

 Open access • Posted Content • DOI:10.1101/2020.03.20.000737

gapseq: Informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models — [Source link](#)

Johannes Zimmermann, Christoph Kaleta, Silvio Waschina

Institutions: University of Kiel

Published on: 23 Mar 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Metabolic network, MetaCyc, Systems biology, Flux (metabolism) and UniProt

Related papers:

- [A protocol for generating a high-quality genome-scale metabolic reconstruction.](#)
- [Ontology-based metabolic pathway prediction using saccharomyces cerevisiae data from genbank, ecocyc and kegg](#)
- [A Distance-Based Framework for the Characterization of Metabolic Heterogeneity in Large Sets of Genome-Scale Metabolic Models](#)
- [Genome-Scale In Silico Models](#)
- [Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/gapseq-informed-prediction-of-bacterial-metabolic-pathways-4b868rrbm4>

SOFTWARE

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

Johannes Zimmermann¹, Christoph Kaleta¹ and Silvio Waschina^{1,2*}

*Correspondence:

s.waschina@nutrinf.uni-kiel.de

²Christian-Albrechts-University
Kiel, Institute of Human Nutrition
and Food Science,
Nutriinformatics,
Heinrich-Hecht-Platz 10, 24118
Kiel, Germany

Full list of author information is
available at the end of the article

Abstract

Microbial metabolic processes greatly impact ecosystem functioning and the physiology of multi-cellular host organisms. The inference of metabolic capabilities and phenotypes from genome sequences with the help of reference biomolecular knowledge stored in online databases remains a major challenge in systems biology. Here, we present gapseq: a novel tool for automated pathway prediction and metabolic network reconstruction from microbial genome sequences. gapseq combines databases of reference protein sequences (UniProt, TCDB), in tandem with pathway and reaction databases (MetaCyc, KEGG, ModelSEED). This enables the prediction of an organism's metabolic capabilities from sequence homology and pathway topology criteria. By incorporating a novel LP-based gap-filling algorithm, gapseq facilitates the construction of genome-scale metabolic models that are suitable for metabolic phenotype predictions by using constraint-based flux analysis. We validated gapseq by comparing predictions to experimental data for more than 3,000 bacterial organisms comprising 14,895 phenotypic traits that include enzyme activity, energy sources, fermentation products, and gene essentiality. This large-scale phenotypic trait prediction test showed, that gapseq yields an overall accuracy of 81% and thereby outperforms other commonly used reconstruction tools. Furthermore, we illustrate the application of gapseq-reconstructed models to simulate biochemical interactions between microorganisms in multi-species communities. Altogether, gapseq is a new method that improves the predictive potential of automated metabolic network reconstructions and further increases their applicability in biotechnological, ecological, and medical research. gapseq is available at <https://github.com/jotech/gapseq>.

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

1 Background

Anything you have to do repeatedly may be ripe for automation.

— Doug McIlroy

- 4 Metabolism is central for organismal life. It provides metabolites and energy for
- 5 all cellular processes. A majority of metabolic reactions are catalysed by enzymes,
- 6 which are encoded in the genome of the respective organism. Those catalysed reac-
- 7 tions form a complex metabolic network of numerous biochemical transformations,
- 8 which the organism is presumably able to perform [1].
- 9 In systems biology, the reconstruction of metabolic networks plays an essential role,

10 as the network represents an organism's capabilities to interact with its biotic and
11 abiotic environment and to transform nutrients into biomass. Mathematical analysis
12 has shown great potential for dissecting the functioning of metabolic networks on
13 the level of topological, stoichiometric, and kinetic models [2], which together pro-
14 vide a wide array of methods [3]. Although different microbial metabolic modelling
15 approaches exist, they can be summarised by a theoretical framework that provides
16 a unifying view on microbial growth [4]. Metabolic models not only have demon-
17 strated their ability to predict phenotypes on the level of cellular growth and gene
18 knockouts, but also provide potential molecular mechanisms in form of gene and
19 reaction activities, which can be validated experimentally [5]. Due to this predictive
20 potential, genome-scale metabolic models have been applied to identify metabolic
21 interactions between different organisms [6, 7, 8, 9, 10], to study host-microbiome
22 interactions [11, 12, 13], to predict novel drug targets to fight microbial pathogens
23 [14, 15], and for the rational design of microbial genotypes and growth-media con-
24 ditions for the industrial production or degradation of biochemicals [16, 17]. Further-
25 more, recent advances in DNA-sequencing technologies have led to a vast increase
26 in available genomic- and metagenomic sequences in databases [18], which further
27 expands the applicability of genome-scale metabolic network reconstructions.

28 The reconstruction of metabolic networks links genomic content with biochemical
29 reactions and therefore depends on sequence annotations and reaction databases,
30 which are both crucial for overall network quality [19, 20]. A general problem in
31 reconstructing metabolic networks occurs by an incorrect representation of the or-
32 ganism's physiology. First, inconsistencies in databases can lead to an incorporation
33 of imbalanced reactions into the metabolic network, which may become responsible
34 for incorrect energy production by futile cycles [20]. Second, many genes are lack-
35 ing a functional annotation due to a lack of knowledge [21] and, thus, also the gene
36 products cannot be integrated into the metabolic networks, which potentially lead
37 to gaps in pathways. Third, the gap-filling of metabolic networks is frequently done
38 by adding a minimum number of reactions from a reference database that facilitate
39 growth under a chemically defined growth medium [22, 23, 24]. Such approaches
40 miss further evidences potentially hidden in sequences and are biased towards the
41 growth medium used for gap-filling. And fourth, the validation of predictions made
42 by metabolic networks is so far only performed with smaller experimental data sets
43 from model laboratory strains such as *Escherichia coli* K12 or *Bacillus subtilis* 168
44 and therefore the overall performance of many metabolic models is insufficiently
45 assured.

46 Genome-scale metabolic network reconstructions are increasingly applied to simu-
47 late complex metabolic processes in microbial communities [25]. Such simulations
48 are highly sensitive to the quality of the individual metabolic networks of the com-
49 munity members. This is because the accurate prediction of fermentation products
50 and carbon source utilisation is crucial for the correct prediction of metabolic in-
51 teractions since the substances produced by one organism may serve as resource for
52 others [26]. Thus, in multi-species communities, the metabolic fluxes of organisms
53 are intrinsically connected, which can lead to error propagation when one defective
54 model affects otherwise correctly working models. As a consequence, the feasibility
55 of community modeling intrinsically depends on the accuracy of the individual or-
56 ganismal models.

57 In this work, we present **gapseq** a novel software for pathway analysis and metabolic
58 network reconstruction. The pathway prediction is based on multiple biochem-
59 istry databases that comprise information on pathway structures, the pathways'
60 key enzymes, and reaction stoichiometries. Moreover, **gapseq** constructs genome-
61 scale metabolic models that enable metabolic phenotype predictions as well as the
62 application in simulations of community metabolism. Models are constructed using
63 a manually curated reaction database that is free of energy-generating thermody-
64 namically infeasible reaction cycles. As input, **gapseq** takes the organism's genome
65 sequence in FASTA format, without the need for an additional annotation file.
66 Topology as well as sequence homology to reference proteins inform the filling of
67 network gaps, and the screening for potential carbon sources and metabolic products
68 is done in a way that reduces the impact of growth medium definitions. Finally, we
69 used large-scale experimental data sets to validate enzyme activity, carbon source
70 utilisation, fermentation products, gene essentiality, and metabolite-cross feeding
71 interactions in microbial communities.

72 **2 Results**

73 **2.1 Biochemistry database and universal model**

74 The pathway-, transporter, and complex prediction is based on a protein sequence
75 database that is derived from UniProt as well as TCDB and consists in total of
76 130,671 unique sequences (111,542 reviewed unipac 0.9 clusters and 19,129 TCDB
77 transporter) and also 1,131,132 unreviewed unipac 0.5 cluster that can be included
78 optionally. In addition, the protein sequence database in **gapseq** can be updated to
79 include new sequences from Uniprot and TCDB. For the construction of genome-
80 scale metabolic network models we have built a biochemistry database, that is
81 derived from the ModelSEED biochemistry database. In total, the resulting cu-
82 rated **gapseq** metabolism database comprises 14.287 reactions (including trans-
83 porters) and 7.570 metabolites. All metabolites and reactions from the biochem-
84 istry database are incorporated in the universal model that **gapseq** utilises for the
85 gap-filling algorithm. When removing all dead-end metabolites and corresponding
86 reactions, the universal model comprises 10.194 reactions and 3.337 metabolites. It
87 needs to be noted, that the current biochemistry database and the derived universal
88 model represents bacterial metabolic functions and that, at the current version of
89 **gapseq**, the database does not include archaea-specific reactions. However, those
90 reactions and, thus, also the possibility to use **gapseq** for the reconstruction of
91 archaeal models will be included in an later version of the software.

92 **2.2 Agreement with enzymatic data (BacDive)**

93 We used experimental data of active metabolic enzymes to compare the accuracy of
94 model generation pipelines. In total, we compared 10,538 enzyme activities, com-
95 prising 30 unique enzymes, in 3,017 organisms. For all organisms, genome-scale
96 metabolic models were constructed using three different pipelines (CarveMe[39],
97 **gapseq**, ModelSEED[24]). **gapseq** models had with 6% the lowest false-negative
98 rate compared to CarveMe (32%) and ModelSEED (28%). Correspondingly, **gapseq**
99 showed with 53% also highest true positive rate compared to CarveMe (27%) and
100 ModelSEED (30%), while the rates of false positive and true negative predictions

101 were comparable (Figure 1A). For this test, the most prominent EC numbers were
102 the catalase, 1.11.1.6, accounting for 26% of the comparisons and the cytochrome
103 oxidase, 1.9.3.1, accounting for 22%.

104 2.3 Validation of carbon source usage (ProTraits)

105 Growth predictions are essential for metabolic models. We checked the quality of
106 model generation pipelines to predict the growth on different carbon sources. In
107 summary, we compared 1,795 different growth prediction for 526 organism and 48
108 carbon sources (Figure 1B). *gapseq* outperformed the other methods in terms of
109 false negatives (14% compared with 29% ModelSEED and 37% CarveMe) and true
110 positives (45% compared with 31% ModelSEED and 23% CarveMe). ModelSEED
111 showed fewer false positives (5% compared with 10% *gapseq* and 11% CarveMe) and
112 more true negatives (35% compared with 30% *gapseq* and 30% CarveMe). *gapseq*,
113 predicted most false positives for formate (29 times). This overestimate of formate
114 as potential carbon source is likely due to the fact that we tested carbon source
115 utilisation on the basis of electron transfer from the source to electron carriers (i.e.
116 ubiquinol, menaquinol, or NADH), which is analogous to the experimental carbon
117 source test of BIOLOG plates [46]. However, while it is known that formate can
118 serve in fact as electron donor in a number of different bacteria [84], the role as
119 source of carbon atoms for the synthesis of biomass components is limited to a few
120 known methylotrophs [85].

121 Across all methods, the most accurately predicted carbon sources, with more than
122 100 tested organisms, were fructose (91% correct predictions), mannose (89%), or
123 arginine (84%), whereby less good predictions were obtained for arabinose (29%
correct predictions), dextrin (40%), or acetate (42%).

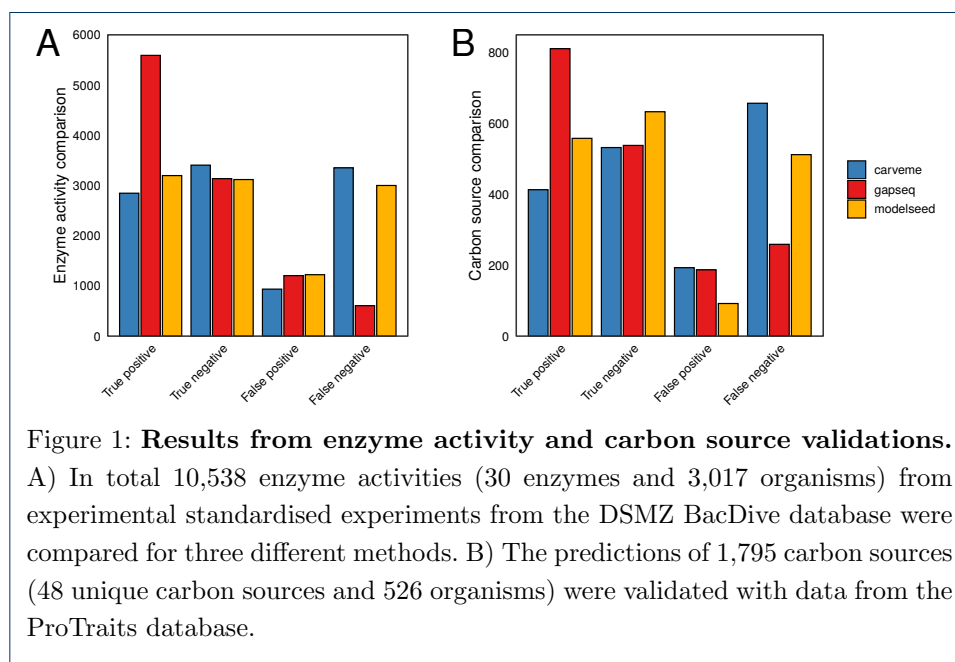


Figure 1: Results from enzyme activity and carbon source validations.

A) In total 10,538 enzyme activities (30 enzymes and 3,017 organisms) from experimental standardised experiments from the DSMZ BacDive database were compared for three different methods. B) The predictions of 1,795 carbon sources (48 unique carbon sources and 526 organisms) were validated with data from the ProTraits database.

