Genome Biology

SOFTWARE

Open Access

gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models

Johannes Zimmermann[1], Christoph Kaleta[1] and Silvio Waschina[1,2*]

*Correspondence:
s.waschina@nutrinf.uni-kiel.de
[1]Christian-Albrechts-University Kiel, Institute of Experimental Medicine, Research Group Medical Systems Biology, Michaelis-Str. 5, 24105 Kiel, Germany
[2]Christian-Albrechts-University Kiel, Institute of Human Nutrition and Food Science, Nutriinformatics, Heinrich-Hecht-Platz 10, 24118 Kiel, Germany

Abstract

Genome-scale metabolic models of microorganisms are powerful frameworks to predict phenotypes from an organism's genotype. While manual reconstructions are laborious, automated reconstructions often fail to recapitulate known metabolic processes. Here we present `gapseq` (https://github.com/jotech/gapseq), a new tool to predict metabolic pathways and automatically reconstruct microbial metabolic models using a curated reaction database and a novel gap-filling algorithm. On the basis of scientific literature and experimental data for 14,931 bacterial phenotypes, we demonstrate that gapseq outperforms state-of-the-art tools in predicting enzyme activity, carbon source utilisation, fermentation products, and metabolic interactions within microbial communities.

Keywords: Metabolic pathway analysis, Metabolic networks, Genome-scale metabolic models, Benchmark, Community simulation, Microbiome, Metagenome

Background

> Anything you have to do repeatedly may be ripe for automation.
>
> — Doug McIlroy

Metabolism is central for organismal life. It provides metabolites and energy for all cellular processes. A majority of metabolic reactions are catalysed by enzymes, which are encoded in the genome of the respective organism. Those catalysed reactions form a complex metabolic network of numerous biochemical transformations, which the organism is presumably able to perform [1].

In systems biology, the reconstruction of metabolic networks plays an essential role, as the network represents an organism's capabilities to interact with its biotic and abiotic environment and to transform nutrients into biomass. Mathematical analysis has shown great potential for dissecting the functioning of metabolic networks on the level

of topological, stoichiometric, and kinetic models [2], which together provide a wide array of methods [3]. Although different microbial metabolic modelling approaches exist, they can be summarised by a theoretical framework that provides a unifying view on microbial growth [4]. Metabolic models not only have demonstrated their ability to predict phenotypes on the level of cellular growth and gene knockouts, but also provide potential molecular mechanisms in form of gene and reaction activities, which can be validated experimentally [5–7]. Due to this predictive potential, metabolic models have been applied to identify metabolic interactions between different organisms [8–13], to study host-microbiome interactions [14–16], to predict novel drug targets to fight microbial pathogens [17, 18], and for the rational design of microbial genotypes and growth-media conditions for the industrial production or degradation of biochemicals [19, 20]. Furthermore, recent advances in DNA-sequencing technologies have led to a vast increase in available genomic- and metagenomic sequences in databases [21], which further expands the applicability of genome-scale metabolic network reconstructions.

In the process of genome-scale metabolic network reconstruction, the genomic content of an organism is linked to biochemical processes, including enzymatic reactions and cross-membrane metabolite transport [22]. Therefore, the quality and integrity of network models depend on the genome sequence annotation and the underlying reaction and transporter database [22, 23]. Advances in the computational annotation of genomes and the massive increase of biochemical knowledge stored in online databases [24–26] have prompted the development of several software approaches to automate the reconstruction process [27]. A recent study by Mendoza et al. comprehensively compared seven current genome-scale metabolic reconstruction tools [28], namely AuReMe [29], CarveMe [30], Merlin [31], MetaDraft [32], ModelSEED [33], Pathway Tools [34], and RAVEN [35]. On the basis of 18 specific criteria, Mendoza et al. concluded that each tool displayed strengths and shortcomings in different aspects [28]. One of the comparison criterion was the ability of the software to provide a 'ready-to-use' model as output, where the 'use' refers to the possibility to perform flux balance analysis (FBA [36]) or FBA-derived simulation techniques to predict the organism's metabolic physiology, including biomass production, under a given chemical environment. This criterion was fully met only by CarveMe and ModelSEED [28].

The feature to directly obtain network models that can be used for FBA-based growth simulations is especially powerful in situations where large numbers of new microbial genomes are assembled from high-throughput metagenomic datasets [37]. In such studies, the models can be used to predict physiological properties of the sampled microbial community, including metabolite cross-feeding interactions between species. However, a fundamental issue with automatically reconstructed genome-scale models is that their physiological predictions (e.g. using FBA) are often inaccurate [38]. Since the reconstruction process involves various steps, the causes for false metabolic flux predictions from automatic reconstructions can be manifold: First, inconsistencies in databases can lead to an incorporation of imbalanced reactions into the metabolic network, which may become responsible for incorrect energy production by futile cycles [22]. Second, many genes are lacking a functional annotation due to a lack of knowledge [39] and, thus, also the gene products cannot be integrated into the metabolic networks, which potentially lead to gaps in pathways. And third, the gap-filling of metabolic networks is frequently done by adding a minimum number of reactions from a reference database that facilitate growth under a

chemically defined growth medium [33, 40, 41]. Such approaches miss further evidences potentially hidden in sequences and are biased towards the growth medium used for gap-filling.

The potential of automated reconstruction tools to directly predict metabolic-physiological properties of organisms based on their genome sequence was so far only evaluated on the basis of smaller experimental data sets from model laboratory strains such as *Escherichia coli* K12 or *Bacillus subtilis* 168. The overall performance of reconstruction tools, particularly for non-model organisms, is therefore insufficiently assured. Yet, accurate phenotype predictions for a wide range of organisms is crucial for the broad application of automated network reconstruction pipelines in research. For instance, genome-scale metabolic network reconstructions are increasingly applied to simulate complex metabolic processes in microbial communities [42, 43]. Such simulations are highly sensitive to the quality of the individual metabolic networks of the community members. This is because the accurate prediction of by-products and carbon source utilisation is crucial for the correct prediction of metabolic interactions since the substances produced by one organism may serve as resource for others [44]. Thus, in multi-species communities, the metabolic fluxes of organisms are intrinsically connected, which can lead to error propagation when one defective model affects otherwise correctly working models. As a consequence, the feasibility of community modelling fundamentally depends on the accuracy of the individual organismal models.

In this work, we present `gapseq` a novel software for pathway analysis and metabolic network reconstruction. The pathway prediction is based on multiple biochemistry databases that comprise information on pathway structures, the pathways' key enzymes, and reaction stoichiometries. Moreover, `gapseq` constructs genome-scale metabolic models that enable FBA-based metabolic phenotype predictions as well as the application in simulations of community metabolism. Models are constructed using a manually curated reaction database that is free of energy-generating thermodynamically infeasible reaction cycles. As input, `gapseq` takes the organism's genome sequence in FASTA format, without the need for an additional annotation file. Network topology as well as sequence homology to reference proteins inform the filling of network gaps. A novel Linear Programming (LP)-based gap-filling algorithm identifies and resolves gaps in order to enable biomass formation on a given medium. In addition, the algorithm also identifies and fills gaps in metabolic functions, whose presence in the network is supported by sequence homology to reference proteins and which are likely to be relevant for growth in environments that are different to the chosen gap-filling medium. This approach reduces the gap-filling medium-specific effects on the final network structures and thereby increases the versatility of `gapseq` models for subsequent physiological predictions under various chemical growth environments. Finally, we use large-scale phenotype data sets to validate enzyme activity, carbon source utilisation, fermentation products, gene essentiality, and metabolite cross-feeding interactions in microbial communities. The results obtained with `gapseq` are benchmarked against CarveMe [30] and ModelSEED [33], as these tools also provide the full procedure to construct models, which can directly be employed for FBA-based metabolic flux simulations of microbial growth.

Results

Biochemistry database and universal model

The pathway, transporter, and complex prediction is based on a protein sequence database that is derived from UniProt as well as TCDB and consists in total of 131,207 unique sequences (112,056 reviewed unipac 0.9 clusters and 19,151 TCDB transporter) and also 1,138,176 unreviewed unipac 0.5 cluster that can be included optionally. The reference protein sequences are regularly updated by the `gapseq` maintainers using the latest UniProt and TCDB releases. `gapseq` automatically checks for updates and retrieves the latest reference sequences upon start of the software. For the construction of genome-scale metabolic network models we have built a biochemistry database, that is derived from the ModelSEED biochemistry database. In total, the resulting curated `gapseq` metabolism database comprises 15,150 reactions (including transporters) and 8446 metabolites. All metabolites and reactions from the biochemistry database are incorporated in the universal model that `gapseq` utilises for the gap-filling algorithm. If all dead-end metabolites and corresponding reactions would be removed, the universal model comprises 10,792 reactions and 3885 metabolites. However, since genome-scale metabolic networks are also used as structured knowledge-bases, no dead ends are removed from the universal model. It needs to be noted, that the current biochemistry database and the derived universal model represents mainly bacterial metabolic functions and that, at the current version of `gapseq`, the database does not include all archaea-specific nor eukaryotic-specific reactions. However, those reactions and, thus, also the possibility to use `gapseq` for the reconstruction of archaeal and eukaryotic models will be included in a later version of the software.

Enzymatic data

Microbial isolates are commonly subject to laboratory enzyme activity tests for strain characterisation and identification. The Bacterial Diversity Metadatabase (BacDive) provides results from enzyme activity tests spanning a wide taxonomic range and different enzymes [45]. This data represents highly valuable phenotypic information that can be used to scrutinise whether metabolic network models of microorganisms also harbour the enzymatic reaction that was experimentally tested. Here, we performed this evaluation for automated network reconstructions obtained with the tools CarveMe [30], ModelSEED [33], and our `gapseq` approach.

In total, we compared 10,538 enzyme activities, which consists of data for 3017 organisms and 30 unique enzymes. For all organisms, genome-scale metabolic models were constructed using the three different software tools. `gapseq` models had with 6% the lowest false negative rate compared to CarveMe (32%) and ModelSEED (28%). Correspondingly, `gapseq` showed with 53% also the highest true positive rate compared to CarveMe (27%) and ModelSEED (30%), while the rates of false positive and true negative predictions were comparable (Fig. 1a). For this test, the most prominent EC numbers were the catalase, 1.11.1.6, accounting for 26% of the comparisons and the cytochrome oxidase, 1.9.3.1, accounting for 22%, which reflects the ecological importance of cytochrome oxidases and catalases as proxy for an aerobic lifestyle. The overall results remain stable when sampling equal numbers of test data for each EC number and thereby controlling for a potential bias by the over-representation of these EC numbers (Additional file 1: Fig. S4).

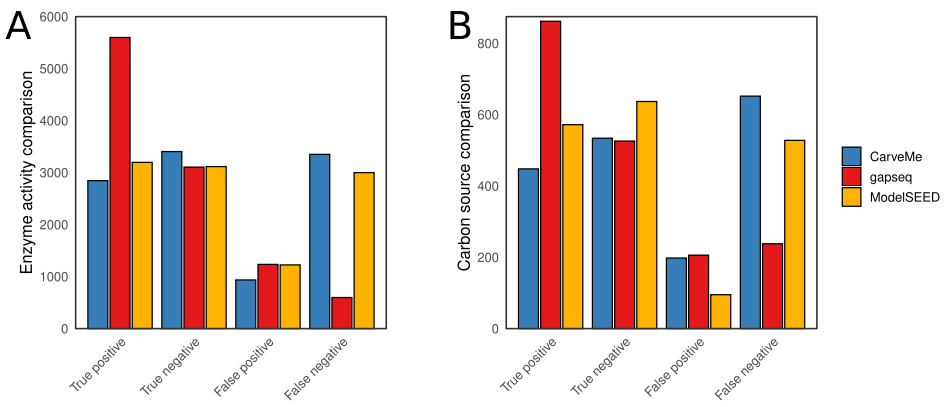

Fig. 1 Results from enzyme activity and carbon source validations. **a** In total 10,538 enzyme activities (30 enzymes and 3017 organisms) of experimental data from the DSMZ BacDive database [45] were compared for three different methods. **b** The predictions of 1795 carbon sources (48 unique carbon sources and 526 organisms) were evaluated with data from the ProTraits database [46]

Carbon source utilisation

The bacterial kingdom comprises a tremendous diversity in carbon source utilisation strategies. In the context of genome-scale metabolic modelling, a major challenge is the accurate prediction of carbon source utilisation phenotypes from an organism's genome sequence. In order to evaluate `gapseq`'s potential to predict carbon source utilisation capabilities we retrieved data on bacterial phenotypes from the ProTraits resource [46]. In brief, ProTraits provides information on phenotypic traits, including carbon source utilisation, of individual microorganisms, where the phenotypic trait data is inferred from scientific literature and comparative genomics. Here, we evaluated the quality of automated model reconstruction pipelines by testing if the models are able to recapitulate carbon source utilisation phenotypes as indicated in ProTraits.

In summary, we compared 1795 different carbon source utilisation predictions for 526 organism and 48 carbon sources (Fig. 1b). `gapseq` outperformed the other methods in terms of false negatives (13% compared with 29% ModelSEED and 36% CarveMe) and true positives (47% compared with 31% ModelSEED and 24% CarveMe). ModelSEED showed fewer false positives (5% compared with 11% `gapseq` and 11% CarveMe) and more true negatives (35% compared with 29% `gapseq` and 29% CarveMe). `gapseq`, predicted most false positives for formate (29 times). This overestimate of formate as potential carbon source is likely due to the fact that we tested carbon source utilisation on the basis of electron transfer from the source to electron carriers (i.e. ubiquinol, menaquinol, or NADH), which is analogous to the experimental carbon source test of BIOLOG plates [47]. However, while it is known that formate can serve in fact as electron donor in a number of different bacteria [48], the role as source of carbon atoms for the synthesis of biomass components is limited to a few known methylotrophs [49]. Across all methods, the most accurately predicted carbon sources, with more than 100 tested organisms, were fructose (92% correct predictions), mannose (91%), or arginine (82%), whereby the predictions were less accurate for arabinose (29% correct predictions), dextrin (41%), or acetate (51%).

In general, we note that testing carbon source utilisation via the proxy of electron transfer from the substrate to reducing equivalents has the advantage that one can test a vast

number of model reconstructions without the need to define a complete chemical growth environment that contains besides the carbon source also all other compounds required for growth (e.g. specific amino acids in case of auxotrohies). However, this approach has the shortcoming that in some cases, the ability of an organism to use a substance as electron donor does not always imply that the substance can also be used as source of carbon. Nevertheless, we argue that the implemented carbon source utilisation prediction is pertinent as it reflects the same approach as BIOLOG plates, which is an established system for carbon source utilisation profiling.

