software available for GSMN reconstruction in addition to Pathway Tools such as Kbase [3], ModelSEED [21] or CarveMe [32]. Metabolic networks resulting from such platforms can be used as inputs to M2M for metabolic analysis as the reconstruction step can be bypassed and the tool accepts GSMNs in SBML format [23], widely used for exchanging and distributing models.

Functional analysis of metabolic networks using the qualitative criterion of the scope has been demonstrated to be robust and relevant [20, 10, 43]. The scope provides a snapshot of producible metabolites, thus identifying the subnetwork that can be activated under given nutrient conditions. It is computed with the network expansion algorithm. A main difficulty can be to build the set of seeds that are available as nutrients for the analysis and on which the computation of network expansion relies. In particular, the algorithm has been demonstrated to be sensitive to cycles in GSMNs and it is therefore relevant to include some cofactors (e.g. ATP) in the seeds to activate such cycles, the way many studies proceed [12, 19, 15, 25]. Network expansion is a good trade-off with respect to quantitative constraint-based methods such as flux balance analysis [40] as it does not require biomass reactions nor accurate stoichiometry, and it can easily scale to thousands of networks considered in interaction. The cost of

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

exchanges is not taken into account in the M2M pipeline as transport reactions are hardly recovered by automatic methods [5]. Yet, the standalone MiSCoTo package used in M2M has an option for taking into account exchanges. It can compute communities while minimising the cost of exchanges, and suggest them although it comes with a computational cost.

The number of curated metabolic networks for species found in microbiotas is in growing evolution [33]. They are a highly valuable resource for the study of interactions between members of communities. Yet the variety of (reference) genomes obtained from shotgun metagenomic experiments is such than species and strains vary a lot and may not belong to the ones for which a curated GSMN is available. In addition, the rise of methods for assembling genomes directly from metagenomes (MAGs) leads to incomplete genomes for possibly unknown species on which one may still want to get metabolic insights [2]. Therefore, building de novo automatic metabolic network drafts for each genome or MAG is a relevant solution to globally analyse the functions of a metagenome, especially in the rapidly evolving context of the gut microbiota [59].

A critical issue in the identification of minimal communities is the large number of equivalent solutions and the fact that many species can play an equivalent role in these communities. Three strategies can be investigated. Targets can be divided into families to better understand the role of species with respect to their production. This can be achieved by relying on ontologies to classify metabolites. A second strategy is to identify keystone species and carefully distinguish essential and alternative symbionts. Finally, we presented an original method to visualise and summarise the large number of minimal communities given a metabolic objective by building power graphs [44]. It evidenced a taxonomic homogeneity in the clusters of species identified for groups of targets in the gut microbiota dataset. Altogether, Studying the metabolic potential of large communities is an iterative process that still requires biological expertise.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Conclusions

In this paper we present a new software for the functional analysis of genomes and MAGs from metagenomic experiments by reconstructing and comparing metabolic networks. M2M automatically builds GSMNs starting from annotated genomes. The pipeline is scalable and efficient by performing the reconstruction in a multiprocess framework. The resulting metabolic networks are analysed to capture the set of metabolites they are expected to reach, either individually or collectively through metabolic cooperation. The added-value of cooperation in terms of increase of the producible metabolites set is computed, and minimal communities for this objective are calculated. The large combinatorics of minimal communities due to functional redundancy in microbiotas is addressed by providing the alternative and essential symbionts of all solutions without the need for a complete enumeration. The groups of species identified are relevant for deeper analyses and their size makes the following studies more tractable than when considering the whole dataset of initial genomes.

