Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species

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

Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species

Arnaud Belcour[1†], **Clémence Frioux**[2,1,3,4*†], **Méziane Aite**[1], **Anthony Bretaudeau**[1,5,6], **Falk Hildebrand**[3,4], **Anne Siegel**[1]

***For correspondence:** clemence.frioux@inria.fr (CF)

[†]These authors contributed equally to this work

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France; [2]Inria Bordeaux Sud-Ouest, France; [3]Gut Microbes and Heath, Quadram Institute, NR4 7GJ, Norwich, United Kingdom; [4]Digital Biology, Earlham Institute, NR4 7GJ, Norwich, United Kingdom; [5]INRAE, UMR IGEPP, BioInformatics Platform for Agroecosystems Arthropods (BIPAA), 35042 Rennes, France; [6]Inria, IRISA, GenOuest Core Facility, 35042 Rennes, France

Abstract To capture the functional diversity of microbiota, one must identify metabolic functions and species of interest within hundreds or thousands of microorganisms. We present Metage2Metabo (M2M) a resource that meets the need for de-novo functional screening of genome-scale metabolic networks (GSMNs) at the scale of a metagenome, and the identification of critical species with respect to metabolic cooperation. M2M comprises a flexible pipeline for the characterisation of individual metabolisms and collective metabolic complementarity. In addition, M2M identifies key species, that are meaningful members of the community for functions of interest. We demonstrate that M2M is applicable to collections of genomes as well as metagenome-assembled genomes, permits an efficient GSMN reconstruction with Pathway Tools, and assesses the cooperation potential between species. M2M identifies key organisms by reducing the complexity of a large-scale microbiota into minimal communities with equivalent properties, suitable for further analyses.

Introduction

Understanding the interactions between organisms within microbiomes is crucial for ecological (***Sunagawa et al., 2015***) and health (***The Integrative HMP Research Network Consortium, 2014***) applications. Improvements in metagenomics, and in particular the development of methods to assemble individual genomes from metagenomes, have given rise to unprecedented amounts of data which can be used to elucidate the functioning of microbiomes. Hundreds or thousands of genomes can now be reconstructed from various environments (***Pasolli et al., 2019***; ***Forster et al., 2019***; ***Zou et al., 2019***; ***Stewart et al., 2018***; ***Almeida et al., 2020***), either with the help of reference genomes or through metagenome-assembled genomes (MAGs), paving the way for numerous downstream analyses. Some major interactions between species occur at the metabolic level. This is the case for negative interactions such as exploitative competition (e.g. for nutrient resources), or for positive interactions such as cross-feeding or syntrophy (***Coyte and Rakoff-Nahoum, 2019***) that we will refer to with the generic term of cooperation . In order to unravel such interactions between species, it is necessary to go beyond functional annotation of individual genomes and connect metagenomic data to metabolic modelling. The main challenges impeding mathematical and computational analysis and simulation of metabolism in microbiomes are the scale of metagenomic

datasets and the incompleteness of their data.

Genome-scale metabolic networks (GSMNs) integrate all the expected metabolic reactions of an organism. ***Thiele and Palsson*** (***2010***) defined a precise protocol for their reconstruction, associating the use of automatic methods and thorough curation based on expertise, literature and mathematical analyses. There now exist a variety of GSMN reconstruction implementations: all-in-one platforms such as Pathway Tools (***Karp et al., 2016***), CarveMe (***Machado et al., 2018***) or KBase that provides narratives from metagenomic datasets analysis up to GSMN reconstruction with ModelSEED (***Henry et al., 2010a***; ***Seaver et al., 2020***). In addition, a variety of toolboxes (***Aite et al., 2018***; ***Wang et al., 2018***; ***Schellenberger et al., 2011***), or individual tools perform targeted refinements and analyses on GSMNs (***Prigent et al., 2017***; ***Thiele et al., 2014***; ***Vitkin and Shlomi, 2012***). Reconstructed GSMNs are a resource to analyse the metabolic complementarity between species, which can be seen as a representation of the putative cooperation within communities (***Opatovsky et al., 2018***). SMETANA (***Zelezniak et al., 2015***) estimates the cooperation potential and simulates fluxes exchanges within communities. MiSCoTo (***Frioux et al., 2018***) computes the metabolic potential of interacting species and performs community reduction. NetCooperate (***Levy et al., 2015***) predicts the metabolic complementarity between species.

In addition, a variety of toolboxes have been proposed to study communities of organisms using GSMNs (***Kumar et al., 2019***; ***Sen and Orešič, 2019***), most of them relying on constraint-based modelling (***Chan et al., 2017***; ***Zomorrodi and Maranas, 2012***; ***Khandelwal et al., 2013***). However, these tools can only be applied to communities with few members, as the computational cost scales exponentially with the number of included members (***Kumar et al., 2019***). Only recently has the computational bottleneck started to be addressed (***Diener et al., 2020***). In addition, current methods require GSMNs of high-quality in order to produce accurate mathematical predictions and quantitative simulations. Reaching this level of quality entails manual modifications to the models using human expertise, which is not feasible at a large scale in metagenomics. Automatic reconstruction of GSMNs scales to metagenomic datasets, but it comes with the cost of possible missing reactions and inacurate stoichiometry that impede the use of constraint-based modelling (***Bernstein et al., 2019***). Therefore, development of tools tailored to the analysis of large communities is needed.

Here we describe Metage2Metabo (M2M), a software system for the characterisation of metabolic complementarity starting from annotated individual genomes. M2M capitalises on the parallel reconstruction of GSMNs and a relevant metabolic modelling formalism to scale to large microbiotas. It comprises a pipeline for the individual and collective analysis of GSMNs and the identification of communities and key species ensuring the producibility of metabolic compounds of interest. M2M automates the systematic reconstruction of GSMNs using Pathway Tools or relies on GSMNs provided by the user. The software system uses the algorithm of network expansion (***Ebenhöh et al., 2004***) to capture the set of producible metabolites in a GSMN. This choice answers the needs for stoichiometry inaccuracy handling and the robustness of the algorithm was demonstrated by the stability of the set of reachable metabolites despite missing reactions (***Handorf et al., 2005***; ***Kruse and Ebenhöh, 2008***). Consequently, M2M scales metabolic modelling to metagenomics and large collections of (metagenome-assembled) genomes.

We applied M2M on a collection of 1,520 draft bacterial reference genomes from the gut microbiota (***Zou et al., 2019***) to illustrate the range of analyses the tool can produce. It demonstrates that M2M efficiently reconstructs metabolic networks for all genomes, identifies potential metabolites produced by cooperating bacteria, and suggests minimal communities and key species associated to their production. We then compared metabolic network reconstruction applied to the gut reference genomes to the results obtained with a collection of 913 cow rumen MAGs (***Stewart et al., 2018***). In addition, we tested the robustness of metabolic prediction with respect to genome incompleteness by degrading the rumen MAGs. The comparison of outputs from the pipeline indicates stability of the results with moderately degraded genomes, and the overall suitability of M2M to MAGs. Finally, we demonstrated the applicability of M2M in practice to metagenomic data of individuals.

To that purpose, we reconstructed communities for 170 samples of healthy and diabetic individuals (***Forslund et al., 2015***; ***Diener et al., 2020***). We show how M2M can help connect sequence analyses to metabolic screening in metagenomic datasets

Results

M2M pipeline and key species

M2M is a flexible software solution that performs automatic GSMN reconstruction and systematic screening of metabolic capabilities for up to thousands of species for which an annotated genome is available. The tool computes both the individual and collective metabolic capabilities to estimate the complementarity between the metabolisms of the species. Then based on a determined metabolic objective which can be ensuring the producibility of metabolites that need cooperation, that we call cooperation potential, M2M performs a community reduction step that aims at identifying a minimal community fulfilling the metabolic objective, and the set of associated key species.

M2M's main pipeline (***Figure 1*** a) consists in five main steps that can be performed sequentially or independently: i) reconstruction of metabolic networks for all annotated genomes, ii) computation of individual and iii) collective metabolic capabilities, iv) calculation of the cooperation potential and v) identification of minimal communities and key species for a targeted set of compounds.

Sets of producible metabolites for individual or communities of species are computed using the network expansion algorithm (***Ebenhöh et al., 2004***) that is implemented in Answer Set Programming in dependencies of M2M. Network expansion enables the calculation of the *scope* of one or several metabolic networks in given nutritional conditions, described as seed compounds. The scope therefore represents the metabolic potential or reachable metabolites in these conditions (see Methods). M2M calculates individual scopes for all metabolic networks, and the community scope comprising all reachable metabolites for the interacting species. Network expansion is also used in the community reduction optimisation implemented in MiSCoTo (***Frioux et al., 2018***), the dependency of M2M, as reduced communities are expected to produce the metabolites of interest.

The inputs for the whole workflow are a set of annotated genomes, a list of nutrients representing a growth medium, and optionally a list of targeted compounds to be produced by selected communities that will bypass the default objective of ensuring the producibility of the cooperation potential. Users can use the annotation pipeline of their choice prior running M2M. The whole pipeline is called with the command `m2m workflow` but each step can also be run individually as described in ***Table 1***.

A main characteristic of M2M is to provide at the end of the pipeline a set of key species associated to a metabolic function together with one minimal community predicted to satisfy this function. We define as *key species* organisms whose GSMNs are selected in at least one of the minimal communities predicted to fulfil the metabolic objective. Among key species, we distinguish those that occur in every minimal community, suggesting that they possess key functions associated to the objective, from those that occur only in some communities. We call the former *essential symbionts*, and the latter *alternative symbionts*. These terms were inspired by the terminology used in flux variability analysis (***Orth et al., 2010***) for the description of reactions in all optimal flux distributions. If interested, one can compute the enumeration of all minimal communities with *m2m_analysis*, which will provide the total number of minimal communities as well as the composition of each. ***Figure 1*** b. illustrates these concepts. The initial community is formed of eight species. There are four minimal communities satisfying the metabolic objective. Each includes three species, and in particular the yellow one is systematically a member. Therefore the yellow species is an essential symbiont whereas the four other species involved in minimal communities constitute the set of alternative symbiont. As key species represent the diversity associated to all minimal communities, it is likely that their number is greater than the size of a minimal community, as this is the case in ***Figure 1*** b.

M2M connects metagenomics to metabolism with GSMN reconstruction, metabolic complementarity screening and community reduction

In order to illustrate its applicability to real data, M2M was applied to a collection of 1,520 bacterial high-quality draft reference genomes from the gut microbiota presented in ***Zou et al.*** (***2019***). The genomes were derived from cultured bacteria, isolated from faecal samples covering typical gut phyla (***Costea et al., 2018***): 796 Firmicutes, 447 Bacteroidetes, 235 Actinobacteria, 36 Proteobacteria and 6 Fusobacteria. The dereplicated genomes represent 338 species. The genomes were already annotated and could therefore directly enter M2M pipeline. The full workflow (from GSMN reconstruction to key species computation) took 155 minutes on a cluster with 72 CPUs and 144 Gb of memory. We illustrate in the next paragraphs the scalability of M2M and the range of analyses it proposes by applying the pipeline to this collection of genomes.

GSMN reconstruction

GSMNs were automatically reconstructed for the 1,520 isolate-based genomes using their published annotation. A total of 3,932 unique reactions and 4,001 metabolites were included in the reconstructed GSMNs (***Table 2***). The reconstructed gut metabolic networks contained on average 1,144 (± 255) reactions and 1,366 (± 262) metabolites per genome. 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).

The metabolic potential, or *scope*, was computed for each individual GSMN (***Table 2***). Nutrients in this experiment were components of a classical diet (see Methods). The union of all individual scopes is of size 828 (21 % of all compounds included in the GSMNs), indicating a small part of the metabolism reachable in the chosen nutritional environment (Supplementary File 1 - Table 1). ***Appendix 1*** (Figure 1 h, i) displays the distributions of the scopes. Across all GSMNs, individual scopes are overall stable in size. The core set of producible metabolites is small and a variety of metabolites are only reachable by a small number of organisms (***Appendix 1*** Figure 1i). The overall small size of metabolic potentials can be explained by the restricted amount of seeds used for computation. Among metabolites that are reachable by all or almost all metabolic networks, the primary metabolism is highly represented, as expected, with metabolites derived from common sugars (glucose, fructose), pyruvate, 2-oxoglutaric acid, amino acids... On the other hand but not surprisingly, metabolites that predicted to be reached by a limited number of individual producers include compounds from secondary metabolism: (fatty) acids (e.g. oxalate, maleate, allantoate, hydroxybutanoate, methylthiopropionate) and derivatives of amino acids, amines (spermidine derivatives)...

Cooperation potential

Metabolic cooperation enables the activation of more reactions in GSMNs than what can be expected when networks are considered in isolation. By taking into account the complementarity between GSMNs in each dataset, it is possible to capture the putative benefit of metabolic cooperation on the diversity of producible metabolites. Running *m2m cscope* predicted 156 new metabolites as producible by the gut collection of GSMNs if cooperation is allowed.

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. 80.1% of them could be grouped 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). The groups were used in the subsequent analyses. The remaining 30 compounds were highly heterogeneous, we therefore restrained our subsequent analyses to subcategories of biochemically homogeneous targets.

We paid a particular attention to the predicted producibility of short chain fatty acids (SCFAs) among genomes of the gut collection. We analysed formate, acetate, propionate and butyrate in individual and collective metabolic potentials (Supplementary File 1 - Table 25). 543 metabolic

networks are predicted to be able to produce all four molecules in a cooperative context, 74% of them belonging the Firmicutes, as expected. Surprisingly, predicted individual producers of the four SCFAs (n = 128) are mostly Bacteroidetes (70%) suggesting the dependency of Firmicutes to interactions in order to permit the producibility of SCFAs in this experimental setting. The same observations are made when focusing on butyrate alone, that has the particularity of belonging to the seeds. As Bacteroidetes are not the main butyrate producers in the gut, the predictions of such producibility is likely an artefact relying on alternative pathways, and further emphasises the fact that owning the genetic material for a function does not entail its expression.

Key species associated to groups of metabolites

M2M proposes by default one community composition for an objective defined by enabling the producibility of metabolic end-products. Given the functional redundancy of gut bacteria (***Moya and Ferrer, 2016***), there could be thousands of bacterial composition combinations, and it is computationally costly to enumerate them. To circumvent this restriction, M2M identifies *key species* without the need for all possible combinations of species to be enumerated, consequentially reducing computational time. Key species include all species occurring in at least one minimal community for the production of chosen end-products. They can be distinguished in two categories: *essential symbionts* occurring in all minimal communities, and *alternative symbionts* occurring in some minimal communities.

To explore the spectrum of possible key species, we ran M2M community reduction step (`m2m mincom` command) with the above 6 metabolic target groups. This allowed us to compute likely key species for each of them (***Table 3***). The contents of key species for each of the six groups of targets as well as for the complete set of targets is displayed in Supplementary File 1 (Tables 6 to 12). To our surprise, the size of the minimal community is relatively small for each group of metabolites (between 4 and 11), compared to the initial community of 1,520 GSMNs. The number of identified key species varies between 59 and 227, which might be closer to the total taxonomic diversity found in the human gut microbiome. This strong reduction compared to the initial number of 1,520 GSMNs used for the analysis illustrates the existence of groups of bacteria with specific metabolic capabilities. In particular, essential symbionts are likely of high importance for the functions as they are found in each solution. More generally, compositions vary across the target categories: a high proportion of key species for the production of lipids targets are Bacteroidetes whereas Firmicutes were more often key species for aminoacids and derivatives production. The propensity of Bacteroidetes to metabolise lipids has been proposed previously, it has for example been observed in the *Bacteroides* enterotype for functions related to lipolysis (***Vieira-Silva et al., 2016***).

Analysis of minimal communities identifies groups of organisms with equivalent roles

To go further, we enumerated all minimal communities for each individual group of targets using *m2m_analysis*. The number of optimal solutions is large, reaching more than 7 million equivalent minimal communities for the sugar-derived metabolites (***Table 3***). Our analysis of key species indicates that the large number of optimal communities is due to combinatorial choices among a rather small number of bacteria (***Table 3***).

In order to visualise the association of GSMNs in minimal communities, we created for each target set a graph whose nodes are the key species (Supplementary File 1 - Tables 6 to 12), and whose edges represent the association between two species if they co-occur in at least one of the enumerated communities. Graphs were very dense: 185 nodes, 6888 edges for the lipids, 142 nodes and 6602 edges for the sugar derivatives. This density is expected given the large number of optimal communities and the comparatively small number of key species. The graphs were compressed into power graphs to capture the combinatorics of association within minimal communities. Power graphs enable a lossless compression of re-occurring motifs within a graph: cliques, bicliques and star patterns (***Royer et al., 2008***). The increased readability of power graphs

permits pinpointing metabolic equivalency between members of the key species with respect to the target compound families.