124

125 2.4 Gene essentiality

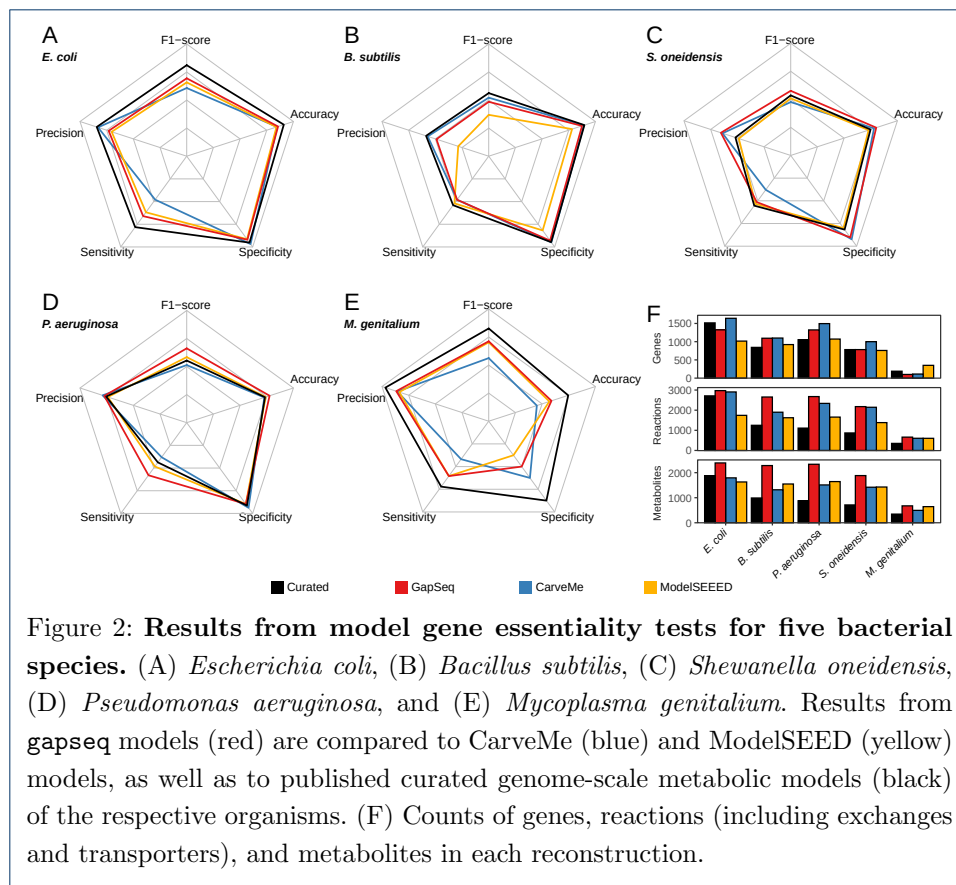


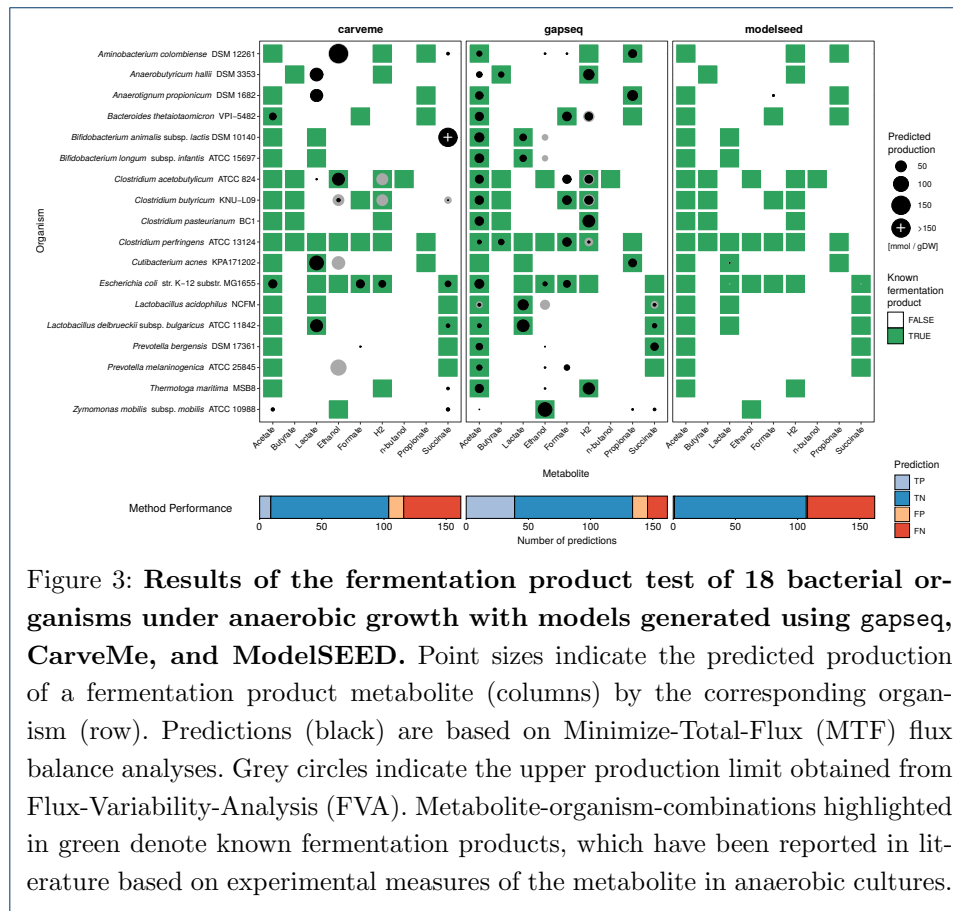
Figure 2: Results from model gene essentiality tests for five bacterial species. (A) *Escherichia coli*, (B) *Bacillus subtilis*, (C) *Shewanella oneidensis*, (D) *Pseudomonas aeruginosa*, and (E) *Mycoplasma genitalium*. Results from *gapseq* models (red) are compared to CarveMe (blue) and ModelSEED (yellow) models, as well as to published curated genome-scale metabolic models (black) of the respective organisms. (F) Counts of genes, reactions (including exchanges and transporters), and metabolites in each reconstruction.

126 We compared the ability of *gapseq* models to predict the essentiality of genes with
127 predictions from ModelSEED and CarveMe reconstructions as well as with curated
128 models for the same organisms (Figure 2). As expected, the curated models out-
129 perform all three automated reconstruction tools for most species and prediction
130 metrics (namely precision, sensitivity, specificity, accuracy, and F1-score). Interest-
131 ingly, for *Pseudomonas aeruginosa* the *gapseq* model shows better gene essentiality
132 predictions in terms of sensitivity, accuracy, and F1-score than the curated model
133 (Figure 2D). Compared to CarveMe, *gapseq* shows generally a higher sensitivity in
134 essentiality predictions but, at the same time, a lower precision rate. This pattern is
135 attributed to the fact, that *gapseq* models tend to predict more genes as essential
136 than CarveMe, leading to a higher number of true positive (TP) predictions but
137 also more false positives (FP). For most organisms and on the basis of most pre-
138 diction metrics, *gapseq* outperforms network models that were reconstructed using
139 ModelSEED.

140 2.5 Fermentation products

141 Anaerobic or facultative anaerobic bacteria utilise different fermentation pathways
142 in order to extract energy from environmental compounds by chemical transfor-
143 mations in the absence of oxygen. We tested if the identity of fermentation prod-
144 ucts can be predicted by metabolic network model constructions obtained from
145 *gapseq*, CarveMe, and ModelSEED for 18 different bacterial organisms (Figure 2).

146 The organisms were selected based on following criteria: (1) the organisms have
 147 a published RefSeq genome sequence [52], (2) are known anaerobic or facultative
 148 anaerobic organisms, and (3) the identity of fermentation products has been exper-
 imentally described and reported in primary literature (Suppl. table S2). Overall,



149
 150 **gapseq** showed the highest number of true positive predictions (TP) with 36 TP
 151 predicted with the Minimize-Total-Flux (MTF) and 37 TP predicted with Flux-
 152 Variability-Analysis (FVA) which is substantially higher compared to CarveMe (8
 153 TP with MTF, 10 TP with FVA) and ModelSEED (1 TP, 3 TP). The produc-
 154 tion of the short-chain-fatty-acids acetate, butyrate, and propionate was correctly
 155 predicted by **gapseq** in 78% of cases and thereby outcompetes CarveMe (9%) and
 156 ModelSEED (0%), which did not predict butyrate or propionate production for any
 157 organism tested. Moreover, **gapseq** correctly predicted homolactic fermentation by
 158 *Lactobacillus delbrueckii* and *Lactobacillus acidophilus*, which is dominated by lac-
 159 tate as fermentation end-product and also predicted heterolactic fermentation by
 160 *Bifidobacterium longum*. However, **gapseq** failed to predict lactate production of or-
 161 ganisms that utilise different fermentation strategies, which also yield lactate (e.g.
 162 mixed-acid fermentation by *Escherichia coli*). Interestingly, the predicted quantities
 163 of fermentation product release is higher for true positive than for false negative
 164 predictions (Figure 3). This further suggests, that **gapseq** is able to predict the
 165 main fermentation products of bacterial organisms during anaerobic growth based
 166 on the organism's genome sequence.

167 2.6 Anaerobic food web of the gut microbiome

168 The prediction of metabolic interactions between microbial organisms is of special
169 interest in ecology, medicine, and biotechnology. So far, we showed the capacity
170 of `gapseq` on the level of individual models. In a next step, we simulated several
171 individual models together as a multi-species community to validate the potential
172 of `gapseq` in microbial community modelling. As sample application we selected
173 representative members of the human gut microbiome that are known to form an
174 anaerobic food web [64, 65]. Altogether, we employed 20 organisms and simulated
175 the combined growth in a shared environment for several time steps using the com-
176 munity modeling framework BacArena [68]. On the community level, simulations
177 using `gapseq` models captured all important substances, which are known to be
178 produced in the context of the food web (Figure 4). This included the production
179 of short chain fatty acids (acetate, propionate, butyrate), lactate, hydrogen, hy-
180 drogen sulfide (H₂S), methane, formate, and succinate. The formation of acetate,
181 formate, and hydrogen was most prevalent, which are also common end-products
182 of fermentation. Lactate, succinate, acetate, hydrogen, formate, and H₂S were fur-
183 ther metabolised by some community members (Figure 4). The predicted identity
184 of fermentation end-products and other by-products of metabolism was found to
185 be in line with literature information [64, 65, 86]. For example, the formation of
186 lactate was observed for *Lactobacillus acidophilus* and *Bifidobacterium longum*, and
187 butyrate was released by known butyrate producers, i.e. *Faecalibacterium praus-*
188 *nitzii*, *Anaerobutyricum hallii*, *Clostridium perfringens*, and *Coprococcus* spp.. Es-
189 pecially the main products of mixed acid fermentation (acetate, formate, hydrogen,
190 ethanol) were predicted for most members of the community which is in agreement
191 with what is known about common metabolic end products of many gut-dwelling
192 microorganisms [86]. Interestingly, for *Faecalibacterium prausnitzii* no acetate pro-
193 duction is reported [86], which was also observed in our simulations. Moreover, H₂S
194 was correctly predicted to be produced by *Desulfovibrio desulfuricans*. In general,
195 the anaerobic oxidation of fatty acids is not favored by the gut environment because
196 the host competes for the uptake of butyrate, propionate, and acetate, which serve
197 as energy source for colonic epithelial cells and are involved in many host functions
198 [87]. Therefore, the gut community lacks syntrophic organisms which are able to
199 anaerobically degrade butyrate [88]. In agreement with this, we found no microbial
200 uptake of butyrate in the community simulation. In contrast, lactate was predicted
201 to be produced and consumed by distinct community members. We found utilis-
202 ation of lactate by *Coprococcus comes*, *Megasphaera elsdenii*, and *Veillonella dispar*,
203 which is a known feature of these organisms [64]. In addition, succinate was cor-
204 rectly predicted to be used by *Bacteroides* species [86]. The formation of methane
205 is known to be limited to methanogenic archaea, and thus *Methanosarcina barkeri*
206 produced methane from acetate and hydrogen during our simulations.
207 For comparison, the community simulation were also performed using models re-
208 constructed with CarveMe and ModelSEED (Figure 4). In both cases, most of the
209 above-mentioned known metabolic cross-feeding interactions and end-products were
210 not predicted, for instance the production of the short chain fatty acids butyrate
211 and propionate was missing. In summary, `gapseq` models were able to recapitulate
212 the major interactions, which are described for microbial communities in the hu-
213 man gut. The overall consumption pattern and individual microbial contributions

214 were found to be in agreement with literature data. Taken together, the community
 215 simulation results illustrate the capacity of **gapseq** to construct predictive models
 216 for complex metabolic interaction networks comprising several different species.

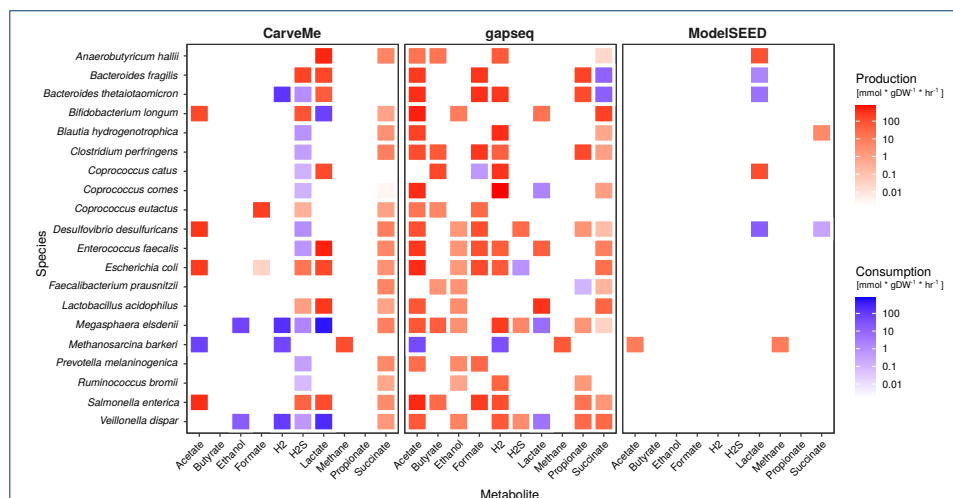


Figure 4: **Predicted metabolic products and food web of a microbial community.** The metabolism of a community consisting of 19 bacterial species commonly found in the human gut and one archaeon (*Methanosarcina barkeri*) was predicted using BacArena [68]. All bacterial models were reconstructed by CarveMe, **gapseq**, or ModelSEED; with the exception of *M. barkeri* for which a published and manually curated model [66] was used.

217 2.7 Pathway prediction of soil and gut microorganisms

218 To demonstrate the pathway prediction capabilities of **gapseq**, we analysed two
 219 communities of soil and gut microorganisms comprising 922 and 822 organisms,
 220 respectively. The two communities could be separated from each other by differ-
 221 ences in energy metabolism (Principal component analysis, Figure 5A). Here, most
 222 variance was explained by subsystems of pathways that are involved in chemoau-
 223 totrophic, respiratory, and fermentative processes including hydrogen production.
 224 Out of 128 energy pathways, the presence of 40 pathways differed significantly
 225 (Kolmogorov-Smirnov test, $P < 0.05$) between soil and gut microorganisms and
 226 could be categorised into 12 subsystems (Figure 5B). In total, gut microorganisms
 227 showed less variety in energy pathways than soil microorganisms. Only pathways
 228 relevant for the formation of acetate, hydrogen, and lactate were predicted to be
 229 enriched. In the case of all other energy subsystems, more pathways were predicted
 230 for soil organisms, most prominently pathways relevant for aerobic and anaerobic
 231 respiration as well as the tricarboxylic acid cycle (TCA). In summary, members of
 232 the soil community showed a more versatile energy metabolisms, which potentially
 233 indicates a higher energetic specialisation of gut microbes. This sample application
 234 demonstrates how **gapseq** can facilitate the characterisation and comparison of mi-
 235 crobial communities based on the analysis of the presence and absence of specific
 236 metabolic pathways.