M2M can be used as a pipeline but each step can also be performed individually thus increasing the flexibility of analysis and the range of its applications. Our method is robust against the uncertainty inherent to metagenomics data and is, to our knowledge, the only one available to predict keystone species at the metabolic level starting from large sets of genomes that can originate from poorly-studied or unknown species. M2M can be contemplated for a use together with abundance data to understand the importance of the identified keystone species. It could also serve as a basis to understand the evolution of microbial composition of patients in longitudinal studies at the metabolic level. We believe this software is of interest as the number of available genomes from metagenomic studies continues to rise, entailing a need for scalable predictive methods that tolerate the incompleteness of data.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Availability and requirements

Project name: metage2metabo

Project home page: https://github.com/AuReMe/metage2metabo

Operating system(s): Linux and MacOS

Programming language: Python 3.6 or higher

License: GNU General Public License v3.0

Any restrictions to use by non-academics: Licence needed for the use of Pathway Tools

List of abbreviations

M2M: metage2metabo

GSMN: Genome-Scale Metabolic Network

MAG: Metagenome-Assembled Genome

GFF: Generic Feature Format

RAM: Random-Access Memory

PGDB: Pathway Genome DataBase

SBML: Systems Biology Markup Language

GUI: Graphical User Interface

PN: Power Node

Availability of data and material

The rumen dataset MAGs used for the experiments were downloaded from `https://www.ncbi.nlm.nih.gov/assembly/?term=PRJEB21624`. The gut microbiota genomes were downloaded from `https://www.ncbi.nlm.nih.gov/assembly/?term=PRJNA482748`

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

ABe and CF developed the pipeline, designed and ran the experiments, performed the analyses. CF wrote the manuscript. MA and ABr designed technical solutions. AS supervised the work. All authors read and reviewed the manuscript.

Funding

This work has been supported by the IDEALG project ANR-10-BTBR-04.

Acknowledgements

The authors acknowledge the GenOuest bioinformatics core facility for providing the computing infrastructure. We also acknowledge P. Karp, S. Paley, M. Krummenacker, R. Billington, A. Kothari from the Bioinformatics Research Group of SRI International for their help regarding Pathway Tools. Finally, we thank Lucas Bourneuf for his help on power graph analyses.

References

[1] Méziane Aite, Marie Chevallier, Clémence Frioux, Camille Trottier, Jeanne Got, María Paz Cortés, Sebastián N. Mendoza, Grégory Carrier, Olivier Dameron, Nicolas Guillaudeux, Mauricio Latorre, Nicolás Loira, Gabriel V. Markov, Alejandro Maass, and Anne Siegel. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. *PLoS Computational Biology*, 14(5):e1006146, may 2018.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[2] Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, page 1, feb 2019.

[3] Adam P Arkin, Robert W Cottingham, Christopher S Henry, Nomi L Harris, Rick L Stevens, Sergei Maslov, Paramvir Dehal, Doreen Ware, Fernando Perez, Shane Canon, Michael W Sneddon, Matthew L Henderson, William J Riehl, Dan Murphy-Olson, Stephen Y Chan, Roy T Kamimura, Sunita Kumari, Meghan M Drake, Thomas S Brettin, Elizabeth M Glass, Dylan Chivian, Dan Gunter, David J Weston, Benjamin H Allen, Jason Baumohl, Aaron A Best, Ben Bowen, Steven E Brenner, Christopher C Bun, John-Marc Chandonia, Jer-Ming Chia, Ric Colasanti, Neal Conrad, James J Davis, Brian H Davison, Matthew DeJongh, Scott Devoid, Emily Dietrich, Inna Dubchak, Janaka N Edirisinghe, Gang Fang, José P Faria, Paul M Frybarger, Wolfgang Gerlach, Mark Gerstein, Annette Greiner, James Gurtowski, Holly L Haun, Fei He, Rashmi Jain, Marcin P Joachimiak, Kevin P Keegan, Shinnosuke Kondo, Vivek Kumar, Miriam L Land, Folker Meyer, Marissa Mills, Pavel S Novichkov, Taeyun Oh, Gary J Olsen, Robert Olson, Bruce Parrello, Shiran Pasternak, Erik Pearson, Sarah S Poon, Gavin A Price, Srividya Ramakrishnan, Priya Ranjan, Pamela C Ronald, Michael C Schatz, Samuel M D Seaver, Maulik Shukla, Roman A Sutormin, Mustafa H Syed, James Thomason, Nathan L Tintle, Daifeng Wang, Fangfang Xia, Hyunseung Yoo, Shinjae Yoo, and Dantong Yu. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology*, 36(7):566–569, jul 2018.