Figure 3 presents the compressed graphs for each set of targets. Graph nodes are the key species, coloured by their phylum. Nodes are included into power nodes that are connected by power edges, illustrating the redundant metabolic function(s) that species provide to the community when considering particular end-products. GSMNs belonging to a power node play the same role in the construction of the minimal communities. In this visualisation, essential symbionts are easily identifiable, either into power nodes with loops (***Figure 3*** a, e) or as individual nodes connected to power nodes (***Figure 3*** a, c, d, f).

We observe that power nodes often contain GSMNs from the same phylum, indicating that phylogenetic groups encode redundant functions. ***Figure 3 a*** has additional comments to guide the reader into analysing the community composition on one example. Each minimal community suitable for the production of the targeted lipids is composed of one Bacteroidetes from power node (PN) 1, one Actinobacteria from PN 2, the Firmicutes member 3, one Proteobacteria from PN 4 and finally the two Firmicutes and the Proteobacteria from PN 5. For all the target groups of this study, the large enumerations can be summarised with a boolean formula derived from the graph compressions. For instance for the lipids of ***Figure 3*** a, the community composition as described above is the following:

$$(\vee PN1) \wedge (\vee PN2) \wedge (PN3) \wedge (\vee PN4) \wedge (\wedge PN5).$$

We further investigated the essential symbionts associated to carbohydrate-derived metabolites in our study: *Paenibacillus polymyxa*, *Lactobacillus lactis*, *Bacillus licheniformis*, *Lactobacillus plantarum*, and *Dorea longicatena*. Interestingly, out of these five species, the first four have already been studied in the context of probiotics, for animals or humans (***Cutting, 2011***; ***Monteagudo-Mera et al., 2012***). In particular, the study of *P. polymyxa* CAZymes demonstrated its ability to assist in digesting complex carbohydrates (***Soni et al., 2020***). *L. plantarum* is also known for its role in carbohydrate acquisition (***Vries et al., 2006***; ***Marco et al., 2010***). The present analysis illustrates that within the full genome collection, these species are likely to exhibit functions related to carbohydrate synthesis and degradation that are not found in other species. *Bacillus licheniformis* is also an essential symbiont for the lipid metabolites. Among essential symbionts for other groups of metabolites, *Burkholderiales bacterium* (Proteobacteria) and *Hungatella hathewayi* (Firmicutes) have the particularity of both occurring in predictions for the lipids, carboxy acids and aromatic metabolites. This suggest a metabolism for these two species that differs from the other species, with non-redundant contributions to some metabolites of these categories. While *Hungatella hathewayi* is a relatively frequent gut commensal, little is known about this species (***Manzoor et al., 2017***). The *Burkholderiales* order is also poorly known, but its ability to degrade a numerous aromatic compounds has been established (***Pérez-Pantoja et al., 2012***). Finally, the only essential symbiont predicted for the coA-related metabolites is *Fusobacterium varium*, a butyrate producer known for its ability to ferment both sugars and amino acids (***Potrykus et al., 2008***).

Altogether, computation of key species coupled to the visualisation of community compositions enables a better understanding of the associations of organisms into the minimal communities. In this genome collection, groups of equivalent GSMNs allow us to identify genomes that are providing specialised functions to the community, enabling metabolic pathways leading to specific end-products.

M2M is suited to the metabolic analysis of metagenome-assembled genomes

Comparison of M2M applications to MAGs and draft reference genomes

In order to compare the effect of genome quality on M2M predictions, we performed analyses on a collection of 913 MAGs binned from cow rumen metagenomes (***Stewart et al., 2018***). These MAGs were predicted to be > 80% complete and < 10% contaminated. The complete M2M workflow ran in 81 minutes on a cluster with 72 CPUs and 144 Gb of memory.

Results of the GSMN reconstruction are presented in ***Table 2***. GSMNs of the cow rumen MAGs dataset consisted in average of 1,155 (± 199) reactions and 1,422 (± 212) metabolites. 73.8% of the reactions could be associated to genes. We compared these numbers with those obtained for the collection of draft reference genomes from the human gut microbiota of the previous subsection. ***Appendix 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 (***Appendix 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. This could be explained by the higher phylogenetic diversity observed in MAGs compared to culturable bacteria, or a higher potential for contaminated genomes, or a difference in average genome size. However, the smallest GSMN size is observed in the rumen (340 reactions vs 617 for the smallest GSMN of the gut dataset). 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. This is consistent with the high quality scores of the MAGs described in the original publication: the 913 MAGs exhibited a CheckM (***Parks et al., 2015***) completeness score between 80% and 100% (average: 90.61%, standart deviation: ±5.26%, median: 91.03%) (***Stewart et al., 2018***).

M2M modelling analyses were run on the reconstructed GSMNs. ***Appendix 1*** (j and k) displays the distributions of the individual scopes for each GSMN. We identified a cooperation potential of 296 metabolic end-products only reachable through the community. The minimal community consisted of 44 GSMNs, sufficient to produce all metabolites reachable through cooperation in the initial community composition. These could be described through 127 key species, consisting of 20 essential symbionts and 107 alternative ones. This indicates that each equivalent minimal community for these compounds would consist in the same 20 GSMNs, associated to 24 others selected within the 107 alternative species, thereby reaching a total of 44 GSMNs. Results are displayed in Supplementary File 1 (Table 4), together with those of an equivalent analysis (default settings of M2M with cooperation potential as targets) for the gut reference genomes.

M2M robustly identifies key species, even with degraded genomes

A recurring concern in metagenomics is the completeness of reconstructed MAGs due to the possible loss of functions during the genome assembly process (***Parks et al., 2015***). Misidentified genes can impede GSMN reconstruction and consequently the contents of the scopes and cooperation potential. To assess the impact of MAG completeness, we altered the rumen MAGs dataset by randomly removing genes. We created four altered datasets by removing: i) 2% of genes in all genomes, ii) 5% of genes in 80% of the genomes, iii) 5% of genes in all genomes and iv) 10% of genes in 70% of the genomes. We analysed these degraded datasets with the same M2M bioinformatic workflow, using a community selection with the metabolic cooperation potential as a community objective.

The metabolic cooperation potential, the global set of reachable metabolites in the community and the key species (essential and alternative symbionts) were computed and compared between the four altered datasets and the original one. Results are depicted in ***Figure 2***. The global set of producible metabolites in the community and the contents of the metabolic cooperation potential remain stable between datasets. In both we observe a single subset of 36 metabolites that is only reachable by the original dataset. It consists in a variety of metabolites mostly from secondary metabolism. Discrepancies appear between datasets when studying key species with respect to the original dataset, with an overall stability of the datasets to 2% degradation and to 5% degradation in 80% of genomes. The main discrepancies are observed for alternative symbionts with an additional small set of symbionts that are selected in altered datasets but not in the original one. For the most degraded genomes (10% degradation in 70% of MAGs), key species composition is altered compared to original genomes: a set of 31 key species is no longer identified. However, producibility

analyses and community selections performed by M2M are stable to small genome degradations of up to 2% of random gene loss in all genomes or 5% in 80% of the genomes. Altogether, ***Figure 2*** illustrates relative stability of the information computed by M2M to missing genes. The criteria typically used for MAG quality (>80% completeness, <10% degradation) are likely sufficient to get a coarse-grained, yet valuable first picture of the metabolism.. This robustness in our algorithms could be explained by the fact that i) missing genes in degraded MAGs may not be related to metabolism, ii) by the reported stability of the network expansion algorithm to missing reactions (***Handorf et al., 2005***), and iii) by the fact that multiple genes can be associated to the same metabolic reaction (redundancy in pathway representation). Additional analyses (***Appendix 1***) enable the refutation of the first hypothesis as the average gene loss in metabolic networks is similar to the genomic loss. Yet, the percentage of reactions associated to genes is similar in every experiments, which goes in the direction of the redundancy loss hypothesis. Likewise, we observed that the loss in reactions for degraded genomes is lower than the loss of genes.

Application of M2M to human shotgun metagenomic data from diabetic and healthy individuals

Protocols and cohort effect

In order to illustrate the applicability of M2M to metagenomic samples and cohorts of individuals, we reused the work presented in (***Diener et al., 2020***) and analysed the gut metagenomes of 170 individuals from a Danish (MHD) cohort and a Swedish (SWE) cohort (***Forslund et al., 2015***) in the context of Type-1 (T1D) and Type-2 (T2D) diabetes. Based on species-level dereplicated MAGs, metagenomic species (MGS), we built GSMNs and bacterial communities for each individual. We relied only on the available metagenomic data to perform analyses, and used qualitative information (presence/absence of MGS or species in the sample) to build the communities as M2M works with qualitative information. Two experiments were performed: M2M was run on each sample firstly using communities of newly reconstructed GSMN from MGS, and secondly using communities consisting of curated GSMNs of the AGORA resources (***Magnúsdóttir et al., 2016***) mapped to OTUs at the species level as described in ***Diener et al.*** (***2020***). 778 MGS were retrieved from the dataset and used to build GSMNs (***Table 2***), whereas when using the mapping of OTUs to existing curated GSMNs, only 289 GSMNs were used. The distribution of phyla in the two cases is illustrated in ***Appendix 2*** (Figure 2). We first focus on the results obtained with the MGS-based protocol.

In average, communities were composed of 108 ($\pm$ 29) GSMNs. The median community size was 111 (Supplementary File 1 - Table 20). Diversity and richness analyses are available in ***Appendix 2*** The effect of the cohort (MHD, SWE) was strong in the analyses performed with MGS, impacting community sizes, size and composition (in families of metabolites) of the set of metabolites producible by the community, as well as the cooperation potential. Results are depicted in ***Appendix 2***(Figure 3). A classification experiment using the composition of the community scope or the composition of the cooperation potential can efficiently determine the cohort of the samples (***Appendix 2*** figure 3 panels c and g). Similar differences between cohorts were also observed in (***Diener et al., 2020***) and in (***Forslund et al., 2015***) based on functional or taxonomic annotations, likely driven by the different sampling protocols used in the two datasets (***Forslund et al., 2015***). This indicates that the commonly observed cohort effect in metagenomics is also reflected at the metabolic modelling scale, which could be explained by the observed GSMN redundancies shared within phyla.

Impact of the disease status

We studied the impact of the disease status on the community metabolism for the 115 samples of the MHD cohort. The community diversity varied between disease statuses, with a significantly higher number of MGS observed in T1D individuals forming the initial communities (anova $F(2,112) = 8.346$, $p < 0.01$, eta-squared = 0.13, Tukey HSD test $p < 0.01$ vs control). We observe that the distribution of the community sizes is broader for control individuals. The higher diversity for diseased individuals is reflected at the metabolic level through the putative producibility of a wider

set of metabolites for T1D (anova $F(2,112) = 6.606$, $p < 0.01$, eta-squared = 0.11, Tukey HSD $p < 0.01$ vs control) and to a lesser extent for T2D communities (Tukey HSD $p = 0.05$ vs control). The putative producibility of some families of metabolites (alcohols, esters, carbohydrates, amino-acids, acids) in the community scopes also differed between metabolic communities derived from diseased and healthy individuals (anova $p < 0.05$), whereas other metabolic families like lipids remained stable between cohorts. This can be at least partly explained by the number of metabolites matching these categories according to the Metacyc database (e.g. 191 metabolites tagged "All-carbohydrates" in average in community scopes, and only 10 tagged "Lipids" as the remaining of them are scattered in other categories). No clear difference appears between the three statuses (***Figure 4*** e) in terms of community scope composition. Regarding the cooperation potential, two groups tend to appear, separated due to diverse secondary metabolites, but they are not driven by the disease status of the individual (***Figure 4*** f). A classification experiment on the composition of the community scope can, to some extent, (AUC = 0.75 ± 0.15) decipher between healthy or diabetes statuses (***Figure 4*** d) but classification between T1D and T2D was not achievable (***Appendix 2*** Figure 4). Although metagenomic data would more precisely perform such a separation, it is informative to observe that despite metabolic redundancy in the gut microbiota, there are differences at the metabolic modelling level. Qualitative differences are noticeable between healthy and diabetic individuals: it is possible to distinguish them to some extent using the set of metabolites predicted to be producible by the microorganisms found in their faeces.

We then computed for each sample the key species (essential and alternative symbionts) associated to the cooperation potential. The ratio of key species (KS), essential symbionts (ES) and alternative symbionts (AS) with respect to the initial community size did not vary altogether between statuses. The exception was the ratio of AS (and of KS, which include AS) when comparing diabetes individuals and controls, differences that were not significant when distinguishing the two types of diabetes. Comparing the phylum-level taxonomy of these putative key species, in the initial communities, we noted that the occurence of Firmicutes was broader compared to other phyla (Bacteroidota, Proteobacteria, Actinobacteria). Firmicutes are known to be phylogenetically diverse (***Costea et al., 2018***), and therefore their combined metabolism could also be more diverse. A notable change is the narrower distribution of Bacteroidota in the initial communities as well as in selected symbionts in diseased individuals compared to control (***Figure 4*** g). Altogether, no clear trend was observable from metabolic modelling analyses between disease states, but we observed some difference for the taxonomic composition of minimal communities, which could be explained by the diversity discrepancies in microbiome compositions (***Forslund et al., 2015***).

Focus on short chain fatty acids production

Given the importance of short chain fatty acids (SCFAs) in human health (***Baxter et al., 2019***), we focused on the production of butyrate, propionate and acetate in communities for each sample of the dataset. A small number of MGS (N = 11) GSMNs were predicted to be able to individually produce butyrate from the nutrients. All 778 MGS were capable to ferment acetate and most of them propionate (N = 515). The putative production of butyrate in the 170 communities when allowing cooperation between GSMNs was systematic. As expected (***Rivière et al., 2016***), a majority (54.1%) of the unique MGS predicted as possible butyrate producers in communities (GSMNs comprising a reaction producing butyrate that could be activated in a community) belonged to the Firmicutes phylum. Altogether, in 62.6% of cases, the putative butyrate producers observed in the communities were Firmicutes. We compared the number of putative butyrate producers in communities from MHD samples according to the disease status of the individuals. Their number was significantly higher in the communities of T1D individuals compared to control and T2D (anova $F(2,111) = 9.27$, $p < 0.01$, eta-squared = 0.14, and Tukey HSD test $p < 0.01$ vs control and $p = 0.02$ vs T2D) which could be explained by the higher MGS diversity observed in T1D communities compared to the others. We then analysed the difference between using GSMNs of MGS reconstructed from metagenomic data and using curated GSMNs mapped at the species level to OTUs as performed

in *Diener et al.* (*2020*). The same increase in butyrate producers was observed when running M2M on MHD communities consisting of the mapped AGORA GSMNs (anova $F(2,112) = 5.368$, $p < 0.01$, eta-squared = 0.11, and Tukey HSD test $p < 0.01$ vs control). To conclude, similar to the analyses of *Diener et al.* (*2020*), we observe that the producibility of SCFAs, particularly butyrate, is highly driven by cooperation in the microbial communities of individuals and can be performed by heterogenous sets of commensal species. The MGS-driven approach and the systematic GSMN reconstruction permit taking advantage of the whole metagenomic information and capturing the metabolic complementarity in each sample.

Discussion

Metage2Metabo is a new software system for the functional analysis of metagenomic datasets at the metabolic level. M2M can be used as an all-in-one pipeline or as independent steps to depict an initial picture of metabolic complementarity within a community. It connects directly to metagenomics through the automation of GSMN reconstruction, and integrates in this collective analysis community reduction with respect to targeted functions. M2M was applied to a large collection of gut microbiota reference genomes, demonstrating the scalability of the methods and how it can help identifying equivalence classes among species for the producibility of metabolite families. We showed that metabolic networks reconstructed from reference genomes and MAGs display similar characteristics and that M2M modelling predictions are robust to missing genes in the original genomes. Finally, application to real metagenomic samples of individuals demonstrated that qualitative modelling of metabolism retrieves known features from metagenomics and quantitative modelling analyses. M2M provides a first order analysis with a minimal cost in terms of required data and computational effort.

The identification of cornerstone taxa in microbiota is a challenge with many applications, for instance restoring balance in dysbiotic environments. Keystone species, a concept introduced in ecology, are particularly looked for as they are key drivers of communities with respect to functions of interest (*Banerjee et al., 2018*). There is a variety of techniques to identify them (*Carlström et al., 2019*; *Floc'h et al., 2020*), and computational biology has a major role in it (*Fisher and Mehta, 2014*; *Berry and Widder, 2014*). The identification of alternative and essential symbionts by M2M is an additional solution to help identify these critical species. In particular, essential symbionts are close to the concept of keystone species as they are predicted to have a role in every minimal community associated to a function. Additionally, alternative species and the study of their combinations in minimal communities, for example with power graphs, are also informative as they reveal equivalence groups among species.