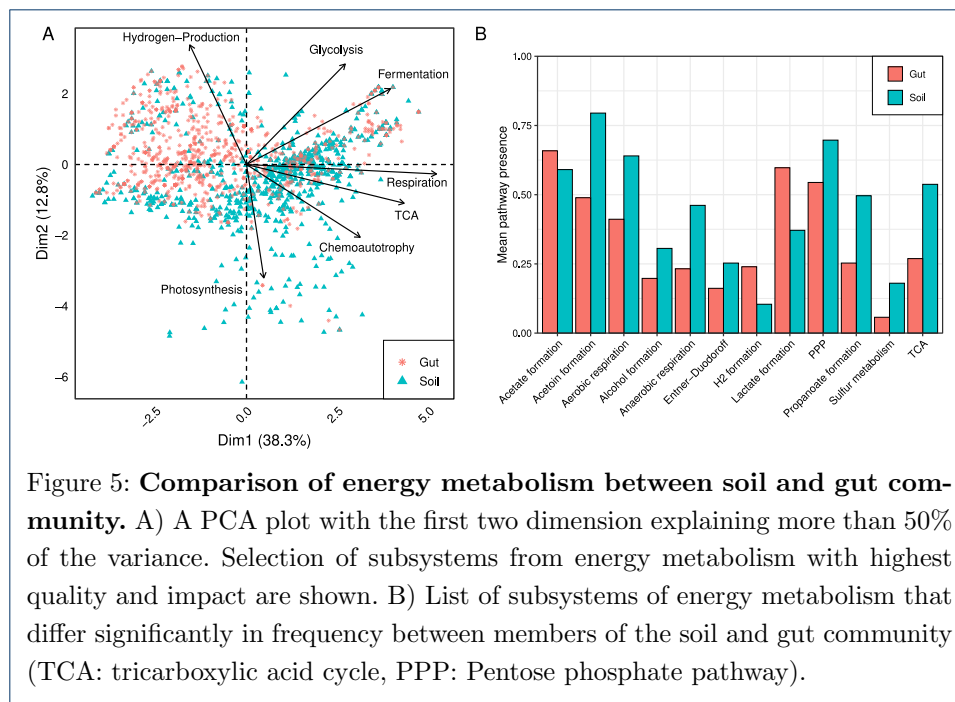


Figure 5: **Comparison of energy metabolism between soil and gut community.** A) A PCA plot with the first two dimension explaining more than 50% of the variance. Selection of subsystems from energy metabolism with highest quality and impact are shown. B) List of subsystems of energy metabolism that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

237 2.8 Model reconstructions for metagenomic assemblies

238 Genome-scale metabolic models can also be reconstructed on the basis of species-
239 level genome bins (SGBs, [69]) assembled from shotgun metagenomic sequencing
240 reads. Yet, genome assemblies from metagenomic material are more prone to errors,
241 fragmentation, and sequence gaps than assemblies of isolated genomes [89], which
242 can potentially cause gaps in the metabolic network reconstructions. We tested
243 whether **gapseq** is able to identify and fill such gaps by comparing the models re-
244 constructed for 127 SGBs from the human microbiome[69] to corresponding models
245 of closely-related reference genomes that were assembled from DNA-sequencing of
246 pure cultures (Figure S2).

247 As expected, we found a strong positive correlation between the SGBs' genome com-
248 pletion and their model similarity to their respective reference models (Spearman's
249 rank correlation, $n = 127$, $P < 10^{-9}$). To estimate the quantitative effect of genome
250 completion on the model similarity, a logarithmic function ($y(x) = c + b * \log(x)$)
251 was fitted to the data ($R^2 = 0.71$, Figure S2). The fitted model indicated, that
252 **gapseq** is able to reconstruct the underlying metabolic network of an organism even
253 on the basis of incomplete and fragmented genomes. For instance, **gapseq** was on
254 average able to recover 90% of the enzymatic reactions that are found in the refer-
255 ence models for SGBs with a predicted genome completion of only 80% (Figure
256 S2).

257 2.9 Summary of validation tests

258 For each validation approach, predictions were compared to experimental data
259 obtained from databases and literature to calculate prediction performance scores.
260 The overall accuracy (proportion all correct prediction in relation to all predic-
261 tions made) of model predictions with experimental data was 66% (CarveMe), 70%

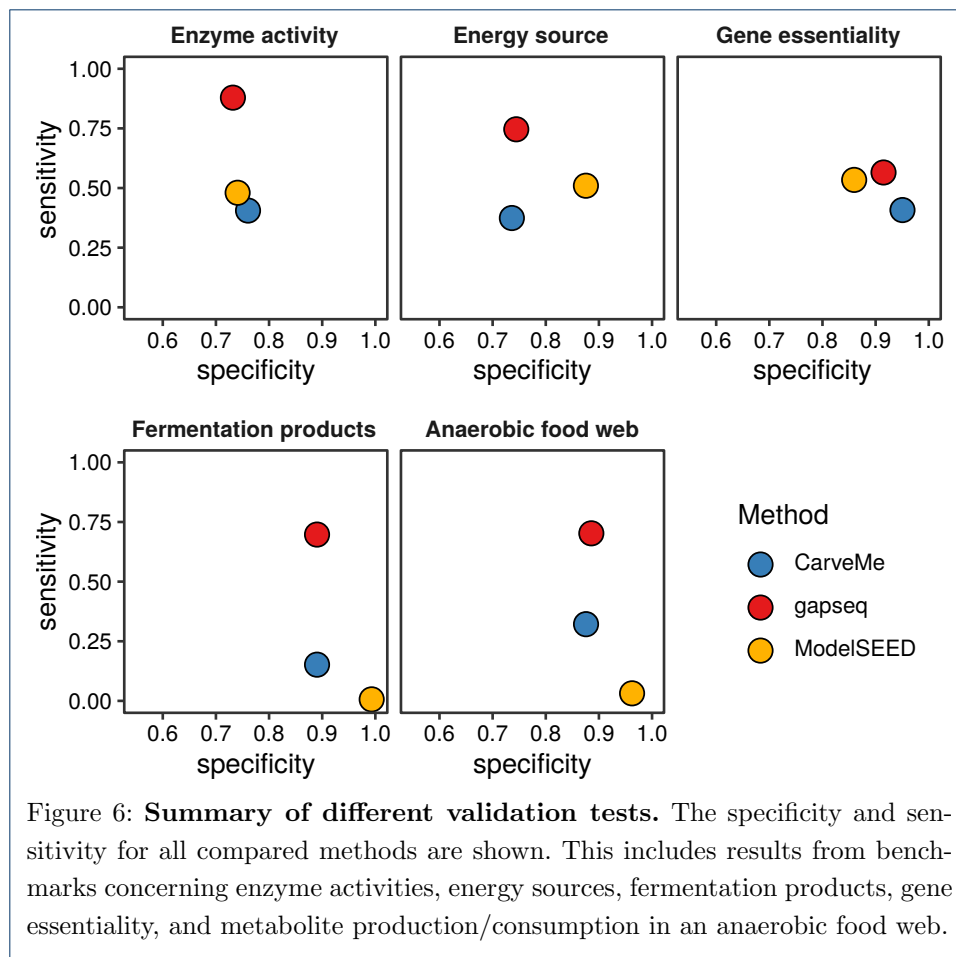


Figure 6: **Summary of different validation tests.** The specificity and sensitivity for all compared methods are shown. This includes results from benchmarks concerning enzyme activities, energy sources, fermentation products, gene essentiality, and metabolite production/consumption in an anaerobic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

262 (ModelSEED), and 81% (**gapseq**)(Table 1). Sensitivity measures the proportion of
263 correctly predicted positives, whereas specificity accounts for the accurate predic-
264 tion of negatives. All approaches showed a high specificity > 0.7 with highest values
265 for fermentation product and gene essentiality tests. Notably, **gapseq** showed the
266 highest sensitivity over all tests (Figure 6). In summary, **gapseq** outperformed other
267 methods in terms of accuracy and sensitivity while showing similar specificity.

268 **3 Discussion**

269 Here, we introduced **gapseq** - a new tool for metabolic pathway analysis and
270 genome-scale metabolic network reconstruction. The novelty of **gapseq** lies in the
271 combination of (i) a novel reaction prediction that is based both on genomic se-
272 quence homology as well as pathway topology, (ii) a profound curation of the re-
273 action and transporter database to prevent thermodynamically infeasible reaction
274 cycles, and (iii) a reaction evidence score-oriented gap-filling algorithm. In order to
275 scrutinise **gapseq** metabolic models, we compared the models' network structures
276 and predictions with large-scale experimental data sets, which were retrieved from
277 publicly available databases. Furthermore, the ability of **gapseq** to predict bacterial
278 phenotypes was compared to two other commonly used automatic reconstruction
279 methods, namely, CarveMe [39] and ModelSEED [24] (Table 1). ModelSEED is also
280 implemented in the KBASE online software platform [90].

Table 1: Summary of different methods that were compared in this work. Accuracy, sensitivity, and specificity scores are based on 14,895 tested phenotypes including energy sources, enzyme activity, fermentation products, gene essentiality, and anaerobic food web structure predictions.

Metric	CarveMe	gapseq	ModelSEED
<i>Implementation</i>			
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
<i>Performance</i>			
Accuracy	0.66	0.81	0.70
Sensitivity	0.34	0.73	0.32
Specificity	0.84	0.83	0.88
Model file quality**	0.32 ± 0.006	0.78 ± 0.004	0.39 ± 0.016

* Solver runs on ModelSEED server. No local solver is required.

** MEMOTE total score (± SD).

281 Crucial large-scale benchmarking of metabolic models

282 The quality of genome-scale metabolic networks can be assessed by comparing model
 283 predictions with experimental physiological data. The protocol by Thiele and Pals-
 284 son (2010) for the reconstruction of genome-scale metabolic networks recommends
 285 the quality assessment and manual network curation using data for (i) known se-
 286 cretion products (e.g. fermentation end-products), (ii) single-gene deletion mutant
 287 growth phenotypes (i.e. gene essentiality), and (iii) the utilisation of carbon/energy
 288 sources [20]. Tools for the automatic reconstruction of metabolic networks should
 289 also make use of such physiological data whenever available for benchmarking. Here,
 290 we tested our **gapseq** approach on the basis of all three recommended phenotypic
 291 data and compared the performance with CarveMe and ModelSEED. Additionally,
 292 we included two novel benchmark tests: The comparison of model predictions with
 293 (iv) the activity of specific enzymes known from experimental studies [49] and (v)
 294 metabolic interactions (food web) among microorganisms in a multi-species com-
 295 munity within an anaerobic environment. Across all five benchmark tests, we could
 296 show that **gapseq** outperformed CarveMe and ModelSEED in terms of sensitivity
 297 while achieving specificity scores that are comparable to the other two tools (Figure
 298 6).

299 Publicly available genome sequences of microorganisms, which can be subject for au-
 300 tomated metabolic network reconstruction are massively increasing in number due
 301 to continuing advances in high-quality and high-throughput sequencing technologies
 302 [18]. This development is further fueled by the the increasing number of genome as-
 303 semblies from metagenomic material [91]. In contrast, standardised phenotypic data
 304 for microorganisms remains a bottleneck for the validation of automated metabolic
 305 network reconstruction pipelines such as **gapseq**. As consequence, it is crucial for
 306 the future development of automated network reconstruction software to include
 307 possibly all available phenotypic data for benchmarking, especially data from non-
 308 model organisms. To benchmark **gapseq** in in relation to CarveMe and ModelSEED
 309 using phenotypic data from mainly non-model organisms, we retrieved phenotypic
 310 data of enzyme activity for more than 3,000 organisms and carbon source utilisation
 311 for more than 500 organisms from online databases, which is, to our knowledge, the

312 yet largest phenotypic data set used for validation of automatically reconstructed
313 metabolic networks. In this validation approach **gapseq** achieved the highest pre-
314 diction accuracy among all three tools tested (Figure 1).

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

320

321 Automated network reconstructions for community modelling

322 While single organisms can be considered as the building blocks of microbial com-
323 munities, individual metabolic models of organisms are the building blocks of *in*
324 *silico* microbial community simulations. Therefore, genome-scale metabolic models
325 are increasingly applied to predict the function of multi-species microbial communi-
326 ties [61, 92, 93]. To correctly infer metabolic interaction networks between different
327 organisms, it is important that individual models accurately predict nutrient util-
328 isation (e.g. carbon source) and metabolic end-products (e.g. fermentation prod-
329 ucts). In this study, the benchmarks for carbon source utilisation and fermentation
330 end-product identity indicated that **gapseq** has the highest prediction performance
331 compared to other reconstruction tools (Figure 1 and Figure 3).

332 To illustrate the applicability of **gapseq**-reconstructed metabolic models for the
333 simulation of multi-species community metabolism, we generated models for micro-
334 bial strains from the human gut microbiota and simulated their growth in a shared
335 environment. Without further curation, the community simulation reproduced all
336 important hallmarks of intestinal anaerobic food webs [64, 86]. Above all, short chain
337 fatty acids (SCFA) were predicted to be the primary end products of fermentation.
338 This prediction is important to represent intestinal metabolism, because SCFA are
339 crucially involved in host physiology by affecting regulatory response in intestinal
340 and immune cells [94, 95]. Furthermore, the simulation accurately predicted the
341 exchange of metabolites between different members of the microbial community
342 (Figure 4). Cross-feeding of metabolites and the formation of anaerobic food chains
343 have been associated with a healthy microbiome [9, 96]. For instance, the cross-
344 feeding of lactate has been reported to be vital for the early establishment of a
345 healthy gut microbiota in infants [96]. Accordingly we observed the exchange of
346 lactate between different bacterial species in the community simulations (Figure 4)
347 and involved known lactate producers (e.g. *Enterococcus faecalis*) and consumers
348 (e.g. *Megasphaera elsdenii*). This example illustrates that we are able to predict
349 key features of the anaerobic food-web within the gastrointestinal microbiota using
350 **gapseq** models. In addition to the ability to accurately model metabolic processes
351 within existing microbial communities, **gapseq** will further promote the potential
352 of metabolic modelling to predict how complex microbial communities can be mod-
353 ulated by targeted interventions. Specific interventions, which could for instance be
354 predicted, are the introduction of new species to the community (i.e. probiotics) or
355 microbiome-modulating compounds (prebiotics) to the environment. Predictions of
356 potential intervention strategies that target the microbiome are of vast relevance

357 for biomedical research. Furthermore, metabolic interactions between microbiome
358 members are difficult to detect *in vivo* due to the simultaneous production and
359 uptake of metabolites. Thus, *in silico* predictions of metabolite cross-feeding in-
360 teractions are highly valuable for hypothesis generation about the function and
361 dynamics of microbial communities.