Gene essentiality

We compared the ability of `gapseq` models to predict the essentiality of genes with predictions from ModelSEED and CarveMe reconstructions as well as with curated models for the same organisms (Fig. 2). As expected, the curated models outperformed all three automated reconstruction tools for most species and prediction metrics (namely precision, sensitivity, specificity, accuracy, and F1-score). Interestingly, for *Shewanella oneidensis* and *Pseudomonas aeruginosa* `gapseq` reconstructions outperformed curated models in most test scores with the exceptions of the sensitivity in the case of *S. oneidensis* and specificity for *P. aeruginosa* (Fig. 2c, d). Compared to CarveMe, `gapseq` showed in four out of five cases a higher sensitivity in essentiality predictions but, at the same time, a slightly lower specificity. This pattern is attributed to the fact that `gapseq` models tend to predict more genes as essential than CarveMe, leading to a higher number of true positive (TP) predictions but also more false positives (FP). For most organisms and on the basis of most prediction metrics, `gapseq` outperformed network models that were

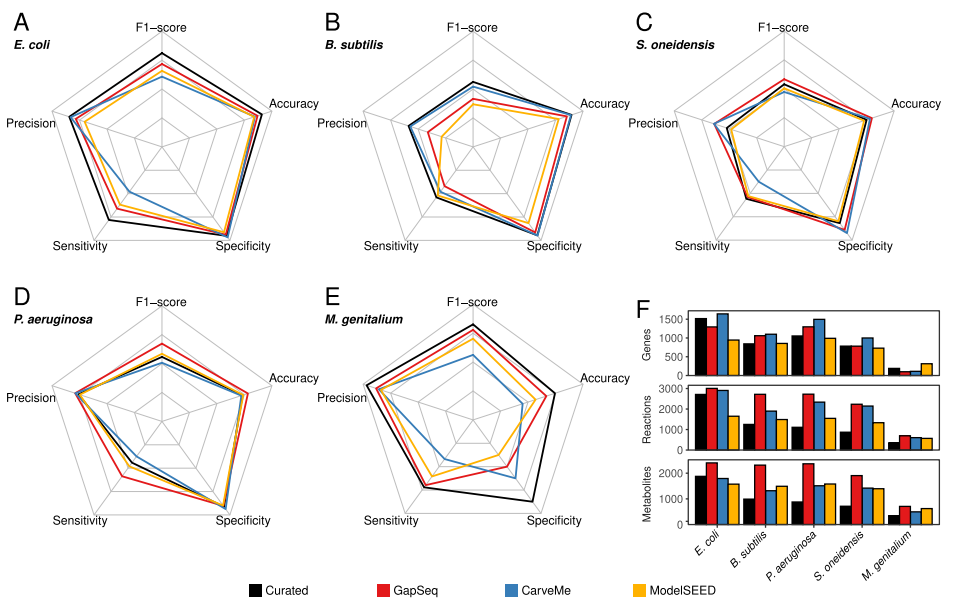

Fig. 2 Results from model gene essentiality tests for five bacterial species. **a** *Escherichia coli*. **b** *Bacillus subtilis*. **c** *Shewanella oneidensis*. **d** *Pseudomonas aeruginosa*. **e** *Mycoplasma genitalium*. Results from `gapseq` models (red) are compared to CarveMe (blue) and ModelSEED (yellow) models, as well as to published curated genome-scale metabolic models (black) of the respective organisms. Radar chart axes scales are linear with 0 in the centre and 1 at the corners. **f** Counts of genes, reactions (including exchanges and transporters), and metabolites in each reconstruction

reconstructed using ModelSEED. The results presented here consider genes as essential, if the predicted growth rate of the focal gene-knockout strain was below 0.01 h^{-1}. However, we note that the results remained virtually unaltered with a higher (0.05 h^{-1}) or lower (0.001 h^{-1}) threshold (Additional file 1: Fig. S1).

Accurate gene essentiality predictions rely on precise gene-protein-reaction (GPR) associations, which are formulated as Boolean expressions to describe the reactions' dependence on proteins and the corresponding protein-encoding genes. The automated prediction of GPR associations is especially challenging for reactions that depend on protein complexes consisting of different protein/peptide subunits. We compared the GPR expressions for such reactions in the metabolic network of *E. coli* between the manually curated network (iML1515) and the automated reconstructions from CarveMe, ModelSEED, and `gapseq` (Additional file 2: Table S6). 59 protein complex-associated reactions were shared among all networks. Considering the GPR associations of the curated network as reference, only 6% were equivalent to those in the CarveMe network, 10% for ModelSEED, and 19% for `gapseq`. These results suggest, that accurate GPR association predictions are still a weakness in the tested automated reconstruction tools and thereby limit the essentiality predictions of individual genes, which encode protein subunits.

Fermentation products

Anaerobic or facultative anaerobic bacteria utilise different fermentation pathways in order to extract energy from environmental compounds by chemical transformations in the absence of oxygen. We tested if fermentation products can be predicted by metabolic reconstructions obtained from `gapseq`, CarveMe, and ModelSEED for 24 different bacterial organisms (Fig. 3). The organisms were selected based on following criteria: (1) the organisms have a published RefSeq genome sequence, (2) are known anaerobic or facultative anaerobic organisms, and (3) the identity of fermentation products has been experimentally described and reported in primary literature (Additional File 2: Table S2). Overall, `gapseq` showed the highest number of true positive predictions (TP) with 50 TP predicted with the Minimise-Total-Flux (MTF) and 51 TP predicted with Flux-Variability-Analysis (FVA) which is substantially higher compared to CarveMe (15 TP with MTF, 16 TP with FVA) and ModelSEED (2 TP, 4 TP). The production of the short-chain fatty acids acetate, butyrate, and propionate was correctly predicted (TP) by `gapseq` in 91% of cases and thereby outcompetes CarveMe (12%) and ModelSEED (0%), which did not predict butyrate or propionate production for any tested organism. Moreover, `gapseq` correctly predicted homolactic fermentation by *Lactobacillus delbrueckii* and *Lactobacillus acidophilus*, which is dominated by lactate as fermentation end product and also predicted known heterolactic fermentation by *Bifidobacterium longum*, *Bifidobacterium animalis*, and *Lactobacillus plantarum*. However, `gapseq` failed to predict lactate production of organisms that utilise different fermentation strategies, which also yield lactate (e.g. mixed-acid fermentation by *Escherichia coli*). Interestingly, the predicted quantities of fermentation product release is higher for true positive than for false negative predictions (Fig. 3). This further suggests, that `gapseq` is able to predict the main fermentation products of bacterial organisms during anaerobic growth.

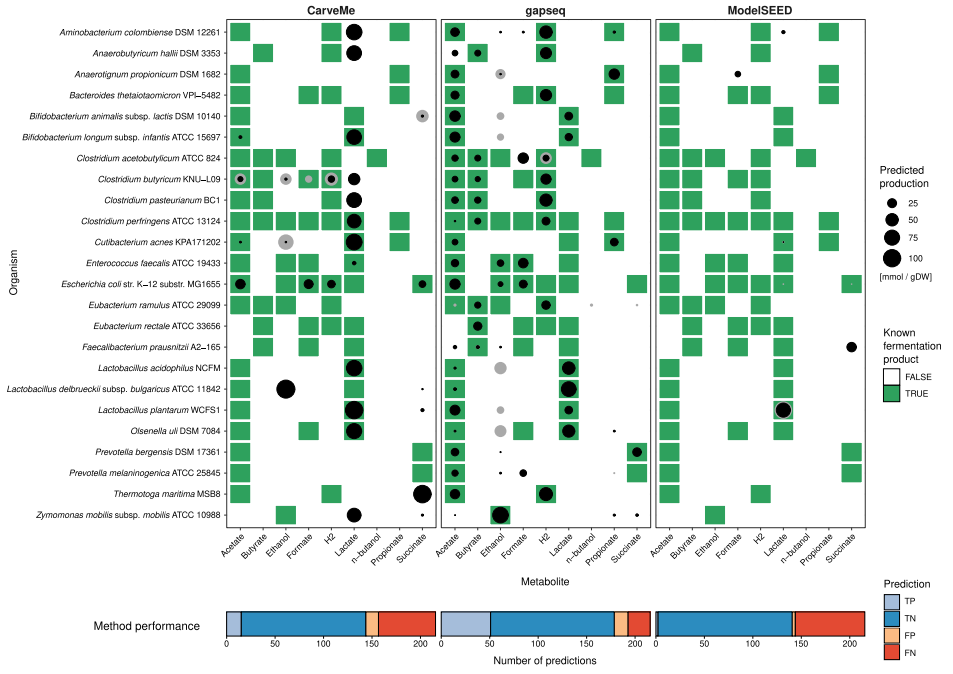

Fig. 3 Results of the fermentation product test of 24 bacterial organisms under anaerobic growth with models generated using `gapseq`, CarveMe, and ModelSEED. Point sizes indicate the predicted production of a fermentation product metabolite (columns) by the corresponding organism (row). Predictions (black) are based on Minimise-Total-Flux (MTF) flux balance analyses. Grey circles indicate the upper production limit obtained from Flux-Variability-Analysis (FVA). Metabolite-organism-combinations highlighted in green denote known fermentation products, which have been reported in literature based on experimental measures of the metabolite in anaerobic cultures

Anaerobic food web of the gut microbiome

The prediction of metabolic interactions between microbial organisms is of special interest in ecology, medicine, and biotechnology. So far, we showed the capacity of `gapseq` on the level of individual models. In a next step, we simulated several individual models together as a multi-species community to validate the potential of `gapseq` in microbial community modelling. As sample application we selected representative members of the gut microbiome that are known to form an anaerobic food web [50, 51]. Altogether, we employed 20 organisms and simulated the combined growth in a shared environment for several time steps using the community modelling framework BacArena [52]. BacArena permits a dynamic and spatial simulation of individual models which are optimised separately in a shared growth environment. Based on metabolic models and environmental substance availability, BacArena predicts growth and nutrient exchanges of individual microorganisms and overall alteration in substance concentrations. Metabolite production and consumption rates by individual community members was analysed at time step 3 for CarveMe and `gapseq` and at time step 5 for ModelSEED, to ensure the community metabolism is captured during the exponential growth phase before the inflection point (Fig. 4).

On the community level, simulations using `gapseq` models captured the central substances, which are known to be produced in the context of the food web (Fig. 4). This included the production of short-chain fatty acids (acetate, propionate, butyrate), lactate,

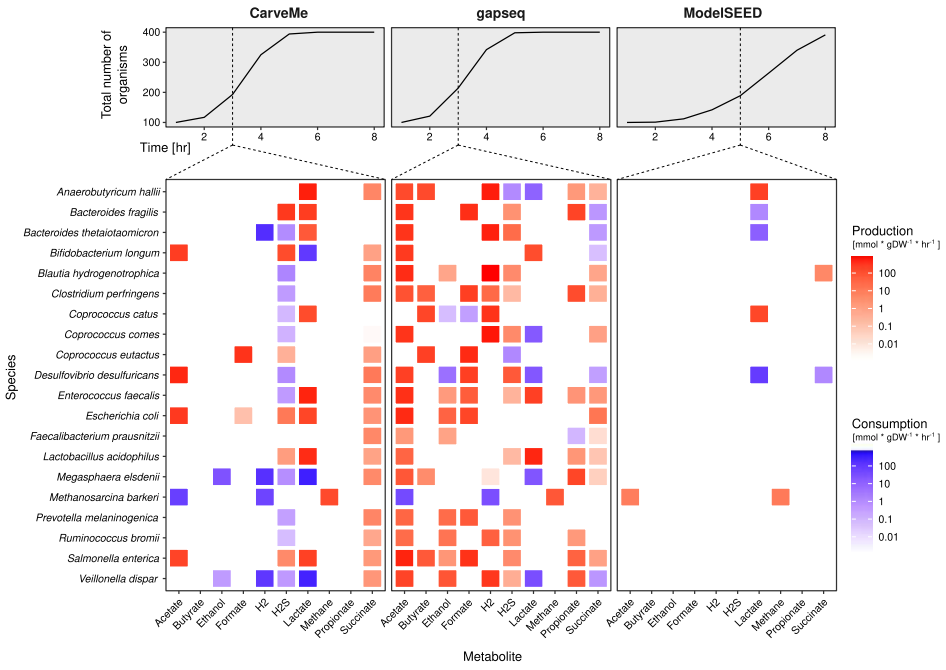

Fig. 4 Predicted metabolic products and food web of an anaerobic microbial community. The metabolism of a community consisting of 19 bacterial species commonly found in the human gut and one archaeon (*Methanosarcina barkeri*) was predicted using BacArena [52]. Growth curves show the sum of organisms in the shared environment over the simulation time. The heatmaps display the predicted metabolite production (red) and consumption (blue) during exponential growth before the inflection point. Production and consumption rates at all time steps are shown in Additional file 1, Fig. S5. All bacterial models were reconstructed by CarveMe, `gapseq`, or ModelSEED, with the exception of *M. barkeri* for which a published and manually curated model [117] was used

hydrogen, hydrogen sulphide (H_2S), methane, ethanol, formate, and succinate. The formation of acetate, formate, and hydrogen was most prevalent, which are also common end products of intestinal fermentation. With the exception of butyrate and methane, parts of the produced fermentation products are further metabolised by some community members (Fig. 4). The predicted identity of fermentation end products and other by-products of metabolism was found in most cases to be closely in line with literature information [50, 51, 53]. For example, the formation of lactate was observed in the simulation for *Lactobacillus acidophilus*, *Enterococcus faecalis*, and *Bifidobacterium longum*, and butyrate was released by known butyrate producers, including *Anaerobutyricum hallii*, *Clostridium perfringens*, *Coprococcus* spp., and *Megasphaera elsdenii*. Yet, the predictions did not include known butyrate production by *Faecalibacterium prausnitzii*. In general, the main products of mixed acid fermentation (acetate, formate, hydrogen, ethanol, succinate) were predicted for diverse members of the community which is in agreement with what is known about common metabolic end products of many gut-dwelling microorganisms [53]. Specifically, high levels of H_2 production was correctly predicted for known hydrogen producers including *A. hallii*, *Bacteroides thetaiotaomicron*, *Coprococcus catus*, *Coprococcus comes*, and *Veillonella dispar*.

In general, the anaerobic oxidation of fatty acids is not favoured by the gut environment because the host competes for the uptake of butyrate, propionate, and acetate, which serve as energy source for colonic epithelial cells and are involved in many host functions [54].