M2M functionally analyses large collections of genomes in order to obtain metabolic insights into the metabolic complementarity between them. While the functionality of metagenomic sequences is commonly analysed at higher levels by directly computing functional profiles from reads (*Franzosa et al., 2018*; *Silva et al., 2016*; *Sharma et al., 2015*; *Petrenko et al., 2015*), the metabolic modelling oriented approach provides more in-depth predictions on reactions and pathways organisms could catalyse in given environmental conditions. M2M answers to the upscaling limitation of individual GSMN reconstruction with Pathway Tools by automating this task using the Mpwt wrapper. GSMNs in SBML format obtained from other platforms such as Kbase *Arkin et al.* (*2018*), ModelSEED (*Henry et al., 2010b*; *Seaver et al., 2020*) or CarveMe (*Machado et al., 2018*), can also be used as inputs to M2M for all metabolic analyses. For instance, we used highly curated models from AGORA in the application of M2M to metagenomic datasets. The above reconstruction platforms already implement solution to facilitate the treatment of large genomic collections. There is no universal implementation for GSMN reconstruction (*Mendoza et al., 2019*); depending on their needs (local run, external platform, curated or non-curated GSMNs...), users can choose either method and connect it to M2M.

Most metabolic modelling methods rely on flux analyses (*Orth et al., 2010*) solved with linear programming, which may turn out to be challenging to implement for simulations of large communi-

ties (**Basile et al., 2020**), although recent efforts in that direction are encouraging (**Popp and Centler, 2020**). M2M uses the network expansion algorithm and solves combinatorial optimisation problems with Answer Set Programming, thereby ensuring fast simulations and community predictions, suitable when performing systematic screening and multiple experiments. Network expansion has been widely used to analyse and refine metabolic networks (**Matthäus et al., 2008**; **Laniau et al., 2017**; **Christian et al., 2009**; **Prigent et al., 2017**), including for microbiota analysis (**Christian et al., 2007**; **Ofaim et al., 2017**; **Opatovsky et al., 2018**; **Frioux et al., 2018**). Network expansion is a complementary alternative to quantitative constraint-based methods (**Ebenhöh et al., 2004**; **Handorf et al., 2005**) such as flux balance analysis as it does not require biomass reactions nor accurate stoichiometry. This algorithm offers a good trade-off between the accuracy of metabolic predictions and the precision required for the input data, adapted to the challenges in studying non-model organisms and their likely incomplete models of metabolism (**Bernstein et al., 2019**).

Answer Set Programming can easily scale the analysis of minimal communities among thousands of networks considered in interaction and ensures with efficient solving heuristics that the whole space of solutions is parsed to retrieve key species for chosen end-products. M2M therefore suggests (metagenomic) species for further analyses such as targeted curation of metabolic networks and deeper analysis of the genomes or quantitative flux predictions. MiSCoTo (**Frioux et al., 2018**), the algorithm for minimal community selection used in M2M, has been recently experimentally demonstrated. It was applied to design bacterial communities to support the growth of a brown alga in nearly axenic conditions (**Burgunter-Delamare et al., 2019**). Despite the difficulty inherent to controlling the communities for a complex alga, the inoculated algae exhibited a significant increase in growth and metabolic profiles that at least partially aligned with the predictions, demonstrating the versatility in application fields of our methods.

There are limitations associated to the software solution described in this paper. One challenge in applying our tool is to accurately estimate the nutrients available in a given environment (seeds), on which the computation of network expansion relies. The algorithm provides a snapshot of producible metabolites, representing the sub-network that can be activated under given nutritional conditions. However, network expansion has been shown to be sensitive to cycles in GSMNs and it is therefore relevant to include some cofactors (or currency metabolites e.g. ADP) in the seeds to activate such cycles, the way many studies proceed (**Cottret et al., 2010**; **Greenblum et al., 2012**; **Eng and Borenstein, 2016**; **Julien-Laferrière et al., 2016**). In addition, it has to be noted that the cost of exchanges or their number are not taken into account in M2M. Transport reactions are hardly recovered by automatic methods (**Bernstein et al., 2019**) and validation of cross-feedings implies an additional work on transporters identification. The standalone MiSCoTo package used in M2M has a solving mode taking into account exchanges: it can compute communities while minimising and suggesting metabolic exchanges, although this comes with additional computational costs and a need for validation. Another limitation of our approach for studying communities of individuals from metagenomic experiments is that we do not take the microbial load or abundance of MGS into account in the pipeline. Considering the presence/absence of MGS might lead to overestimate the production of some metabolites. In addition, we infer phenotypes directly from genotypes, thereby ignoring the possible non-expression of metabolic-related genes in specific conditions and the regulation of those genes. Metaproteomic and metatranscriptomic data could partly overcome these shortcomings but such experiments are not yet routinely performed. Finally, another aspect to be considered in the future is the competition between species, especially for nutrients, as we only focus here on metabolic complementarity and positive interactions.

Despite the above mentioned limitations, M2M has multiple applications for the *de novo* screening of metabolism in microbial communities. The number of curated GSMNs for species found in microbiotas increases (**Magnúsdóttir et al., 2016**), constituting a highly valuable resource for the study of interactions by mapping metagenomic data or OTUs to the taxonomy of genomes associated to these GSMNs. Yet, the variety of (reference) genomes obtained from shotgun metagenomic experiments is such than species and strains may not belong to the ones for which a curated GSMN

is available. In that case, the proportion of reads that are not mapped to a genome with an associated GSMN can be very high (***Diener et al., 2020***). In addition, predictions from GSMN mapping can be misleading as it is known that genomes vary a lot between genera, species, and even strains, (***Ansorge et al., 2019***) and so can the metabolism. Recent methods for assembling genomes directly from metagenomes lead to nearly complete genomes for possibly unknown species on which one may still want to get metabolic insights (***Almeida et al., 2019***). Long-reads sequencing associated with short-reads sequencing can also give access to complete microbial genomes (***Moss et al., 2020***). Finally, single-cell methods can be useful for the acquisition of genomes and metagenomes (***Treitli et al., 2019***). M2M answers to the need for *de novo* metabolic inference and screening, which is likely to become a routine in the rapidly evolving context of microbiota genome sequencing. While studying the metabolic potential of large communities is an iterative process that still requires biological expertise, we provide with this work means to facilitate the screening of metagenomes and reduce these large communities to key members.

Conclusion

Metage2Metabo (M2M) allows metabolic modelling of large-scale communities, based on reference genomes or *de novo* constructed MAGs, inferring metabolic complementarity found within communities. M2M is a flexible framework that automates GSMN reconstruction, individually and collectively analyse GSMNs, and performs community selection for targeted functions. The large combinatorics of minimal communities due to functional redundancy in microbiotas is addressed by providing key species associated to metabolic end-products. This could allow targeting specific members of the community through pro- or prebiotics, to model the metabolites the human host will be exposed to.

We validated the flexibility of the software and the range of analyses it can offer with several datasets, corresponding to multiple use-cases in the microbiome field. This allowed us to characterise metabolic complementarity in a large collection of draft reference genomes. We further assessed the robustness of M2M to data incompleteness by performing analyses on collections of MAGs. Finally, we applied M2M to a common use-case in metagenomics: the study of communities associated to individuals, in a disease context.

Our method is robust against the uncertainty inherent to metagenomics data. It scales to typical microbial communities found in the gut and predicts key species for functions of interest at the metabolic level. Future developments will broaden the range of interactions to be modelled and facilitate the incorporation of abundance data. This software is an answer to the need for scalable predictive methods in the context of metagenomics where the number of available genomes continues to rise.

Material and Methods

Metage2Metabo (M2M) is a Python package. It can be used on a workstation or on a cluster using Docker or Singularity. M2M's source code is available on github.com/AuReMe/metage2metabo, and the package is available though the Python Package Index at pypi.org/project/Metage2Metabo/. A detailed documentation is available on metage2metabo.readthedocs.io.

We detail below the characteristics of M2M through a description of its main steps.

Parallel and large-scale metabolic network reconstruction

M2M can process existing metabolic networks in SBML format or proposes the automatic reconstruction of non-curated metabolic networks (`m2m recon`). As a multi-processing solution, it facilitates the treatment of hundreds or thousands of genomes that can be retrieved from metagenomic experiments. The underlying GSMN reconstruction software is Pathway Tools (***Karp et al., 2016***), a graphical user interface (GUI) based software suite for the generation of individual GSMNs, called Pathway/Genome Databases (PGDBs). Typically, a PGDB is obtained from an annotated

genome using PathoLogic, the software prediction component of Pathway Tools, and curated afterwards.

We developed Mpwt[1] (Multiprocessing Pathway Tools), a command-line Python wrapper for Pathway Tools. Mpwt and M2M i) format the genomic inputs, ii) automate the reconstruction step by initialising a PathoLogic environment for each genome, and iii) extract and converts the resulting GSMNs in PGDB and SBML (***Hucka et al., 2003***, ***2018***) formats using the PADMet library (***Aite et al., 2018***). Mpwt handles three types of genomic inputs (Genbank, Generic Feature Format (GFF) or PathoLogic format) that must contain GO-terms and EC-numbers annotations necessary for Pathway Tools. These annotations are for example found in the Genbank files generated by Prokka (***Seemann, 2014***). In addition, we specifically developed Emapper2gbk, a Python package dedicated to the connection between the Eggnog-mapper annotation tool (***Huerta-Cepas et al., 2017***) and Mpwt in order to generate these inputs.

Analysis of metabolic producibility and calculation of the cooperation potential

This part of the workflow encompasses three steps: computation of the i) individual (`m2m iscope`) and ii) collective (`m2m cscope`) metabolic potentials, and iii) the characterisation of the cooperation potential of the GSMN collection (`m2m addedvalue`). The former two rely on the network expansion algorithm (***Ebenhöh et al., 2004***), the latter being a set difference between the results of the first two steps.

The network expansion algorithm computes the *scope* of a metabolic network from a description of the growth medium called *seeds*. The scope consists in the set of metabolic compounds which are reachable, or producible, according to a boolean abstraction of the network dynamics assuming that cycles cannot be self-activated. More precisely, the algorithm recursively considers products of reactions to be producible if all reactants of the reactions are producible, provided an initiation with a set of seed nutrients. The underlying implementation of the network expansion algorithm used in M2M relies on Answer Set Programming (ASP) (***Schaub and Thiele, 2009***).

We define a metabolic network as a bipartite graph $G = (R \cup M, E)$, where R and M stand for reaction and metabolite nodes. When $(m, r) \in E$ (respectively $(r, m) \in E$), with $m \in M$ and $r \in R$, the metabolite is called a *reactant* (respectively *product*) of the reaction r. The scope of a set of seed compounds S according to a metabolic network G, denoted by $\mathrm{scope}(G, S)$, is iteratively computed until it reaches a fixed point (***Handorf et al., 2005***). It is formally defined by

$$\mathrm{scope}(G, S) = \bigcup_i M_i, \text{ where } M_0 = S \text{ and } M_{i+1} = M_i \cup products(\{r \in R \mid reactants(r) \subseteq M_i\}).$$

Individual metabolic capabilities

The `m2m iscope` command predicts the set of reachable metabolites for each GSMN using the network expansion algorithm and the given nutrients as seeds. The content of each scope is exported to a json file. A summary is also provided to the user comprising the intersection (metabolites reachable by all GSMNs) and the union of all scopes, as well as the average size of the scopes, the minimal size and the maximal size of all. This command extends core functions implemented in Menetools[2], a Python package that was previously used in ***Aite et al.*** (***2018***).

Collective metabolic capabilities

The `m2m cscope` command computes the metabolic capabilities of the whole microbiota by taking into account the complementarity of metabolic pathways between GSMNs. This step simulates the sharing of metabolic biosynthesis through a meta-organism composed of all GSMNs, and assesses the metabolic compounds that can be reached using network expansion. This calculation is an extension of the features of MiSCoTo[3] (***Frioux et al., 2018***) in which the collective scope of a

[1] also available as a standalone tool
[2] also available as a standalone tool for individual GSMNs
[3] also available as a standalone tool

collection of metabolic networks $\{G_1, \dots G_N\}$ is introduced. We define

$$\text{collectiveScope}(G_1..G_N, S) = \text{scope}\left(\left(\bigcup_{i\in\{1..n\}} R_i, \bigcup_{i\in\{1..n\}} M_i, \bigcup_{i\in\{1..n\}} E_i\right), S\right).$$

Target producers

If metabolic compounds of interest or *targets* are provided by the user, a summary of the producers for each target is generated by `m2m workflow`, `m2m metacom`, and `m2m cscope`: it identifies the GSMNs that are predicted to produce the targets, either intrinsically, or through cooperation with other members of the community.

A metabolic network G_i is an *individual target producer* of $t \in T$ if $t \in \text{scope}(G_i, S)$. The metabolic network G_i is a *community target producer* if *(a)* G_i is not an individual target producer of t (i.e. $t \notin \text{scope}(G_i, S)$), but *(b)* G_i contains a reaction $r \in R_i$ which produces t (i.e. $t \in products(r)$) such that *(c)* all reactants are producible by the community (G_i and the other metabolic networks): $reactants(r) \subset \text{collectiveScope}(G_1..G_N, S)$. This means that the metabolic network G_i has the capability of producing t through the reaction r in a cooperation context.

This information can be retrieved in practice in the file "producibility_targets.json" under the keys "individual_producers" and "com_only_producers".

Cooperation potential

Given individual and community metabolic potentials, the *cooperation potential* consists in the set of metabolites whose producibility can only occur if several organisms participate in the biosynthesis. `m2m addedvalue` computes the cooperation potential by performing a set difference between the community scope and the union of individual scopes, and produces a SBML file with the resulting metabolites. This list of compounds is inclusive and could comprise false positives not necessitating cooperation for production, but selected due to missing annotations in the initial genomes. One can modify the SBML file accordingly, prior to the following M2M community reduction step.

The cooperation potential $\text{cooperationPotential}(G_1, .., G_n, S)$ of a collection of metabolic networks $\{G_1..G_n\}$ is defined by

$$\text{cooperationPotential}(G_1, .., G_n, S) = \text{collectiveScope}(G_1, .., G_n, S) \setminus \bigcup_{i\in\{1..n\}} \text{scope}(G_i, S).$$

Computation of minimal communities and identification of key species

A minimal community $\mathcal{C}$ enabling the producibility of a set of targets T from the seeds S is a sub-family of the community $G_1, \dots, G_n$ which is solution of the following optimisation problem:

$$\begin{aligned} &\underset{\{G_{i_1}..G_{i_L}\} \subset \{G_1..G_N\}}{\text{minimize}} && \text{size}(\{G_{i_1}..G_{i_L}\}) \\ &\text{subject to} && T \subset \text{collectiveScope}(G_{i_1}..G_{i_L}), S). \end{aligned}$$

Solutions to this optimisation problem are communities $\mathcal{C} = (G_{i_1} \dots, G_{i_L})$ of minimal size. We define $\text{minimalCommunities}(G_1..G_n, S, T)$ to be the set of all such minimal communities. A first output of the `m2m mincom` command is the (minimal) size L of communities solution of the optimisation problem. The composition of one optimal community is also provided. The targets are by default the components of the cooperation potential, $T = \text{cooperationPotential}(G_1, .., G_n, S)$, but can also be a group of target metabolites defined by the user.

Many minimal communities are expected to be equivalent for a given metabolic objective but their enumeration can be computationally costly. We define *key species* which are organisms occurring in at least one community among all the optimal ones. Key species can be further distinguished into *essential symbionts* and *alternative symbionts*. The former occur in every minimal community whereas the latter occur only in some minimal communities. More precisely, the key species $\text{keySpecies}(G_1..G_n, S, T)$, the essential symbionts $\text{essentialSymbionts}(G_1..G_n, S, T)$, and the

alternative symbionts $\text{alternativeSymbionts}(G_1..G_n, S, T)$ associated to a set of metabolic networks, seeds S and a set of target metabolites T are defined by

$$\text{keySpecies}(G_1..G_n, S, T) = \{G \mid \exists C \in \text{minimalCommunities}(G_1..G_n, S, T),\ G \in C\}.$$
$$\text{essentialSymbionts}(G_1..G_n, S, T) = \{G \mid \forall C \in \text{minimalCommunities}(G_1..G_n, S, T),\ G \in C\}.$$
$$\text{alternativeSymbionts}(G_1..G_n, S, T) = \text{keySpecies}(G_1..G_n, S, T) \setminus \text{essentialSymbionts}(G_1..G_n, S, T).$$

As a strategy layer over MiSCoTo, M2M relies on the Clasp solver (***Gebser et al., 2012***) for efficient resolution of the underlying grounded ASP instances. Although this type of decision problem is NP-hard (***Julien-Laferrière et al., 2016***), as with many real-world optimisation problems worst-case asymptomatic complexity is less informative for applications than practical performance using heuristic methods. The Clasp solver implements a robust collection of heuristics (***Gebser et al., 2007***; ***Andres et al., 2012***) for core-guided weighted MaxSAT (***Manquinho et al., 2009***; ***Morgado et al., 2012***) that provide rapid set-based solutions to combinatorial optimisation problems, much in the same way that heuristic solvers like CPlex provide rapid numerical solutions to mixed integer programming optimisation problems. The kinds of ASP instances constructed by MiSCoTo for M2M are solved in a matter of minutes for the identification of key species and essential/alternative symbionts. Indeed the space of solutions is efficiently sampled using adequate projection modes in ASP, which enables the computation of these groups of species without the need for a full enumeration by taking advantage of the underlying ASP solver and associated projection modes.