[4] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, G Sherlock, and Gavin Sherlock. Gene ontology: tool for the

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, may 2000.

[5] David B Bernstein, Floyd E Dewhirst, and Daniel Segre. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. *eLife*, 8, jun 2019.

[6] Lucas Bourneuf and Jacques Nicolas. FCA in a Logical Programming Setting for Visualization-Oriented Graph Compression. In *ICFCA 2017: Formal Concept Analysis*, pages 89–105. Springer, Cham, 2017.

[7] Ron Caspi, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Research*, oct 2019.

[8] Siu Hung Joshua Chan, Margaret N. Simons, and Costas D. Maranas. SteadyCom: Predicting microbial abundances while ensuring community stability. *PLOS Computational Biology*, 13(5):e1005539, may 2017.

[9] Nils Christian, Thomas Handorf, and Oliver Ebenhöh. Metabolic synergy: increasing biosynthetic capabilities by network cooperation. *Genome informatics International Conference on Genome Informatics*, 18:320–329, 2007.

[10] Nils Christian, Patrick May, Stefan Kempa, Thomas Handorf, and Oliver Ebenhöh. An integrative approach towards completing genome-scale metabolic networks. *Molecular BioSystems*, 5(12):1889–1903, 2009.

[11] The Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease corresponding author. *Cell Host and Microbe*, 16(3):276–289, sep 2014.

[12] Ludovic Cottret, Paulo Vieira Milreu, Vicente Acuña, Alberto Marchetti-Spaccamela, Leen Stougie, Hubert Charles, and Marie-France Sagot.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Graph-Based Analysis of the Metabolic Exchanges between Two Co-Resident Intracellular Symbionts, Baumannia cicadellinicola and Sulcia muelleri, with Their Insect Host, Homalodisca coagulata. *PLoS Computational Biology*, 6(9):e1000904, sep 2010.

[13] Scott Devoid, Ross Overbeek, Matthew DeJongh, Veronika Vonstein, Aaron A. Best, and Christopher Henry. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods in molecular biology (Clifton, N.J.)*, 985:17–45, 2013.

[14] Oliver Ebenhöh, Thomas Handorf, and Reinhart Heinrich. Structural analysis of expanding metabolic networks. *Genome informatics. International Conference on Genome Informatics*, 15(1):35–45, 2004.

[15] Alexander Eng and Elhanan Borenstein. An algorithm for designing minimal microbial communities with desired metabolic capacities. *Bioinformatics*, 32(13):2008–2016, 2016.

[16] Samuel C. Forster, Nitin Kumar, Blessing O. Anonye, Alexandre Almeida, Elisa Viciani, Mark D. Stares, Matthew Dunn, Tapoka T. Mkandawire, Ana Zhu, Yan Shao, Lindsay J. Pike, Thomas Louie, Hilary P. Browne, Alex L. Mitchell, B. Anne Neville, Robert D. Finn, and Trevor D. Lawley. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*, 37(2):186–192, feb 2019.

[17] Eric A. Franzosa, Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, Rob Knight, J. Gregory Caporaso, Nicola Segata, and Curtis Huttenhower. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11):962–968, nov 2018.

[18] Clémence Frioux, Enora Fremy, Camille Trottier, and Anne Siegel. Scalable and exhaustive screening of metabolic functions carried out by microbial consortia. *Bioinformatics*, 34(17):i934–i943, sep 2018.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[19] Sharon Greenblum, Peter J Turnbaugh, and Elhanan Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences of the United States of America*, 109(2):594–9, jan 2012.

[20] Thomas Handorf, Oliver Ebenhöh, and Reinhart Heinrich. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *Journal of Molecular Evolution*, 61(4):498–512, 2005.