Therefore, the gut community lacks syntrophic organisms which are able to anaerobically degrade butyrate [55]. In agreement with this, we found no microbial uptake of butyrate in the community simulation. In contrast, cross-feeding interactions that involve the uptake of metabolites such as acetate, lactate, succinate, and hydrogen are important components in the microbial ecology within the large intestine of humans [56–58]. In our simulations, lactate was predicted to be produced and consumed by distinct community members. We found utilisation of lactate by *A. hallii*, *C. comes*, *Desulfovibrio desulfuricans*, *M. elsdenii*, and *V. dispar*, which is a known feature of these organisms [50, 59]. In addition, succinate was correctly predicted to be utilised by *Bacteroides* species [53]. The formation of methane is known to be limited to methanogenic archaea, and thus *Methanosarcina barkeri* produced methane from acetate and hydrogen during our simulations. It also needs to be noted that certain known cross-feeding interactions were not observed in the community simulations. *A. hallii* and *F. prausnitzii* have been described to consume acetate that is produced by other community members, yet, this cross-feeding is not part of the predicted food web (Fig. 4). Also, no utilisation of gut bacteria-derived hydrogen by *Blautia hydrogenotrophica* as source of energy [60] was predicted. In order to investigate the causes of missing metabolite consumption predictions, the uptake fluxes of *A. hallii*, *F. prausnitzii*, and *B. hydrogenotrophica* were analysed. All three organisms utilised saccharides (i.e. glucose and fructose) as main sources of energy instead of acetate (*A. hallii* and *F. prausnitzii*) or H_2 (*B. hydrogenotrophica*). This suggests, that the correct prediction of the anaerobic utilisation of low energy-yielding substrates, such as acetate, remains a challenge for automatic model reconstructions. Specifically acetate was also identified in the carbon source test (see Carbon source utilisation), whose utilisation predictions failed to recapitulate reported acetate utilisation properties of bacteria in nearly half of the cases.

For comparison, the community simulations were also performed using models reconstructed with CarveMe and ModelSEED (Fig. 4 and Additional file 1: Fig. S2). In both cases, most of the above-mentioned known metabolic cross-feeding interactions and end products were not predicted. For instance the production of the short-chain fatty acids butyrate and propionate was missing. The expected consumption of H_2 by *B. hydrogenotrophica* and acetate by *A. hallii* and *F. prausnitzii*, which were not predicted in the community simulations using gapseq models were also missing in the simulations with CarveMe and ModelSEED reconstructions.

In summary, `gapseq` models were able to recapitulate pivotal interactions, which are described for microbial communities in the human gut. While not all expected cross-feeding interactions were recapitulated in the community simulation, important individual contributions to the production and consumption of microbial metabolites in an anaerobic environment were predominantly found to be in close agreement with literature data. Taken together, the community simulation results illustrate the capacity of `gapseq` to construct predictive models for complex metabolic interaction networks comprising several different species.

Pathway prediction of soil and gut microorganisms

To demonstrate the pathway prediction capabilities of `gapseq`, we analysed two communities of soil and gut microorganisms comprising 922 and 822 organisms, repectively. The two communities could be separated from each other by differences in energy metabolism (principal component analysis, Fig. 5a). Here, most variance was explained by subsystems

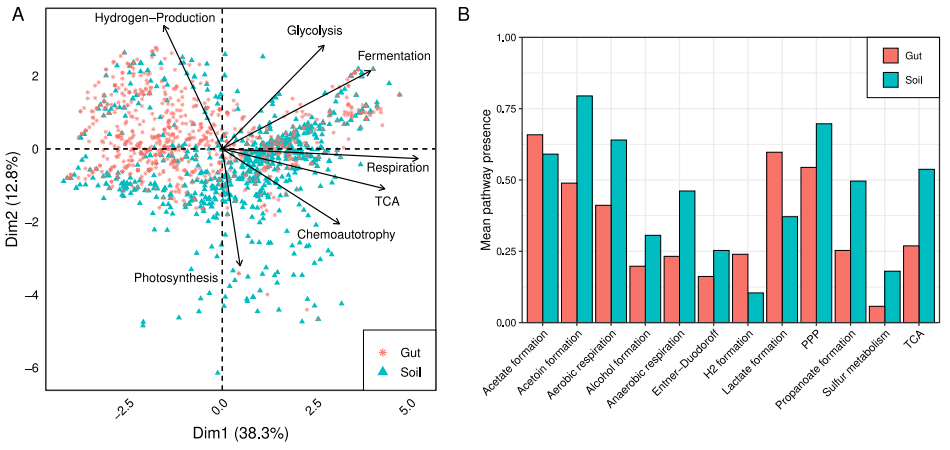

Fig. 5 Comparison of energy metabolism between soil and gut community. **a** A PCA plot with the first two dimension explaining more than 50% of the variance. Selection of subsystems from energy metabolism with highest quality and impact are shown. **b** List of subsystems of energy metabolism that differ significantly in frequency between members of the soil and gut community (TCA, tricarboxylic acid cycle; PPP, Pentose phosphate pathway)

of pathways that are involved in chemoautotrophic, respiratory, and fermentative processes including hydrogen production. Out of 128 energy pathways, the presence of 40 pathways differed significantly (Kolmogorov-Smirnov test, $P < 0.05$) between soil and gut microorganisms and could be categorised into 12 subsystems (Fig. 5b). In total, gut microorganisms showed less variety in energy pathways than soil microorganisms. Only pathways relevant for the formation of acetate, hydrogen, and lactate were predicted to be enriched. In the case of all other energy subsystems, more pathways were predicted for soil organisms, most prominently pathways relevant for aerobic and anaerobic respiration as well as the tricarboxylic acid cycle (TCA). In summary, members of the soil community showed a more versatile energy metabolisms, which potentially indicates a higher energetic specialisation of gut microorganisms. This sample application demonstrates how `gapseq` can facilitate the characterisation and comparison of microbial communities based on the analysis of the presence and absence of specific metabolic pathways.

Model reconstructions for metagenomic assemblies

Genome-scale metabolic models can also be reconstructed on the basis of species-level genome bins (SGBs, [61]) assembled from shotgun metagenomic sequencing reads. Yet, genome assemblies from metagenomic material are more prone to errors, fragmentation, and sequence gaps than assemblies of isolated genomes [62], which can potentially cause gaps in the metabolic network reconstructions. We tested whether `gapseq` is able to identify and fill such gaps by comparing the models reconstructed for 127 SGBs from the human microbiome [61] to corresponding models of closely related reference genomes that were assembled from DNA-sequencing of pure cultures (Additional file 1: Fig. S3).

As expected, we found a strong positive correlation between the SGBs' genome completion and their model similarity to their respective reference models (Spearman's rank correlation, n = 127, $P < 10^{-9}$). To estimate the quantitative effect of genome completion on the model similarity, a logarithmic function ($y(x) = c + b * \log(x)$) was fitted to

the data ($R^2 = 0.71$, Additional file 1: Fig. S3). The fitted model indicated, that gapseq is able to reconstruct the underlying metabolic network of an organism even on the basis of incomplete and fragmented genomes. For instance, `gapseq` was on average able to recover 90% of the enzymatic reactions that are found in the reference models for SGBs with a predicted genome completion of only 80% (Additional file 1, Fig. S3).

Summary of validation tests

In summary, gapseq was evaluated on the basis of five validation tests: (1) The predictions of specific enzymes were compared to experimental data of enzyme activities for a wide range of bacterial strains. The experimental data was retrieved from the BacDive database [45]. (2) The ability of bacterial metabolic models to utilise certain carbon sources was scrutinised by comparing predicted utilisation with data from ProTraits [46], a resource of 424 literature- and genome-inferred prokaryotic phenotypes for more than 3000 organisms. (3) Predicted essentiality of genes was evaluated on the basis of in silico gene-knockout simulations and empirical essentiality data from single gene-knockout studies spanning five bacterial strains. (4) Predicted fermentation products of 24 bacteria in an anaerobic environment were contrasted with fermentation end products reported in scientific literature. (5) An anaerobic microbial community was simulated with reconstructed metabolic models in a shared in silico growth environment. Predicted metabolite production and consumption was compared to those reported in scientific literature.

The overall accuracy (proportion of all correct prediction in relation to all predictions made) of model predictions with empirical data was 66% (CarveMe), 70% (ModelSEED), and 81% (`gapseq`)(Table 1). Sensitivity measures the proportion of correctly predicted positives, whereas specificity accounts for the correct prediction of negatives. All approaches showed a high specificity >0.7 with highest values for fermentation product and gene essentiality tests. Notably, `gapseq` showed the highest sensitivity over all tests (Fig. 6). In summary, `gapseq` outperformed other methods in terms of accuracy and sensitivity while showing similar specificity.

Table 1 Summarised comparison of CarveMe, gapseq, and ModelSEED

Metric	CarveMe	gapseq	ModelSEED
Implementation			
Infrastructure	Local	Local	Web service
Input (FASTA file)	Protein	Nucleotide	Nucleotide
Programming languages	Python	Shell script, R	Perl/javascript
Gap-fill solver	CPLEX	GLPK/CPLEX	Not needed*
Gap-fill problem formulation	MILP	LP	MILP
Performance			
Accuracy	0.66	0.80	0.69
Sensitivity	0.34	0.71	0.33
Specificity	0.85	0.82	0.88
Model file quality**	0.32 ± 0.006	0.78 ± 0.004	0.39 ± 0.016

Accuracy, sensitivity, and specificity scores are based on 14,931 tested phenotypes including energy sources, enzyme activity, fermentation products, gene essentiality, and anaerobic food web structure predictions.
*Solver runs on ModelSEED server. No local solver is required.
**MEMOTE total score mean (± SD).

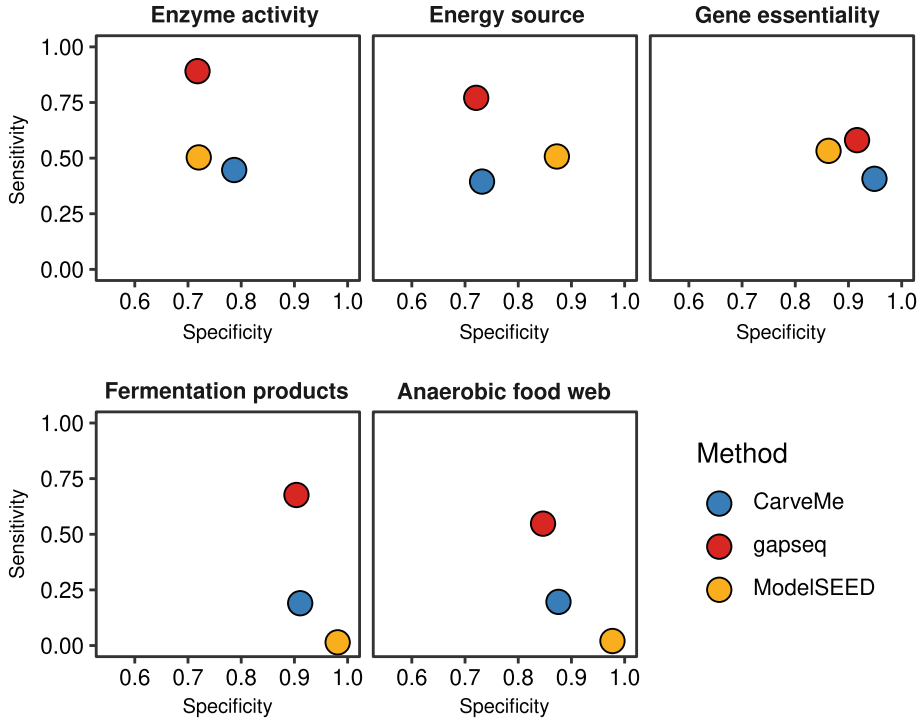

Fig. 6 Summary of different validation tests. The specificity and sensitivity for all compared methods are shown. This includes results from benchmarks concerning enzyme activities, energy sources, fermentation products, gene essentiality, and metabolite production/consumption in an anaerobic food web

Comparison with current genome-scale metabolic network reconstruction tools

The above benchmark tests compared `gapseq` with two other tools (CarveMe and ModelSEED) which are able to reconstruct models that enable FBA-simulations of cell growth to predict reaction activity. In this subsection and on the basis of key criteria defined by Mendoza et al. (2019) [28] for the assessment of reconstruction software, `gapseq` is compared to a broader range of currently available network reconstruction tools (Table 2).

Table 2 Comparison of gapseq (GS) with other reconstruction tools based on criteria defined by Mendoza et al., 2019 [28]

Feature/comparison criterion	AU	CM	ME	MD	MS	PT	RA	GS
Software maintenance/support/updates	●	●	●	●	●	●	●	●
Eukaryotes model support	●	○	●	●	●	●	●	
SBML level 3 as output	●	●	●	●			●	●
User-friendly interface	○	○	●	●	●	●	○	○
Open Source (source code is open to all users)	●	●		●	●		●	●
Automatisation until FBA-functional models*	○	●		○	●	○	○	●
Manual refinement assistance	○		●			●	○	○
Customisable for a high number of genomes		●			●	○	●	●
Traceability	○		○			○		●
Automatic refinement using experimental data	○					○		

Evaluations for AuReMe (AU, [29]), CarveMe (CM, [30]), merlin (ME, [31]), MetaDraft (MD, [32]), ModelSEED (MS, [33]), Pathway Tools (PT, [34]), and RAVEN (RA, [35]) are directly adopted from Mendoza et al., (2019) [28], with the exception of the SBML output, where methods were only classified based on whether the export in SBML level 3 is supported. Legend: ● - outstanding; ○ - good to satisfactory; no circle - poor to unsatisfactory

*Models with FBA-predicted flux through biomass reactions on a given growth medium

The comparison aims to aid potential users to decide when to use `gapseq` and when other tools might be more fitting to their specific research question.

As all other tools listed in Table 2, `gapseq` is maintained by a core team of scientists that provides updates and user support. Issues can directly be reported and additional features requested at the github repository [63], where also the latest version of the software (incl. its source code) can be obtained. `gapseq` exports models in standard SBML level 3 format [64], which enables the integration of `gapseq` in pipelines that further analyse the models with other tools for constraint-based analysis. Additionally, gapseq stores models as R-objects of class *modelorg*, which can be analysed in R using the *sybil* package [65]. As mentioned above, `gapseq` enables the full automatisation of the reconstruction process from the genome to a FBA-functional model that allows growth predictions for the focal organism under a given growth environment. This feature is only shared with CarveMe and ModelSEED and is especially relevant in situations where large numbers of genomes are subject to genome-scale network reconstruction and directly subsequent metabolic flux simulations of microbial growth. Another advantage of `gapseq` is the high traceability of reactions and metabolites throughout the entire reconstruction process: In the final model, `gapseq` adds a flag to each reaction that denotes why the specific reactions have been added to the network, e.g. due to sequence homology to reference proteins, or at which gap-filling step. This information is stored in the reactions' attributes within the model's R-object and could be highly relevant for further manual refinement of the network model.