Analysis of enumerated communities

The `m2m-analysis` command permits the enumeration of minimal communities. If the taxonomy of species associated to the metabolic networks is provided, descriptive statistics are performed. In addition, minimal communities can be visualised as an association graph connecting GSMNs that co-occur in at least one minimal community. The association graph can itself be compressed in a power graph that enables visualising motifs such as cliques, bicliques and stars. Power graphs are generated using PowerGrASP (***Bourneuf and Nicolas, 2017***). In this paper, they were visualised with Cytoscape (v.2.8.3) (***Shannon et al., 2003***) and the CyOog plugin (v.2.8.2) developed by ***Royer et al.*** (***2008***).

Application to datasets

Analysis of human gut and cow rumen published collections of genomes

In order to evaluate the influence of genome collections based on sequencing cultured isolates or metagenomic genome reconstructions, we used 1,520 high-quality draft reference genomes of bacteria from the human gut microbiota retrieved from ***Zou et al.*** (***2019***) and 913 MAGs from the cow rumen published in ***Stewart et al.*** (***2018***). The genomes from the former set were already annotated.

We designed a set of seed metabolites representing a nutritional environment which is required for the metabolic modelling analyses. Seeds (93 metabolites) consist in components of a classical diet for the gut microbiota, EU average from the VMH resource (***Noronha et al., 2018***), and a small number of currency metabolites (***Schilling et al., 2000***) (Supplementary File 1 - Table 1 and Github repository[4] of M2M). M2M was run using version 23.0 of Pathway Tools.

The cow rumen dataset of MAGs was not functionally annotated. Therefore, as a preliminary step of analysis, we annotated the genomic contigs using Prokka (v.1.13.4) (***Seemann, 2014***). M2M was run using version 23.0 of Pathway Tools. The nutritional environment for modelling experiments consisted in basic nutrients: 26 metabolites including inorganic compounds, carbon dioxide, glucose and cellobiose and a small number of currency metabolites (Supplementary File 1 - Table 2).

The rumen MAGs were artificially degraded to assess the robustness of M2M with respect to incomplete MAGs. This was done by randomly removing genes in all or a fraction of genomes. Four

[4] https://github.com/AuReMe/metage2metabo/tree/master/article_data

degradation scenarios were tested: removal of 2% of genes in all MAGs, removal of 5% of genes in 80% of the genomes, removal of 5% of genes in all genomes and removal of 10% of genes in 70% of the genomes. The subsequent parts of the analysis (annotation with Prokka, M2M runs) were done as described above. Supervenn diagrams presented in ***Figure 2*** to compare the results were obtained using the Supervenn Python package [5].

Shotgun metagenomic analysis of individuals

Metagenomic shotgun data from samples previously studied in ***Diener et al.*** (***2020***) from 186 Danish and Swedish individuals (***Forslund et al., 2015***) were used in this paper. Genomes were *de novo* reconstructed from the dataset using the MATAFILER pipeline described in ***Hildebrand et al.*** (***2019***). Briefly, metagenomic samples were quality-filtered using sdm (***Hildebrand et al., 2014***), assembled using MEGAHIT (***Li et al., 2015***), genes were predicted using Prodigal (***Hyatt et al., 2010***), and a non-redundant gene catalogue was constructed across all samples using MMseqs2 (***Steinegger and Söding, 2017***). MAGs were predicted from metagenomic assemblies using MetaBAT2 (***Kang et al., 2019***) and dereplicated into species level metagenomic species (MGS), using a combination of shared genes among MetaBAT2 bins, canopy clustering (***Nielsen et al., 2014***) and custom R scripts (***Hildebrand et al., 2019***). Abundance of MGS was estimated across samples by using the average coverage of 40 conserved, single copy marker genes associated to each MGS (***Mende et al., 2013***). This abundance matrix was further populated with specI species from the proGenomes database (***Mende et al., 2019***), that were not represented by MGS and are high quality genomes from cultured bacteria. This pipeline is described in further detail in ***Hildebrand et al.*** (***2019***).

Samples for which the global estimated abundance of MGS was lower than 1,000 in accumulated coverage of all species were removed. This corresponds to a low number of reads passing the upstream quality checks for these samples (< 9.10e6). 170 samples were kept for analysis. The initial bacterial community of each sample was determined using the estimated abundance matrix provided by the MATAFILER pipeline, following a boolean rule of presence/absence of MGS and specI species in samples. Genomes consisted in MGS obtained with MATAFILER as well as SpecI genomes from the Progenomes database (***Mende et al., 2019***) that were identified in the samples. For the latter case, we downloaded the genes and proteins of the corresponding representative genome from the database (SpecI v3). Functional annotation of genes from both specI genomes and MGS core genomes was performed using EggNOG-mapper v2.0.0 (***Huerta-Cepas et al., 2017***) based on eggNOG orthology data (***Huerta-Cepas et al., 2019***). Sequence searches were performed using Diamond v0.9.24.125 (***Buchfink et al., 2014***). Treatment of EggNOG-mapper annotation and creation of M2M Genbank inputs was done with the package Emapper2gbk[6] that we developed for the project. Seeds describing the nutritional environment were compounds of the western diet as presented in the study by ***Diener et al.*** (***2020***). These metabolites were translated into identifiers from the Metacyc database (***Caspi et al., 2019***), to which were added a small number of currency metabolites. The seeds are available on the Github repository of M2M.

M2M was run for each sample and community selection was performed with different sets of targets (short-chain fatty acids, cooperation potential in each community). The cohort and disease status of each sample was known, enabling the comparison of scopes and cooperation potentials contents between statuses. M2M was also run using the approach presented in MICOM (***Diener et al., 2020***) building sample communities, as presented in the paper and associated data repository, through the attribution of curated GSMN (***Magnúsdóttir et al., 2016***) to operational taxonomic units (OTUs) identified in samples. We reused the mapping at species-level provided in the MICOM paper to build the communities.

Downstream analyses were performed in R (***Team, 2017***) and Python. Figures were produced using the package ggplot2 (***Wickham, 2009***) and diversity measures were computed with the vegan R package (v2.5-6). Classifications of disease statuses or cohorts using sets of predicted producible

[5] https://github.com/gecko984/supervenn
[6] https://github.com/AuReMe/emapper_to_gbk

Table 1. List and description of M2M commands

Command	Action
`m2m workflow`	Runs the whole m2m workflow
`m2m metacom`	Runs the workflow with already-reconstructed metabolic networks
`m2m recon`	Reconstructs metabolic networks using Pathway Tools
`m2m iscope`	Computes scopes for individual metabolic networks
`m2m cscope`	Computes the community scope
`m2m addedvalue`	Computes the cooperation potential
`m2m mincom`	Selects a minimal community and computes key species
`m2m seeds`	Creates a SBML file for nutrients
`m2m test`	Runs m2m workflow on a sample dataset
`m2m-analysis`	Runs additional analyses on community selection

Table 2. Results of the GSMN reconstruction step and metabolic potential analysis for the three datasets presented in the article (Avg = Average, "±" precedes standard deviation)

	Gut dataset	Rumen dataset	Diabetes dataset
initial data	draft reference genomes	MAGs	MAGs
number of genomes	1,520	913	778
GSMN reconstruction			
all reactions	3,932	4,418	5,554
all metabolites	4,001	4,466	5,386
avg reactions per GSMN	1,144 (± 255)	1,155 (± 199)	1,640 (± 368)
avg metabolites per GSMN	1,366 (± 262)	1,422 (± 212)	1,925 (± 361)
avg genes per mn	596 (± 150)	543 (± 107)	1,658 (± 469)
% reactions associated to genes	74.6 (± 2.17)	73.8 (± 2.61)	79.57 (± 1.60)
avg pathways per mn	163 (± 49)	146 (± 32)	220 (± 58)
metabolic potential			
number of seeds	93	26	175
avg scope per mn	286 (± 70)	101 (± 44)	508 (± 83)
union of individual scopes	828	368	1,326

metabolites were made using the Python package Scikit-learn (v0.23.1) (***Pedregosa et al., 2011***). Briefly, redundancy between features were removed with a Multidimensional scaling (MDS), Support Vector Machine (SVM) classifications with Stratified K-Folds cross-validations were performed using the MDS results. Finally, receiver operating characteristic curve (ROC-AUC) were computed and visualised with tools from the package.

Acknowledgments

The authors acknowledge the GenOuest bioinformatics core facility for providing the computing infrastructure. This research was supported in part by the NBI Computing infrastructure for Science (CiS) group. FH and CF's salaries have been funded by the BBSRC Institute Strategic Programme Gut Microbes and Health BB/r012490/1, its constituent project BBS/e/F/000Pr10355. AB's, CF,and MA's salaries have been funded by the ANR project IDEALG (ANR-10-BTBR-04) "Investissements d'Avenir, Biotechnologies-Bioressources". The authors 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. The authors also thank Lucas Bourneuf for his help on power graph analyses, Yann Le Cunff for his help regarding statistical analyses, and Samuel Blanquart and David Sherman for their useful comments on the manuscript.

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: key species, ES: essential symbionts, AS: alternative symbionts, Firm.: Firmicutes, Bact.: Bacteroidetes, Acti.: Actinobacteria, Prot.: Proteobacteria, Fuso.: Fusobacteria.

		Firm.	Bact.	Acti.	Prot.	Fuso.	total
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

References

Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, Mendoza SN, Carrier G, Dameron O, Guillaudeux N, Latorre M, Loira N, Markov GV, Maass A, Siegel A. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLoS Computational Biology. 2018 may; 14(5):e1006146. http://dx.plos.org/10.1371/journal.pcbi.1006146, doi: 10.1371/journal.pcbi.1006146.

Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. Nature. 2019 feb; p. 1. http://www.nature.com/articles/s41586-019-0965-1.

Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature biotechnology. 2020 jul; p. 1–10. http://www.ncbi.nlm.nih.gov/pubmed/32690973, doi: 10.1038/s41587-020-0603-3.

Andres B, Kaufmann B, Matheis O, Schaub T. Unsatisfiability-based optimization in clasp. In: Dovier A, Costa VS, editors. *Technical Communications of the 28th International Conference on Logic Programming (ICLP'12)*, vol. 17 of Leibniz International Proceedings in Informatics (LIPIcs) Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik; 2012. p. 211–221. http://drops.dagstuhl.de/opus/volltexte/2012/3623, doi: 10.4230/LIPIcs.ICLP.2012.211.

Ansorge R, Romano S, Sayavedra L, Porras MÁG, Kupczok A, Tegetmeyer HE, Dubilier N, Petersen J. Functional diversity enables multiple symbiont strains to coexist in deep-sea mussels. Nature Microbiology. 2019 oct; p. 1–11. http://www.nature.com/articles/s41564-019-0572-9, doi: 10.1038/s41564-019-0572-9.

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nature Biotechnology. 2018 jul; 36(7):566–569. http://www.nature.com/doifinder/10.1038/nbt.4163, doi: 10.1038/nbt.4163.

Banerjee S, Schlaeppi K, Heijden MGAvd. Keystone taxa as drivers of microbiome structure and functioning. Nature Reviews Microbiology. 2018; 16(9):567–576. doi: 10.1038/s41579-018-0024-1.

Basile A, Campanaro S, Kovalovszki A, Zampieri G, Rossi A, Angelidaki I, Valle G, Treu L. Revealing metabolic mechanisms of interaction in the anaerobic digestion microbiome by flux balance analysis. Metabolic Engineering. 2020; 62:138–149. doi: 10.1016/j.ymben.2020.08.013.

Baxter NT, Schmidt AW, Venkataraman A, Kim KS, Waldron C, Schmidt TM. Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers. mBio. 2019 jan; 10(1). doi: 10.1128/mBio.02566-18.

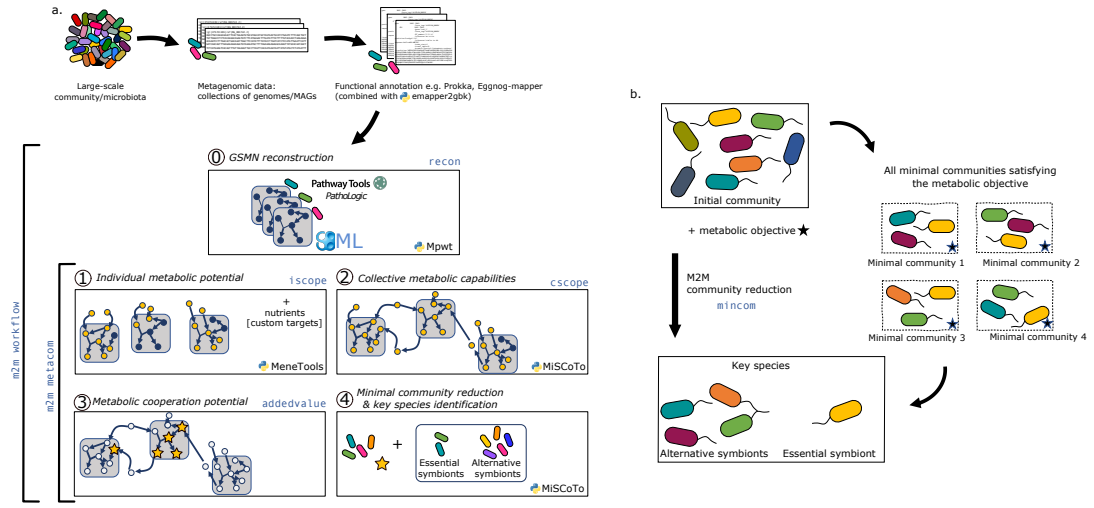


Figure 1. Overview of the M2M pipeline. *a.* Main steps of the M2M pipeline and associated tools. The software's main pipeline (`m2m workflow`) takes as inputs a collection of annotated genomes that can be reference genomes or metagenomics-assembled genomes. The first step of M2M consists in reconstructing metabolic networks with Pathway Tools (step 0). This first step can be bypassed and GSMNs can be directly loaded in M2M. The resulting metabolic networks are analysed to identify individual (step 1) and collective (step 2) metabolic capabilities. The added-value of cooperation is calculated (step 3) and used as a metabolic objective to compute a minimal community and key species (step 4). Optionally, one can customise the metabolic targets for community reduction. The pipeline without GSMN reconstruction can be called with `m2m metacom`, and each step can also be called independently (`m2m iscope`, `m2m cscope`, `m2m addedvalue`, `m2m mincom`). *b.* Description of key species. Community reduction performed at step 4 can lead to multiple equivalent communities. M2M provides one minimal community and efficiently computes the full set of species that occur in all minimal communities, without the need for a full enumeration, thanks to solving heuristics. It is possible to distinguish the species occurring in every minimal community (essential symbionts), from those occurring in some (alternative symbionts). Altogether, these two groups form the key species.

Bernstein DB, Dewhirst FE, Segre D. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. eLife. 2019 jun; 8. https://elifesciences.org/articles/39733, doi: 10.7554/eLife.39733.

Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Frontiers in Microbiology. 2014; 5:219. doi: 10.3389/fmicb.2014.00219.

Bourneuf L, Nicolas J. FCA in a Logical Programming Setting for Visualization-Oriented Graph Compression. In: *ICFCA 2017: Formal Concept Analysis* Springer, Cham; 2017. p. 89–105. http://link.springer.com/10.1007/978-3-319-59271-8{_}6, doi: 10.1007/978-3-319-59271-8_6.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2014 jan; 12(1):59–60. doi: 10.1038/nmeth.3176.

Burgunter-Delamare B, KleinJan H, Frioux C, Fremy E, Wagner M, Corre E, Salver AL, Leroux C, Leblanc C, Boyen C, Siegel A, Dittami SM. Metabolic complementarity between a brown alga and associated cultivable bacteria provide indications of beneficial interactions. bioRxiv. 2019; https://www.biorxiv.org/content/early/2019/10/22/813683, doi: 10.1101/813683.

Carlström CI, Field CM, Bortfeld-Miller M, Müller B, Sunagawa S, Vorholt JA. Synthetic microbiota reveal priority effects and keystone strains in the Arabidopsis phyllosphere. Nature Ecology & Evolution. 2019; 3(10):1445–1454. doi: 10.1038/s41559-019-0994-z.

Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Research. 2019 oct; https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz862/5581728, doi: 10.1093/nar/gkz862.

Chan SHJ, Simons MN, Maranas CD. SteadyCom: Predicting microbial abundances while ensuring community stability. PLOS Computational Biology. 2017 may; 13(5):e1005539. http://dx.plos.org/10.1371/journal.pcbi.1005539, doi: 10.1371/journal.pcbi.1005539.

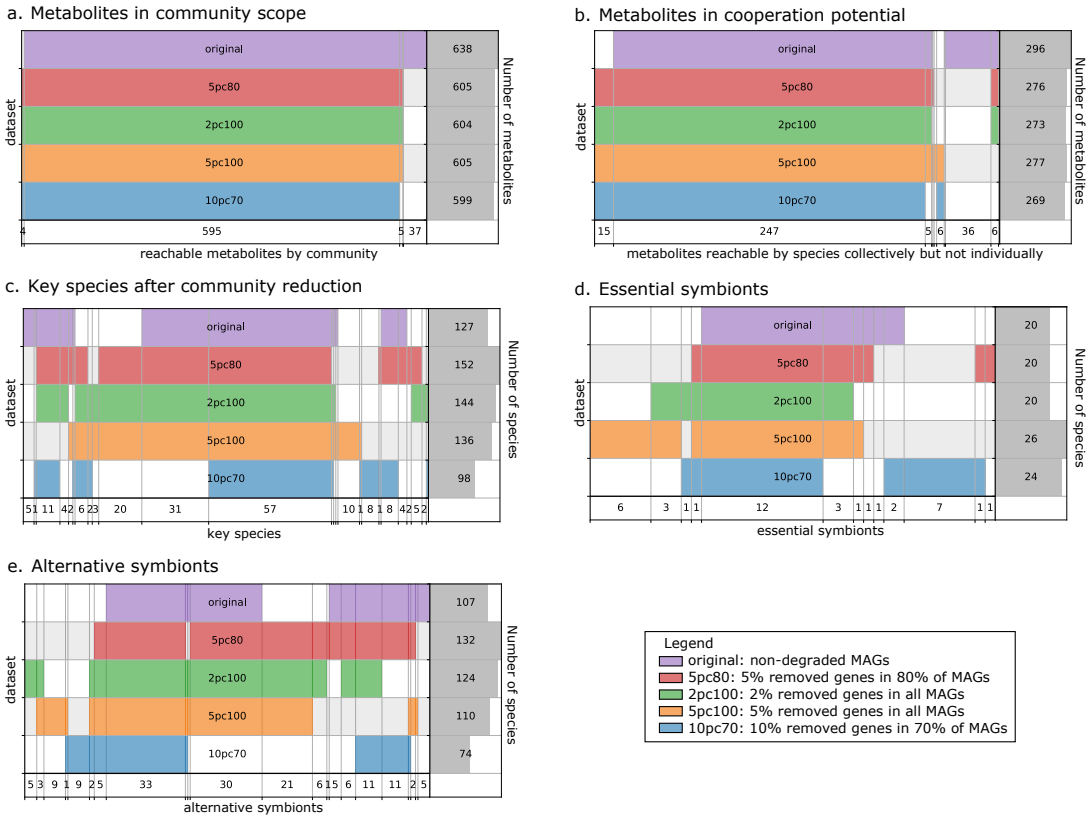


Figure 2. Robustness analysis of M2M results on datasets of altered MAGs. A proportion of genes were randomly removed from all or a random subset of the 913 rumen MAGs: 2% from all genomes (*2pc100*), 5% from 80% of the genomes (*5pc80*), 5% from all genomes (*5pc100*) and 10% from 70% of the genomes (*10pc70*). M2M pipeline was ran on these four datasets and comparison was made with respect to the initial non-altered dataset of MAGs (*original*). Subfigures *a.* to *e* each represent one piece of information computed by M2M and compared between the five experiments. *a*. Set of producible compounds by all metabolic networks in a cooperative system (community scope); supervenn representation. Each dataset of metabolic networks obtained from the original or degraded genomes is represented horizontally, with a unique colour. The right panel of the supervenn diagram indicates the number of metabolites in the community scope of the corresponding dataset. Vertical overlaps between sets represent intersections (e.g groups of metabolites retrieved in several datasets) whose size is indicated on the X axis. For example, there is a set of 37 metabolites that are producible in the original dataset only, and a set of 5 metabolites predicted as producible in all datasets but the one where 70% of genomes were 10%-degraded. A full superimposition of all the coloured bars would indicate a complete stability of the community scope between datasets. *b*. Comparison of the cooperation potential between the five experiments. *c*. Comparison of key species that gather essential symbionts (*d*.) and alternative symbionts (*e*.).

798 **Christian N**, Handorf T, Ebenhöh O. Metabolic synergy: increasing biosynthetic capabilities by network cooper-
799 ation. Genome informatics International Conference on Genome Informatics. 2007; 18:320–329.

800 **Christian N**, May P, Kempa S, Handorf T, Ebenhöh O. An integrative approach towards completing genome-
801 scale metabolic networks. Molecular BioSystems. 2009; 5(12):1889–1903. http://pubs.rsc.org/en/content/
802 articlehtml/2009/mb/b915913b, doi: 10.1039/B915913b.

803 **Costea PI**, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, de Vos WM, Ehrlich SD, Fraser CM,
804 Hattori M, Huttenhower C, Jeffery IB, Knights D, Lewis JD, Ley RE, Ochman H, O'Toole PW, Quince C, Relman DA,
805 Shanahan F, et al. Enterotypes in the landscape of gut microbial community composition. Nature microbiology.
806 2018 jan; 3(1):8–16. http://www.ncbi.nlm.nih.gov/pubmed/29255284http://www.pubmedcentral.nih.gov/
807 articlerender.fcgi?artid=PMC5832044, doi: 10.1038/s41564-017-0072-8.

808 **Cottret L**, Milreu PV, Acuña V, Marchetti-Spaccamela A, Stougie L, Charles H, Sagot MF. Graph-Based Analysis
809 of the Metabolic Exchanges between Two Co-Resident Intracellular Symbionts, Baumannia cicadellinicola
810 and Sulcia muelleri, with Their Insect Host, Homalodisca coagulata. PLoS Computational Biology. 2010 sep;
811 6(9):e1000904. http://dx.plos.org/10.1371/journal.pcbi.1000904, doi: 10.1371/journal.pcbi.1000904.

812 **Coyte KZ**, Rakoff-Nahoum S. Understanding Competition and Cooperation within the Mammalian Gut Mi-
813 crobiome. Current Biology. 2019 jun; 29(11):R538–R544. https://www.sciencedirect.com/science/article/pii/
814 S0960982219304154?via{%}3Dihub, doi: 10.1016/J.CUB.2019.04.017.

815 **Cutting SM**. Bacillus probiotics. Food Microbiology. 2011; 28(2):214–220. doi: 10.1016/j.fm.2010.03.007.

816 **Diener C**, Gibbons SM, Resendis-Antonio O. MICOM: Metagenome-Scale Modeling To Infer Metabolic Inter-
817 actions in the Gut Microbiota. mSystems. 2020 jan; 5(1). http://msystems.asm.org/lookup/doi/10.1128/
818 mSystems.00606-19, doi: 10.1128/mSystems.00606-19.

819 **Ebenhöh O**, Handorf T, Heinrich R. Structural analysis of expanding metabolic networks. Genome informatics
820 International Conference on Genome Informatics. 2004; 15(1):35–45. http://www.ncbi.nlm.nih.gov/pubmed/
821 15712108.

822 **Eng A**, Borenstein E. An algorithm for designing minimal microbial communities with desired metabolic
823 capacities. Bioinformatics. 2016; 32(13):2008–2016. http://bioinformatics.oxfordjournals.org/content/32/13/
824 2008.long, doi: 10.1093/BIOINFORMATICS/BTW107.

825 **Fisher CK**, Mehta P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries
826 Using Sparse Linear Regression. PLoS ONE. 2014; 9(7):e102451. doi: 10.1371/journal.pone.0102451.

827 **Floc'h JB**, Hamel C, Harker KN, St-Arnaud M. Fungal Communities of the Canola Rhizosphere: Keystone Species
828 and Substantial Between-Year Variation of the Rhizosphere Microbiome. Microbial Ecology. 2020; p. 1–16.
829 doi: 10.1007/s00248-019-01475-8.

830 **Forslund K**, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir
831 V, Krogh Pedersen H, Arumugam M, Kristiansen K, Yvonne Voigt A, Vestergaard H, Hercog R, Igor Costea P,
832 Roat Kultima J, Li J, Jørgensen T, Levenez F, et al. Disentangling type 2 diabetes and metformin treatment
833 signatures in the human gut microbiota. Nature. 2015 dec; 528(7581):262–266. http://www.nature.com/
834 articles/nature15766, doi: 10.1038/nature15766.

835 **Forster SC**, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y,
836 Pike LJ, Louie T, Browne HP, Mitchell AL, Neville BA, Finn RD, Lawley TD. A human gut bacterial genome
837 and culture collection for improved metagenomic analyses. Nature Biotechnology. 2019 feb; 37(2):186–192.
838 http://www.nature.com/articles/s41587-018-0009-7, doi: 10.1038/s41587-018-0009-7.

839 **Franzosa EA**, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Capo-
840 raso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscrip-
841 tomes. Nature Methods. 2018 nov; 15(11):962–968. http://www.nature.com/articles/s41592-018-0176-y, doi:
842 10.1038/s41592-018-0176-y.

843 **Frioux C**, Fremy E, Trottier C, Siegel A. Scalable and exhaustive screening of metabolic functions carried out by
844 microbial consortia. Bioinformatics. 2018 sep; 34(17):i934–i943. https://academic.oup.com/bioinformatics/
845 article/34/17/i934/5093211, doi: 10.1093/bioinformatics/bty588.

846 **Gebser M**, Kaminski R, Kaufmann B, Schaub T. Answer Set Solving in Practice. Synthesis lectures on artificial
847 intelligence and machine learning. 2012; .

Gebser M, Kaufmann B, Neumann A, Schaub T. Conflict-Driven Answer Set Solving. In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence* IJCAI'07, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2007. p. 386–392.

Greenblum S, Turnbaugh PJ, Borenstein E. 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. 2012 jan; 109(2):594–9. http://www.ncbi.nlm.nih.gov/pubmed/22184244http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3258644, doi: 10.1073/pnas.1116053109.

Handorf T, Ebenhöh O, Heinrich R. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. Journal of Molecular Evolution. 2005; 61(4):498–512. https://link.springer.com/content/pdf/10.1007{%}2Fs00239-005-0027-1.pdf, doi: 10.1007/s00239-005-0027-1.

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology. 2010; 28(9):977–982. doi: 10.1038/nbt.1672.

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology. 2010 sep; 28(9):977–982. http://www.nature.com/nbt/journal/v28/n9/abs/nbt.1672.html, doi: 10.1038/nbt.1672.

Hildebrand F, Moitinho-Silva L, Blasche S, Jahn MT, Gossmann TI, Huerta-Cepas J, Hercog R, Luetge M, Bahram M, Pryszlak A, Alves RJ, Waszak SM, Zhu A, Ye L, Costea PI, Aalvink S, Belzer C, Forslund SK, Sunagawa S, Hentschel U, et al. Antibiotics-induced monodominance of a novel gut bacterial order. Gut. 2019 feb; p. gutjnl–2018–317715. https://gut.bmj.com/content/early/2019/02/01/gutjnl-2018-317715, doi: 10.1136/GUTJNL-2018-317715.

Hildebrand F, Tadeo R, Voigt A, Bork P, Raes J. LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome. 2014 sep; 2(1):30. http://www.microbiomejournal.com/content/2/1/30, doi: 10.1186/2049-2618-2-30.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, et al. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics. 2003 mar; 19(4):524–531. http://www.ncbi.nlm.nih.gov/pubmed/12611808, doi: 10.1093/bioinformatics/btg015.

Hucka M, Bergmann FT, Dräger A, Hoops S, Keating SM, Le Novère N, Myers CJ, Olivier BG, Sahle S, Schaff JC, Smith LP, Waltemath D, Wilkinson DJ. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core. Journal of integrative bioinformatics. 2018 mar; 15(1). http://www.ncbi.nlm.nih.gov/pubmed/29522418http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6167032, doi: 10.1515/jib-2017-0081.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular Biology and Evolution. 2017 aug; 34(8):2115–2122. https://academic.oup.com/mbe/article/34/8/2115/3782716, doi: 10.1093/molbev/msx148.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen L, von Mering C, Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2019 jan; 47(D1):D309–D314. https://academic.oup.com/nar/article/47/D1/D309/5173662, doi: 10.1093/nar/gky1085.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 mar; 11:119. doi: 10.1186/1471-2105-11-119.

Julien-Laferrière A, Bulteau L, Parrot D, Marchetti-Spaccamela A, Stougie L, Vinga S, Mary A, Sagot MF. A Combinatorial Algorithm for Microbial Consortia Synthetic Design. Scientific Reports. 2016 jul; 6:29182. http://www.nature.com/articles/srep29182, doi: 10.1038/srep29182.

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019 jul; 7(7):e7359. https://peerj.com/articles/7359, doi: 10.7717/peerj.7359.

900 **Karp PD**, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, Kothari A, Weaver D, Lee T,
901 Subhraveti P, Spaulding A, Fulcher C, Keseler IM, Caspi R. Pathway tools version 19.0 update: Software
902 for pathway/genome informatics and systems biology. Briefings in Bioinformatics. 2016 sep; 17(5):877–
903 890. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5036846/pdf/bbv079.pdfhttps://arxiv.org/pdf/1510.
904 03964.pdfhttps://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv079, doi: 10.1093/bib/bbv079.

905 **Khandelwal RA**, Olivier BG, Röling WFM, Teusink B, Bruggeman FJ. Community flux balance analysis for mi-
906 crobial consortia at balanced growth. PloS one. 2013 may; 8(5):e64567. http://dx.plos.org/10.1371/journal.
907 pone.0064567http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064567, doi: 10.1371/jour-
908 nal.pone.0064567.

909 **Kruse K**, Ebenhöh O. Comparing flux balance analysis to network expansion: producibility, sustainability and
910 the scope of compounds. Genome Informatics. 2008; 20:91–101. http://www.worldscientific.com/doi/pdf/10.
911 1142/9781848163003{_}0008http://www.ncbi.nlm.nih.gov/pubmed/19425125, doi: 9781848163003_0008
912 [pii].

913 **Kumar M**, Ji B, Zengler K, Nielsen J. Modelling approaches for studying the microbiome. Nature Microbiology.
914 2019 aug; 4(8):1253–1267. http://www.nature.com/articles/s41564-019-0491-9, doi: 10.1038/s41564-019-
915 0491-9.

916 **Laniau J**, Frioux C, Nicolas J, Baroukh C, Cortes MPMP, Got J, Trottier C, Eveillard D, Siegel A. Combining graph
917 and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. PeerJ.
918 2017 oct; 5(10):e3860. https://peerj.com/articles/3860, doi: 10.7717/peerj.3860.

919 **Levy R**, Carr R, Kreimer A, Freilich S, Borenstein E. NetCooperate: a network-based tool for inferring host-
920 microbe and microbe-microbe cooperation. BMC bioinformatics. 2015 may; 16(1):164. http://www.ncbi.
921 nlm.nih.gov/pubmed/25980407http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4434858, doi:
922 10.1186/s12859-015-0588-y.

923 **Li D**, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and com-
924 plex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015; 31(10):1674–1676. doi:
925 10.1093/bioinformatics/btv033.

926 **Machado D**, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic
927 models for microbial species and communities. Nucleic Acids Research. 2018 sep; 46(15):7542–7553. https:
928 //academic.oup.com/nar/article/46/15/7542/5042022, doi: 10.1093/nar/gky537.

929 **Magnúsdóttir S**, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jäger C, Baginska J, Wilmes
930 P, Fleming RMT, Thiele I. Generation of genome-scale metabolic reconstructions for 773 members of the
931 human gut microbiota. Nature Biotechnology. 2016 nov; 35(1):81–89. http://www.nature.com/doifinder/10.
932 1038/nbt.3703, doi: 10.1038/nbt.3703.

933 **Manquinho V**, Marques-Silva J, Planes J. Algorithms for Weighted Boolean Optimization. In: Kullmann O, editor.
934 *Theory and Applications of Satisfiability Testing - SAT 2009* Berlin, Heidelberg: Springer Berlin Heidelberg; 2009.
935 p. 495–508.