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

366 Pathway analysis of microbial communities

367 The construction of genome-scale metabolic models is based on metabolic networks
368 that are inferred from genomic sequences in the context of biochemical databases
369 [20]. Although, the reconstruction of metabolic networks is closely related to the
370 prediction of metabolic pathways, metabolic modelling and pathway analysis are
371 often treated separately [97]. In **gapseq**, the prediction of metabolic pathways is
372 intrinsically tied to the reconstruction of metabolic networks and gap-filling. In ad-
373 dition, reaction, transporter, and pathway predictions can also be used to evaluate
374 the functional capacities of microorganisms without the need of metabolic mod-
375 elling. As an example for metabolic pathway analysis, we compared the predicted
376 energy metabolism of two large microbial communities that occur in soil and the hu-
377 man gut. We could show that the predicted distribution of pathways differ between
378 both communities based on the habitat, which usually accommodates the members
379 of the respective community. Gut microorganisms showed a less versatile energy
380 metabolism and a specialisation towards fermentation pathways, which lead to the
381 formation of acetate, hydrogen, and lactate. Variations in pathways distributions
382 between both communities may be explained by distinct evolutionary histories. The
383 habitat of the diverse group of soil microorganisms more likely represents an open
384 ecosystem, whereas the gut microbiome is directly constraint by a multi-cellular
385 host that potentially affect microbial phenotypic traits [98]. In general, metabolic
386 modelling should be accompanied by the analysis of pathways based on statistical
387 methods [97] to compensate for additional assumptions, which are introduced in
388 constraint-based metabolic flux modelling [4].

389 Limitations and outlook

390 **gapseq** requires 1-2h for the reconstruction of a single model, whereas ModelSEED
391 and CarveMe operate faster (10min) on a standard desktop computer. Nonetheless,
392 CarveMe needs as input gene sequences (protein or nucleotide), which has to be
393 predicted first, and ModelSEED works as a web service, which can complicate the
394 handling of large-scale reconstruction projects. In **gapseq**, pathways were predicted
395 based on topology and sequence homology searches. However, the assignment of
396 enzymatic function from sequence comparisons has been shown to potentially miss
397 protein domain structures and thus can cause false annotations [99, 100]. In ad-
398 dition, **gapseq** uses many resources to find potential sequences for reactions in
399 pathway databases. Together this might explain why although **gapseq** performed
400 better than other methods on predicting positive phenotypes (function present),

401 it went head to head with regard to negative phenotype predictions (function not
402 present). CarveMe takes a different approach when inferring function by taking care
403 of functional regions (protein domains) [101], resulting in orthologous groups [102],
404 which results in a slightly better specificity (true negative phenotype predictions)
405 in benchmarks (Figure 6). Future developments of `gapseq` will address orthologous
406 groups by using multiple inference methods. Furthermore, the integration of func-
407 tional predictions coming from phylogenetic inference without the need of genomic
408 sequences [103] might also be promising for further developments of `gapseq`.

409 Conclusion

410 We provide a new software tool called `gapseq` that is suitable for metabolic net-
411 work analysis and metabolic model reconstruction. To enhance phenotype predic-
412 tions, `gapseq` employs various data sources and a novel gap-filling procedure that
413 reduces the impact of arbitrary growth medium requirements. We further brought
414 together the so far largest benchmarking of genome-scale metabolic models, in which
415 `gapseq` outperformed comparable alternative tools. With the increased model qual-
416 ity of automated network reconstructions, `gapseq` will provide new insights into the
417 metabolic phenotypes of non-model and yet-uncultured bacteria whose genomes are
418 assembled from metagenomic material. In this way, the models and their simulations
419 allow predictions on the organisms' ecological role in their natural environments.
420 Taken together, we consider `gapseq` as important contribution to the modelling of
421 microbial communities in the age of the microbiome.

422 4 Methods

423 4.1 Program overview & source code availability

424 The source code is accessible and maintained at [https://github.com/jotech/](https://github.com/jotech/gapseq)
425 `gapseq`. The program is called by `./gapseq`, which is a wrapper script for the
426 main modules. Important program calls are `./gapseq find` (pathway and reac-
427 tion finder), `./gapseq find-transport` (transporter detection), `./gapseq draft`
428 (draft model creation), `./gapseq fill` (gap-filling), or `./gapseq doall` to per-
429 form all in line. When ever necessary, method sections directly refer to config,
430 data and source code files from the `gapseq` package, which contains the main sub-
431 directory `src/` with source code files and `dat/`, which contains databases and also
432 the sequence files in `dat/seq/`. Figure 7 shows an overview of the different `gapseq`
433 modules.

434 4.2 Pathway and sequence databases

435 Pathways are considered as a list of reactions with enzyme names and EC numbers.
436 Pathway definition were obtained from MetaCyc [27], KEGG [28], and ModelSEED
437 [24]. For MetaCyc, PathwayTools [29] was used in combination with Python-
438 Cyc to obtain pathway definitions [30] (`src/meta2pwy.py`). Information on Kegg
439 pathways were retrieved directly from the KEGG homepage: reactions ([http://](http://rest.kegg.jp/list/reaction)
440 rest.kegg.jp/list/reaction), and EC numbers ([http://rest.kegg.jp/link/](http://rest.kegg.jp/link/pathway/ec)
441 [pathway/ec](http://rest.kegg.jp/link/pathway/ec)) and further processed (`src/kegg_pwy.R`). In case of ModelSEED,
442 subsystem definition were obtained from the homepage: [http://modelseed.org/](http://modelseed.org/genomes/Annotations)
443 [genomes/Annotations](http://modelseed.org/genomes/Annotations) (`src/seed_pwy.R`). In addition, manual defined and revised

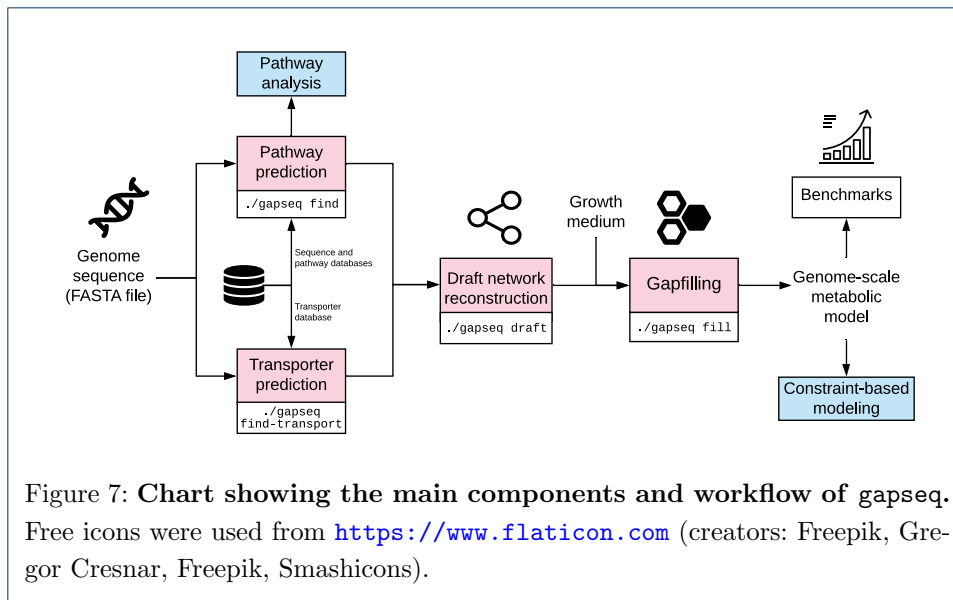


Figure 7: Chart showing the main components and workflow of gapseq. Free icons were used from <https://www.flaticon.com> (creators: Freepik, Gregor Cresnar, Freepik, Smashicons).

444 pathways are stored in the file `dat/custom_pwy.tbl`.

445 Sequence data needed for pathway prediction were downloaded from UniProt [31] for
446 each reaction identified by EC number, enzyme name, or gene name. Both reviewed
447 and unreviewed sequences are considered and stored as clustered UniPac sequences
448 (`src/uniprot.sh`). To increase the sequence pool for a given reaction, alternative
449 EC numbers from BRENDA [32] and from the Enzyme Nomenclature Committee
450 <https://www.qmul.ac.uk/sbcs/iubmb/enzyme/> are integrated (`src/altec.R`,
451 `dat/brenda_ec.csv`).

452 4.3 Pathway prediction

453 For each pathway selected from a pathway database (MetaCyc, KEGG, ModelSEED, custom), gapseq searches for sequence evidence and a pathway is defined
454 as present if enough of its reactions were found to have sequence evidence. In more
455 detail, sequence data (section 4.2) is used for homology search by *tblastn* [33] with
456 the protein sequence as query and the genome as database. By default, a bitscore
457 ≥ 200 and a coverage of at least 75% is needed for a match. For certain reactions,
458 the user can define additional criteria, for example an identity of $\geq 75\%$
459 (`dat/exception.tbl`). In case of protein complexes with subunits, a more complex
460 procedure is followed (section 4.4). Spontaneous reactions, which do not need an
461 enzyme, were set to be present in any case. In general, a pathway is considered to
462 be present if at least 80% of the reactions are found (`completenessCutoffNoHints`
463 `threshold`). This pathway completeness threshold is lowered for pathways in follow-
464 ing cases:
465

- 466 1 If the pathway contains key reactions, as it is defined for some pathways in
467 MetaCyc, and all key reactions are found, then `completenessCutoff` of the
468 total reactions needed to be found. We used a value of 2/3 for this threshold.
- 469 2 In cases in which no sequence data is available for specific reactions, the status
470 of the reactions is set to "vague" and these reactions do not count as missing

471 if they account for less than `vagueCutoff` of the total reactions of a pathway.
472 We used a value of 1/3 for this threshold.
473 The pathway prediction algorithm is implemented in the bash shell script
474 `src/gapseq_find.sh`, which uses GNU parallel [34] and fastaindex/fastafetch from
475 exonerate [35].

476 4.4 Protein complex prediction

477 A problem with automatic sequence download for reactions (as FASTA files) comes
478 with protein complexes, for which a single blast hit may be not sufficient to predict
479 enzyme presence. In `gapseq`, subunits are detected by text matching in the FASTA
480 headers. Search terms are: "subunit", "chain", "polypeptide", "component", and
481 different numbering systems (roman, arabic, greek) are homogenised. To avoid ar-
482 tifacts in text matching, subunits that occur less than five times in the sequence
483 file are not considered, and in cases in which a subunit occurs almost exclusively
484 ($\geq 66\%$) the other entries are not taken into account. All FASTA entries, which
485 could not be matched by text mining, or which are excluded because of the coverage,
486 are labeled 'undefined subunit' and do not add to the total amount of subunits.
487 For each recognised subunit, a blast search is done. A protein complex counts as
488 present if more than 50% of the subunits could be found, whereby the presence of
489 'undefined subunits' tip the balance if exactly 50% of the subunits were found. The
490 text matching with regular expressions is done with R's `stringr` [36] and `biostrings`
491 [37] as defined in `src/complex_detection.R`. The script is called from within the
492 shell script `src/gapseq_find.sh`.

493 4.5 Transporter prediction

494 For transporter search, sequence data from the Transporter Classification Database
495 is employed [38]. In addition, manual defined sequences can be defined in
496 `dat/seq/transporter.fasta`. The sequence set is reduced to a subset of trans-
497 porters that involve metabolites known to be produced or consumed by microor-
498 ganisms (`dat/sub2pwy.csv`). Subsequently, the genome is queried by the reduced
499 sequences using `tblastn` [33]. For each hit (default cutoffs: bitscore ≥ 200 and cov-
500 erage $\geq 75\%$), the transporter type (1. Channels and pores, 2. Electrochemical
501 potential-driven transporter, 3. Primary active transporters, 4. Group transloca-
502 tors) is determined using the TC number mentioned in the FASTA header. A
503 suitable candidate reaction is searched in the reaction database. If there is a hit for
504 a transporter of a substance but no candidate reaction for the respective transporter
505 type can be found, then other transporter types are considered. The transporter
506 search is done by the shell script `src/transporter.sh` that uses GNU parallel [34]
507 and fastaindex/fastafetch from exonerate [35].

508 Candidate transporters are selected from the reaction database by transporter type
509 and substance name. This is done by text search and is currently implemented
510 only for the ModelSEED namespace. From the ModelSEED reaction database
511 all reactions with the flag `is_transport = 1` are taken and the transporter type
512 is predicted by keywords: "channel", "pore" (1. Channels and pores); "uniport",
513 "symport", "antiport", "permease", "gradient" (2. Electrochemical potential-driven
514 transporters); "ABC", "ATPase", "ATP" (3. Primary active transporters); "PTS"

515 (4. Group translocators). If no transporter type could be identified by keywords,
516 additional string matching is done for ATPases, proton/sodium antiporter, and
517 PTS by considering the stoichiometry of the involved metabolites. The transported
518 substance is identified as the substance that occurs on both sides of the reaction. In
519 addition, reactions from the reaction database can be linked manually to substances
520 and transporter types (`dat/seed_transporter_custom.tbl`). The text matching
521 with regular expressions is done with `stringr` [36] (`src/seed_transporter.R`).

522 4.6 Biochemistry database curation and construction of universal metabolic model

523 For the construction of genome-scale metabolic network models, `gapseq` uses a re-
524 actions and metabolite database that is derived from the ModelSEED database [24]
525 as from January 2018. In addition, 30 new reactions and 2 new metabolites were
526 introduced to the `gapseq` biochemistry database (see suppl. table S1). All reactions
527 and metabolites from the database were included for the construction of a full uni-
528 versal metabolic network model; an approach that is also used in CarveMe [39]. We
529 curated the underlying biochemistry database in order to correct inconsistencies in
530 reaction stoichiometries and reversibilities. Inconsistencies were identified by opti-
531 mising the universal network model for ATP-production without any nutritional
532 input to the model using flux balance analysis. In case of ATP-production, the flux
533 distributions of such thermodynamically infeasible reaction cycles were investigated
534 by cross-checking the involved reactions with literature information, the BRENDA
535 database for enzymes [32], and the MetaCyc database [27]. Stoichiometries and
536 reversibilities of erroneous reactions were corrected accordingly. This curation pro-
537 cedure was repeated until no thermodynamically infeasible and ATP-generating
538 reaction cycles were observed.

539 Hits from the pathway prediction (4.3) and transporter prediction (4.5) are mapped
540 to the `gapseq` reaction database using different common identifiers. A majority of
541 reactions are directly matched via their corresponding Enzyme Commission (EC)
542 system identifier [40] and Transporter Classification (TC) system identifier [38], re-
543 spectively. For this mapping, also alternative EC-numbers for enzymatic reactions
544 as defined in the BRENDA database [32] are considered. Moreover, the databases
545 used for pathway and transporter predictions often provide cross-links to the reac-
546 tion's KEGG ID, which is also assigned to most reactions in the `gapseq` database
547 and used to match reactions. Additionally, the MNXref database [41] provides cross
548 links between several biochemistry databases, which `gapseq` also utilises to trans-
549 late hits from the pathway predictions to model reactions. Finally, a manual trans-
550 lation of enzyme names to model reactions is done for some reactions, which we
551 identified as important reactions but which failed to match between the pathway
552 databases (4.3) and the `gapseq` model reactions using other reaction identifiers
553 (`dat/seed_Enzyme_Name_Reactions_Aliases.tsv`). The overall mapping is done
554 by the function `getDBhit()` as defined in `./src/gapseq_find.sh`.