For certain metabolic network reconstruction projects, other tools than `gapseq` might be more fitting: (i) `gapseq` does not support the construction of genome-scale models for eukaryotic organisms. The tools AuReMe, merlin, MetaDraft, ModelSEED, Pathway Tools, and RAVEN explicitly provide this feature (Table 2). (ii) `gapseq` does not offer a graphical user interface, which might be a hurdle for users less accustomed to command line software tools. (iii) In its current version, `gapseq` does provide a function (`./gapseq adapt`) that allows users to manually add or remove reactions or pathways from a reconstructed network model. However, Pathway Tools and merlin offer workspaces with extended functions and assistance for manual refinement, including network visualisation. (iv) Finally, options to automatically refine models based on experimental data is not yet implemented in `gapseq`, while the tools AuReMe and Pathway Tools provide this feature for certain data types.

Discussion

Here, we introduced `gapseq`—a new tool for metabolic pathway analysis and genome-scale metabolic network reconstruction. The novelty of `gapseq` lies in the combination of (i) a novel reaction prediction that is based both on genomic sequence homology as well as pathway topology, (ii) a profound curation of the reaction and transporter database to prevent thermodynamically infeasible reaction cycles, and (iii) a reaction evidence score-oriented gap-filling algorithm. In order to scrutinise `gapseq` metabolic models, we compared the models' network structures and predictions with large-scale experimental data sets, which were retrieved from publicly available databases. Furthermore, the ability of `gapseq` to predict bacterial phenotypes was compared to two other commonly used automatic reconstruction methods, namely, CarveMe [30] and ModelSEED [33] (Table 1). ModelSEED is also implemented in the KBASE online software platform [66].

Crucial large-scale benchmarking of metabolic models

The quality of genome-scale metabolic networks can be assessed by comparing model predictions with experimental physiological data. The protocol by Thiele and Palsson (2010) for the reconstruction of genome-scale metabolic networks recommends the quality assessment and manual network curation using data for (i) known secretion products (e.g. fermentation end products), (ii) single gene deletion mutant growth phenotypes (i.e. gene essentiality), and (iii) the utilisation of carbon/energy sources [22]. Tools for the automatic reconstruction of metabolic networks should also make use of such physiological data whenever available for benchmarking. Here, we tested our `gapseq` approach on the basis of all three recommended phenotypic data and compared the performance with CarveMe and ModelSEED. Additionally, we included two novel benchmark tests: The comparison of model predictions with (iv) the activity of specific enzymes known from experimental studies [45] and (v) metabolic interactions among microorganisms in a multi-species community within an anaerobic environment (food web). Across all five benchmark tests, we could show that `gapseq` outperformed CarveMe and ModelSEED in terms of sensitivity while achieving specificity scores that are comparable to the other two tools (Fig. 6).

Publicly available genome sequences of microorganisms, which can be subject for automated metabolic network reconstruction are massively increasing in number due to continuing advances in high-quality and high-throughput sequencing technologies [21]. This development is further fuelled by the increasing number of genome assemblies from metagenomic material [67]. In contrast, standardised phenotypic data for microorganisms remains a bottleneck for the validation of automated metabolic network reconstruction pipelines such as `gapseq`. As consequence, it is crucial for the future development of automated network reconstruction software to include possibly all available phenotypic data for benchmarking, especially data from non-model organisms. To benchmark `gapseq` in relation to CarveMe and ModelSEED using phenotypic data from mainly non-model organisms, we retrieved phenotypic data of enzyme activity for more than 3000 organisms and carbon source utilisation for more than 500 organisms from online databases, which is, to our knowledge, the yet largest phenotypic data set used for validation of automatically reconstructed metabolic networks. In this validation approach `gapseq` achieved the highest prediction accuracy among all three tools tested (Fig. 1, Table 1).

Hence, those results suggest that `gapseq` is a powerful new tool for the automated reconstruction of genome-scale metabolic network models. Moreover, the underlying reference protein sequences as well as the pathway database can readily be updated using online resources, which makes `gapseq` flexible to include future developments and findings in microbial metabolic physiology.

Automated network reconstructions for community modelling

While single organisms can be considered as the building blocks of microbial communities, individual metabolic models of organisms are the building blocks of in silico microbial community simulations. Therefore, genome-scale metabolic models are increasingly applied to predict the function of multi-species microbial communities [68–70]. To correctly infer metabolic interaction networks between different organisms, it is important that individual models accurately predict nutrient utilisation (e.g. carbon

source) and metabolic end products (e.g. fermentation products). In this study, the benchmarks for carbon source utilisation and fermentation end product identity indicated that `gapseq` has the highest prediction performance compared to other reconstruction tools (Figs. 1 and 3).

To illustrate the applicability of `gapseq`-reconstructed metabolic models for the simulation of multi-species community metabolism, we generated models for microbial strains from the gut microbiota and simulated their growth in a shared environment. Without further curation, the community simulation reproduced important hallmarks of intestinal anaerobic food webs [50, 53]. Above all, short-chain fatty acids (SCFA) were predicted to be the primary end products of fermentation. This prediction is important to represent intestinal metabolism, because SCFA are crucially involved in host physiology by affecting regulatory response in intestinal and immune cells [71, 72]. Furthermore, the simulation accurately predicted the exchange of metabolites between different members of the microbial community (Fig. 4). Cross-feeding of metabolites and the formation of anaerobic food chains have been associated with a healthy microbiome [11, 73]. For instance, the cross-feeding of lactate has been reported to be vital for the early establishment of a healthy gut microbiota in infants [73]. Accordingly we observed the exchange of lactate between different bacterial species in the community simulations (Fig. 4) and involved known lactate producers (e.g. *Enterococcus faecalis*) and consumers (e.g. *Megasphaera elsdenii*). This example illustrates that we are able to predict key features of the anaerobic food web within the gastrointestinal microbiota using `gapseq` models. In addition to the ability to accurately model metabolic processes within existing microbial communities, `gapseq` will further promote the potential of metabolic modelling to predict how complex microbial communities can be modulated by targeted interventions. Specific interventions, which could for instance be predicted, are the introduction of new species to the community (i.e. probiotics) or microbiome-modulating compounds (prebiotics) to the environment. Predictions of potential intervention strategies which target the microbiome are of vast relevance for biomedical research. Furthermore, metabolic interactions between microbiome members are difficult to detect in vivo due to the simultaneous production and uptake of metabolites. Thus, in silico predictions of metabolite cross-feeding interactions are highly valuable for hypothesis generation about the function and dynamics of microbial communities.

Taken together, the results obtained with `gapseq` suggest, that metabolic models which are reconstructed using `gapseq` are promising starting points to construct ecosystem-scale models of inter-species biochemical processes and to predict targeted strategies to modulate microbiome structure and function.

Pathway analysis of microbial communities

The construction of genome-scale metabolic models is based on metabolic networks that are inferred from genomic sequences in the context of biochemical databases [22]. Although the reconstruction of metabolic networks is closely related to the prediction of metabolic pathways, metabolic modelling and pathway analysis are often treated separately [74]. In `gapseq`, the prediction of metabolic pathways is intrinsically tied to the reconstruction of metabolic networks and gap-filling. In addition, reaction, transporter, and pathway predictions can also be used to evaluate the functional capacities of microorganisms without the need of metabolic modelling. As an example for metabolic pathway

analysis, we compared the predicted energy metabolism of two large microbial communities that occur in soil and the human gut. We could show that the predicted distribution of pathways differ between both communities based on the habitat, which usually accommodates the members of the respective community. Gut microorganisms showed a less versatile energy metabolism and a specialisation towards fermentation pathways, which lead to the formation of acetate, hydrogen, and lactate. Variations in pathways distributions between both communities may be explained by distinct evolutionary histories. The habitat of the diverse group of soil microorganisms more likely represents an open ecosystem, whereas the gut microbiome is directly constraint by a multi-cellular host that potentially affect microbial phenotypic traits [75]. In general, metabolic modelling should be accompanied by the analysis of pathways based on statistical methods [74] to compensate for additional assumptions, which are introduced in constraint-based metabolic flux modelling [4].

Limitations and outlook

`gapseq` requires 1–2 h for the reconstruction of a single model, whereas ModelSEED and CarveMe operate faster (10 min) on a standard desktop computer. Nonetheless, CarveMe needs as input gene sequences (protein or nucleotide), which has to be predicted first, and ModelSEED works as a web service, which can complicate the handling of large-scale reconstruction projects. In `gapseq`, pathways were predicted based on topology and sequence homology searches. However, the assignment of enzymatic function from sequence comparisons has been shown to potentially miss protein domain structures and thus can cause false annotations [76, 77]. In addition, `gapseq` employs many resources to find potential sequences for reactions in pathway databases. Together this might explain why although `gapseq` performed better than other methods on predicting positive phenotypes (function present), it went head to head with regard to negative phenotype predictions (function not present). CarveMe takes a different approach when inferring function by taking care of functional regions (protein domains) [78], resulting in orthologous groups [79], which results in a slightly better specificity (true negative phenotype predictions) in benchmarks (Fig. 6). Future developments of `gapseq` will address orthologous groups by using multiple inference methods. The integration of functional predictions coming from phylogenetic inference without the need of genomic sequences [80] might also be promising for further developments of `gapseq`. Moreover, future versions of `gapseq` will address challenges that we identified in the course of the evaluation tests presented in this study. For instance, Gene-Protein-Reaction (GPR) association predictions will be improved by incorporating new computational methods in protein complex detection [81].

Conclusion

We provide a new software tool called `gapseq` that is suitable for metabolic network analysis and metabolic model reconstruction. To enhance phenotype predictions, `gapseq` employs various data sources and a novel gap-filling procedure that reduces the impact of arbitrary growth medium requirements. We further brought together the so far largest benchmarking of genome-scale metabolic models, in which `gapseq` outperformed comparable alternative tools. With the increased model quality of automated network reconstructions, `gapseq` will provide new insights into the metabolic

phenotypes of non-model and yet-uncultured bacteria whose genomes are assembled from metagenomic material. In this way, the models and their simulations allow predictions on the organisms' ecological role in their natural environments. Taken together, we consider `gapseq` as important contribution to the modelling of microbial communities in the age of the microbiome.

Methods

Program overview and source code availability

The source code is accessible and maintained at https://github.com/jotech/gapseq. The program is called by `./gapseq`, which is a wrapper script for the main modules. Important program calls are `./gapseq find` (pathway and reaction finder), `./gapseq find-transport` (transporter detection), `./gapseq draft` (draft model generation), `./gapseq fill` (gap-filling), or `./gapseq doall` to perform all in line. When ever necessary, method sections directly refer to config, data and source code files from the `gapseq` package, which contains the main sub-directory `src/` with source code files and `dat/`, which contains databases and also the sequence files in `dat/seq/`. Figure 7 shows an overview of the different `gapseq` modules. Documentation and tutorials for gapseq can be found at https://gapseq.readthedocs.io.

Pathway, subsystem, and sequence databases

It is crucial to link pathways and subsystems to protein sequences of the involved enzymes, which can be employed for homology search. Pathways are considered as a list of reactions with enzyme names and EC numbers. In addition, pathways can be generalised as subsystems, which are sets of 'functional roles that together implement a specific biological process' [82]. Pathway and subsystem definitions were obtained from MetaCyc [26], KEGG [24], and ModelSEED [33]. For MetaCyc, PathwayTools [34] was used in combination with PythonCyc to obtain pathway definitions [83] (`src/meta2pwy.py`). Information on Kegg pathways were retrieved directly from the KEGG homepage: reactions (http://rest.kegg.jp/list/reaction), and EC numbers (http://rest.kegg.jp/link/

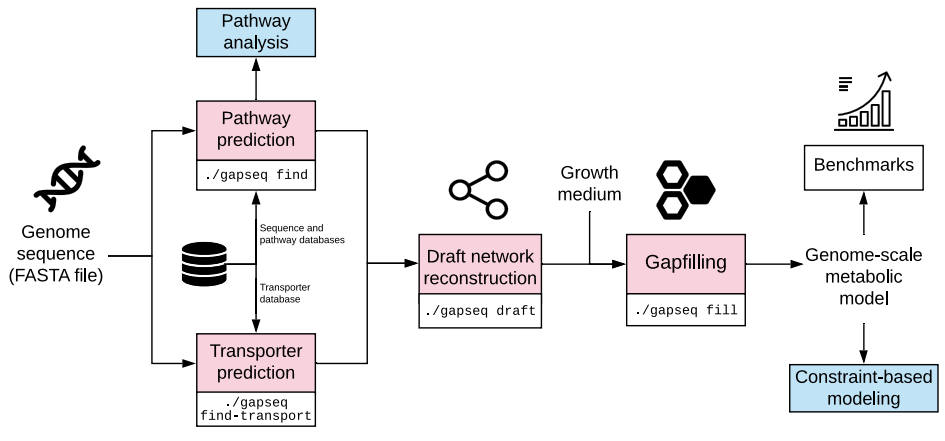

Fig. 7 Chart showing the main components and workflow of `gapseq`. Free icons were used from https://www.flaticon.com (creators: Freepik, Gregor Cresnar, Freepik, Smashicons)

pathway/ec) and further processed (`src/kegg_pwy.R`). In case of ModelSEED, subsystem definition were obtained from the homepage: http://modelseed.org/genomes/Annotations (`src/seed_pwy.R`). In addition, manual defined and revised pathways are stored in the file `dat/custom_pwy.tbl`.

Amino acid sequence (protein) data required for pathway prediction were retrieved from UniProt [84] for each reaction identified by EC number, enzyme name, or cross-references (curated UniProt IDs stored in other databases). Both reviewed and unreviewed sequences are considered and stored as clustered UniPac sequences (`src/uniprot.sh`). To increase the sequence pool for a given reaction, alternative EC numbers from BRENDA [85] and from the Enzyme Nomenclature Committee https://www.qmul.ac.uk/sbcs/iubmb/enzyme/ are integrated (`src/altec.R`, `dat/brenda_ec.csv`). For the download from UniProt, EC numbers and database cross-references are prioritised over enzyme names because the matching is often ambiguous. For a default gapseq run, 95% of the reactions have associated EC numbers and for 75% of the reactions without EC number cross-references to UniProt IDs are available. From the available EC numbers, 66% are specific (i.e. have a full four-level number code) and cross-references to UniProt IDs exist for 86% of the unspecific EC numbers. In 1.8% of cases, multiple EC numbers belong to one reaction. In those cases, each EC number is considered as an individual reaction.