936 **Manzoor SE**, McNulty CAM, Nakiboneka-Ssenabulya D, Lecky DM, Hardy KJ, Hawkey PM. Investigation of
937 community carriage rates of Clostridium difficile and Hungatella hathewayi in healthy volunteers from four
938 regions of England. Journal of Hospital Infection. 2017; 97(2):153 – 155. http://www.sciencedirect.com/
939 science/article/pii/S0195670117302840, doi: https://doi.org/10.1016/j.jhin.2017.05.014.

940 **Marco ML**, Vries MCd, Wels M, Molenaar D, Mangell P, Ahrne S, Vos WMd, Vaughan EE, Kleerebezem M.
941 Convergence in probiotic Lactobacillus gut-adaptive responses in humans and mice. The ISME Journal. 2010;
942 4(11):1481–1484. doi: 10.1038/ismej.2010.61.

943 **Matthäus F**, Salazar C, Ebenhöh O. Biosynthetic Potentials of Metabolites and Their Hierarchical Organization.
944 PLoS Computational Biology. 2008 apr; 4(4):e1000049. https://dx.plos.org/10.1371/journal.pcbi.1000049, doi:
945 10.1371/journal.pcbi.1000049.

946 **Mende DR**, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, Hernández-Plaza A, Orakov AN, Forslund
947 SK, Sunagawa S, Zeller G, Huerta-Cepas J, Coelho LP, Bork P. proGenomes2: an improved database for
948 accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids
949 Research. 2019 oct; https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1002/5606617, doi:
950 10.1093/nar/gkz1002.

951 **Mende DR**, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nature
952 Methods. 2013 sep; 10(9):881–884. http://www.nature.com/articles/nmeth.2575, doi: 10.1038/nmeth.2575.

953 **Mendoza SN**, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic
954 reconstruction tools. Genome Biology. 2019; 20(1):158. doi: 10.1186/s13059-019-1769-1.

955 **Monteagudo-Mera A**, Rodríguez-Aparicio L, Rúa J, Martínez-Blanco H, Navasa N, García-Armesto MR, Ferrero
956 MA. In vitro evaluation of physiological probiotic properties of different lactic acid bacteria strains of dairy
957 and human origin. Journal of Functional Foods. 2012; 4(2):531–541. doi: 10.1016/j.jff.2012.02.014.

958 **Morgado A**, Heras F, Marques-Silva J. Improvements to Core-Guided Binary Search for MaxSAT. In: Cimatti A,
959 Sebastiani R, editors. *Theory and Applications of Satisfiability Testing – SAT 2012* Berlin, Heidelberg: Springer
960 Berlin Heidelberg; 2012. p. 284–297.

961 **Moss EL**, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequenc-
962 ing. Nature biotechnology. 2020 June; 38(6):701—707. https://www.nature.com/articles/s41587-020-0422-6,
963 doi: 10.1038/s41587-020-0422-6.

964 **Moya A**, Ferrer M. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance.
965 Trends in Microbiology. 2016 may; 24(5):402–413. https://www.sciencedirect.com/science/article/pii/
966 S0966842X16000263, doi: 10.1016/j.tim.2016.02.002.

967 **Nielsen HB**, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le
968 Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto JM, Quintanilha Dos Santos
969 MB, Blom N, Borruel N, Burgdorf KS, et al. Identification and assembly of genomes and genetic elements
970 in complex metagenomic samples without using reference genomes. Nature Biotechnology. 2014 aug;
971 32(8):822–828. http://www.nature.com/articles/nbt.2939, doi: 10.1038/nbt.2939.

972 **Noronha A**, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, Daníelsdóttir AD, Krecke M, Merten D,
973 Haraldsdóttir HS, Heinken A, Heirendt L, Magnúsdóttir S, Ravcheev DA, Sahoo S, Gawron P, Friscioni L, Garcia
974 B, Prendergast M, Puente A, et al. The Virtual Metabolic Human database: integrating human and gut
975 microbiome metabolism with nutrition and disease. Nucleic Acids Research. 2018; 47(D1):D614–D624. doi:
976 10.1093/nar/gky992.

977 **Ofaim S**, Ofek-Lalzar M, Sela N, Jinag J, Kashi Y, Minz D, Freilich S. Analysis of Microbial Functions in the
978 Rhizosphere Using a Metabolic-Network Based Framework for Metagenomics Interpretation. Frontiers in
979 Microbiology. 2017 aug; 8:1606. http://journal.frontiersin.org/article/10.3389/fmicb.2017.01606/full, doi:
980 10.3389/fmicb.2017.01606.

981 **Opatovsky I**, Santos-Garcia D, Ruan Z, Lahav T, Ofaim S, Mouton L, Barbe V, Jiang J, Zchori-Fein E, Freilich
982 S. Modeling trophic dependencies and exchanges among insects' bacterial symbionts in a host-simulated
983 environment. BMC Genomics. 2018 dec; 19(1):402. https://bmcgenomics.biomedcentral.com/articles/10.
984 1186/s12864-018-4786-7, doi: 10.1186/s12864-018-4786-7.

985 **Orth JD**, Thiele I, Palsson BØ. What is Flux Balance Analysis ? Nature biotechnology. 2010 mar; 28(3):245–
986 248. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3108565/http://www.ncbi.nlm.nih.gov/pmc/articles/
987 PMC3108565/pdf/nihms299330.pdf, doi: 10.1038/nbt.1614.

988 **Parks DH**, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial
989 genomes recovered from isolates, single cells, and metagenomes. Genome research. 2015 jul; 25(7):1043–
990 55. http://www.ncbi.nlm.nih.gov/pubmed/25977477http://www.pubmedcentral.nih.gov/articlerender.fcgi?
991 artid=PMC4484387, doi: 10.1101/gr.186072.114.

992 **Pasolli E**, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado
993 MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. Extensive Unexplored
994 Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geog-
995 raphy, and Lifestyle. Cell. 2019 jan; 176(3):649–662.e20. https://www.sciencedirect.com/science/article/pii/
996 S0092867419300017, doi: 10.1016/J.CELL.2019.01.001.

997 **Pedregosa F**, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg
998 V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Édouard Duchesnay. Scikit-learn: Machine
999 Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–2830. http://jmlr.org/papers/
1000 v12/pedregosa11a.html.

1001 **Petrenko P**, Lobb B, Kurtz DA, Neufeld JD, Doxey AC. MetAnnotate: function-specific taxonomic profiling and
1002 comparison of metagenomes. BMC Biology. 2015 dec; 13(1):92. http://bmcbiol.biomedcentral.com/articles/
1003 10.1186/s12915-015-0195-4, doi: 10.1186/s12915-015-0195-4.

1004 **Popp D**, Centler F. μBialSim: Constraint-Based Dynamic Simulation of Complex Microbiomes. Frontiers in
1005 Bioengineering and Biotechnology. 2020; 8:574. doi: 10.3389/fbioe.2020.00574.

1006 **Potrykus J**, White RL, Bearne SL. Proteomic investigation of amino acid catabolism in the indigenous gut
1007 anaerobe Fusobacterium varium. PROTEOMICS. 2008; 8(13):2691–2703. doi: 10.1002/pmic.200700437.

1008 **Prigent S**, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, Gutknecht F, Got J, Eveillard D, Bourdon J, Plewniak
1009 F, Tonon T, Siegel A. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide
1010 Metabolic Networks. PLOS Computational Biology. 2017 jan; 13(1):e1005276. http://dx.plos.org/10.1371/
1011 journal.pcbi.1005276, doi: 10.1371/journal.pcbi.1005276.

1012 **Pérez-Pantoja D**, Donoso R, Agulló L, Córdova M, Seeger M, Pieper DH, González B. Genomic analysis of the
1013 potential for aromatic compounds biodegradation in Burkholderiales. Environmental Microbiology. 2012;
1014 14(5):1091–1117. doi: 10.1111/j.1462-2920.2011.02613.x.

1015 **Rivière A**, Selak M, Lantin D, Leroy F, De Vuyst L. Bifidobacteria and butyrate-producing colon bacteria: Impor-
1016 tance and strategies for their stimulation in the human gut. Frontiers in Microbiology. 2016; 7(JUN). doi:
1017 10.3389/fmicb.2016.00979.

1018 **Royer L**, Reimann M, Andreopoulos B, Schroeder M. Unraveling Protein Networks with Power Graph
1019 Analysis. PLoS Comput Biol. 2008 jul; 4(7):e1000108. http://dx.plos.org/10.1371/journal.pcbi.1000108http:
1020 //dx.plos.org/10.1371/journal.pcbi.1000108{%}5Cnhttp://www.ploscompbiol.org/article/fetchObject.action?
1021 uri=info{%}3Adoi{%}2F10.1371{%}2Fjournal.pcbi.1000108{&}representation=PDF, doi: 10.1371/jour-
1022 nal.pcbi.1000108.

1023 **Schaub T**, Thiele S. Metabolic network expansion with answer set programming. In: *Lecture Notes in Computer*
1024 *Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5649
1025 LNCS Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 312–326. http://link.springer.com/10.1007/
1026 978-3-642-02846-5{_}27http://www.springerlink.com/index/9368L534R7V10671.pdf, doi: 10.1007/978-3-642-
1027 02846-5_27.

1028 **Schellenberger J**, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A,
1029 Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BO. Quantitative prediction of cel-
1030 lular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature proto-
1031 cols. 2011 sep; 6(9):1290–307. http://www.nature.com/articles/nprot.2011.308http://www.ncbi.nlm.nih.
1032 gov/pubmed/21886097{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3319681, doi:
1033 10.1038/nprot.2011.308.

1034 **Schilling CH**, Letscher D, Palsson BO. Theory for the systemic definition of metabolic pathways and their use in
1035 interpreting metabolic function from a pathway-oriented perspective. Journal of Theoretical Biology. 2000
1036 apr; 203(3):229–248. doi: 10.1006/jtbi.2000.1073.

1037 **Seaver SMD**, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, Mundy M, Chia N, Noor E, Beber M, Best AA,
1038 DeJongh M, Kimbrel JA, D'haeseleer P, McCorkle SR, Bolton JR, Pearson E, Canon S, Wood-Charlson EM,
1039 Cottingham RW, et al. The ModelSEED Biochemistry Database for the integration of metabolic annotations
1040 and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic
1041 Acids Research. 2020; p. gkaa746–. doi: 10.1093/nar/gkaa746.

1042 **Seemann T**. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014 jul; 30(14):2068–2069. https:
1043 //academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu153, doi: 10.1093/bioin-
1044 formatics/btu153.

1045 **Sen P**, Orešič M. Metabolic Modeling of Human Gut Microbiota on a Genome Scale: An Overview. Metabolites.
1046 2019 jan; 9(2). http://www.ncbi.nlm.nih.gov/pubmed/30695998, doi: 10.3390/metabo9020022.

1047 **Shannon P**, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski BB, Ideker T. Cytoscape: A
1048 software Environment for integrated models of biomolecular interaction networks. Genome Research. 2003
1049 nov; 13(11):2498–2504. http://www.ncbi.nlm.nih.gov/pubmed/14597658http://www.pubmedcentral.nih.gov/
1050 articlerender.fcgi?artid=PMC403769, doi: 10.1101/gr.1239303.

1051 **Sharma AK**, Gupta A, Kumar S, Dhakan DB, Sharma VK. Woods: A fast and accurate functional annotator and
1052 classifier of genomic and metagenomic sequences. Genomics. 2015 jul; 106(1):1–6. https://www.sciencedirect.
1053 com/science/article/pii/S0888754315000543, doi: 10.1016/J.YGENO.2015.04.001.

1054 **Silva GGZ**, Green KT, Dutilh BE, Edwards RA. SUPER-FOCUS: a tool for agile functional analysis of shotgun
1055 metagenomic data. Bioinformatics. 2016 feb; 32(3):354–361. https://academic.oup.com/bioinformatics/
1056 article-lookup/doi/10.1093/bioinformatics/btv584, doi: 10.1093/bioinformatics/btv584.

1057 **Soni R**, Nanjani S, Keharia H. Genome analysis reveals probiotic prop-ensities of Paenibacillus polymyxa HK4.
1058 Genomics. 2020; doi: 10.1016/j.ygeno.2020.10.017.

1059 **Steinegger M**, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive
1060 data sets. Nature Biotechnology. 2017 nov; 35(11):1026–1028. doi: 10.1038/nbt.3988.

1061 **Stewart RD**, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker
1062 AW, Roehe R, Watson M. Assembly of 913 microbial genomes from metagenomic sequencing of the cow
1063 rumen. Nature communications. 2018; 9(1):870. http://www.ncbi.nlm.nih.gov/pubmed/29491419http://www.
1064 pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5830445, doi: 10.1038/s41467-018-03317-6.

1065 **Sunagawa S**, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti
1066 A, Cornejo-Castillo FM, Costea PI, Cruaud C, D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F,
1067 Kokoszka F, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science (New
1068 York, NY). 2015 may; 348(6237):1261359. http://www.ncbi.nlm.nih.gov/pubmed/25999513, doi: 10.1126/sci-
1069 ence.1261359.

1070 **Team RC**. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2017. https://www.
1071 R-project.org/.

1072 **The Integrative HMP Research Network Consortium i**. The integrative human microbiome project: Dynamic
1073 analysis of microbiome-host omics profiles during periods of human health and disease corresponding
1074 author. Cell Host and Microbe. 2014 sep; 16(3):276–289. https://www.sciencedirect.com/science/article/pii/
1075 S1931312814003060?via{%}3Dihub, doi: 10.1016/j.chom.2014.08.014.

1076 **Thiele I**, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature
1077 protocols. 2010 jan; 5(1):93–121. http://www.ncbi.nlm.nih.gov/pubmed/20057383http://www.pubmedcentral.
1078 nih.gov/articlerender.fcgi?artid=PMC3125167, doi: 10.1038/nprot.2009.203.

1079 **Thiele I**, Vlassis N, Fleming RMT. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics (Oxford,
1080 England). 2014 sep; 30(17):2529–2531. doi: 10.1093/bioinformatics/btu321.

1081 **Treitli SC**, Kolisko M, Husník F, Keeling PJ, Hampl V. Revealing the metabolic capacity of Streblomastix strix and
1082 its bacterial symbionts using single-cell metagenomics. Proceedings of the National Academy of Sciences.
1083 2019; 116(39):19675–19684. https://www.pnas.org/content/116/39/19675, doi: 10.1073/pnas.1910793116.

1084 **Vieira-Silva S**, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, Vandeputte D, Valles-Colomer M,
1085 Hildebrand F, Chaffron S, Raes J. Species–function relationships shape ecological properties of the human gut
1086 microbiome. Nature Microbiology. 2016 aug; 1(8):16088. http://www.nature.com/articles/nmicrobiol201688,
1087 doi: 10.1038/nmicrobiol.2016.88.

1088 **Vitkin E**, Shlomi T. MIRAGE: a functional genomics-based approach for metabolic network model reconstruction
1089 and its application to cyanobacteria networks. Genome biology. 2012; 13(11):R111. doi: 10.1186/gb-2012-13-
1090 11-r111.

1091 **Vries MCd**, Vaughan EE, Kleerebezem M, Vos WMd. Lactobacillus plantarum—survival, functional and potential
1092 probiotic properties in the human intestinal tract. International Dairy Journal. 2006; 16(9):1018–1028. doi:
1093 10.1016/j.idairyj.2005.09.003.

1094 **Wang H**, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, Kerkhoven EJ. RAVEN
1095 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor.
1096 PLOS Computational Biology. 2018 oct; 14(10):e1006541. http://dx.plos.org/10.1371/journal.pcbi.1006541,
1097 doi: 10.1371/journal.pcbi.1006541.

1098 **Wickham H**. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2009. http://ggplot2.org.

1099 **Zelezniak A**, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. Metabolic dependencies drive species
1100 co-occurrence in diverse microbial communities. Proceedings of the National Academy of Sciences of the
1101 United States of America. 2015 may; 112(20):6449–6454. doi: 10.1073/pnas.1421834112.

Zomorrodi AR, Maranas CD. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. PLoS Computational Biology. 2012; 8(2):e1002363. http://dx.doi.org/10.1371/journal.pcbi.1002363http://www.ploscompbiol.org/article/fetchObject.action?uri=info{%}3Adoi{%}2F10.1371{%}2Fjournal.pcbi.1002363{&}representation=PDF, doi: 10.1371/journal.pcbi.1002363.

Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, Wan D, Jiang R, Su L, Feng Q, Jie Z, Guo T, Xia Z, Liu C, Yu J, Lin Y, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nature Biotechnology. 2019 feb; 37(2):179–185. http://www.nature.com/articles/s41587-018-0008-8, doi: 10.1038/s41587-018-0008-8.