555 4.7 Model draft generation

556 A draft genome-scale metabolic model is constructed based on the results from the
557 pathway and transporter predictions (see above). A reaction is added to the draft
558 model if the corresponding enzyme/transporter was directly found or if the pathway

559 was predicted to be present (i.e. due to pathway completeness and key enzymes) in
560 which the reaction participates. Additionally, spontaneous reactions as defined in
561 the MetaCyc database as well as transport reaction of compounds, which are know
562 to be able to cross cell membranes by means of diffusion (e.g. H_2), are directly
563 added to every draft model. As part of the draft model construction `gapseq` adds a
564 biomass reaction to the network that aims to describe the composition of molecular
565 constituents that the organism needs to produce in order to form 1 g dry weight (1
566 gDW) of bacterial biomass. `gapseq` uses the biomass composition definition from
567 the ModelSEED database for Gram-positive (`dat/seed_biomass.DT_gramPos.tsv`)
568 and Gram-negative bacteria (`dat/seed_biomass.DT_gramNeg.tsv`). If no Gram-
569 staining property is specified by the user, `gapseq` predicts the Gram-staining-
570 dependent biomass reactions by finding the closest 16S-rRNA-gene neighbor us-
571 ing a `blastn` search against reference 16S-rRNA gene sequences from 4647 bac-
572 terial species with known Gram-staining properties that are obtained from the
573 PROTRAITS database [42]. The model draft generation is done by the R script
574 `src/generate_GSdraft.R`.

575 4.8 Gap-filling algorithm

576 `gapseq` provides a gap-filling algorithm that adds reactions to the model in order to
577 enable biomass production (i.e. growth) and likely anabolic and catabolic capabili-
578 ties. The algorithm uses the alignment statistics (i.e. the bitscore) from the pathway-
579 and transporter prediction steps of `gapseq` (see above) to preferentially add reac-
580 tions to the network, which have the highest genetic evidence. This approach is
581 especially relevant in cases where the sequence similarity to known enzyme-coding
582 reference genes was close to but did not reach the cutoff value b , which is required
583 for a reaction to be included directly into the draft network. In contrast to the gap-
584 filling algorithms described in previous works [43] and [39], which also use genetic
585 evidence-weighted gap-filling, the gap-filling problem in `gapseq` is not formulated as
586 Mixed Integer Linear Program (MILP) but as Linear Program (LP), and is derived
587 from the parsimonious enzyme usage Flux Balance Analysis (pFBA) algorithm de-
588 veloped by Lewis *et al.*, 2010 [3]. Therefore, the alignment statistics (i.e. bitscore)
589 are translated into weights for the corresponding model reactions and incorporated
590 into the problem formulation:

$$\text{max: } v_j - c \sum_{i \in R_{all}} w_i |v_i|, \quad (1)$$

$$w_i = \begin{cases} w_{min} & b_i \geq u \quad | \quad i \in R_{draft} \\ (b_i - u) \left(\frac{w_{min} - w_{max}}{u - l} \right) + w_{min} & l \leq b_i < u \\ w_{max} & b_i < l \end{cases}$$

s.t.

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

$$\mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub}$$

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

602 Two other LP-based gap-filling algorithms that incorporate reaction evidence scores
603 have been formulated by Dreyfuss *et al.* (2013) [44] and Medlock *et al.* (2020) [45],
604 respectively. These approaches require a definition of a minimum flux through the
605 biomass reaction to ensure growth. The pFBA-derived LP formulation of **gapseq**
606 (equation 1) includes the flux through the biomass/objective reaction v_j together
607 with the reaction evidence scores in a single objective function.

608 In **gapseq** and following the solution of the LP (1), reactions carrying a flux and
609 which are not part of the draft model are added to the network model. The algo-
610 rithm is implemented in `src/gapfill4.R`.

611 4.9 Gap-filling of biomass, carbon sources, and fermentation products

612 Gap-filling of a draft model in **gapseq** requires only for the first step a user-defined
613 growth medium that is ideally known to support growth of the organism of interest
614 *in vivo*. If no growth medium is specified by the user, a complete medium (ALLmed)
615 is chosen by **gapseq** (as done for the large-scale benchmarks of enzyme activity and
616 carbon sources, cf. 4.11, 4.12). A set of common microbial growth media (e.g. LB,
617 TSB, M9) is provided in the **gapseq** software directory `dat/medium/`. In addition,
618 the user can provide a custom growth medium definition. The above described
619 gap-filling algorithm is used to improve the generated draft model in four steps.

- 620 1 **Biomass production:** To ensure that the model is able to produce biomass
621 under the given nutritional input (medium) the gap-filling algorithm is applied
622 while the objective is defined as the flux through the biomass reaction. This
623 step will add all missing reactions that are essential for *in silico* growth.
- 624 2 **Individual biomass components:** It is checked whether the model supports
625 the biosynthesis of biomass components. Therefore, model is re-constrained
626 to a M9-like minimal medium with a carbon source for which an exchange
627 reactions is found (default: glucose if available). The objective function is
628 set to the production of one biomass component at a time and the gap-fill
629 algorithm is performed. This gap-filling step is repeated for each biomass
630 component metabolite twice, with and without oxygen to potentially allow
631 aerobic and anaerobic growth for facultative anaerobe species.
- 632 3 **Alternative energy sources:** **gapseq** attempts to gap-fill likely metabolic
633 pathways, which enable the utilisation of alternative energy sources, which
634 might not be part of the defined growth medium from step (1). To this end,
635 the model is re-constrained to a M9-like minimal medium containing a single

636 carbon source of interest at the time. As objective function, the summed flux
637 of artificial reactions that accept electrons from the electron carriers ubiquinol,
638 menaquinol, or NADH is defined. This test can be considered as an *in silico*
639 simulation of the commonly used BIOLOG carbon source utilisation test ar-
640 rays [46] in which the colometric effect is coupled to a dehydrogenase [47]. This
641 gap-filling step is performed for all metabolites defined in `dat/sub2pwy.csv`.

642 **4 Metabolic products:** Finally, the same list of compounds as for step (3), is
643 used to check whether the network can be gap-filled to allow the formation
644 of these metabolites given the original medium. For each compound the gap-
645 filling algorithm is applied with the production of the focal compound as
646 objective function.

647 While step (1) considers all reaction from the universal model as potential candi-
648 date reactions for gap-filling, steps (2-4) allow only the addition of candidate
649 reactions to the model with a corresponding bitscore from the pathway prediction
650 (4.3) higher than a threshold value b (default: 50). Thus, these so-termed '*core re-*
651 *actions*' represent only reactions, for which `gapseq` has found genomic sequence or
652 pathway evidence. This approach for steps (2-4) is chosen to avoid the addition
653 of biosynthetic capabilities to the model, which the organism presumably does not
654 possess.

655 4.10 Formal and functional model file testing

656 The validity of genome-scale metabolic model files was checked with MEMOTE
657 (0.10.2) [48]. For all models used in the anaerobic food web (4.16), the total MEM-
658 OTE score was computed for the respective SBML-Model files. MEMOTE was exe-
659 cuted using the parameter `--skip test_find_metabolites_not_produced_with_open_bounds`
660 and `--skip test_find_metabolites_not_consumed_with_open_bounds` since these
661 tests do not contribute to the total MEMOTE score but require long computation
662 time.

663 4.11 Validation with enzymatic data (BacDive)

664 The Bacterial Diversity Metadatabase (BacDive) [49] was used to obtain enzy-
665 matic activity data. For this purpose, a list of type strains IDs were downloaded
666 using the advanced search. Afterwards the IDs were used to query the database
667 via the R package BacDiveR (0.9.1) to obtain the data [50]. If the stored data
668 contained non-zero entries for enzymatic activity and if a genome assembly was
669 available on NCBI, the type strain was considered for the validation analysis.
670 The respective genome assemblies were downloaded with `ncbi-genome-download`
671 (<https://github.com/kblin/ncbi-genome-download>). If multiple genomes were
672 available for one type strain, '*representative*' and '*complete*' (NCBI tags) genomes
673 were preferred and, in case there were still multiple candidate genomes available,
674 the most complete genome was selected. Genome completeness was estimated by
675 employing the software BUSCO (3.0.2) [51]. In total, 3017 type strain genomes
676 were taken as input for ModelSEED (2.5.1), CarveMe (1.2.2), and `gapseq` to create
677 metabolic models. The gap-filling parameters were set to default values for each
678 program, i.e. a complete medium was assumed. The final test whether a reaction
679 activity is covered by a model was done by checking if the corresponding reaction

680 is present in the model. This was done by matching enzymes and reactions via EC
681 numbers. For CarveMe the vmh (<https://www.vmh.life>) and for ModelSEED and
682 `gapseq` the ModelSEED (<http://modelseed.org>) reaction database was used to
683 match reactions and EC numbers. For the EC numbers 3.1.3.1, 3.1.3.2, the corre-
684 sponding reactions were the same, and thus unspecific, so that both EC numbers
685 were not considered for the validation analysis. In general, the enzyme activities
686 in the BacDive database have the form active ("+") or not active ("-") but some
687 entries were ambiguous (e.g.: "+/-"). The ambiguous entries were omitted from the
688 analysis.

689 4.12 Validation with carbon sources data (ProTraits)

690 Data for the validation of carbon source utilisation was obtained from the "atlas
691 of prokaryotic traits" database (ProTraits) [42]. A tab-separated table with bina-
692 rised predictions with a stringent threshold of precision of ≥ 0.95 were downloaded
693 from <http://protraits.irb.hr/data.html>. For organisms which had at least one
694 carbon source prediction, the corresponding genome was obtained from NCBI Ref-
695 Seq [52] if available. In cases where a genome assembly was found, it was taken
696 as input for ModelSEED, CarveMe, and `gapseq` to create metabolic models. The
697 number of potential carbon sources was reduced to a subset for which a map-
698 ping from substance name to ModelSEED and CarveMe model namespace existed
699 (`dat/sub2pwy.csv`). The tests for D-lyxose were removed because it was listed as
700 all negative in ProTraits and also all compared pipelines predicted no utilisation.
701 The main test whether a carbon source can be used by a model was done in a
702 BIOLOG-like manner as described above (see 4.9). To this end, temporary reac-
703 tions to recycle reduced electron carriers as carbon source utilisation indicators were
704 added to the respective model. The objective for optimisation was set to maximise
705 the flux through these recycling reactions. The exchange reactions were limited to a
706 minimal medium with minerals and the focal potential carbon source. This theoret-
707 ical approach tested, whether the model is able to pass electrons from the potential
708 carbon source to electron carrier metabolites. A carbon source was predicted to be
709 able to serve as energy source if the recycle reactions carried a positive flux.

710 4.13 Prediction of gene essentiality

711 To predict the essentiality of genes we performed *in silico* single gene deletion
712 phenotype analysis for the network reconstructions of *Escherichia coli* str. K-
713 12 substr. MG1655 (RefSeq assembly accession: GCF_000005845.2), *Bacillus sub-*
714 *tilis* substr. *subtilis* str. 168 (GCF_000789275.1), *Shewanella oneidensis* MR-1
715 (GCF_000146165.2), *Pseudomonas aeruginosa* PAO1 (GCF_000006765.1), and *My-*
716 *coplasma genitalium* G37 (GCF_000027325.1). The analysis was performed on the
717 basis of the models' Gene-Protein-Reaction (GPR) mappings and according to the
718 protocol by Thiele and Palsson, 2010 [20]. To this end, the contingency tables of pre-
719 dicted growth/no growth phenotypes from the network models and experimentally
720 determined growth phenotypes of gene deletion mutants were constructed. Genes
721 were predicted to be conditionally essential under the given growth environment if
722 the predicted growth rates of the models were below 0.01 hr^{-1} . The growth media
723 compositions for growth predictions were defined as M9 with glucose as carbon-

724 and energy source for *E. coli*, lysogeny broth (LB) for *B. subtilis* and *S. oneidensis*,
725 M9 with succinate as carbon and energy-source for *P. aeruginosa*, and a complete
726 medium (all external metabolites available for uptake) for *M. genitalium*. Experi-
727 mental data for gene essentiality was obtained from [53, 54, 55, 56, 57].

728 4.14 Fermentation product tests

729 The release of by-products from anaerobic metabolism was predicted using Flux
730 Balance Analysis (FBA) coupled with a minimisation of total flux [58] to avoid
731 fluxes that do not contribute to the objective function of the biomass production.
732 In addition, Flux-Variability-Analysis (FVA) [59] was applied to predict the maxi-
733 mum fermentation product release of individual metabolites across all possible FBA
734 solutions. Metabolites with a positive exchange flux (i.e. outflow) were considered
735 as fermentation products. The analysis was performed for 18 different bacterial or-
736 ganisms, which (1) have a genome assembly available in the RefSeq database [52],
737 (2) are known to grow in anaerobic environments, and (3) for which the fermenta-
738 tion products have been described in the literature based on anaerobic cultivation
739 experiments (suppl. table S2). The gap-filling of the network models using `gapseq`,
740 `CarveMe`, and `ModelSEED` as well as the simulations of anaerobic growth were
741 all performed assuming the same growth medium that comprised several organic
742 compounds (i.e. carbohydrates, polyols, nucleotides, amino acids, organic acids) as
743 potential energy sources and nutrients for growth (see media file `dat/media/FT.csv`
744 at the `gapseq` github repository).

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

751 4.15 Pathway prediction of soil and gut microorganisms

752 The pathway analysis was done by comparing predicted pathways of soil and gut
753 microorganisms. For this means, genomes were downloaded from a resource of ref-
754 erence soil organisms [60] and gut microbes [61]. The default parameter of `gapseq`
755 were used for pathway prediction. The principal component analysis was done in
756 R using the `factoextra` package [62]. For predicted pathways for soil and gut mi-
757 croorganisms, it was checked if samples belong to different distributions using a
758 bootstrap version of the Kolmogorov-Smirnov test [63].

759 4.16 Anaerobic food web of the human gut microbiome

760 Representative bacterial organisms known to be relevant in the human intestinal
761 cross-feeding of metabolites were selected based on the proposed food webs by Louis
762 *et al.*, 2014 [64] and Rivera-Chavez *et al.*, 2015 [65]. The genomes of organisms
763 were obtained from NCBI RefSeq [52] and metabolic models reconstructed using
764 `gapseq`, `carveme`, and `modelseed`. A medium containing minerals, vitamins, amino
765 acids, fermentation- and metabolic by-products (namely acetate, formate, lactate,
766 butyrate, propionate, H_2 , CH_4 , ethanol, H_2S , succinate), and carbohydrates (glu-
767 cose, fructose, arabinose, ribose, fucose, rhamnose, lactose) was used for gap-filling.

768 Furthermore, a published model of *Methanosarcina barkeri* was added to the com-
769 munity [66] to represent archaea that are also known to be part of anaerobic food
770 webs [67]. All organisms of the modeled community and their respective genome
771 assembly accession numbers are listed in supplementary table S3. All metabolic
772 models were then simulated with BacArena [68] by using the described medium
773 but without the fermentation and by-products, plus sulfite and 4-aminobenzoate
774 which were needed for growth by the *M. barkeri* model. The community was sim-
775 ulated for five time steps (corresponding to 5 hours simulated time). The analysis
776 of metabolite uptake and production were done after the third time step, for which
777 all organisms were still growing exponentially.