Pathway and subsystem prediction

As a first step, `gapseq` predicts the presence of metabolic pathways based on the organism's genome sequence. For each pathway or subsystem selected from a pathway database (MetaCyc, KEGG, ModelSEED, custom), `gapseq` searches for sequence evidence and a pathway or subsystem is defined as present if enough of its reactions were found to have sequence evidence. In more detail, sequence data (see "Pathway, subsystem, and sequence databases" subsection) is used for homology search by *tblastn* [86] with the protein sequence as query and the genome as database. By default, a bitscore $\geq$ 200 and a coverage of at least 75% is needed for a match. For certain reactions, the user can define additional criteria, for example an identity of $\geq$ 75% (`dat/exception.tbl`). In case of protein complexes with subunits, a more complex procedure is followed ('Protein complex prediction' section). Spontaneous reactions, which do not need an enzyme, were set to be present in any case. In general, a pathway or subsystem is considered to be present if at least 80% of the reactions are found (`completenessCutoffNoHints` threshold). This completeness threshold is lowered for pathways or subsystems in following cases:

1. If the pathway or subsystem contains key reactions, as it is defined for a large number of pathways in MetaCyc, and all key reactions are found, then `completenessCutoff` of the total reactions needed to be found. We used a value of 2/3 for this threshold.
2. In cases in which no sequence data is available for specific reactions, the status of the reactions is set to 'vague' and these reactions do not count as missing if they account for less than `vagueCutoff` of the total reactions of a pathway or subsystem. We used a value of 1/3 for this threshold.

It is important to note that `gapseq` uses MetaCyc's *base pathway* as default pathway structures and we highly recommend to use this default setting for genome-scale

metabolic model reconstructions. This is because MetaCyc provides *base pathway* definitions that follow strict criteria: First, these pathways represent experimentally determined metabolic routes for small molecules (metabolites), which are curated on the basis of scientific literature [87]. Second, the *base pathway* definition includes that these pathways are composed of reactions only, 'where no portion of the pathway is designated as a subpathway' [88]. These criteria results in a larger number of pathways of smaller size compared to KEGG pathways or ModelSEED subsystems. It was previously emphasised that the smaller pathway structures of MetaCyc allow more focused predictions of pathway existence from sequenced genomes [89]. Alternatively, users may choose to use KEGG pathways or ModelSEED subsystems for pathway predictions. This option could be of interest for users, who intend to use `gapseq` solely to investigate enzyme presence and pathway/subsystem coverage from a genome on the basis of reference pathways other than MetaCyc pathways.

The pathway and subsystem prediction algorithm is implemented in the bash shell script `src/gapseq_find.sh`, which uses GNU parallel [90] and fastaindex/fastafetch from exonerate [91].

Protein complex prediction

A problem with automatic sequence download for reaction-associated reference proteins (as FASTA files) comes with protein complexes, for which a single blast hit may be not sufficient to predict enzyme presence. In `gapseq`, the subunit information of protein complex components is extracted from the sequence FASTA headers of the reaction-associated protein sequences obtained as references from UniProt. Search terms are: 'subunit', 'chain', 'polypeptide', 'component', and different numbering systems (roman, arabic, greek) are homogenised. To avoid artefacts in text matching, subunits that occur less than five times in the sequence file are not considered, and in cases in which a subunit occurs almost exclusively ($\geq$ 66%) the other entries are not taken into account. All FASTA entries, which could not be matched by text mining, or which were excluded because of the coverage, are labelled 'undefined subunit' and do not add to the total amount of subunits. For each recognised subunit, a blast search is performed. A protein complex is considered present if more than 50% of the subunits could be found, whereby the presence of 'undefined subunits' tip the balance if exactly 50% of the subunits were found. The text matching with regular expressions is done with R's stringr [92] and Biostrings [93] as defined in `src/complex_detection.R`. The script is called from within the shell script `src/gapseq_find.sh`.

Transporter prediction

Transport reactions govern the exchange of metabolites with the environment and are therefore essential for phenotype predictions. For transporter search, sequence data from the Transporter Classification Database (TCDB) is employed [94]. In addition, manual defined sequences can be defined in `dat/seq/transporter.fasta`. The sequence set is reduced to a subset of transporters that involve metabolites known to be produced or consumed by microorganisms (`dat/sub2pwy.csv`). Subsequently, the genome is queried by the reduced sequences using *tblastn* [86]. For each hit (default cutoffs: bitscore $\geq$ 200 and coverage $\geq$ 75%), the transporter type (1. Channels and pores, 2. Electrochemical potential-driven transporter, 3. Primary active transporters, 4. Group translocators) is

determined using the TC number mentioned in the FASTA header of the source sequence from TCDB. A suitable candidate reaction is searched in the reaction database. If there is a hit for a transporter of a substance but no candidate reaction for the respective transporter type can be found, then other transporter types are considered. The transporter search is done by the shell script `src/transporter.sh` that uses GNU parallel [90] and fastaindex/fastafetch from exonerate [91].

Candidate transporters are selected from the reaction database by transporter type and substance name. This is done by text search and is currently implemented only for the ModelSEED namespace. From the ModelSEED reaction database all reaction with the flag *is_transport* = 1 are taken into account and the transporter type is predicted by keywords: 'channel', 'pore' (1. Channels and pores); 'uniport', 'symport', 'antiport', 'permease', 'gradient' (2. Electrochemical potential-driven transporters); 'ABC', 'ATPase', 'ATP' (3. Primary active transporters); and 'PTS' (4. Group translocators). If no transporter type could be identified by keywords, additional string matching is done for ATPases, proton/sodium antiporter, and PTS by considering the stoichiometry of the involved metabolites. The transported substance is identified as the substance that occurs on both sides of the reaction. In addition, reactions from the reaction database can be linked manually to substances and transporter types (`dat/seed_transporter_custom.tbl`). The text matching with regular expressions is done with stringr [92] (`src/seed_transporter.R`).

Biochemistry database curation and construction of universal metabolic model

For the construction of genome-scale metabolic network models, `gapseq` uses a reaction and metabolite database that is derived from the ModelSEED database [33] as from January 2018. In addition, 67 new reactions and 12 new metabolites were introduced to the `gapseq` biochemistry database (see Additional file 2: Table S1). All reactions and metabolites from the database were included for the construction of a full universal metabolic network model; an approach that is also used in CarveMe [30]. We curated the underlying biochemistry database in order to correct inconsistencies in reaction stoichiometries and reversibilities. Inconsistencies were identified by optimising the universal network model for ATP-production without any nutritional input to the model using flux balance analysis. In case of ATP-production, the flux distributions of such thermodynamically infeasible reaction cycles were investigated by cross-checking the involved reactions with literature information, the BRENDA database for enzymes [85], and the MetaCyc database [26]. Stochiometries and reversibilities of erroneous reactions were corrected accordingly. This curation procedure was repeated until no theromodynamically infeasible and ATP-generating reaction cycles were observed. In total, more than 1500 reactions were subject to corrections in their stoichiometry and/or reversibility.

It needs to be noted, that since we have retrieved the biochemistry database from ModelSEED for the development of `gapseq`, the ModelSEED database has been comprehensively improved and extended [95]. During the development of `gapseq` we have transferred corrections made by the ModelSEED community also to our database. The `gapseq`-developer team will continue this process together with future developments of the ModelSEED database and our software.

Hits from the pathway prediction (Pathway and subsystem prediction) and transporter prediction (Transporter prediction) are mapped to the `gapseq` reaction database using

different common identifiers. A majority of reactions are directly matched via their corresponding Enzyme Commission (EC) system identifier [96] and Transporter Classification (TC) system identifier [94], respectively. For this mapping, also alternative EC numbers for enzymatic reactions as defined in the BRENDA database [85] are considered. Moreover, the databases used for pathway and transporter predictions often provide cross-links to the reaction's KEGG ID, which is also assigned to most reactions in the `gapseq` database and used to match reactions. Additionally, the MNXref database [97] provides cross-links between several biochemistry databases, which `gapseq` also utilises to translate hits from the pathway predictions to model reactions. Finally, a manual translation of enzyme names to model reactions is done for some reactions, which we identified as important reactions but which failed to match between the pathway databases (Pathway and subsystem prediction) and the `gapseq` model reactions using other reaction identifiers (`dat/seed_Enzyme_Name_Reactions_Aliases.tsv`). The overall mapping is done by the function `getDBhit()` as defined in `./src/gapseq_find.sh`.

Model draft generation

A draft genome-scale metabolic model is constructed based on the results from the pathway and transporter predictions (see above). A reaction is added to the draft model if the corresponding enzyme/transporter was directly found (i.e. the blast hit reached the bitscore cutoff value) or if the pathway was predicted to be present (i.e. due to pathway completeness and key enzymes) in which the reaction participates. Additionally, spontaneous reactions as defined in the MetaCyc database as well as transport reaction of compounds, which are know to be able to cross cell membranes by means of diffusion (e.g. H_2), are directly added to every draft model. As part of the draft model construction `gapseq` adds a biomass reaction to the network that aims to describe the composition of molecular constituents that the organism needs to produce in order to form 1 g dry weight (1 gDW) of bacterial biomass. `gapseq` uses the biomass composition definition from the ModelSEED database for Gram-positive (`dat/biomass/seed_biomass.DT_gramPos.tsv`) and Gram-negative bacteria (`dat/biomass/seed_biomass.DT_gramNeg.tsv`). If no Gram-staining property is specified by the user, `gapseq` predicts the Gram-staining-dependent biomass reactions by finding the closest 16S-rRNA-gene neighbour using a `blastn` search against reference 16S-rRNA gene sequences from 4647 bacterial species with known Gram-staining properties that are obtained from the ProTraits database [46]. The model draft generation is done by the R script `src/generate_GSdraft.R`.

Gap-filling algorithm

`gapseq` provides a gap-filling algorithm that adds reactions to the model in order to enable biomass production (i.e. growth) and likely anabolic and catabolic capabilities. The algorithm uses the alignment statistics (i.e. the bitscore) from the pathway- and transporter prediction steps of `gapseq` (see above) to preferentially add reactions to the network, which have the highest genetic evidence. This approach is especially relevant in cases where the sequence similarity to known enzyme-coding reference genes was close to but did not reach the cutoff value b, which is required for a reaction to be included directly into the draft network. In contrast to the gap-filling algorithms described in previous works [98] and [30], which also use genetic evidence-weighted gap-filling, the gap-filling

problem in `gapseq` is not formulated as Mixed Integer Linear Program (MILP) but as Linear Program (LP), and is derived from the parsimonious enzyme usage Flux Balance Analysis (pFBA) algorithm developed by Lewis et al., 2010 [3]. Therefore, the alignment statistics (i.e. bitscore) are translated into weights for the corresponding model reactions and incorporated into the problem formulation:

$$\begin{aligned} \text{max: } & v_j - c \sum_{i \in R_{\text{all}}} w_i |v_i| \ , \\ & w_i = \begin{cases} w_{\min} & b_i \geq u \quad | \quad i \in R_{\text{draft}} \\ (b_i - u)\left(\frac{w_{\min} - w_{\max}}{u - l}\right) + w_{\min} & l \leq b_i < u \\ w_{\max} & b_i < l \end{cases} \\ \text{s.t.} & \\ & \boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \\ & \boldsymbol{lb} \leq \boldsymbol{v} \leq \boldsymbol{ub} \end{aligned} \tag{1}$$

where R_{all} is the set of all reaction in the universal model, R_{draft} are the reactions, which are already part of the draft network before gap-filling, v_j is the flux through the objective reactions (e.g. biomass production), v_i the flux through reaction i, w_i the weight for reaction i, v the flux vector for all reactions, and c a scalar factor that determines the contribution of the absolute reduction of weighted fluxes to the overall FBA solution (default: $c = 0.001$). Moreover, a maximum weight value $w_{\max}$ (default: 100) is assigned if the reaction's highest bitscore is smaller than a threshold l (default: 50). A minimum reaction weight $w_{\min}$ (default: 0.005) is assigned to reactions with a bitscore higher than u (default: 200) or if the reactions are already part of the draft model. S is the stoichiometric matrix and lb and ub the lower and upper flux bound vectors.

Two other LP-based gap-filling algorithms that incorporate reaction evidence scores have been formulated by Dreyfuss et al. (2013) [99] and Medlock et al. (2020) [100], respectively. These approaches require a definition of a minimum flux through the biomass reaction to ensure growth. The pFBA-derived LP formulation of `gapseq` (Eq. 1) includes the flux through the biomass/objective reaction v_j together with the reaction evidence scores in a single objective function.

In `gapseq` and following the solution of the LP (Eq. 1), reactions carrying a flux and which are not part of the draft model are added to the network model. The algorithm is implemented in `src/gapfill4.R`.

Gap-filling of biomass, carbon sources, and fermentation products

Gap-filling of a draft model in `gapseq` requires only for the first step a user-defined growth medium that is ideally known to support growth of the organism of interest in vivo. If no growth medium is specified by the user, a complete medium (dat/media/ALLmed.csv) is chosen by `gapseq` (as done for the large-scale benchmarks of enzyme activity and carbon sources, cf. Validation with enzymatic data (BacDive), Validation with carbon sources data (ProTraits)). A set of common microbial growth media (e.g. LB, TSB, M9) is provided in the gapseq software directory `dat/medium/`. In addition, the user can provide a custom growth medium definition. The above described gap-filling algorithm is used to improve the generated draft model in four steps. Importantly, steps

2-4 only add reactions having sequence support and aim for improve the model quality without reliance on a specific gap-filling medium.

1. **Biomass production**: To ensure that the model is able to produce biomass under the given nutritional input (gap-filling medium) the gap-filling algorithm is applied while the objective is defined as the flux through the biomass reaction. This step will add all missing reactions that are essential for in silico growth and are not part of the model yet.
2. **Individual biomass components**: It is checked whether the model supports the biosynthesis of individual biomass components. Therefore, the model is re-constrained to a M9-like minimal medium with a carbon source for which an exchange reaction is found (default: glucose if available). The objective function is set to the production of one biomass component at a time and the gap-fill algorithm is performed using only reactions with sequence support as source. This gap-filling step is repeated for each biomass component metabolite twice, with and without oxygen to potentially allow aerobic and anaerobic growth for facultative anaerobes.
3. **Alternative energy sources**: `gapseq` attempts to gap-fill likely metabolic pathways, which enable the utilisation of alternative energy sources, which might not be part of the defined growth medium from step (1). To this end, the model is re-constrained to a M9-like minimal medium containing a single carbon source of interest at the time. Next, three temporary reactions were inserted into the model that recycle common reducing equivalent carriers (ESP1: $\text{menaquinol} \rightarrow \text{menaquinone} + 2H^+$; ESP2: $\text{quinol} \rightarrow \text{quinone} + 2H^+$; ESP3: $NADH \rightarrow NAD^+ + H^+$). As objective function, the summed flux of the temporary reactions ESP1, ESP2, and ESP3 is used. Again, the gap-filling of this step only employs those reactions having sequence support. By this, the capacity of a potential carbon source to function as electron donor can be evaluated. This test can be considered as an in silico simulation of the commonly used BIOLOG carbon source utilisation test arrays [47] in which the colometric effect is coupled to a dehydrogenase [101]. This gap-filling step is performed for all metabolites defined in `dat/sub2pwy.csv`.
4. **Metabolic products**: Finally, the same list of compounds (`dat/sub2pwy.csv`), is used to check whether the network can be gap-filled to allow the formation of these metabolites given the original medium. For each compound the gap-filling algorithm is applied with the production of the focal compound as objective function. As for step 2–3, only reactions with sequence evidence are considered for gap-filling.