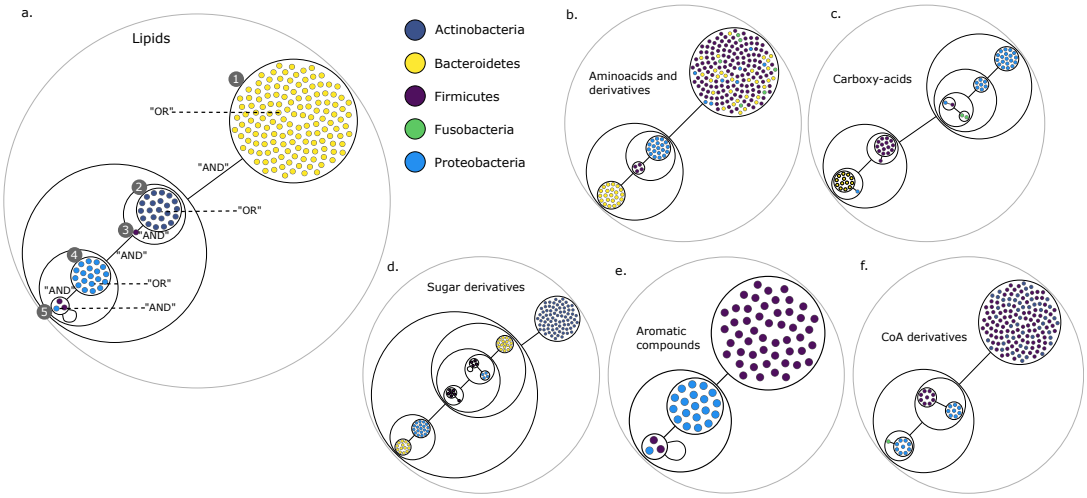


Figure 3. Power graph analysis of predicted microbial associations within communities for the human gut dataset. Each category of metabolites predicted as newly producible in the gut was defined as a target set for community selection among the 1,520 GSMNs from the gut microbiota reference genomes dataset. For each metabolic group, key 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 and groups of equivalence within key species. Power graphs a., b., c., d., e., f., g. were generated for the sets of lipids, aminoacids and derivatives, carboxy-acids, sugar derivatives, aromatic compounds, and coenzyme A derivative compounds respectively. Node colour describes the phylum associated to the GSMN. Figure a. has an additional description to ease readability. Edges symbolise conjunctions ("AND") and the co-occurrences of nodes in regular power nodes (as in power node 1, 2, 4) symbolise disjunctions ("OR") related to alternative symbionts. Power nodes with a loop (e.g. power node 5) indicate conjunctions. Therefore, each enumerated minimal community for lipid production is composed of the two Firmicutes and the Proteobacteria from power node 5, the Firmicutes node 3 (the four of them being the essential symbionts), and one Proteobacteria from power node 4, one Actinobacteria from power node 2 and 1 Bacteroidetes from power node 1. Members from an inner power node are interchangeable with respect to the metabolic objective. A version of the figures with species identification is available in ***Figure 3–Figure Supplement 1***, ***Figure 3–Figure Supplement 2***, ***Figure 3–Figure Supplement 3***, ***Figure 3–Figure Supplement 4***, ***Figure 3–Figure Supplement 5***, ***Figure 3–Figure Supplement 6*** (see. Supplementary File 1 - Table 4 for a mapping between identifiers and taxonomy). Power graphs can be generated with `m2m_analysis`. The figures display one visual representation for each power graph although such representations are not unique. The number of power edges is minimal, which leads to nesting of (power) nodes.

Figure 3–Figure supplement 1. Sugars derivatives power graph

Figure 3–Figure supplement 2. Lipids derivatives power graph

Figure 3–Figure supplement 3. Amino-acids and derivatives power graph

Figure 3–Figure supplement 4. Aromatic compounds power graph

Figure 3–Figure supplement 5. Carboxy-acids compounds power graph

Figure 3–Figure supplement 6. Coenzyme A derivatives power graph

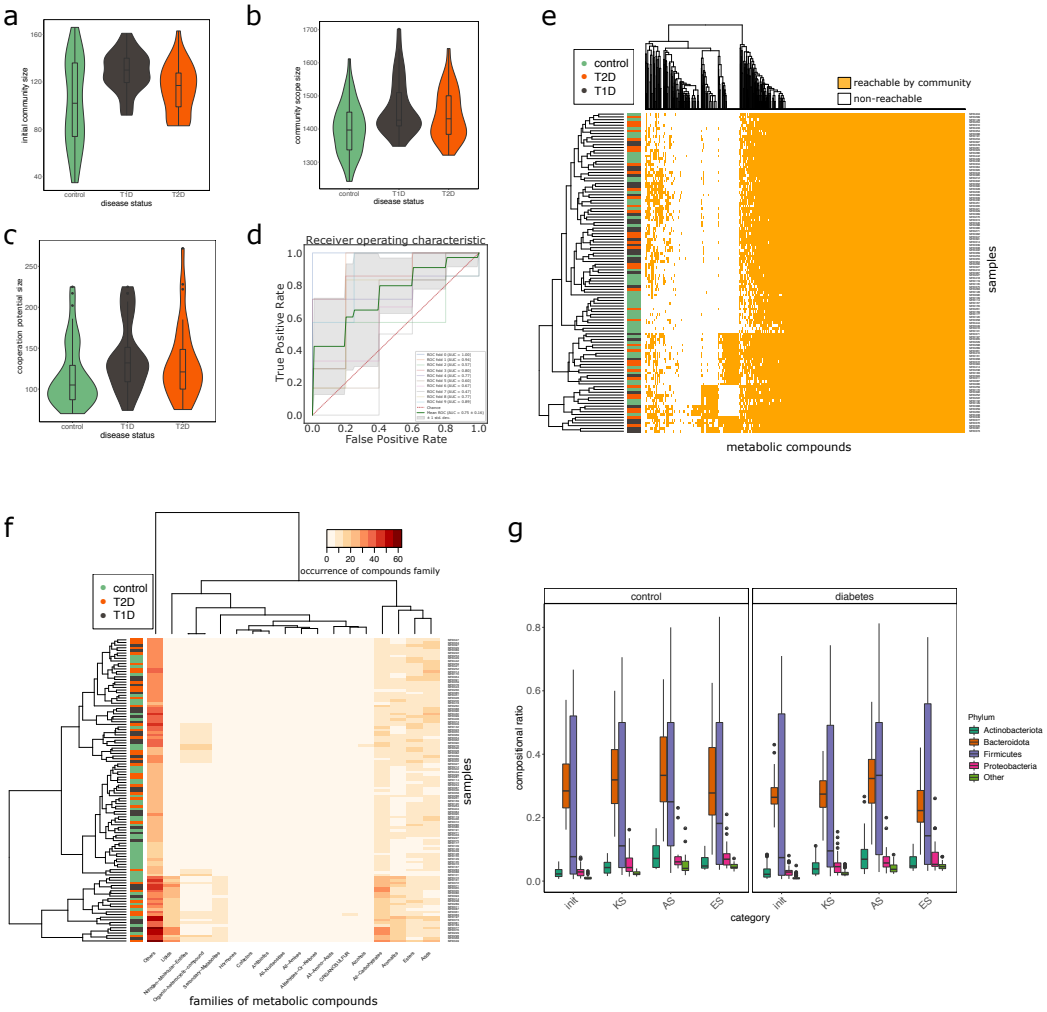
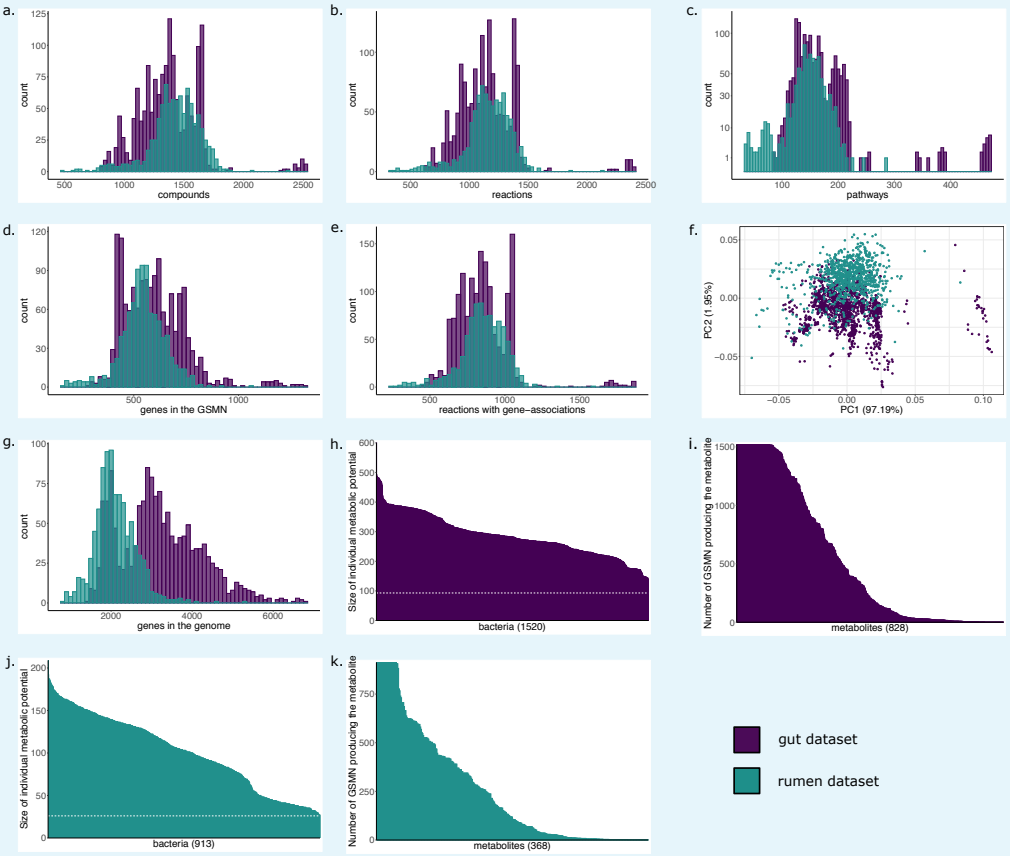


Figure 4. Effect of the disease status on the metabolism of communities in MHD samples. M2M was run on collections of GSMNs associated to MAGs identified in metagenomic samples from a cohort of healthy and diabetic individuals. Figure *a* describes the distributions of community sizes for all metagenomic samples according to the disease status: T1D: Type-1 Diabetes, T2D: Type-2 Diabetes. Figures *b* and *c* show the distribution of the community scope sizes and cooperation potential sizes respectively, according to the disease status. Figure *d* is the receiver operating curve (ROC) of a SVM classification experiment aiming at predicting the disease status for the MHD cohort (control n=49 or diabetes n=66) based on the community scope composition. Figure *e* illustrates the community scope composition in terms of metabolites for all samples. Disease status is indicated by the colour at the left side of each row. Figure *f* illustrates the composition of the cooperation potential according to the belonging of metabolites to Metacyc families of compounds. Disease status is indicated by the colour at the left side of each row. Figure *g* describes the taxonomic distribution at the phylum level of groups of species before and after community reduction, according to the disease status. Selection of communities was performed with the objective of making producible by the reduced communities the set of metabolites in the cooperation potential. init: initial composition of communities, KS: key species, AS: alternative symbionts, ES: essential symbionts. T1D: Type-1 Diabetes, T2D: Type-2 Diabetes.

Appendix 1

Analysis of GSMNs from the human gut reference genomes and rumen MAGs collections

Comparison of GSMNs reconstructed from MAGs and from reference genomes



Appendix 1 Figure 1. Characteristics of the metabolic networks built for the gut and the rumen datasets. a. Distribution of the number of metabolic compounds in GSMNs reconstructed for the gut dataset (purple) and the rumen dataset (green). b. Distribution of the number of metabolic reactions. c. Distribution of the number of complete pathways according to the MetaCyc database. d. Distribution of the number of genes included into the GSMNs. e. Distribution of the number of reactions associated to genes. f. Principal component analysis of the GSMNs reconstructions based on the previous characteristics (a. to e.). g. Distribution of the number of genes (not necessarily related to metabolism) in the initial genomes/MAGs. h. Individual metabolic potentials (scopes) for the gut bacteria, dotted line represents the number of seeds (nutrients) used in the algorithm. i. Reachability of metabolites by gut bacteria. j. Individual metabolic potentials (scopes) for the rumen bacteria, dotted line represents the number of seeds (nutrients) used in the algorithm. k. Reachability of metabolites by rumen bacteria.

Robustness analysis of GSMN reconstruction with MAGs

MAGs from the rumen dataset were degraded by randomly removing contigs. The following degradations were tested: removal of 2% of genes in all MAGs, removal of 5% of genes in 80% of MAGs, removal of 5% of genes in all MAGs, removal of 10% of genes in 70% of MAGs.

Table 1 summarises the characteristics of the genomes and GSMNs for all experiments. The average gene loss in genomes is similar to the average gene loss in metabolic networks. However, the average loss of metabolites and reactions is lower than the genetic loss: it increases more slowly than the loss of genes. For instance, the 2-percent degradation of MAGs leads to a nearly 2 percent decrease in reaction numbers in GSMNs. However, the

10-percent degradation in 70% of genomes (average gene loss of 7% in the initial community) only leads to a 5% decrease in reaction numbers. One notable observation is the stability in the percentage of reactions associated to genes, suggesting that the loss of reactions in degraded genomes mainly occurs among reactions that are not associated to genes. It is also possible that the loss of genes in GSMNs is due to the redundancy loss: some reactions associated to several genes before degradation lose some of these gene associations after degradation. Data for each genome and GSMN is available in Supplementary File 1 - Tables 16, 21-24.

Appendix 1 Table 1. Effect of MAG degradation on GSMN reconstructions. Numbers are averages. "±" precedes standard deviation values. "original": initial MAGs prior degradation, "2pc100": 2% gene removal in all MAGs, "5pc80": 5% gene removal in 80% of MAGs, "5pc100": 5% gene removal in all MAGs, "10pc70": 10% gene removal in 70% of MAGs.

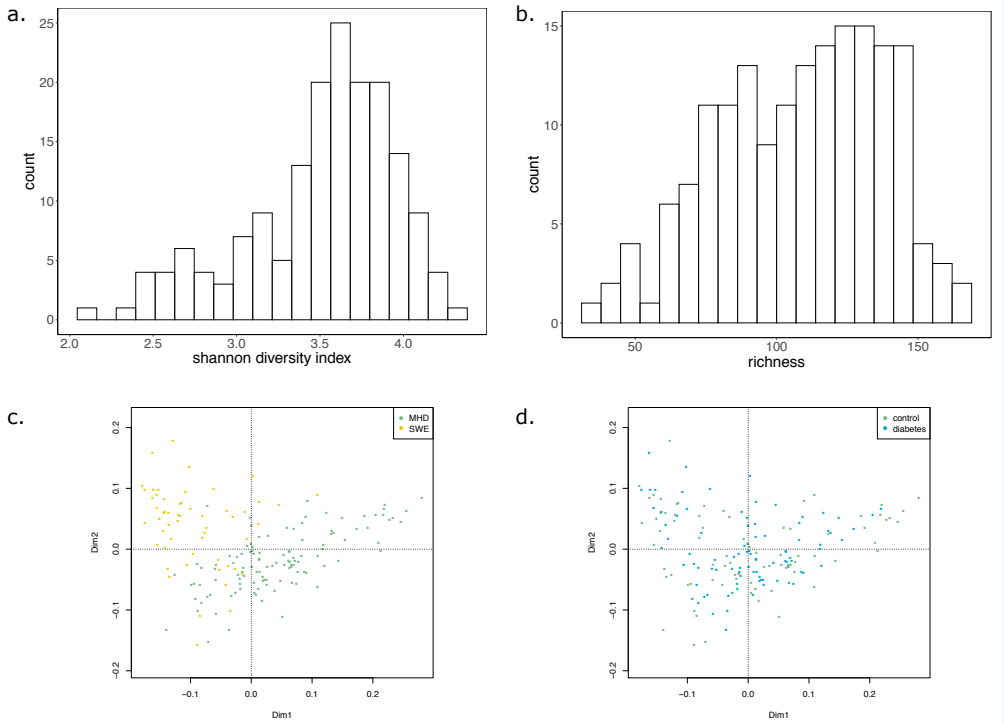
	original	2pc100	5pc80	5pc100	10pc70
Genes in MAGs	2,100 (± 501)	2,058 (± 491)	2,016 (± 484)	1,994 (± 478)	1,954 (± 480)
Reactions in GSMNs	1,155 (± 199)	1,131 (± 192)	1,116 (± 192)	1,108 (± 190)	1,094 (± 192)
Metabolites in GSMNs	1,422 (± 212)	1,402 (± 207)	1,388 (± 208)	1,381 (± 206)	1,366 (± 208)
Genes in GSMNs	543 (± 108)	532 (± 106)	521 (± 105)	515 (± 103)	505 (± 105)
% reactions with genes	73.84%	74.05%	73.82%	73.72%	73.61%
Gene loss in MAGs	—	1.98%	4.01%	5.03%	6.94%
Reaction loss in GSMNs	—	1.96%	3.30%	3.89%	5.17%
Metabolite loss in GSMNs	—	1.37%	2.41%	2.91%	3.92%
Gene loss in GSMNs	—	2.09%	4.17%	5.11%	7.02%