778 4.17 Model reconstructions from metagenomic assemblies

779 4,930 species-level genome bins (SGBs) assembled from shotgun metagenome se-
780 quencing reads were obtained from the study of Pasolli *et al.*, 2019 [69]. Only those
781 SGBs were considered for further analysis, which were already classified as bacteria
782 on a species-level in the original publication by Pasolli *et al.* For each SGB, closely
783 related reference assemblies from the RefSeq database [52] were identified by con-
784 structing a multi-locus phylogenetic tree using autoMLST (version as of April 7th
785 2020, [70]). RefSeq assemblies were considered as genomes from the same species-
786 level taxonomic group as the focal SGB if their predicted MASH distance (D)[71]
787 were below or equal to 0.05. This threshold was shown before to cluster bacterial
788 genomes at the taxonomic level of species [71]. Only SGBs with 10 or more assigned
789 reference assemblies were considered for further analysis, which yielded in total 127
790 SGBs. Metabolic models were reconstructed using `gapseq` for each SGB and their
791 10 closest reference assemblies (Suppl. Table S5).

792 Next, similarity of SGB models with their respective reference models was calcu-
793 lated using the following metabolic network similarity score T_{SGB} :

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

(2)

with

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

794 R_{SGB_Ref} is the union set of reactions with associated genes that are part of the
795 network models reconstructed for the ten reference genome assemblies of the focal
796 SGB. R_{SGB} is the set of reactions part of the SGB's model reconstruction. b_i is the
797 frequency of reaction i among the ten SGB's reference models.

798 Completion of the genome sequence of SGBs was estimated by using BUSCO (ver-
799 sion 4.0.6, [51]) using the specific completion score.

800 4.18 Technical details

801 The pathway prediction part of `gapseq` is implemented as Bash shell script and
802 the metabolic model generation part is written in R. Linear optimisation can be

803 performed with a different solvers (GLPK or CPLEX). Other requirements are
804 exonerate, bedtools, and barrnap. In addition, the following R packages are needed:
805 data.table [72], stringr [73], sybil [74], getopt [75], reshape2 [76], doParallel [77],
806 foreach [78], R.utils [79], stringi [80], glpkAPI [81], and BioStrings [82]. Models can
807 be exported as SBML [83] file using sybilSBML [74] or R data format (RDS) for
808 further analysis in R, for example with sybil [74] or BacArena [68].

809 **Competing interests**

810 The authors declare that they have no competing interests.

811 **Author's contributions**

812 JZ, CK, and SW conceptualized gapseq. JZ and SW developed the software and did the analysis. JZ, CK, and SW
813 wrote the manuscript.

814 **Acknowledgements**

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

817 **Availability of data and materials**

818 gapseq is implemented in R and python and is freely available under the GNU General Public License (v3.0) on
819 GitHub (<https://github.com/jotech/gapseq/>). All results presented in this manuscript were produced using the
820 specific gapseq version 1.0 as archived on GitHub. The datasets used for model construction and validation purposes
821 were obtained from publicly available databases and publications as cited at the respective parts of the manuscript.

822 **Funding**

823 CK and SW acknowledges support by the Collaborative Research Centre 1182 - "Origin and Function of
824 Metaorganisms" - Deutsche Forschungsgemeinschaft and by the Cluster of Excellence 2167 - "Precision medicine in
825 chronic inflammation" - Deutsche Forschungsgemeinschaft. The funders had no role in study design, data collection
826 and analysis, decision to publish, or preparation of the manuscript.

827 **Author details**

828 ¹Christian-Albrechts-University Kiel, Institute of Experimental Medicine, Research Group Medical Systems Biology,
829 Michaelis-Str. 5, 24105 Kiel, Germany. ²Christian-Albrechts-University Kiel, Institute of Human Nutrition and Food
830 Science, Nutriinformatics, Heinrich-Hecht-Platz 10, 24118 Kiel, Germany.

831 **References**

- 832 1. Fell, D.A.: Systems properties of metabolic networks. In: Bar-Yam, Y. (ed.) Unifying Themes In Complex
833 Systems, Volume 1, pp. 163–178. CRC Press, ??? (2003)
- 834 2. Steuer, R.: Computational approaches to the topology, stability and dynamics of metabolic networks.
835 Phytochemistry **68**(16), 2139–2151 (2007). doi:[10.1016/j.phytochem.2007.04.041](https://doi.org/10.1016/j.phytochem.2007.04.041). Dynamic Metabolic
836 Networks
- 837 3. Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N.,
838 Schramm, G., Purvine, S.O., Lopez-Ferrer, D., Weitz, K.K., Eils, R., König, R., Smith, R.D., Palsson, B.O.:
839 Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models.
840 Molecular Systems Biology **6**(1), 390 (2010). doi:[10.1038/msb.2010.47](https://doi.org/10.1038/msb.2010.47).
841 <https://www.embopress.org/doi/pdf/10.1038/msb.2010.47>
- 842 4. de Jong, H., Casagrande, S., Giordano, N., Cinquemani, E., Ropers, D., Geiselmann, J., Gouzé, J.-L.:
843 Mathematical modeling of microbes: Metabolism, gene expression, and growth. Journal of the Royal Society
844 Interface **14**(20170502) (2017). doi:[10.1098/rsif.2017.0502](https://doi.org/10.1098/rsif.2017.0502)
- 845 5. Varma, A., Palsson, B.O.: Metabolic capabilities of escherichia coli ii. optimal growth patterns. Journal of
846 Theoretical Biology **165**(4), 503–522 (1993). doi:[10.1006/jtbi.1993.1203](https://doi.org/10.1006/jtbi.1993.1203)
- 847 6. Stolyar, S., Van Dien, S., Hillesland, K.L., Pinel, N., Lie, T.J., Leigh, J.A., Stahl, D.A.: Metabolic modeling of
848 a mutualistic microbial community. Molecular Systems Biology **3**(1), 92 (2007). doi:[10.1038/msb4100131](https://doi.org/10.1038/msb4100131).
849 <https://www.embopress.org/doi/pdf/10.1038/msb4100131>
- 850 7. Zomorodi, A.R., Islam, M.M., Maranas, C.D.: d-optcom: Dynamic multi-level and multi-objective metabolic
851 modeling of microbial communities. ACS Synthetic Biology **3**(4), 247–257 (2014). doi:[10.1021/sb4001307](https://doi.org/10.1021/sb4001307).
852 PMID: 24742179. <https://doi.org/10.1021/sb4001307>
- 853 8. Harcombe, W., Riehl, W., Dukovski, I., Granger, B., Betts, A., Lang, A., Bonilla, G., Kar, A., Leiby, N.,
854 Mehta, P., Marx, C., Segrè, D.: Metabolic resource allocation in individual microbes determines ecosystem
855 interactions and spatial dynamics. Cell Reports **7**(4), 1104–1115 (2014). doi:[10.1016/j.celrep.2014.03.070](https://doi.org/10.1016/j.celrep.2014.03.070)
- 856 9. Aden, K., Rehman, A., Waschina, S., Pan, W.-H., Walker, A., Lucio, M., Nunez, A.M., Bharti, R.,
857 Zimmerman, J., Bethge, J., Schulte, B., Schulte, D., Franke, A., Nikolaus, S., Schroeder, J.O., Vandeputte,
858 D., Raes, J., Szymczak, S., Waetzig, G.H., Zeuner, R., Schmitt-Kopplin, P., Kaleta, C., Schreiber, S.,
859 Rosenstiel, P.: Metabolic functions of gut microbes associate with efficacy of tumor necrosis factor antagonists
860 in patients with inflammatory bowel diseases. Gastroenterology (2019). doi:[10.1053/j.gastro.2019.07.025](https://doi.org/10.1053/j.gastro.2019.07.025)
- 861 10. Koch, S., Kohrs, F., Lahmann, P., Bissinger, T., Wendschuh, S., Benndorf, D., Reichl, U., Klamt, S.: Redcom:
862 A strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing
863 experimental datasets from anaerobic digestion. PLOS Computational Biology **15**(2), 1–32 (2019).
864 doi:[10.1371/journal.pcbi.1006759](https://doi.org/10.1371/journal.pcbi.1006759)

- 865 11. Heinken, A., Thiele, I.: Systematic prediction of health-relevant human-microbial co-metabolism through a
866 computational framework. *Gut Microbes* **6**(2), 120–130 (2015). doi:[10.1080/19490976.2015.1023494](https://doi.org/10.1080/19490976.2015.1023494). PMID:
867 25901891. <https://doi.org/10.1080/19490976.2015.1023494>
- 868 12. Pryor, R., Norvaisas, P., Marinos, G., Best, L., Thingholm, L.B., Quintaneiro, L.M., Haes, W.D., Esser, D.,
869 Waschina, S., Lujan, C., Smith, R.L., Scott, T.A., Martinez-Martinez, D., Woodward, O., Bryson, K., Laudes,
870 M., Lieb, W., Houtkooper, R.H., Franke, A., Temmerman, L., Bjedov, I., Cochemé, H.M., Kaleta, C.,
871 Cabreiro, F.: Host-microbe-drug-nutrient screen identifies bacterial effectors of metformin therapy. *Cell* **178**(6),
872 1299–131229 (2019). doi:[10.1016/j.cell.2019.08.003](https://doi.org/10.1016/j.cell.2019.08.003)
- 873 13. Zimmermann, J., Obeng, N., Yang, W., Pees, B., Petersen, C., Waschina, S., Kissoyan, K.A., Aidley, J.,
874 Hoepfner, M.P., Bunk, B., Spröer, C., Leippe, M., Dierking, K., Kaleta, C., Schulenburg, H.: The functional
875 repertoire contained within the native microbiota of the model nematode *Caenorhabditis elegans*. *The ISME*
876 *Journal* **14**(1), 26–38 (2019). doi:[10.1038/s41396-019-0504-y](https://doi.org/10.1038/s41396-019-0504-y)
- 877 14. Oberhardt, M.A., Yizhak, K., Ruppin, E.: Metabolically re-modeling the drug pipeline. *Current Opinion in*
878 *Pharmacology* **13**(5), 778–785 (2013). doi:[10.1016/j.coph.2013.05.006](https://doi.org/10.1016/j.coph.2013.05.006). Anti-infectives • New technologies
- 879 15. Trawick, J.D., Schilling, C.H.: Use of constraint-based modeling for the prediction and validation of
880 antimicrobial targets. *Biochemical Pharmacology* **71**(7), 1026–1035 (2006). doi:[10.1016/j.bcp.2005.10.049](https://doi.org/10.1016/j.bcp.2005.10.049).
881 Special Issue on Antibacterials
- 882 16. Rau, M.H., Zeidan, A.A.: Constraint-based modeling in microbial food biotechnology. *Biochem. Soc. Trans.*
883 **46**, 249–260 (2018). doi:[10.1042/BST20170268](https://doi.org/10.1042/BST20170268)
- 884 17. Park, J.H., Lee, S.Y.: Towards systems metabolic engineering of microorganisms for amino acid production.
885 *Current Opinion in Biotechnology* **19**(5), 454–460 (2008). doi:[10.1016/j.copbio.2008.08.007](https://doi.org/10.1016/j.copbio.2008.08.007). Tissue, cell and
886 pathway engineering
- 887 18. Loman, N.J., Pallen, M.J.: Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*
888 **13**(12), 787–794 (2015). doi:[10.1038/nrmicro3565](https://doi.org/10.1038/nrmicro3565)
- 889 19. Wittig, U., De Beuckelaer, A.: Analysis and comparison of metabolic pathway databases. *Briefings in*
890 *bioinformatics* **2**(2), 126–142 (2001)
- 891 20. Thiele, I., Palsson, B.O.: A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat*
892 *Protoc* **5**(1), 93–121 (2010). doi:[10.1038/nprot.2009.203](https://doi.org/10.1038/nprot.2009.203)
- 893 21. Blaby-Haas, C.E., de Crécy-Lagard, V.: Mining high-throughput experimental data to link gene and function.
894 *Trends in Biotechnology* **29**(4), 174–182 (2011). doi:[10.1016/j.tibtech.2011.01.001](https://doi.org/10.1016/j.tibtech.2011.01.001)
- 895 22. Thiele, I., Vlassis, N., Fleming, R.M.T.: fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*
896 **30**(17), 2529–2531 (2014). doi:[10.1093/bioinformatics/btu321](https://doi.org/10.1093/bioinformatics/btu321)
- 897 23. Prigent, S., Frioux, C., Dittami, S.M., Thiele, S., Larhlimi, A., Collet, G., Gutknecht, F., Got, J., Eveillard, D.,
898 Bourdon, J., Plewniak, F., Tonon, T., Siegel, A.: Meneco, a topology-based gap-filling tool applicable to
899 degraded genome-wide metabolic networks. *PLOS Computational Biology* **13**(1), 1005276 (2017).
900 doi:[10.1371/journal.pcbi.1005276](https://doi.org/10.1371/journal.pcbi.1005276)
- 901 24. Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., Stevens, R.L.: High-throughput generation,
902 optimization and analysis of genome-scale metabolic models. *Nature Biotechnology* **28**(9), 977–982 (2010).
903 doi:[10.1038/nbt.1672](https://doi.org/10.1038/nbt.1672)
- 904 25. Kumar, M., Ji, B., Zengler, K., Nielsen, J.: Modelling approaches for studying the microbiome. *Nature*
905 *Microbiology* **4**(8), 1253–1267 (2019). doi:[10.1038/s41564-019-0491-9](https://doi.org/10.1038/s41564-019-0491-9)
- 906 26. Phelan, V.V., Liu, W.-T., Pogliano, K., Dorrestein, P.C.: Microbial metabolic exchange—the
907 chemotype-to-phenotype link. *Nat. Chem. Biol.* **8**, 26–35 (2012). doi:[10.1038/nchembio.739](https://doi.org/10.1038/nchembio.739)
- 908 27. Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M.,
909 Midford, P.E., Ong, Q., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D.: The metacyc database of metabolic
910 pathways and enzymes. *Nucleic Acids Research* **46**(D1), 633–639 (2018). doi:[10.1093/nar/gkx935](https://doi.org/10.1093/nar/gkx935)
- 911 28. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., Tanabe, M.: New approach for understanding genome
912 variations in KEGG. *Nucleic Acids Research* **47**(D1), 590–595 (2018). doi:[10.1093/nar/gky962](https://doi.org/10.1093/nar/gky962).
913 <http://oup.prod.sis.lan/nar/article-pdf/47/D1/D590/27436321/gky962.pdf>
- 914 29. Karp, P.D., Latendresse, M., Paley, S.M., Ong, M.K.Q., Billington, R., Kothari, A., Weaver, D., Lee, T.,
915 Subhraveti, P., Spaulding, A., Fulcher, C., Keseler, I.M., Caspi, R.: Pathway tools version 19.0: Integrated
916 software for pathway/genome informatics and systems biology (2015). [1510.03964](https://doi.org/10.1510/03964)
- 917 30. SRI International: PythonCyc (2014). <https://github.com/latendresse/PythonCyc>
- 918 31. The UniProt Consortium: UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**(D1),
919 158–169 (2016). doi:[10.1093/nar/gkw1099](https://doi.org/10.1093/nar/gkw1099).
920 <http://oup.prod.sis.lan/nar/article-pdf/45/D1/D158/23819877/gkw1099.pdf>
- 921 32. Jeske, L., Placzek, S., Schomburg, I., Chang, A., Schomburg, D.: BRENDA in 2019: a European ELIXIR core
922 data resource. *Nucleic Acids Research* **47**(D1), 542–549 (2018). doi:[10.1093/nar/gky1048](https://doi.org/10.1093/nar/gky1048).
923 <http://oup.prod.sis.lan/nar/article-pdf/47/D1/D542/27437170/gky1048.pdf>
- 924 33. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: Blast+:
925 architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). doi:[10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421)
- 926 34. Tange, O.: Gnu parallel - the command-line power tool. *login: The USENIX Magazine*, 42–47 (2011)
- 927 35. Slater, G.S.C., Birney, E.: Automated generation of heuristics for biological sequence comparison. *BMC*
928 *Bioinformatics* **6**, 31 (2005). doi:[10.1186/1471-2105-6-31](https://doi.org/10.1186/1471-2105-6-31)
- 929 36. Wickham, H.: Stringr: Simple, Consistent Wrappers for Common String Operations. (2019). R package
930 version 1.4.0. <https://CRAN.R-project.org/package=stringr>
- 931 37. Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S.: Biostrings: Efficient Manipulation of Biological Strings.
932 (2019). R package version 2.50.2
- 933 38. Saier, M.H., Reddy, V.S., Tamang, D.G., Vastermark, A.: The transporter classification database. *Nucleic*
934 *Acids Research* **42**(D1), 251–258 (2013). doi:[10.1093/nar/gkt1097](https://doi.org/10.1093/nar/gkt1097)
- 935 39. Machado, D., Andrejev, S., Tramontano, M., Patil, K.R.: Fast automated reconstruction of genome-scale
936 metabolic models for microbial species and communities. *Nucleic Acids Research* **46**(15), 7542–7553 (2018).