While step (1) considers all reaction from the universal model as potential candidate reactions for gap-filling, steps (2–4) allow only the addition of candidate reactions to the model with a corresponding bitscore from the pathway prediction (Pathway and subsystem prediction) higher than a threshold value b (default: 50). Thus, these so-termed core reactions represent only reactions, for which `gapseq` has found genomic sequence evidence. Gaps in the specific metabolic functions (individual biomass component formation, alternative energy source utilisation, by-product formation) are only resolved if all required additional reactions are core reactions. The gap-filling steps (2–4) are implemented to reduce the impact of a gap-filling medium on the final metabolic model. Thus,

these steps aim to increase the versatility of `gapseq` reconstructions for downstream metabolic simulations of the models in growth environments that are potentially different to the chemical composition of the actual gap-filling medium. This could be especially of relevance for dynamic community metabolism simulations, where the nutritional environment changes over time due to the release of by-products by certain community members that serve as resources for others. In case `gapseq` users prefer to perform gap-filling solely for biomass production on a defined gap-filling medium, the argument `-q` can be passed on to the gap-filling command `./gapseq fill`. The number of reactions added during gap-filling is given as output during runtime. In addition, detailed information on why a specific reaction was added to the model is provided in the reaction attributes table (*@react_attr*) of the model's S4 R-object of class *sybil::modelorg* [65]. In this table, the column *'gs.origin'* states an integer number between 0 and 10 where 0 indicates that the focal reaction was directly included in the draft model due to predicted sequence homology (see Pathway and subsystem prediction and Transporter prediction); 1–4 correspond to the four gap-filling steps as described above; 6 indicates the biomass reaction; 7 and 8 refer to exchange and diffusion reactions, respectively; 9 refers to reactions that were added due to pathway completion; code 10 indicates reactions that are added after using the optional function `gapseq adapt`). The code 5 is currently not used, but might be assigned in a futur e version of `gapseq`.

Formal and functional model file testing

The validity of genome-scale metabolic model files was checked with MEMOTE (0.10.2) [102]. For all models used in the anaerobic food web (Anaerobic food web of the human gut microbiome), the total MEMOTE score was computed for the respective SBML-Model files. MEMOTE was executed using the parameter `-skip test_find_metabolites_not_produced_with_open_bounds` and `-skip test_find_metabolites_not_consumed_with_open_bounds` since these tests do not contribute to the total MEMOTE score but require long computation time.

Validation with enzymatic data (BacDive)

Enzyme activity tests are commonly performed for characterisation and identification of microbial isolates. In those tests, microbial cell cultures or extracts from the cultures are exposed to the substrate of the focal enzyme and it is measured whether the substrate is transformed. While the experimental culture and test conditions can vary between microorganisms tested and the specific enzymes of interest, the experiments are generally designed to invoke the expression of the specific enzyme, if the organism harbours the respective gene(s). The Bacterial Diversity Metadatabase (BacDive) is a large curated database for, among other data, results from laboratory enzyme activity tests [45]. We used this information to benchmark automated model reconstructions by scrutinising if the model reconstructions possess the enzymatic reactions, whose activities were tested in laboratory experiments. For this purpose, a list of type strain IDs where downloaded using the advanced search within the BacDive website. Subsequently, the IDs were used to retrieve the strains' data stored at BacDive via the R package BacDiveR (version 0.9.1, [103]). If the stored data contained non-zero entries for enzymatic activity and if a genome assembly was available from NCBI RefSeq, the type strain was considered for the validation analysis (Additional file 2: Table S7). The respective genome assemblies were

downloaded with `ncbi-genome-download` (https://github.com/kblin/ncbi-genome-download). If multiple genomes were available for one type strain, 'representative' and 'complete' (NCBI tags) genomes were preferred and, in case there were still multiple candidate genomes available, the most complete genome was selected. Genome completeness was estimated using the software BUSCO (3.0.2) [104]. In total, 3017 type strain genomes were subject for automated model reconstructions using ModelSEED (2.5.1), CarveMe (1.2.2), and `gapseq`. The gap-filling parameters were set to default values for each program, i.e. a complete medium was assumed. A reaction activity was predicted if the corresponding reaction was found to be present in the model. This was done by matching enzymes and reactions via EC numbers. For CarveMe the vmh (https://www.vmh.life) and for ModelSEED and `gapseq` the ModelSEED (http://modelseed.org) reaction database was used to match reactions and EC numbers. Only those EC numbers were considered for testing for which a matching from EC number to reaction IDs exist. For the EC numbers 3.1.3.1, 3.1.3.2, the corresponding reactions were the same, and thus unspecific, so that both EC numbers were from the validation analysis. In general, the enzymatic data in the BacDive database has the entries *enzyme name*, *EC number*, and experimentally measured *enzyme activity* (active: '+'; not active: '-') but some entries were ambiguous (e.g. '+/-'). These ambiguous entries were omitted from the analysis. If an enzyme was measured to be active according to the BacDive database and the corresponding reaction also present in the metabolic network, then the enzymatic test was called a *true positive*. In contrast, if the reaction was not present in the network the test was called *false negative* (vice versa for false positive, true negative). Sampling of enzymatic data was performed in order to preclude a potential bias due to over-representation of certain EC numbers. All EC numbers with at least 100 tests were considered for sampling. For each EC number, 100 tests were randomly chosen. The re-sampling was repeated 500 times to estimate the variation of true positives, true negatives, false positives, and false negatives.

Validation with carbon sources data (ProTraits)

In order to predict accurate growth phenotypes using genome-scale metabolic models, it is crucial that the network reconstructions possess the metabolic capability to utilise the specific carbon sources, which the organism can also use in their natural environment. Here, we used microbial phenotypic trait data for the validation of carbon source utilisation from the 'atlas of prokaryotic traits' database (ProTraits) [46]. In brief, ProTraits is an online resource that provides phenotypic trait data spanning over 3000 microorganisms. The data stored in ProTraits represent phenotypes that are inferred from scientific literature and comparative genomics [46]. Each phenotype in ProTraits is provided with a confidence score between 0 and 100%, whereas 100% denotes the highest confidence for the respective phenotype of a specific organisms. Here, we used only the carbon source utilisation phenotypes with the stringent confidence threshold of $\geq$ 95%. The data was directly downloaded from the ProTraits website (http://protraits.irb.hr/data.html) as a tab-separated table. For organisms which had at least one high-confidence carbon source utilisation trait, the corresponding genome assembly was obtained from NCBI RefSeq [105] if available. In cases where a genome assembly was found, it was applied as input for ModelSEED, CarveMe, and `gapseq` to reconstruct metabolic models. The number of potential carbon sources was reduced to a subset for which a mapping from substance name to ModelSEED and CarveMe model namespace existed (`dat/sub2pwy.csv`).

The tests for D-lyxose were removed because it was listed as all negative in ProTraits and also all compared pipelines predicted no utilisation. The main test whether a carbon source can be used by a model was done in a BIOLOG-like manner as described above (see Gap-filling of biomass, carbon sources, and fermentation products). To this end, temporary reactions to recycle reduced electron carriers as carbon source utilisation indicators were added to the respective model. The objective for optimisation was set to maximise the flux through these recycling reactions. The exchange reactions were limited to a minimal medium with minerals and the focal potential carbon source. This theoretical approach tested, whether the model is able to pass electrons from the potential carbon source to electron carrier metabolites. A carbon source was predicted to be able to serve as energy source if at least one of the recycle reactions carried a positive flux.

Prediction of gene essentiality

Genome-scale metabolic models are commonly used to predict if essentiality of genes for cellular growth. In order to further evaluate our `gapseq` approach, we compared gene essentiality predictions with previously reported growth phenotypes of single gene-knockout experiments. To predict the essentiality of genes, we performed in silico single gene deletion phenotype analysis for the network reconstructions of *Escherichia coli* str. K-12 substr. MG1655 (RefSeq assembly accession: GCF_000005845.2), *Bacillus subtilis* substr. *subtilis* str. 168 (GCF_000789275.1), *Shewanella oneidensis* MR-1 (GCF_000146165.2), *Pseudomonas aeruginosa* PAO1 (GCF_000006765.1), and *Mycoplasma genitalium* G37 (GCF_000027325.1). The analysis was performed on the basis of the models' Gene-Protein-Reaction (GPR) mappings and according to the protocol by Thiele and Palsson, 2010 [22]. To this end, the contingency tables of predicted growth/no-growth phenotypes from the network models and experimentally determined growth phenotypes of gene deletion mutants were constructed. Genes were predicted to be conditionally essential under the given growth environment if the predicted growth rates of the models were below 0.01 h^{-1}. The growth-media compositions for growth predictions were defined as M9 with glucose as carbon- and energy source for *E. coli*, lysogeny broth (LB) for *B. subtilis* and *S. oneidensis*, M9 with succinate as carbon and energy source for *P. aeruginosa*, and a complete medium (all external metabolites available for uptake) for *M. genitalium*. Experimental data for gene essentiality was obtained from [106–110]. In order to compare GPR associations between reconstructions, it was tested if the GPR boolean expressions from two models for the same enzymatic reaction return identical results for all possible combinations of gene presence/knockout of the involved genes.

Fermentation product tests

To evaluate the potential of our approach to predict bacterial metabolism in anaerobic environments, we simulated the anoxic growth of bacterial model reconstructions and compared the predictions with fermentation end products reported in primary literature. The release of by-products from anaerobic metabolism was predicted using Flux Balance Analysis (FBA) coupled with a minimisation of total flux [111] to avoid fluxes that do not contribute to the objective function of the biomass production. In addition, Flux-Variability-Analysis (FVA) [112] was applied to predict the maximum fermentation product release of individual metabolites across all possible FBA solutions.

Metabolites with a positive exchange flux (i.e. outflow) were considered as fermentation products. The analysis was performed for 24 different bacterial organisms, which (1) have a genome assembly available in the RefSeq database [105], (2) are known to grow in anaerobic environments, and (3) for which the fermentation products have been described in the literature based on anaerobic cultivation experiments (Additional file 2: Table S2). The gap-filling of the network models using `gapseq`, CarveMe, and ModelSEED as well as the simulations of anaerobic growth were all performed assuming the same growth medium that comprised several organic compounds (i.e. carbohydrates, polyols, nucleotides, amino acids, organic acids) as potential energy sources and nutrients for growth (see media file `dat/media/FT.csv` at the gapseq github repository).

Since the amount of fermentation product release depends on the organism's growth rate, we normalised the outflow of the individual fermentation products, which has the unit mmol $*$ gDW^{-1} $*$ h^{-1}, by the predicted growth rate of the respective organism which has the unit h^{-1}. Thus, we report the amount of fermentation product production in the quantity of the metabolite that is produced per unit of biomass: mmol $*$ gDW^{-1}.

Pathway prediction of soil and gut microorganisms

The pathway analysis was done by comparing predicted pathways of soil and gut microorganisms. For this means, genomes were downloaded from a resource of reference soil organisms [113] and gut microorganisms [68]. The default parameter of `gapseq` were used for pathway prediction. The principal component analysis was done in R using the factoextra package [114]. For predicted pathways for soil and gut microorganisms, it was checked if samples belong to different distributions using a bootstrap version of the Kolmogorov-Smirnov test [115].

Anaerobic food web of the human gut microbiome

Microbial strains rarely live in isolation but usually coexists with other strains in multispecies communities. In such communities, metabolic processes in one organism may affect the metabolism of other cells and vice versa [116]. It is an ambitious goal in systems biology to apply genome-scale metabolic models in simulations of community metabolism, including metabolite exchanges between cells of different species. Here, we evaluated the potential of automatically reconstructed models to predict metabolic interactions in an anaerobic microbial community. As a test case, representative bacterial organisms known to be relevant in the human intestinal cross-feeding of metabolites were selected based on the proposed food webs by Louiset al. (2014) [50] and Rivera-Chavez et al. (2015) [51]. The genomes of organisms were obtained from NCBI RefSeq [105] and metabolic models reconstructed using `gapseq`, carveme, and modelseed. A medium containing minerals, vitamins, amino acids, fermentation- and metabolic byproducts (namely acetate, formate, lactate, butyrate, propionate, H_2, CH_4, ethanol, H_2S, succinate), and carbohydrates (glucose, fructose, arabinose, ribose, fucose, rhamnose, lactose) was used for gap-filling. Furthermore, a published model of *Methanosarcina barkeri* was added to the community [117] to represent archaea that are also known to be part of anaerobic food webs [118]. All organisms of the modelled community and their respective genome assembly accession numbers are listed in Additional file 2: Table S3. In detail, all metabolic models were simulated as microbial community using the R-package 'BacArena', which allows an individual-based dynamic simulation of metabolic

models that are optimised separately in a shared growth environment [52]. A virtual growth environment ('arena') of the size of 20 × 20 grid cells was defined. For each organism of the microbial community, five random grid cells within the arena were populated with the model of the focal strain to define the initial community composition. The gap-filling medium described above, but without the fermentation and by-products was used to determine the initial arena substance concentrations. In addition, sulfite and 4-aminobenzoate were added in 1 mM each to the growth environment as these metabolites are essential for the growth of the *M. barkeri* model. Subsequently, the community was simulated for seven time steps, which corresponds to seven hours simulation time. The metabolite uptake and production rates were analysed after the third time step for CarveMe and `gapseq` models and after the fifth time step for ModelSEED models, in which all organisms were growing exponentially and reached similar total population density.

Model reconstructions from metagenomic assemblies

Genomes assembled from metagenomic data via 'binning' are often fragmented and incomplete. For the reconstruction of metabolic models from such genomes, it is important to estimate how missing genetic fragments may affect final model quality. 4930 species-level genome bins (SGBs) assembled from shotgun metagenome sequencing reads were obtained from the study of Pasolli et al. (2019) [61]. Only those SGBs were considered for further analysis, which were already classified as bacteria on a species-level in the original publication. For each SGB, closely related reference assemblies from the RefSeq database [105] were identified by constructing a multi-locus phylogenetic tree using autoMLST (version as of April 7, 2020, [119]). RefSeq assemblies were considered as genomes from the same species-level taxonomic group as the focal SGB if their predicted MASH distance (D) [120] were below or equal to 0.05. This threshold was shown before to cluster bacterial genomes at the taxonomic level of species [120]. Only SGBs with 10 or more assigned reference assemblies were considered for further analysis, which yielded in total 127 SGBs. Metabolic models were reconstructed using `gapseq` for each SGB and their 10 closest reference assemblies (Additional file 2: Table S5). Next, similarity of SGB models with their respective reference models was calculated using the following metabolic network similarity score T_{SGB}:

$$T_{\mathrm{SGB}} = \frac{\sum_i a_i b_i}{\sum_i b_i} \quad , \quad i \in R_{\mathrm{SGB\_Ref}} \ , \ 0 \leq b_i \leq 1$$

with

$$a_i = \begin{cases} 0 & \text{if } i \notin R_{\mathrm{SGB}} \\ 1 & \text{if } i \in R_{\mathrm{SGB}} \end{cases} \tag{2}$$

$R_{\mathrm{SGB\_Ref}}$ is the union set of reactions with associated genes that are part of the network models reconstructed for the ten reference genome assemblies of the focal SGB. R_{SGB} is the set of reactions part of the SGB's model reconstruction. b_i is the frequency of reaction i among the ten SGB's reference models. Completion of the genome sequence of SGBs was estimated by using BUSCO (version 4.0.6, [104]) using the lineage-specific completeness score.