[21] Christopher S. Henry, Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982, sep 2010.

[22] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Nov??re, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, mar 2003.

[23] Michael Hucka, Frank T Bergmann, Andreas Dräger, Stefan Hoops, Sarah M Keating, Nicolas Le Novère, Chris J Myers, Brett G Olivier, Sven Sahle, James C Schaff, Lucian P Smith, Dagmar Waltemath, and Darren J Wilkinson. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core. *Journal of integrative bioinformatics*, 15(1), mar 2018.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[24] Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian Von Mering, and Peer Bork. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(D1):D286–D293, jan 2016.

[25] Alice Julien-Laferrière, Laurent Bulteau, Delphine Parrot, Alberto Marchetti-Spaccamela, Leen Stougie, Susana Vinga, Arnaud Mary, and Marie-France Sagot. A Combinatorial Algorithm for Microbial Consortia Synthetic Design. *Scientific Reports*, 6:29182, jul 2016.

[26] Peter D. Karp, Richard Billington, Ron Caspi, Carol A. Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M. Keseler, Markus Krummenacker, Peter E. Midford, Quang Ong, Wai Kit Ong, Suzanne M. Paley, and Pallavi Subhraveti. The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 28(12):1–6, aug 2017.

[27] Peter D. Karp, Mario Latendresse, Suzanne M. Paley, Markus Krummenacker, Quang D. Ong, Richard Billington, Anamika Kothari, Daniel Weaver, Thomas Lee, Pallavi Subhraveti, Aaron Spaulding, Carol Fulcher, Ingrid M. Keseler, and Ron Caspi. Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 17(5):877–890, sep 2016.

[28] Peter D. Karp, Daniel Weaver, and Mario Latendresse. How accurate is automated gap filling of metabolic models? *BMC Systems Biology*, 12(1):73, dec 2018.

[29] Ruchir A. Khandelwal, Brett G. Olivier, Wilfred F. M. Röling, Bas Teusink, and Frank J. Bruggeman. Community flux balance analysis for microbial consortia at balanced growth. *PloS one*, 8(5):e64567, may 2013.

[30] Manish Kumar, Boyang Ji, Karsten Zengler, and Jens Nielsen. Modelling

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

approaches for studying the microbiome. *Nature Microbiology*, 4(8):1253–1267, aug 2019.

[31] Julie Laniau, Clémence Frioux, Jacques Nicolas, Caroline Baroukh, M.-P. Maria-Paz Cortes, Jeanne Got, Camille Trottier, Damien Eveillard, and Anne Siegel. Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. *PeerJ*, 5(10):e3860, oct 2017.

[32] Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46(15):7542–7553, sep 2018.

[33] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, nov 2016.

[34] Julian R Marchesi and Jacques Ravel. The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1):31, 2015.

[35] Franziska Matthäus, Carlos Salazar, and Oliver Ebenhöh. Biosynthetic Potentials of Metabolites and Their Hierarchical Organization. *PLoS Computational Biology*, 4(4):e1000049, apr 2008.

[36] J Jeffrey Morris, Richard E Lenski, and Erik R Zinser. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio*, 3(2):e00036–12, may 2012.

[37] Andrés Moya and Manuel Ferrer. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends in Microbiology*, 24(5):402–413, may 2016.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[38] Shany Ofaim, Maya Ofek-Lalzar, Noa Sela, Jiandong Jinag, Yechezkel Kashi, Dror Minz, and Shiri Freilich. Analysis of Microbial Functions in the Rhizosphere Using a Metabolic-Network Based Framework for Metagenomics Interpretation. *Frontiers in Microbiology*, 8:1606, aug 2017.

[39] Itai Opatovsky, Diego Santos-Garcia, Zhepu Ruan, Tamar Lahav, Shany Ofaim, Laurence Mouton, Valérie Barbe, Jiandong Jiang, Einat Zchori-Fein, and Shiri Freilich. Modeling trophic dependencies and exchanges among insects' bacterial symbionts in a host-simulated environment. *BMC Genomics*, 19(1):402, dec 2018.