- 937 doi:[10.1093/nar/gky537](https://doi.org/10.1093/nar/gky537). <http://oup.prod.sis.lan/nar/article-pdf/46/15/7542/25689981/gky537.pdf>
- 938 40. Webb, E.C., *et al.*: Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the
939 International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of
940 Enzymes. vol. Ed. 6. Academic Press, ??? (1992)
- 941 41. Bernard, T., Bridge, A., Morgat, A., Moretti, S., Xenarios, I., Pagni, M.: Reconciliation of metabolites and
942 biochemical reactions for metabolic networks. *Brief Bioinform* **15**(1), 123–135 (2014). doi:[10.1093/bib/bbs058](https://doi.org/10.1093/bib/bbs058)
- 943
- 944 42. Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., Supek, F.: The landscape of microbial phenotypic
945 traits and associated genes. *Nucleic Acids Research*, 964 (2016). doi:[10.1093/nar/gkw964](https://doi.org/10.1093/nar/gkw964)
- 946 43. Benedict, M.N., Mundy, M.B., Henry, C.S., Chia, N., Price, N.D.: Likelihood-based gene annotations for gap
947 filling and quality assessment in genome-scale metabolic models. *PLOS Computational Biology* **10**(10), 1–14
948 (2014). doi:[10.1371/journal.pcbi.1003882](https://doi.org/10.1371/journal.pcbi.1003882)
- 949 44. Dreyfuss, J.M., Zucker, J.D., Hood, H.M., Ocasio, L.R., Sachs, M.S., Galagan, J.E.: Reconstruction and
950 validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM.
951 *PLoS Computational Biology* **9**(7), 1003126 (2013). doi:[10.1371/journal.pcbi.1003126](https://doi.org/10.1371/journal.pcbi.1003126)
- 952 45. Medlock, G.L., Papin, J.A.: Guiding the refinement of biochemical knowledgebases with ensembles of
953 metabolic networks and machine learning. *Cell Systems* **10**(1), 109–1193 (2020).
954 doi:[10.1016/j.cels.2019.11.006](https://doi.org/10.1016/j.cels.2019.11.006)
- 955 46. Smalla, K., Wachtendorf, U., Heuer, H., Liu, W.-t., Forney, L.: Analysis of biologic substrate utilization
956 patterns by microbial communities. *Applied and Environmental Microbiology* **64**(4), 1220–1225 (1998).
957 <https://aem.asm.org/content/64/4/1220.full.pdf>
- 958 47. Bochner, B.R.: Global phenotypic characterization of bacteria. *FEMS Microbiology Reviews* **33**(1), 191–205
959 (2009). doi:[10.1111/j.1574-6976.2008.00149.x](https://doi.org/10.1111/j.1574-6976.2008.00149.x)
- 960 48. Lieven, C., Beber, M.E., Olivier, B.G., Bergmann, F.T., Ataman, M., Babaei, P., Bartell, J.A., Blank, L.M.,
961 Chauhan, S., Correia, K., Diener, C., Dräger, A., Ebert, B.E., Edirisinghe, J.N., Faria, J.P., Feist, A.M.,
962 Fengers, G., Fleming, R.M.T., García-Jiménez, B., Hatzimanikatis, V., van Helvoirt, W., Henry, C.S.,
963 Hermjakob, H., Herrgård, M.J., Kaafarani, A., Kim, H.U., King, Z., Klamt, S., Klipp, E., Koehorst, J.J.,
964 König, M., Lakshmanan, M., Lee, D.-Y., Lee, S.Y., Lee, S., Lewis, N.E., Liu, F., Ma, H., Machado, D.,
965 Mahadevan, R., Maia, P., Mardinoglu, A., Medlock, G.L., Monk, J.M., Nielsen, J., Nielsen, L.K., Nogales, J.,
966 Nookaew, I., Palsson, B.O., Papin, J.A., Patil, K.R., Poolman, M., Price, N.D., Resendis-Antonio, O., Richelle,
967 A., Rocha, I., Sánchez, B.J., Schaap, P.J., Sheriff, R.S.M., Shoaie, S., Sonnenschein, N., Teusink, B., Vilaça,
968 P., Vik, J.O., Wodke, J.A.H., Xavier, J.C., Yuan, Q., Zakhartsev, M., Zhang, C.: Memote for standardized
969 genome-scale metabolic model testing. *Nat. Biotechnol.* **38**, 272–276 (2020). doi:[10.1038/s41587-020-0446-y](https://doi.org/10.1038/s41587-020-0446-y)
- 970 49. Reimer, L.C., Vetcinova, A., Carbasse, J.S., Söhngen, C., Gleim, D., Ebeling, C., Overmann, J.: BacDive in
971 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Research* **47**(D1),
972 631–636 (2018). doi:[10.1093/nar/gky879](https://doi.org/10.1093/nar/gky879).
973 <http://oup.prod.sis.lan/nar/article-pdf/47/D1/D631/27436018/gky879.pdf>
- 974 50. Leinweber, K.: BacDiveR: A Programmatic Interface For BacDive, The DSMZ's Bacterial Diversity
975 Metadatabase. (2019). R package version 0.6.0. <https://github.com/TIBHannover/BacDiveR>
- 976 51. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: Busco: assessing genome
977 assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–2 (2015).
978 doi:[10.1093/bioinformatics/btv351](https://doi.org/10.1093/bioinformatics/btv351)
- 979 52. Sayers, E.W., Beck, J., Brister, J.R., Bolton, E.E., Canese, K., Comeau, D.C., Funk, K., Ketter, A., Kim, S.,
980 Kimchi, A., Kitts, P.A., Kuznetsov, A., Lathrop, S., Lu, Z., McGarvey, K., Madden, T.L., Murphy, T.D.,
981 O'Leary, N., Phan, L., Schneider, V.A., Thibaud-Nissen, F., Trawick, B.W., Pruitt, K.D., Ostell, J.: Database
982 resources of the national center for biotechnology information. *Nucleic Acids Research* **48**(D1), 9–16 (2019).
983 doi:[10.1093/nar/gkz899](https://doi.org/10.1093/nar/gkz899)
- 984 53. Zhu, B., Stülke, J.: SubtiWiki in 2018: from genes and proteins to functional network annotation of the model
985 organism *Bacillus subtilis*. *Nucleic Acids Research* **46**(D1), 743–748 (2017). doi:[10.1093/nar/gkx908](https://doi.org/10.1093/nar/gkx908).
986 <https://academic.oup.com/nar/article-pdf/46/D1/D743/23162685/gkx908.pdf>
- 987 54. Monk, J.M., Lloyd, C.J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori,
988 H., *et al.*: iml1515, a knowledgebase that computes *Escherichia coli* traits. *Nature biotechnology* **35**(10), 904
989 (2017)
- 990 55. Turner, K.H., Wessel, A.K., Palmer, G.C., Murray, J.L., Whiteley, M.: Essential genome of *Pseudomonas*
991 *aeruginosa* in cystic fibrosis sputum. *Proceedings of the National Academy of Sciences* **112**(13), 4110–4115
992 (2015)
- 993 56. Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V., Melnyk, R.A.,
994 Lamson, J.S., Suh, Y., *et al.*: Mutant phenotypes for thousands of bacterial genes of unknown function.
995 *Nature* **557**(7706), 503 (2018)
- 996 57. Glass, J.I., Assad-García, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith,
997 H.O., Venter, J.C.: Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences*
998 **103**(2), 425–430 (2006)
- 999 58. Holzhütter, H.-G.: The principle of flux minimization and its application to estimate stationary fluxes in
1000 metabolic networks. *European journal of biochemistry* **271**(14), 2905–2922 (2004)
- 1001 59. Mahadevan, R., Schilling, C.H.: The effects of alternate optimal solutions in constraint-based genome-scale
1002 metabolic models. *Metab Eng* **5**(4), 264–276 (2003). doi:[10.1016/j.ymben.2003.09.002](https://doi.org/10.1016/j.ymben.2003.09.002)
- 1003 60. Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Flater, J., Tiedje, J.M.,
1004 Hofmockel, K.S., Gelder, B., Howe, A.: Strategies to improve reference databases for soil microbiomes. *The*
1005 *ISME Journal* **11**(4), 829–834 (2016). doi:[10.1038/ismej.2016.168](https://doi.org/10.1038/ismej.2016.168)
- 1006 61. Magnusdóttir, S., Heinken, A., Kutt, L., Ravcheev, D.A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C.,
1007 Baginska, J., Wilmes, P., Fleming, R.M.T., Thiele, I.: Generation of genome-scale metabolic reconstructions
1008 for 773 members of the human gut microbiota. *Nature Biotechnology* (2016). doi:[10.1038/nbt.3703](https://doi.org/10.1038/nbt.3703)