Technical details

The pathway prediction part of `gapseq` is implemented as Bash shell script and the metabolic model generation part is written in R. Linear optimisation can be performed with a different solvers (GLPK or CPLEX). Other requirements are exonerate, bedtools, and barrnap. In addition, the following R packages are needed: data.table [121], stringr [122], sybil [65], getopt [123], reshape2 [124], doParallel [125], foreach [126], R.utils [127], stringi [128], glpkAPI [129], and BioStrings [130]. Models can be exported as SBML [131] file using sybilSBML [65] or R data format (RDS) for further analysis in R, for example with sybil [65] or BacArena [52].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02295-1.

Additional file 1: Supplementary figures S1-S5.
Additional file 2: Supplementary tables S1-S7. OpenDocument spreadsheet (ODS) file
Additional file 3: Review history.

Acknowledgements
We thank Martin Sperfeld for fruitful comments and discussions during the developmental phase. The software was thankfully tested by Georgios Marinos, Shan Zhang, and Lena Best.

Review history
The review history is available as Additional file 3.

Peer review information
Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
All authors conceptualised `gapseq` and wrote the manuscript. JZ and SW developed the software and did the analysis. The authors read and approved the final manuscript.

Authors' information
Twitter handles: @_jozimmermann (Johannes Zimmermann); @KaletaLab (Christoph Kaleta); @SWaschina (Silvio Waschina).

Funding
CK and SW acknowledges support by the Collaborative Research Centre 1182 - 'Origin and Function of Metaorganisms' - Deutsche Forschungsgemeinschaft and by the Cluster of Excellence 2167 - 'Precision medicine in chronic inflammation' - Deutsche Forschungsgemeinschaft. In addition, CK acknowledges support by the German Ministry for Education and Research within the context of iTREAT (BMBF support code 01ZX1902A). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
`gapseq` is implemented as bash shell script and in R and is freely available under the GNU General Public License (v3.0) on GitHub (https://github.com/jotech/gapseq/). Documentation and tutorials for gapseq can be found at https://gapseq.readthedocs.io. All results presented in this manuscript were produced using the specific `gapseq` version 1.1 as archived on GitHub [63] and is available from Zenodo [132]. The datasets used for model construction and validation purposes were obtained from publicly available databases and publications as cited at the respective parts of the manuscript. Scripts and data used for the benchmarking tests in this study are available from the GitHub repository, https://github.com/Waschina/gapseqEval.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 May 2020 Accepted: 10 February 2021
Published online: 10 March 2021

References

1. Fell DA. Systems properties of metabolic networks. In: Bar-Yam Y, editor. Unifying Themes In Complex Systems. Boca Raton, Florida: CRC Press; 2003. p. 163–78.
2. Steuer R. Computational approaches to the topology, stability and dynamics of metabolic networks. Phytochemistry. 2007;68(16):2139–51. https://doi.org/10.1016/j.phytochem.2007.04.041.
3. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, Weitz KK, Eils R, König R, Smith RD, Palsson BO. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6(1):390. https://doi.org/10.1038/msb.2010.47.
4. de Jong H, Casagranda S, Giordano N, Cinquemani E, Ropers D, Geiselmann J, Gouzé J-L. Mathematical modeling of microbes: Metabolism, gene expression, and growth. J R Soc Interface. 2017;14:20170502. https://doi.org/10.1098/rsif.2017.0502.
5. Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. PLoS Comput Biol. 2013;9(6):1003091. https://doi.org/10.1371/journal.pcbi.1003091.
6. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 2007;3(1):119. https://doi.org/10.1038/msb4100162.
7. Lularevic M, Racher AJ, Jaques C, Kiparissides A. Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. Biotechnol Bioeng. 2019;116(9):2339–52. https://doi.org/10.1002/bit.27025.
8. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA. Metabolic modeling of a mutualistic microbial community. Mol Syst Biol. 2007;3(1):92. https://doi.org/10.1038/msb4100131.
9. Zomorrodi AR, Islam MM, Maranas CD. d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. ACS Synth Biol. 2014;3(4):247–57. https://doi.org/10.1021/sb4001307.
10. Harcombe W, Riehl W, Dukovski I, Granger B, Betts A, Lang A, Bonilla G, Kar A, Leiby N, Mehta P, Marx C, Segrè D. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. Cell Rep. 2014;7(4):1104–15. https://doi.org/10.1016/j.celrep.2014.03.070.
11. Aden K, Rehman A, Waschina S, Pan W-H, Walker A, Lucio M, Nunez AM, Bharti R, Zimmerman J, Bethge J, Schulte B, Schulte D, Franke A, Nikolaus S, Schroeder JO, Vandeputte D, Raes J, Szymczak S, Waetzig GH, Zeuner R, Schmitt-Kopplin P, Kaleta C, Schreiber S, Rosenstiel P. Metabolic functions of gut microbes associate with efficacy of tumor necrosis factor antagonists in patients with inflammatory bowel diseases. Gastroenterology. 2019;157(5):1279–92. https://doi.org/10.1053/j.gastro.2019.07.025.
12. Koch S, Kohrs F, Lahmann P, Bissinger T, Wendschuh S, Benndorf D, Reichl U, Klamt S. RedCom: A strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing experimental datasets from anaerobic digestion. PLoS Comput Biol. 2019;15(2):1–32. https://doi.org/10.1371/journal.pcbi.1006759.
13. 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. Metab Eng. 2020;62:138–49. https://doi.org/10.1016/j.ymben.2020.08.013.
14. Heinken A, Thiele I. Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. Gut Microbes. 2015;6(2):120–30. https://doi.org/10.1080/19490976.2015.1023494.
15. Pryor R, Norvaisas P, Marinos G, Best L, Thingholm LB, Quintaneiro LM, Haes WD, Esser D, Waschina S, Lujan C, Smith RL, Scott TA, Martinez-Martinez D, Woodward O, Bryson K, Laudes M, Lieb W, Houtkooper RH, Franke A, Temmerman L, Bjedov I, Cochemé HM, Kaleta C, Cabreiro F. Host-microbe-drug-nutrient screen identifies bacterial effectors of metformin therapy. Cell. 2019;178(6):1299–312. https://doi.org/10.1016/j.cell.2019.08.003.
16. Zimmermann J, Obeng N, Yang W, Pees B, Petersen C, Waschina S, Kissoyan KA, Aidley J, Hoeppner MP, Bunk B, Spröer C, Leippe M, Dierking K, Kaleta C, Schulenburg H. The functional repertoire contained within the native microbiota of the model nematode Caenorhabditis elegans. ISME J. 2019;14(1):26–38. https://doi.org/10.1038/s41396-019-0504-y.
17. Oberhardt MA, Yizhak K, Ruppin E. Metabolically re-modeling the drug pipeline. Curr Opin Pharmacol. 2013;13(5):778–85. https://doi.org/10.1016/j.coph.2013.05.006.
18. Trawick JD, Schilling CH. Use of constraint-based modeling for the prediction and validation of antimicrobial targets. Biochem Pharmacol. 2006;71(7):1026–35. https://doi.org/10.1016/j.bcp.2005.10.049.
19. Rau MH, Zeidan AA. Constraint-based modeling in microbial food biotechnology. Biochem Soc Trans. 2018;46:249–60. https://doi.org/10.1042/BST20170268.
20. Park JH, Lee SY. Towards systems metabolic engineering of microorganisms for amino acid production. Curr Opin Biotechnol. 2008;19(5):454–60. https://doi.org/10.1016/j.copbio.2008.08.007.
21. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015;13(12):787–94. https://doi.org/10.1038/nrmicro3565.
22. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5(1):93–121. https://doi.org/10.1038/nprot.2009.203.
23. Wittig U, De Beuckelaer A. Analysis and comparison of metabolic pathway databases. Brief Bioinform. 2001;2(2):126–42.
24. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2018;47(D1):590–5. https://doi.org/10.1093/nar/gky962.
25. Alcántara R, Axelsen KB, Morgat A, Belda E, Coudert E, Bridge A, Cao H, de Matos P, Ennis M, Turner S, Owen G, Bougueleret L, Xenarios I, Steinbeck C. Rhea–a manually curated resource of biochemical reactions. Nucleic Acids Res. 2011;40(D1):754–60. https://doi.org/10.1093/nar/gkr1126.
26. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Midford PE, Ong Q, Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. 2018;46(D1):633–9. https://doi.org/10.1093/nar/gkx935.

27. Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. Biochem Soc Trans. 2018;46(4):931–6. https://doi.org/10.1042/bst20170246.
28. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biol. 2019;20:158. https://doi.org/10.1186/s13059-019-1769-1.
29. 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 Comput Biol. 2018;14(5):1006146. https://doi.org/10.1371/journal.pcbi.1006146.
30. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018;46(15):7542–53. https://doi.org/10.1093/nar/gky537.
31. Dias O, Rocha M, Ferreira EC, Rocha I. Reconstructing high-quality large-scale metabolic models with merlin. In: Methods in Molecular Biology. New York: Springer; 2017. p. 1–36. https://doi.org/10.1007/978-1-4939-7528-0_1.
32. Hanemaaijer M, Olivier BG, Röling WFM, Bruggeman FJ, Teusink B. Model-based quantification of metabolic interactions from dynamic microbial-community data. PLoS ONE. 2017;12(3):0173183. https://doi.org/10.1371/journal.pone.0173183.
33. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010;28(9):977–82. https://doi.org/10.1038/nbt.1672.
34. Karp PD, Latendresse M, Paley SM, Ong MKQ, Billington R, Kothari A, Weaver D, Lee T, Subhraveti P, Spaulding A, Fulcher C, Keseler IM, Caspi R. Pathway tools version 19.0: Integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 2015;17(5):877–890. https://doi.org/10.1093/bib/bbv079, https://doi.org/10.1093/bib/bbv079.
35. Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, Kerkhoven EJ. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on streptomyces coelicolor. PLoS Comput Biol. 2018;14(10):1006541. https://doi.org/10.1371/journal.pcbi.1006541.
36. Varma A, Palsson BO. Metabolic flux balancing: Basic concepts, scientific and practical use. Bio/Technology. 1994;12(10):994–8. https://doi.org/10.1038/nbt1094-994.
37. Bauer E, Thiele I. From metagenomic data to personalized in silico microbiotas: predicting dietary supplements for Crohn's disease. NPJ Syst Biol Appl. 2018;4:27. https://doi.org/10.1038/s41540-018-0063-2.
38. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol. 2019;20:121. https://doi.org/10.1186/s13059-019-1730-3.
39. Blaby-Haas CE, de Crécy-Lagard V. Mining high-throughput experimental data to link gene and function. Trends Biotechnol. 2011;29(4):174–82. https://doi.org/10.1016/j.tibtech.2011.01.001.
40. Thiele I, Vlassis N, Fleming RMT. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014;30(17):2529–31. https://doi.org/10.1093/bioinformatics/btu321.
41. Prigent S, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, Gutknecht F, Got J, Eveillard D, Bourdon J, Plewniak F, Tonon T, Siegel A. Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. PLoS Comput Biol. 2017;13(1):1005276. https://doi.org/10.1371/journal.pcbi.1005276.
42. Karlsen E, Schulz C, Almaas E. Automated generation of genome-scale metabolic draft reconstructions based on KEGG. BMC Bioinformatics. 2018;19:467. https://doi.org/10.1186/s12859-018-2472-z.
43. Kumar M, Ji B, Zengler K, Nielsen J. Modelling approaches for studying the microbiome. Nat Microbiol. 2019;4(8):1253–67. https://doi.org/10.1038/s41564-019-0491-9.
44. Phelan VV, Liu W-T, Pogliano K, Dorrestein PC. Microbial metabolic exchange–the chemotype-to-phenotype link. Nat Chem Biol. 2012;8:26–35. https://doi.org/10.1038/nchembio.739.
45. Reimer LC, Vetcininova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, Overmann J. BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. Nucleic Acids Res. 2018;47(D1):631–6. https://doi.org/10.1093/nar/gky879.
46. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 2016;44(21):10074–10090. https://doi.org/10.1093/nar/gkw964.
47. Smalla K, Wachtendorf U, Heuer H, Liu W-T, Forney L. Analysis of BIOLOG GN substrate utilization patterns by microbial communities. Appl Environ Microbiol. 1998;64(4):1220–5. https://doi.org/http://arxiv.org/abs/https://aem.asm.org/content/64/4/1220.full.pdf.
48. Cook GM, Greening C, Hards K, Berney M. Chapter one - energetics of pathogenic bacteria and opportunities for drug development. In: Poole RK, editor. Advances in Bacterial Pathogen Biology. Cambridge, Massachusetts: Academic Press; 2014. p. 1–62. https://doi.org/10.1016/bs.ampbs.2014.08.001. http://www.sciencedirect.com/science/article/pii/S0065291114000022.
49. Goldberg I, Rock J, Ben-Bassat A, Mateles R. Bacterial yields on methanol, methylamine, formaldehyde, and formate. Biotechnol Bioeng. 1976;18(12):1657–68.
50. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol. 2014;12:661–72. https://doi.org/10.1038/nrmicro3344.
51. Rivera-Chávez F, Bäumler AJ. The pyromaniac inside you: Salmonella metabolism in the host gut. Ann Rev Microbiol. 2015;69(1):31–48. https://doi.org/10.1146/annurev-micro-091014-104108.
52. Bauer E, Zimmermann J, Baldini F, Thiele I, Kaleta C. BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. PLoS Comput Biol. 2017;13(5):1–22. https://doi.org/10.1371/journal.pcbi.1005544.
53. Oliphant K, Allen-Vercoe E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. Microbiome. 2019;7:91. https://doi.org/10.1186/s40168-019-0704-8.
54. Ríos-Covián D, Ruas-Madiedo P, Margolles A, Gueimonde M, de Los Reyes-Gavilán CG, Salazar N. Intestinal short chain fatty acids and their link with diet and human health. Front Microbiol. 2016;7:185.
55. Ziels RM, Nobu MK, Sousa DZ. Elucidating syntrophic butyrate-degrading populations in anaerobic digesters using stable-isotope-informed genome-resolved metagenomics. mSystems. 2019;4(4):e00159-19. https://doi.org/10.1128/mSystems.00159-19.