[40] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is Flux Balance Analysis ? *Nature biotechnology*, 28(3):245–248, mar 2010.

[41] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L. Rice, Casey DuLong, Xochitl C. Morgan, Christopher D. Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649–662.e20, jan 2019.

[42] Pavel Petrenko, Briallen Lobb, Daniel A. Kurtz, Josh D. Neufeld, and Andrew C. Doxey. MetAnnotate: function-specific taxonomic profiling and comparison of metagenomes. *BMC Biology*, 13(1):92, dec 2015.

[43] Sylvain Prigent, Clémence Frioux, Simon M. Dittami, Sven Thiele, Abdelhalim Larhlimi, Guillaume Collet, Fabien Gutknecht, Jeanne Got, Damien Eveillard, Jérémie Bourdon, Frédéric Plewniak, Thierry Tonon, and Anne Siegel. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. *PLOS Computational Biology*, 13(1):e1005276, jan 2017.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[44] Loïc Royer, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder. Unraveling Protein Networks with Power Graph Analysis. *PLoS Comput Biol*, 4(7):e1000108, jul 2008.

[45] Torsten Schaub and Sven Thiele. Metabolic network expansion with answer set programming. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5649 LNCS, pages 312–326, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[46] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, Joseph Kang, Daniel R Hyduke, and Bernhard O Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, 6(9):1290–307, sep 2011.

[47] Torsten Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, jul 2014.

[48] Partho Sen and Matej Orešič. Metabolic Modeling of Human Gut Microbiota on a Genome Scale: An Overview. *Metabolites*, 9(2), jan 2019.

[49] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Benno Schwikowski, and Trey Ideker. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, nov 2003.

[50] Ashok K. Sharma, Ankit Gupta, Sanjiv Kumar, Darshan B. Dhakan, and Vineet K. Sharma. Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics*, 106(1):1–6, jul 2015.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[51] Genivaldo Gueiros Z. Silva, Kevin T. Green, Bas E. Dutilh, and Robert A. Edwards. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics*, 32(3):354–361, feb 2016.

[52] Robert D Stewart, Marc D Auffret, Amanda Warr, Andrew H Wiser, Maximilian O Press, Kyle W Langford, Ivan Liachko, Timothy J Snelling, Richard J Dewhurst, Alan W Walker, Rainer Roehe, and Mick Watson. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature communications*, 9(1):870, 2018.

[53] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesco D'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Ocean plankton. Structure and function of the global ocean microbiome. *Science (New York, N.Y.)*, 348(6237):1261359, may 2015.

[54] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, jan 2010.

[55] Ines Thiele, Nikos Vlassis, and Ronan M. T. Fleming. fastGapFill: efficient gap filling in metabolic networks. *Bioinformatics (Oxford, England)*, 30(17):2529–2531, sep 2014.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[56] Edward Vitkin and Tomer Shlomi. MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome biology*, 13(11):R111, 2012.

[57] Hao Wang, Simonas Marcišauskas, Benjamín J. Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, and Eduard J. Kerkhoven. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Computational Biology*, 14(10):e1006541, oct 2018.

[58] Aleksej Zelezniak, Sergej Andrejev, Olga Ponomarova, Daniel R. Mende, Peer Bork, and Kiran Raosaheb Patil. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 112(20):6449–6454, may 2015.

[59] Shijie Zhao, Tami D Lieberman, Mathilde Poyet, Kathryn M Kauffman, Sean M Gibbons, Mathieu Groussin, Ramnik J Xavier, and Eric J Alm. Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell host & microbe*, 25(5):656–667.e8, may 2019.

[60] Ali R. Zomorrodi and Costas D. Maranas. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Computational Biology*, 8(2):e1002363, 2012.