Appendix 2

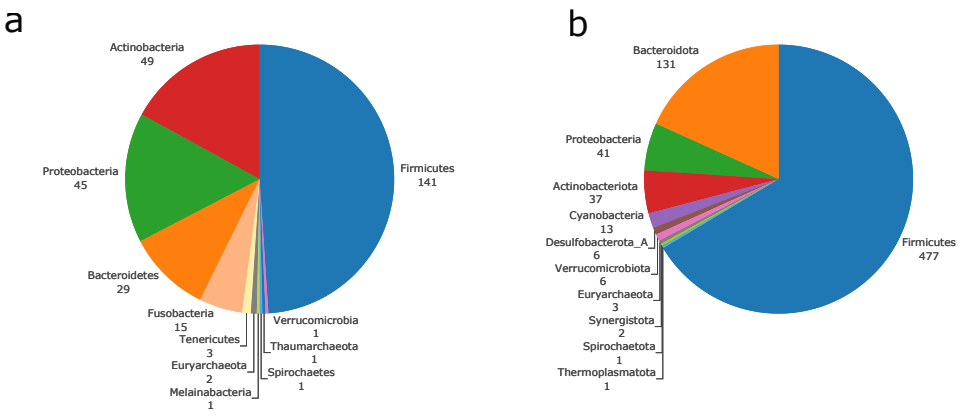
Supplementary information to the Diabetes experiment

Diversity and richness of the samples

The Shannon diversity index and richness of the 170 samples is illustrated in ***Figure 1*** a. and b. We relied on species present in the abundance matrix to define communities for the metabolic analysis. The average size of the community was 108 GSMNs. Their median size was 111. In order to compute a metabolic distance between samples, we retrieved for each genome its KO annotations obtained with Eggnog-Mapper. Using the abundance (normalised by sample) of the genomes in each sample, we were able to retrieve the KO content of samples. We then calculated the Bray-Curtis distance between samples before computing a PCoA (***Figure 1*** c., d.). The PCoA shows a clear distinction between the two datasets, thus motivating their distinct analysis, as performed in the main results of the article. However, there are no distinction between the control and diabetes status of the samples.

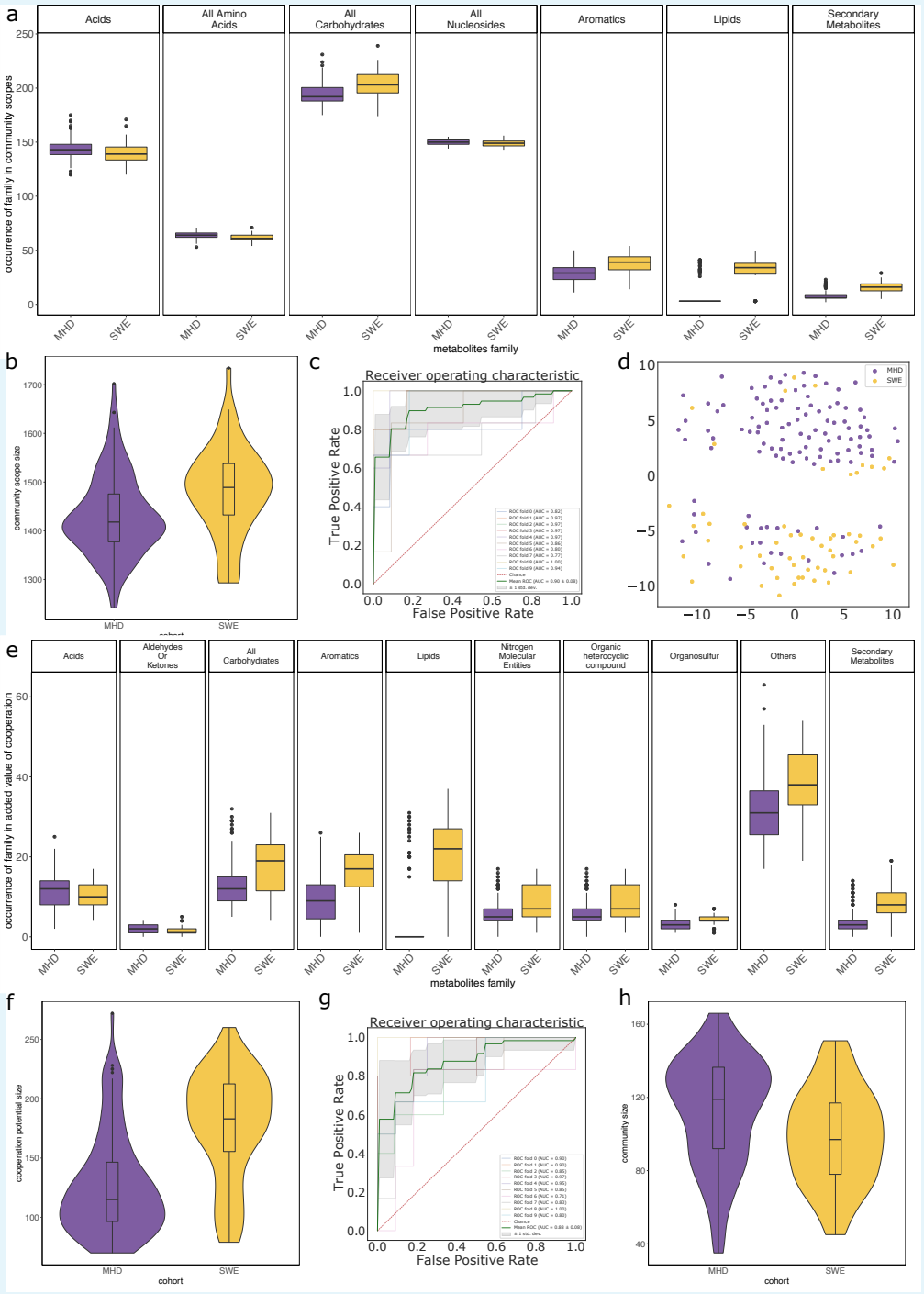


Appendix 2 Figure 1. Shannon diversity index, richness and metabolic distance of the samples. a. Histogram depicting the Shannon diversity index of the samples. b. Histogram depicting the richness of the samples. c. and d.: Principal component analysis (PCoA) of the Bray-Curtis distance calculated on the KO composition of samples coloured by dataset (c.) or disease status (d.)



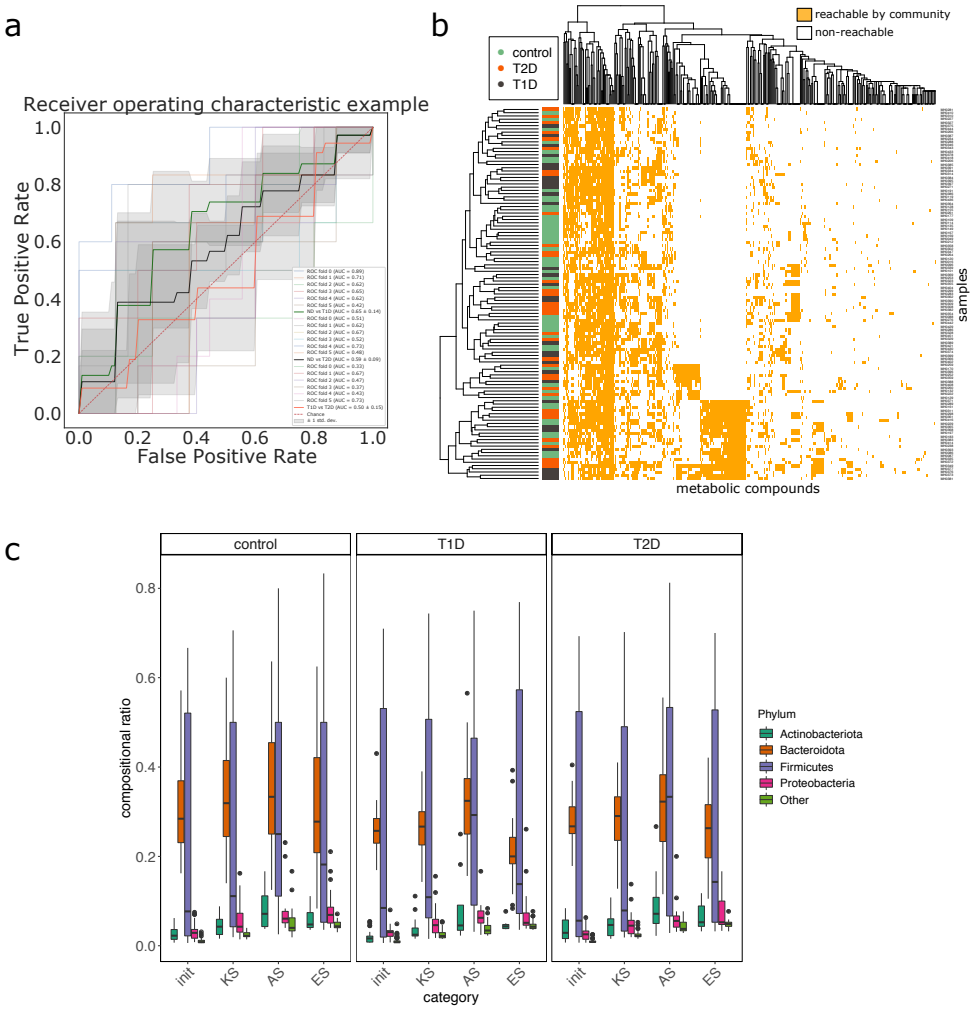
Appendix 2 Figure 2. Taxonomic diversity of the genomes used for GSMNs reconstruction using MGS or OTU mapping (at species level) to curated metabolic models. Phyla composition of the genomes, and number of distinct representatives for each phylum.

Cohort effect at the metabolic level



Appendix 2 Figure 3. Impact of the cohort when studying the metabolisms of individuals from the metagenomic dataset. Panels *a* to *d* focus on the community scope, that is the set of metabolites reachable by the community associated to a sample. Panel *d* shows the representation of a multidimensional scaling (MDS) on the community scope composition between cohorts. Panels *e* to *g* focus on the cooperation potential, that is the set of metabolites that are not expected to be produced by individual members of communities and instead require cooperation. Panel *a* and *e* describe families of metabolites whose occurrences significantly differ between cohorts in the corresponding group (community scope or cooperation potential). Panels *b* and *e* illustrate the size of the community scope and cooperation potential respectively in samples from the two cohorts. Panel *c* (resp. *f*) are receiving operating curves (ROC) of a classification experiment aiming at separating the cohort (MHD n=115, SWE n=55) based on the occurrences of metabolites in the community scope (resp. cooperation potential). Panel *h* describes the size of the initial community associated to samples of both cohorts according to abundance data of MGS.

Status effect at the metabolic level



Appendix 2 Figure 4. Impact of the status when studying the metabolisms of individuals from the MHD metagenomic dataset. Panel *a* is the receiver operating curve (ROC) of the classification experiment aiming at deciphering the disease status for the MHD cohort (control n=49, Type-1 Diabetes n=31 or Type-2 Diabetes n=35) based on the cooperation potential composition. Panel *b* illustrates the cooperation potential composition in terms of metabolites for all samples. Disease status is indicated by the colour at the left side of each row. Panel *c* describes the taxonomic distribution at the phylum level of groups of species before and after community reduction, according to the disease status. Selection of communities was performed with the objective of making producible by the reduced communities the set of metabolites in the cooperation potential. init: initial composition of communities, KS: key species, AS: alternative symbionts, ES: essential symbionts. T1D: Type-1 Diabetes, T2D: Type-2 Diabetes.

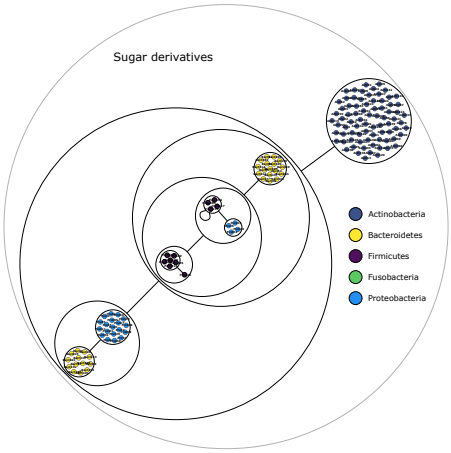


Figure 3–Figure supplement 1. Power graph associated to the minimal communities producing the sugars derivatives group of targets

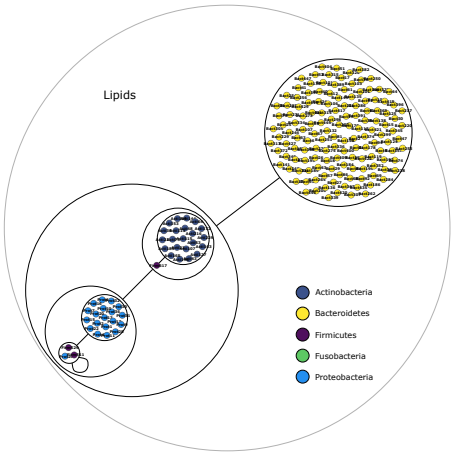


Figure 3–Figure supplement 2. Power graph associated to the minimal communities producing the lipids derivatives group of targets

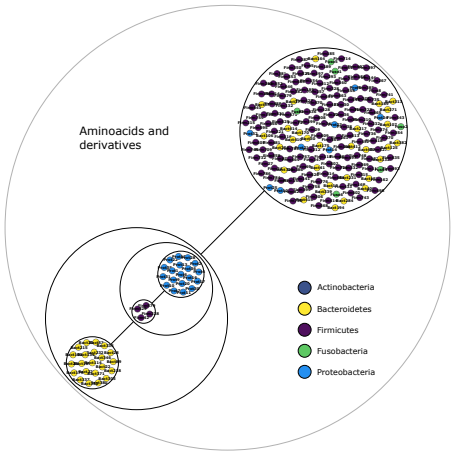


Figure 3–Figure supplement 3. Power graph associated to the minimal communities producing the amino-acids and derivatives group of targets

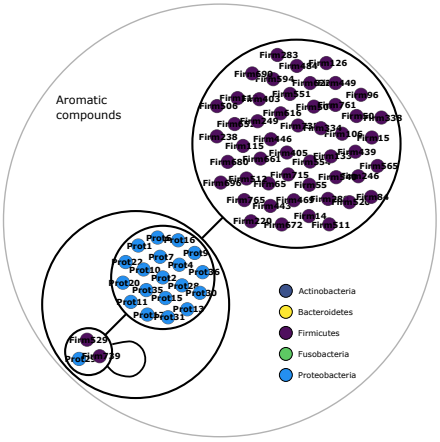


Figure 3–Figure supplement 4. Power graph associated to the minimal communities producing the aromatic compounds group of targets

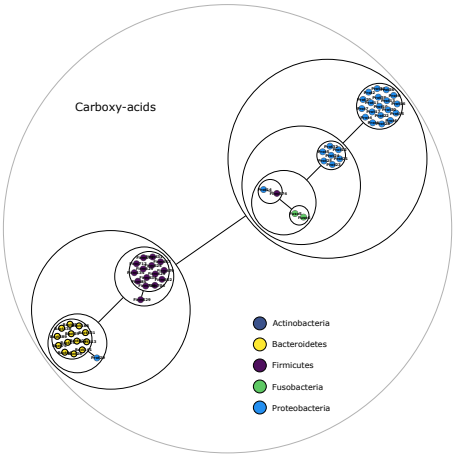


Figure 3–Figure supplement 5. Power graph associated to the minimal communities producing the carboxy-acids group of targets

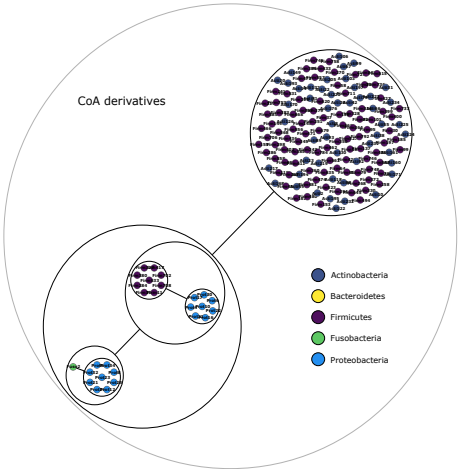


Figure 3–Figure supplement 6. Power graph associated to the minimal communities producing the coenzyme A derivatives group of targets