56. Rivière A, Gagnon M, Weckx S, Roy D, Vuyst LD. Mutual cross-feeding interactions between Bifidobacterium longum subsp. longum NCC2705 and Eubacterium rectale ATCC 33656 explain the bifidogenic and butyrogenic effects of arabinoxylan oligosaccharides. Appl Environ Microbiol. 2015;81(22):7767–81. https://doi.org/10.1128/aem.02089-15.
57. Bunesova V, Lacroix C, Schwab C. Mucin cross-feeding of infant bifidobacteria and Eubacterium hallii. Microb Ecol. 2017;75(1):228–38. https://doi.org/10.1007/s00248-017-1037-4.
58. Fernández-Veledo S, Vendrell J. Gut microbiota-derived succinate: Friend or foe in human metabolic diseases? Rev Endocr Metab Disord. 2019;20(4):439–47. https://doi.org/10.1007/s11154-019-09513-z.
59. Stams AJM, Hansen TA. Oxygen-labile l(+) lactate dehydrogenase activity in Desulfovibrio desulfuricans. FEMS Microbiol Lett. 1982;13(4):389–94. https://doi.org/10.1111/j.1574-6968.1982.tb08293.x.
60. Rey FE, Faith JJ, Bain J, Muehlbauer MJ, Stevens RD, Newgard CB, Gordon JI. Dissecting the in vivo metabolic potential of two human gut acetogens. J Biol Chem. 2010;285(29):22082–90. https://doi.org/10.1074/jbc.m110.117713.
61. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. Extensive unexplored human microbiome diversity revealed by over 150, 000 genomes from metagenomes spanning age, geography, and lifestyle. Cell. 2019;176(3):649–62. https://doi.org/10.1016/j.cell.2019.01.001.
62. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824–34. https://doi.org/10.1101/gr.213959.116.
63. Zimmermann J, Kaleta C, Waschina S. Informed prediction and analysis of bacterial metabolic pathways and genome-scale networks. 2020. github repository. https://github.com/jotech/gapseq.
64. Keating S, Waltemath D, König M, Zhang F, Dräger A, Chaouiya C, Bergmann F, Finney A, Gillespie C, Helikar T, Hoops S, Malik-Sheriff R, Moodie S, Moraru I, Myers C, Naldi A, Olivier B, Sahle S, Schaff J, Smith L, Swat M, Thieffry D, Watanabe L, Wilkinson D, Blinov M, Begley K, Faeder J, Gómez H, Hamm T, Inagaki Y, Liebermeister W, Lister A, Lucio D, Mjolsness E, Proctor C, Raman K, Rodriguez N, Shaffer C, Shapiro B, Stelling J, Swainston N, Tanimura N, Wagner J, Meier-Schellersheim M, Sauro H, Palsson B, Bolouri H, Kitano H, Funahashi A, Hermjakob H, Doyle J, Hucka M. SBML Level 3 Community members. SBML Level 3: an extensible format for the exchange and reuse of biological models. Mol Syst Biol. 2020;16(8):e9110. https://doi.org/10.15252/msb.20199110.
65. Gelius-Dietrich G, Desouki AA, Fritzemeier CJ, Lercher MJ. Sybil–efficient constraint-based modelling in R. BMC Syst Biol. 2013;7:125. https://doi.org/10.1186/1752-0509-7-125.
66. 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, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia J-M, Chia J-M, Colasanti R, Conrad N, Davis JJ, Davison BH, DeJongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S, Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B, Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC, Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang D, Xia F, Yoo H, Yoo S, Yu D. KBase: the United States department of energy systems biology knowledgebase. Nat Biotechnol. 2018;36(7):566–9. https://doi.org/10.1038/nbt.4163.
67. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44. https://doi.org/10.1038/nbt.3935.
68. Magnusdottir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jäger C, Baginska J, Wilmes P, Fleming RMT, Thiele I. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol. 2016;35(1):81–89. https://doi.org/10.1038/nbt.3703.
69. Kim WJ, Kim HU, Lee SY. Current state and applications of microbial genome-scale metabolic models. Curr Opin Syst Biol. 2017;2:10–8. https://doi.org/10.1016/j.coisb.2017.03.001.
70. Graspeuntner S, Waschina S, Künzel S, Twisselmann N, Rausch TK, Cloppenborg-Schmidt K, Zimmermann J, Viemann D, Herting E, Göpel W, Baines JF, Kaleta C, Rupp J, Härtel C, Pagel J. Gut dysbiosis with bacilli dominance and accumulation of fermentation products precedes late-onset sepsis in preterm infants. Clin Infect Dis. 2018;69(2):268–77. https://doi.org/10.1093/cid/ciy882.
71. Byndloss MX, Olsan EE, Rivera-Chávez F, Tiffany CR, Cevallos SA, Lokken KL, Torres TP, Byndloss AJ, Faber F, Gao Y, et al. Microbiota-activated PPAR-γ signaling inhibits dysbiotic Enterobacteriaceae expansion. Science. 2017;357(6351):570–5. https://doi.org/10.1126/science.aam9949.
72. Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly-y M, Glickman JN, Garrett WS. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. Science. 2013;341(6145):569–73. https://doi.org/10.1126/science.1241165.
73. Pham VT, Lacroix C, Braegger CP, Chassard C. Early colonization of functional groups of microbes in the infant gut. Environ Microbiol. 2016;18(7):2246–58. https://doi.org/10.1111/1462-2920.13316.
74. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. Front Physiol. 2015;6:383. https://doi.org/10.3389/fphys.2015.00383.
75. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. The evolution of the host microbiome as an ecosystem on a leash. Nature. 2017;548(7665):43–51. https://doi.org/10.1038/nature23292.
76. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL. The Pfam protein families database. Nucleic Acids Res. 2000;28(1):263–6. http://dx.doi.org/10.1093/nar/28.1.26310.1093/nar/28.1.263.
77. Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol. 1998;1(1):55–67.
78. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer EL, Hirsh L, Paladin L, Piovesan D, Tosatto SC, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2018;47(D1):427–32. https://doi.org/10.1093/nar/gky995.

79. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, 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 Res. 2019;47:309–14. https://doi.org/10.1093/nar/gky1085.
80. Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2: An improved and extensible approach for metagenome inference. bioRxiv. 2020. https://doi.org/10.1101/672295, https://www.biorxiv.org/content/early/2020/03/20/672295.
81. Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Tehrani HS, Mirzaie M, Fakheri BA, Mohammad-Noori M. Protein complex prediction: A survey. Genomics. 2020;112(1):174–83. https://doi.org/10.1016/j.ygeno.2019.01.011.
82. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–702. https://doi.org/10.1093/nar/gki866.
83. Mario Latendresse and Peter Midford. PythonCyc. 2020. github release 1.1. https://github.com/ecocyc/PythonCyc.
84. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2016;45(D1):158–69. https://doi.org/10.1093/nar/gkw1099.
85. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 2018;47(D1):542–9. https://doi.org/10.1093/nar/gky1048.
86. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. Blast+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
87. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2011;40(D1):742–53. https://doi.org/10.1093/nar/gkr1014.
88. 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 Res. 2019;48(D1):445–53. https://doi.org/10.1093/nar/gkz862.
89. Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013;14(1):112. https://doi.org/10.1186/1471-2105-14-112.
90. Tange O. Gnu parallel - the command-line power tool. login: The USENIX Magazine. 2018. https://doi.org/10.5281/zenodo.1146014, https://doi.org/10.5281/zenodo.1146014.
91. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31. https://doi.org/10.1186/1471-2105-6-31.
92. Wickham H. Stringr: Simple, consistent wrappers for common string operations. R package version 1.4.0. 2019. https://CRAN.R-project.org/package=stringr.
93. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.58.0. 2020. https://bioconductor.org/packages/Biostrings.
94. Saier MH, Reddy VS, Tamang DG, Vastermark A. The transporter classification database. Nucleic Acids Res. 2013;42(D1):251–8. https://doi.org/10.1093/nar/gkt1097.
95. Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, Mundy M, Chia N, Noor E, Beber ME, Best AA, DeJongh M, Kimbrel JA, D'haeseleer P, McCorkle SR, Bolton JR, Pearson E, Canon S, Wood-Charlson EM, Cottingham RW, Arkin AP, Henry CS. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res. 2021;49(D1):575–88. https://doi.org/10.1093/nar/gkaa746.
96. Webb EC, et al, Vol. Ed. 6. Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. Cambridge, Massachusetts: Academic Press; 1992.
97. Bernard T, Bridge A, Morgat A, Moretti S, Xenarios I, Pagni M. Reconciliation of metabolites and biochemical reactions for metabolic networks. Brief Bioinform. 2014;15(1):123–35. https://doi.org/10.1093/bib/bbs058.
98. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. PLoS Comput Biol. 2014;10(10):1–14. https://doi.org/10.1371/journal.pcbi.1003882.
99. Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE. Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM. PLoS Comput Biol. 2013;9(7):1003126. https://doi.org/10.1371/journal.pcbi.1003126.
100. Medlock GL, Papin JA. Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. Cell Syst. 2020;10(1):109–19. https://doi.org/10.1016/j.cels.2019.11.006.
101. Bochner BR. Global phenotypic characterization of bacteria. FEMS Microbiol Rev. 2009;33(1):191–205. https://doi.org/10.1111/j.1574-6976.2008.00149.x.
102. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, Diener C, Dräger A, Ebert BE, Edirisinghe JN, Faria JP, Feist AM, Fengos G, Fleming RMT, García-Jiménez B, Hatzimanikatis V, van Helvoirt W, Henry CS, Hermjakob H, Herrgård MJ, Kaafarani A, Kim HU, King Z, Klamt S, Klipp E, Koehorst JJ, König M, Lakshmanan M, Lee D-Y, Lee SY, Lee S, Lewis NE, Liu F, Ma H, Machado D, Mahadevan R, Maia P, Mardinoglu A, Medlock GL, Monk JM, Nielsen J, Nielsen LK, Nogales J, Nookaew I, Palsson BO, Papin JA, Patil KR, Poolman M, Price ND, Resendis-Antonio O, Richelle A, Rocha I, Sánchez BJ, Schaap PJ, Sheriff RSM, Shoaie S, Sonnenschein N, Teusink B, Vilaça P, Vik JO, Wodke JAH, Xavier JC, Yuan Q, Zakhartsev M, Zhang C. Memote for standardized genome-scale metabolic model testing. Nat Biotechnol. 2020;38:272–6. https://doi.org/10.1038/s41587-020-0446-y.

103. Leinweber K. TIBHannover/BacDiveR: Maintenance release (Version 0.9.1). Zenodo. 2019. http://doi.org/10.5281/zenodo.3362500.
104. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351.
105. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2019;48(D1):9–16. https://doi.org/10.1093/nar/gkz899.
106. Zhu B, Stülke J. SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis. Nucleic Acids Res. 2017;46(D1):743–8. https://doi.org/10.1093/nar/gkx908.
107. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, et al. i ML1515, a knowledgebase that computes Escherichia coli traits. Nat Biotechnol. 2017;35(10):904–8.
108. Turner KH, Wessel AK, Palmer GC, Murray JL, Whiteley M. Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum. Proc Natl Acad Sci. 2015;112(13):4110–5.
109. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 2018;557(7706):503.
110. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, Smith HO, Venter JC. Essential genes of a minimal bacterium. Proc Natl Acad Sci. 2006;103(2):425–30.
111. Holzhütter H-G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. Eur J Biochem. 2004;271(14):2905–22.
112. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003;5(4):264–76. https://doi.org/10.1016/j.ymben.2003.09.002.
113. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, Flater J, Tiedje JM, Hofmockel KS, Gelder B, Howe A. Strategies to improve reference databases for soil microbiomes. ISME J. 2016;11(4):829–34. https://doi.org/10.1038/ismej.2016.168.
114. Kassambara A, Mundt F. Factoextra: extract and visualize the results of multivariate data analyses. R package version 1.0.6. 2019. https://CRAN.R-project.org/package=factoextra.
115. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. J Stat Softw. 2011;42(7):1–52.
116. D'Souza G, Shitut S, Preussger D, Yousif G, Waschina S, Kost C. Ecology and evolution of metabolic cross-feeding interactions in bacteria. Nat Prod Rep. 2018;35(5):455–88. https://doi.org/10.1039/c8np00009c.
117. Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. Mol Syst Biol. 2006;2:2006–4. https://doi.org/10.1038/msb4100046.
118. Sieber JR, McInerney MJ, Gunsalus RP. Genomic insights into syntrophy: The paradigm for anaerobic metabolic cooperation. Ann Rev Microbiol. 2012;66(1):429–52. https://doi.org/10.1146/annurev-micro-090110-102844.
119. Alanjary M, Steinke K, Ziemert N. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. Nucleic Acids Res. 2019;47(W1):276–82. https://doi.org/10.1093/nar/gkz282.
120. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132. https://doi.org/10.1186/s13059-016-0997-x.
121. Dowle M, Srinivasan A. Data.table: Extension of 'data.frame'. 2021. R package version 1.14.0. https://CRAN.R-project.org/package=data.table.
122. Wickham H. Stringr: Simple, consistent wrappers for common string operations. R package version 1.4.0. 2019. https://CRAN.R-project.org/package=stringr.
123. Davis TL, Day A. Getopt: C-Like 'getopt' behavior. R package version 1.20.3. 2019. https://CRAN.R-project.org/package=getopt.
124. Wickham H. Reshaping data with the reshape package. J Stat Softw. 2007;21(12):1–20.
125. Corporation M, Weston S. doParallel: Foreach parallel adaptor for the 'parallel' package. R package version 1.0.16. 2020. https://CRAN.R-project.org/package=doParallel.
126. Microsoft, Weston S. Foreach: Provides Foreach Looping Construct. R package version 1.5.1. 2019. https://CRAN.R-project.org/package=foreach.
127. Bengtsson H. R.utils: Various Programming Utilities. R package version 2.10.1. 2019. https://CRAN.R-project.org/package=R.utils.
128. Gagolewski M. R Package Stringi: Character String Processing Facilities. 2020. http://www.gagolewski.com/software/stringi/.
129. Gelius-Dietrich G. glpkAPI: R Interface to C API of GLPK. R package version 1.3.2. 2020. https://CRAN.R-project.org/package=glpkAPI.
130. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient Manipulation of Biological Strings. R package version 2.54.0. 2019.
131. Bornstein BJ, Keating SM, Jouraku A, Hucka M. LibSBML: an API library for SBML. Bioinformatics. 2008;24(6):880–1. https://doi.org/10.1093/bioinformatics/btn051.
132. Zimmermann J, Kaleta C, Waschina S. Gapseq Source Code. 2020. Source code of the version of gapseq used in the computations of the manuscript. https://doi.org/10.5281/zenodo.4199599.

Publisher's Note