[61] Yuanqiang Zou, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, Yan Xia, Suisha Liang, Ying Dai, Daiwei Wan, Rongrong Jiang, Lili Su, Qiang Feng, Zhuye Jie, Tongkun Guo, Zhongkui Xia, Chuan Liu, Jinghong Yu, Yuxiang Lin, Shanmei Tang, Guicheng Huo, Xun Xu, Yong Hou, Xin Liu, Jian Wang, Huanming Yang, Karsten Kristiansen, Junhua Li, Huijue Jia, and Liang Xiao. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 37(2):179–185, feb 2019.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Figures

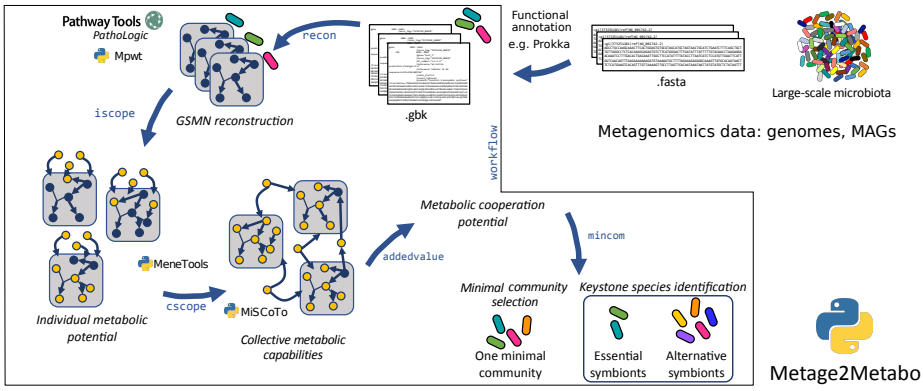

Figure 1: **Overview of the M2M pipeline.** The pipeline takes as inputs a set of annotated genomes that can be reference genomes of metagenomics-assembled genomes. Starting from these annotated genomes, Pathway Tools can be run in parallel for a large number of organisms. The resulting metabolic networks are analysed individually and collectively to identify metabolic capabilities, added-value of cooperation, and species of interest. The pipeline can be run as a whole or the steps can be performed individually.