- 1009 62. Kassambara, A., Mundt, F.: Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.
1010 (2019). R package version 1.0.6. <https://CRAN.R-project.org/package=factoextra>
- 1011 63. Sekhon, J.S.: Multivariate and propensity score matching software with automated balance optimization: The
1012 Matching package for R. *Journal of Statistical Software* **42**(7), 1–52 (2011)
- 1013 64. Louis, P., Hold, G.L., Flint, H.J.: The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev.*
1014 *Microbiol.* **12**, 661–72 (2014). doi:[10.1038/nrmicro3344](https://doi.org/10.1038/nrmicro3344)
- 1015 65. Rivera-Chávez, F., Bäuml, A.J.: The pyromaniac inside you: Salmonella metabolism in the host gut. *Annual*
1016 *Review of Microbiology* **69**(1), 31–48 (2015). doi:[10.1146/annurev-micro-091014-104108](https://doi.org/10.1146/annurev-micro-091014-104108). PMID: 26002180
- 1017 66. Feist, A.M., Scholten, J.C.M., Palsson, B.O., Brockman, F.J., Ideker, T.: Modeling methanogenesis with a
1018 genome-scale metabolic reconstruction of methanosarcina barkeri. *Mol Syst Biol* **2**, 2006–0004 (2006).
1019 doi:[10.1038/msb4100046](https://doi.org/10.1038/msb4100046)
- 1020 67. Sieber, J.R., McInerney, M.J., Gunsalus, R.P.: Genomic insights into syntrophy: The paradigm for anaerobic
1021 metabolic cooperation. *Annual Review of Microbiology* **66**(1), 429–452 (2012).
1022 doi:[10.1146/annurev-micro-090110-102844](https://doi.org/10.1146/annurev-micro-090110-102844). PMID: 22803797.
1023 <https://doi.org/10.1146/annurev-micro-090110-102844>
- 1024 68. Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., Kaleta, C.: Bacarena: Individual-based metabolic modeling
1025 of heterogeneous microbes in complex communities. *PLOS Computational Biology* **13**(5), 1–22 (2017).
1026 doi:[10.1371/journal.pcbi.1005544](https://doi.org/10.1371/journal.pcbi.1005544)
- 1027 69. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A.,
1028 Ghensi, P., Collado, M.C., Rice, B.L., DuLong, C., Morgan, X.C., Golden, C.D., Quince, C., Huttenhower, C.,
1029 Segata, N.: Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from
1030 metagenomes spanning age, geography, and lifestyle. *Cell* **176**(3), 649–66220 (2019).
1031 doi:[10.1016/j.cell.2019.01.001](https://doi.org/10.1016/j.cell.2019.01.001)
- 1032 70. Alanjary, M., Steinke, K., Ziemert, N.: AutoMLST: an automated web server for generating multi-locus
1033 species trees highlighting natural product potential. *Nucleic Acids Research* **47**(W1), 276–282 (2019).
1034 doi:[10.1093/nar/gkz282](https://doi.org/10.1093/nar/gkz282)
- 1035 71. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash:
1036 fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**(1) (2016).
1037 doi:[10.1186/s13059-016-0997-x](https://doi.org/10.1186/s13059-016-0997-x)
- 1038 72. Dowle, M., Srinivasan, A.: Data.table: Extension of 'data.frame'. (2019). R package version 1.12.6.
1039 <https://CRAN.R-project.org/package=data.table>
- 1040 73. Wickham, H.: Stringr: Simple, Consistent Wrappers for Common String Operations. (2019). R package
1041 version 1.4.0. <https://CRAN.R-project.org/package=stringr>
- 1042 74. Gelius-Dietrich, G., Desouki, A.A., Fritzsche, C.J., Lercher, M.J.: Sybil—efficient constraint-based modelling
1043 in R. *BMC Syst Biol* **7**, 125 (2013). doi:[10.1186/1752-0509-7-125](https://doi.org/10.1186/1752-0509-7-125)
- 1044 75. Davis, T.L., Day, A.: Getopt: C-Like 'getopt' Behavior. (2019). R package version 1.20.3.
1045 <https://CRAN.R-project.org/package=getopt>
- 1046 76. Wickham, H.: Reshaping data with the reshape package. *Journal of Statistical Software* **21**(12), 1–20 (2007)
- 1047 77. Corporation, M., Weston, S.: doParallel: Foreach Parallel Adaptor for the 'parallel' Package. (2019). R
1048 package version 1.0.15. <https://CRAN.R-project.org/package=doParallel>
- 1049 78. Microsoft, Weston, S.: Foreach: Provides Foreach Looping Construct. (2019). R package version 1.4.7.
1050 <https://CRAN.R-project.org/package=foreach>
- 1051 79. Bengtsson, H.: R.utils: Various Programming Utilities. (2019). R package version 2.9.2.
1052 <https://CRAN.R-project.org/package=R.utils>
- 1053 80. Gagolewski, M.: R Package Stringi: Character String Processing Facilities. (2019).
1054 <http://www.gagolewski.com/software/stringi/>
- 1055 81. Gelius-Dietrich, G.: glpkAPI: R Interface to C API of GLPK. (2018). R package version 1.3.1.
1056 <https://CRAN.R-project.org/package=glpkAPI>
- 1057 82. Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S.: Biostings: Efficient Manipulation of Biological Strings.
1058 (2019). R package version 2.54.0
- 1059 83. Bornstein, B.J., Keating, S.M., Jouraku, A., Hucka, M.: Libsbnl: an api library for sbml. *Bioinformatics* **24**(6),
1060 880–881 (2008). doi:[10.1093/bioinformatics/btn051](https://doi.org/10.1093/bioinformatics/btn051)
- 1061 84. Cook, G.M., Greening, C., Hards, K., Berney, M.: Chapter one - energetics of pathogenic bacteria and
1062 opportunities for drug development. In: Poole, R.K. (ed.) *Advances in Bacterial Pathogen Biology. Advances*
1063 *in Microbial Physiology*, vol. 65, pp. 1–62. Academic Press, ??? (2014). doi:[10.1016/bs.ampbs.2014.08.001](https://doi.org/10.1016/bs.ampbs.2014.08.001).
1064 <http://www.sciencedirect.com/science/article/pii/S0065291114000022>
- 1065 85. Goldberg, I., Rock, J., Ben-Bassat, A., Mateles, R.: Bacterial yields on methanol, methylamine, formaldehyde,
1066 and formate. *Biotechnology and bioengineering* **18**(12), 1657–1668 (1976)
- 1067 86. Oliphant, K., Allen-Vercoe, E.: Macronutrient metabolism by the human gut microbiome: major fermentation
1068 by-products and their impact on host health. *Microbiome* **7**, 91 (2019). doi:[10.1186/s40168-019-0704-8](https://doi.org/10.1186/s40168-019-0704-8)
- 1069 87. Ríos-Covián, D., Ruas-Madiedo, P., Margolles, A., Gueimonde, M., de Los Reyes-Gavilán, C.G., Salazar, N.:
1070 Intestinal short chain fatty acids and their link with diet and human health. *Front Microbiol* **7**, 185 (2016)
- 1071 88. Ziels, R.M., Nobu, M.K., Sousa, D.Z.: Elucidating syntrophic butyrate-degrading populations in anaerobic
1072 digesters using stable-isotope-informed genome-resolved metagenomics. *mSystems* **4**(4) (2019).
1073 doi:[10.1128/mSystems.00159-19](https://doi.org/10.1128/mSystems.00159-19)
- 1074 89. Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A.: metaSPAdes: a new versatile metagenomic
1075 assembler. *Genome Research* **27**(5), 824–834 (2017). doi:[10.1101/gr.213959.116](https://doi.org/10.1101/gr.213959.116)
- 1076 90. Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., Dehal, P., Ware, D.,
1077 Perez, F., Canon, S., Sneddon, M.W., Henderson, M.L., Riehl, W.J., Murphy-Olson, D., Chan, S.Y.,
1078 Kamimura, R.T., Kumari, S., Drake, M.M., Brettin, T.S., Glass, E.M., Chivian, D., Gunter, D., Weston, D.J.,
1079 Allen, B.H., Baumohl, J., Best, A.A., Bowen, B., Brenner, S.E., Bun, C.C., Chandonia, J.-M., Chia, J.-M.,
1080 Colasanti, R., Conrad, N., Davis, J.J., Davison, B.H., DeJongh, M., Devoid, S., Dietrich, E., Dubchak, I.,

- 1081 Edirisinghe, J.N., Fang, G., Faria, J.P., Frybarger, P.M., Gerlach, W., Gerstein, M., Greiner, A., Gurtowski, J.,
1082 Haun, H.L., He, F., Jain, R., Joachimiak, M.P., Keegan, K.P., Kondo, S., Kumar, V., Land, M.L., Meyer, F.,
1083 Mills, M., Novichkov, P.S., Oh, T., Olsen, G.J., Olson, R., Parrello, B., Pasternak, S., Pearson, E., Poon, S.S.,
1084 Price, G.A., Ramakrishnan, S., Ranjan, P., Ronald, P.C., Schatz, M.C., Seaver, S.M.D., Shukla, M., Sutormin,
1085 R.A., Syed, M.H., Thomason, J., Tintle, N.L., Wang, D., Xia, F., Yoo, H., Yoo, S., Yu, D.: KBase: The united
1086 states department of energy systems biology knowledgebase. *Nature Biotechnology* **36**(7), 566–569 (2018).
1087 doi:[10.1038/nbt.4163](https://doi.org/10.1038/nbt.4163)
- 1088 91. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N.: Shotgun metagenomics, from sampling to
1089 analysis. *Nature Biotechnology* **35**(9), 833–844 (2017). doi:[10.1038/nbt.3935](https://doi.org/10.1038/nbt.3935)
- 1090 92. Kim, W.J., Kim, H.U., Lee, S.Y.: Current state and applications of microbial genome-scale metabolic models.
1091 *Current Opinion in Systems Biology* **2**, 10–18 (2017). doi:[10.1016/j.coisb.2017.03.001](https://doi.org/10.1016/j.coisb.2017.03.001). Regulatory and
1092 metabolic networks • Cancer and systemic diseases
- 1093 93. Graspentner, S., Waschina, S., Künzel, S., Twisselmann, N., Rausch, T.K., Cloppenborg-Schmidt, K.,
1094 Zimmermann, J., Viemann, D., Herting, E., Göpel, W., Baines, J.F., Kaleta, C., Rupp, J., Härtel, C., Pagel,
1095 J.: Gut Dysbiosis With Bacilli Dominance and Accumulation of Fermentation Products Precedes Late-onset
1096 Sepsis in Preterm Infants. *Clinical Infectious Diseases* **69**(2), 268–277 (2018). doi:[10.1093/cid/ciy882](https://doi.org/10.1093/cid/ciy882).
1097 <https://academic.oup.com/cid/article-pdf/69/2/268/28893438/ciy882.pdf>
- 1098 94. Byndloss, M.X., Olsan, E.E., Rivera-Chávez, F., Tiffany, C.R., Cevallos, S.A., Lokken, K.L., Torres, T.P.,
1099 Byndloss, A.J., Faber, F., Gao, Y., *et al.*: Microbiota-activated ppar- γ signaling inhibits dysbiotic
1100 enterobacteriaceae expansion. *Science* **357**(6351), 570–575 (2017). doi:[10.1126/science.aam9949](https://doi.org/10.1126/science.aam9949)
- 1101 95. Smith, P.M., Howitt, M.R., Panikov, N., Michaud, M., Gallini, C.A., Bohlooly-y, M., Glickman, J.N., Garrett,
1102 W.S.: The microbial metabolites, short-chain fatty acids, regulate colonic treg cell homeostasis. *Science*
1103 **341**(6145), 569–573 (2013). doi:[10.1126/science.1241165](https://doi.org/10.1126/science.1241165)
- 1104 96. Pham, V.T., Lacroix, C., Braegger, C.P., Chassard, C.: Early colonization of functional groups of microbes in
1105 the infant gut. *Environmental Microbiology* **18**(7), 2246–2258 (2016). doi:[10.1111/1462-2920.13316](https://doi.org/10.1111/1462-2920.13316)
- 1106 97. García-Campos, M.A., Espinal-Enríquez, J., Hernández-Lemus, E.: Pathway analysis: state of the art. *Frontiers*
1107 *in physiology* **6**, 383 (2015). doi:[10.3389/fphys.2015.00383](https://doi.org/10.3389/fphys.2015.00383)
- 1108 98. Foster, K.R., Schluter, J., Coyte, K.Z., Rakoff-Nahoum, S.: The evolution of the host microbiome as an
1109 ecosystem on a leash. *Nature* **548**(7665), 43–51 (2017). doi:[10.1038/nature23292](https://doi.org/10.1038/nature23292)
- 1110 99. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L.L.: The Pfam Protein Families
1111 Database. *Nucleic Acids Research* **28**(1), 263–266 (2000). doi:[10.1093/nar/28.1.263](https://doi.org/10.1093/nar/28.1.263).
1112 <http://oup.prod.sis.lan/nar/article-pdf/28/1/263/9895152/280263.pdf>
- 1113 100. Galperin, M.Y., Koonin, E.V.: Sources of systematic error in functional annotation of genomes: domain
1114 rearrangement, non-orthologous gene displacement and operon disruption. *In silico biology* **1**(1), 55–67 (1998)
- 1115 101. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J.,
1116 Salazar, G.A., Smart, A., Sonnhammer, E.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C., Finn, R.D.:
1117 The Pfam protein families database in 2019. *Nucleic Acids Research* **47**(D1), 427–432 (2018).
1118 doi:[10.1093/nar/gky995](https://doi.org/10.1093/nar/gky995). <http://oup.prod.sis.lan/nar/article-pdf/47/D1/D427/27436497/gky995.pdf>
- 1119 102. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R.,
1120 Letunic, I., Rattei, T., Jensen, L.J., von Mering, C., Bork, P.: eggNOG 5.0: a hierarchical, functionally and
1121 phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*
1122 **47**, 309–314 (2019). doi:[10.1093/nar/gky1085](https://doi.org/10.1093/nar/gky1085)
- 1123 103. Douglas, G.M., Maffei, V.J., Zaneveld, J., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., Langille,
1124 M.G.I.: PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv* (2019).
1125 doi:[10.1101/672295](https://doi.org/10.1101/672295)

1126 **Additional Files**

1127 Table S1 — New reactions and metabolites added to biochemistry database.
1128 see file: [Table_S1.xlsx](#)

1129 Table S2 — Organisms included in fermentation product validation test.
1130 see file: [Table_S2.xlsx](#)

1131 Table S4 — References for substance production and consumption in anaerobic gut communities (see
1132 Supplementary Figure S1).
1133 see file: [Table_S4.ods](#)

1134 Table S5 — 127 Species-level genome bins (SGBs) from Pasolli *et al.*, 2019 [69] and 1270 mapped reference
1135 genome assembles from RefSeq.
1136 see file: [Table_S5.ods](#)

Table S3 — Organisms used in modelling of the anaerobic food web of the human gut microbiome.

RefSeq Assembly	Organism name	Reconstruction method
GCF_000173975.1	<i>Anaerobutyricum hallii</i> DSM 3353	gapseq / modelseed / carveme
GCF_000025985.1	<i>Bacteroides fragilis</i> NCTC 9343	gapseq / modelseed / carveme
GCF_001314975.1	<i>Bacteroides thetaiotaomicron</i>	gapseq / modelseed / carveme
GCF_000196555.1	<i>Bifidobacterium longum</i> subsp. <i>longum</i> JCM 1217	gapseq / modelseed / carveme
GCF_000157975.1	<i>Blautia hydrogenotrophica</i> DSM 10507	gapseq / modelseed / carveme
GCF_000013285.1	<i>Clostridium perfringens</i> ATCC 13124	gapseq / modelseed / carveme
GCF_003434235.1	<i>Coprococcus catus</i>	gapseq / modelseed / carveme
GCF_000155875.1	<i>Coprococcus comes</i> ATCC 27758	gapseq / modelseed / carveme
GCF_000154425.1	<i>Coprococcus eutactus</i> ATCC 27759	gapseq / modelseed / carveme
GCF_000189295.2	<i>Desulfovibrio desulfuricans</i> ND132	gapseq / modelseed / carveme
GCF_000391485.2	<i>Enterococcus faecalis</i> EnGen0107	gapseq / modelseed / carveme
GCF_000005845.2	<i>Escherichia coli</i> str. <i>K-12</i> substr. <i>MG1655</i>	gapseq / modelseed / carveme
GCF_000162015.1	<i>Faecalibacterium prausnitzii</i> A2-165	gapseq / modelseed / carveme
GCF_003047065.1	<i>Lactobacillus acidophilus</i>	gapseq / modelseed / carveme
GCF_001304715.1	<i>Megasphaera elsdenii</i> 14-14	gapseq / modelseed / carveme
GCF_000195895.1	<i>Methanosarcina barkeri</i> str. <i>Fusaro</i>	manually curated (BiGG-ID: iAF692)[66]
GCF_000144405.1	<i>Prevotella melaninogenica</i> ATCC 25845	gapseq / modelseed / carveme
GCF_900101355.1	<i>Ruminococcus bromii</i>	gapseq / modelseed / carveme
GCF_000006945.2	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhimurium</i> str. <i>LT2</i>	gapseq / modelseed / carveme
GCF_900637515.1	<i>Veillonella dispar</i>	gapseq / modelseed / carveme

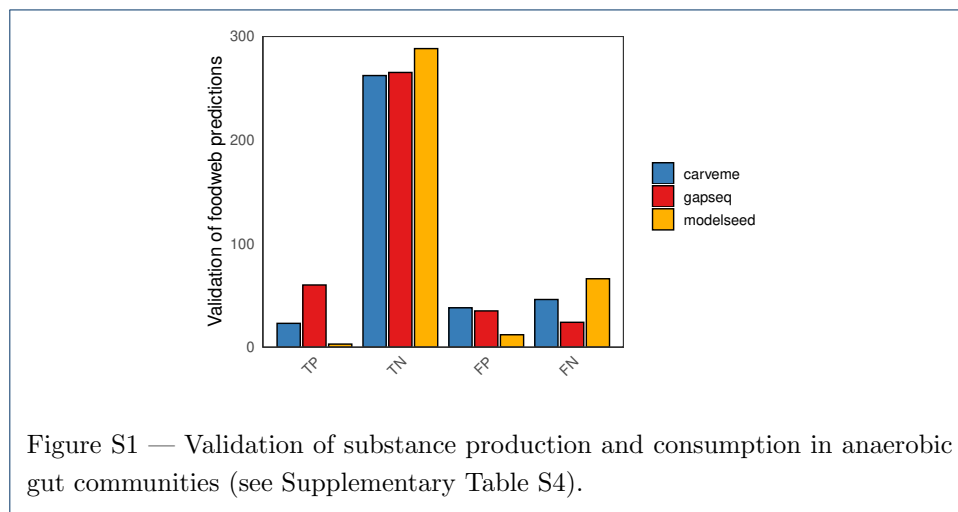


Figure S1 — Validation of substance production and consumption in anaerobic gut communities (see Supplementary Table S4).