Tables

Additional Files

SuppFigures.pdf — Additional figures

SuppTables.xlsx — Additional tables

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

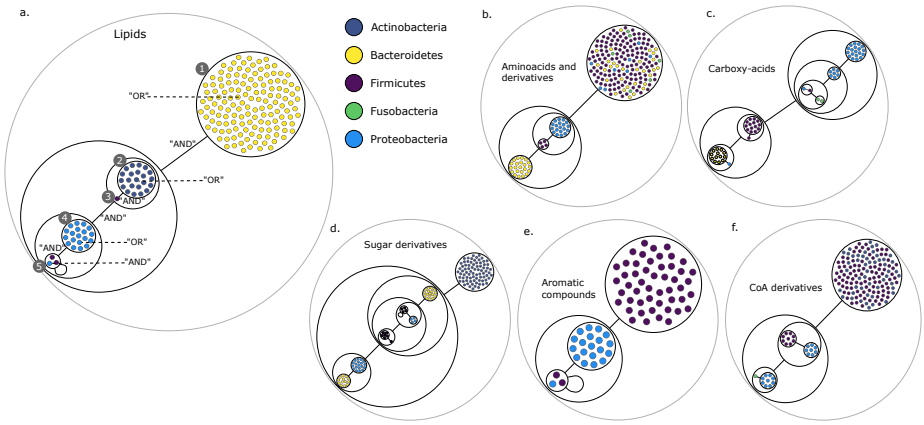

Figure 2: **Network analysis of microbial associations within communities for the gut dataset.** Each category of metabolites predicted as newly producible in the gut was defined as a target set for community selection among the 1,520 GSMNs from the gut dataset. For each metabolic group, keystone species and the full enumeration of all minimal communities were computed. Association graphs were built to associate members that are found together in at least one minimal community among the enumeration. These graphs were compressed as power graphs to identify patterns of associations. Power graphs a., b., c., d., e., f., g. present the patterns of associations for lipids, aminoacids and derivatives, carboxy-acids, sugar derivatives, aromatic compounds, and coenzyme A derivative compounds respectively. Nodes colours describe the phylum the initial genomes belong to. Figure a. has an additional description to ease readability. Edges symbolise conjunctions ("AND"), the co-occurrences of nodes in regular powernodes (as in powernode 1, 2, 4) symbolise disjunctions ("OR") related to alternative symbionts. Powernodes with a loop (e.g. powernode 5) indicate conjunctions. Therefore, each enumerated minimal community for lipids production is composed of the two Firmicutes and the Proteobacteria from powernode 5, the Firmicutes node 3 (the four of them being the essential symbionts), and one Proteobacteria from powernode 4, one Actinobacteria from powernode 2 and 1 Bacteroidetes from powernode 1.

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 1: Results of the GSMN reconstruction step and metabolic potential analysis for two datasets (Avg = Average, "±" precedes standard deviation)

	Gut dataset	Rumen dataset
initial data	draft reference genomes	MAGs
number of genomes	1520	913
GSMN reconstruction		
all reactions	3932	4418
all metabolites	4001	4466
avg reactions per GSMN	1,144 (± 255)	1,155 (± 199)
avg metabolites per GSMN	1,366 (± 262)	1,422 (± 212)
avg genes per mn	596 (± 150)	543 (± 107)
% reactions associated to genes	74.6 (± 2.17)	73.8 (± 2.61)
avg pathways per mn	163 (± 49)	146 (± 32)
metabolic potential		
number of seeds	93	26
avg scope per mn	286 (± 70)	101 (± 44)
union of individual scopes	828	368

Table 2: Results of the community reduction step for the gut and rumen datasets. Keystone species occur in at least one of all equivalent minimal communities. Essential symbionts belong to all the minimal communities. Alternative symbionts belong to some but not all minimal communities

	Gut dataset	Rumen dataset
minimal size of community	25	44
keystone species	205	127
essential symbionts	11	20
alternative symbionts	194	107

bioRxiv preprint first posted online Oct. 15, 2019; doi: http://dx.doi.org/10.1101/803056. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 3: Community reduction analysis of the target categories in the gut. All minimal communities were enumerated, starting from the set of 1,520 GSMNs. KS: keystone species, ES: essential symbionts, AS: alternative symbionts, Firm.: Firmicutes, Bact.: Bacteroidetes, Acti.: Actinobacteria, Prot.: Proteobacteria, Fuso.: Fusobacteria.

		Firm.	Bact.	Acti.	Prot.	Fuso.	total
all (156 targets)	KS	41	47	103	12	2	205
25 bact. per community	ES	8	0	1	2	0	11
not enumerated	AS	33	47	102	10	2	194
aminoacids and derivatives (5 targets)	KS	142	52	0	27	6	227
4 bact. per community	ES	0	0	0	0	0	0
120,329 communities	AS	142	52	0	27	6	227
aromatic compounds (11 targets)	KS	52	0	0	20	0	72
5 bact. per community	ES	2	0	0	1	0	3
950 communities	AS	50	0	0	19	0	69
carboxyacids (14 targets)	KS	16	13	0	28	2	59
9 bact. per community	ES	2	0	0	2	0	4
48,412 communities	AS	14	13	0	26	2	55
coA derivatives (10 targets)	KS	106	0	50	17	1	174
5 bact. per community	ES	0	0	0	0	1	1
95,256 communities	AS	106	0	50	17	0	173
lipids (28 targets)	KS	3	140	22	20	0	185
7 bact. per community	ES	3	0	0	1	0	4
58,520 communities	AS	0	140	22	19	0	181
sugar derivatives (58 targets)	KS	11	30	78	23	0	142
11 bact. per community	ES	5	0	0	0	0	5
7,860,528 communities	AS	6	30	78	23	0	137