PrFBA: Probabilistic Flux Balance Analysis modeling for enhanced coverage of MAG, ASV, and clade-level metabolic models

May 7, 2024

Andrew Freiburger
afreiburger@anl.gov
Northwestern University

Filipe Liu
fliu@anl.gov
Argonne National Laboratory

Kathleen Beilsmith
kbeilsmith@anl.gov
Argonne National Laboratory

Keith Tyo
k-tyo@northwestern.edu
Northwestern University

Chris Henry
chenry@anl.gov
Argonne National Laboratory

Abstract

This is the abstract.

1. Introduction

Microbial communities are ubiquitous [1] and serve fundamental roles as eukaryotic symbionts [2], [3], [4] and pathogens [5], [6], industrial assets [7], [8] and antagonists [9], and ecological agents of biogeochemical cycling [10], [11], [12] and climate regulation [13], [14], [15]. Microbial communities are therefore vital for research fields as diverse as medicine, basic and synthetic biology [16], materials science, and civil engineering. The principles and member interactions that manifest in the stable abundances and community behaviors, however, remain unknown [17], which hinders research in the relevant ecologies in each of these domains [18]. Understanding a microbiome requires knowledge of both which biological functions are occurring within it, to reveal the ecological role of the community, and how those functions are delegated amongst its members, to predict responses of its functions to changing community structure (e.g. a new species replacing an old species) and responses of member interactions to changing environmental conditions.

The functional distribution in a stable microbiome across its members, while difficult to experimentally interrogate, can be elucidated computationally through metabolic modeling [19]. This process, however, requires sequenced genomes of each community member, which is not feasible for particularly natural microbial communities where many members are unculturable. Shotgun metagenome sequencing followed by assembly and binning is often used to genomically capture microbiome members [20], which yields partial and often contaminated metagenome assembled genomes (MAGs) [21]. Alternative approaches such as direct annotation and long-read sequencing [22] are impeded respectively by short reads and no chromosomal context that create unreliable annotations and low read counts. MAGs are often sufficient to obtain a reliable phylogenetic placement in genomic space using average nucleotide identity (ANI) approaches like GTDBtk [23], but they are nearly never complete [24] and the proper sequencing depth for characterizing a community remains an open question [25]. This impairs the reliability of MAGs for characterizing microbiome functions; nevertheless, MAGs are frequently used for this purpose [26]. MAGs have several other limitations for microbiome research, including bias towards reconstructing only the most abundant members into even low-quality MAGs, and deep shotgun sequencing

of microbiome samples is prohibitively both costly for a single experiment (e.g. large temporal sequences of data or exhaustive spatial variation) and for demanding a lot of extracted DNA.

Amplicon sequencing addresses has a complementary set of strengths and weaknesses as the aforementioned shotgun-sequencing-based approaches. Amplicon sequencing amplify a marker gene (typically 16s for microbes) from surrounding DNA and thereby advantageously capture more species from the same quantity of DNA [27]. Amplicons can then be clustered into exact amplicon sequence variants (ASVs) and annotated using taxonomical classifiers [28], which are typically coarser than ANI-based approaches for even a low-quality MAG. ASVs provide no direct information of organismal functions, but can infer functions from those of the reference genomes in the taxa to which the amplicon sequence was mapped [29]. This method exchanges a broader species capture from a greater species uncertainty relative to MAGs analysis.

A final approach to understand functional delegation within a complex microbiome evaluates higher-order clades instead of inferring strain-level resolution from MAGs and ASVs, which may be preferred for ecosystems with key-stone clades instead of examining redundant roles among several strains [30], [31]. A clade-level study of a microbiome, at the investigator's specified taxonomic level, pools metabolic functions of the clade's organisms or reference genomes, at the expense of potentially creating monolithic models from functionally diverse clades.

All of the aforementioned methods for studying functional delegation within a microbiome require a model that captures all of the functional diversity while compensating uncertainties from incompleteness and contamination of MAGs; inferred functions from reference genomes close to ASVs; and underdetermined clade models. Genome-scale metabolic models (GEMs) [32], [33], [34] are a valuable framework for encapsulating metabolic functions and pathways in a linear optimization problem that can be tailored through constraints and objective functions to recapitulate cellular behavior in a given nutritional environment. Flux balance analysis (FBA) [35], [36] is a prominent method of simulating GEMs through linear optimization towards a prescribed objective function, which by default is cellular growth to capture the baseline assumption of cells striving for personal growth [37]. Community metabolic models [38] can be created by combining member GEMs into a community GEM and simulating its metabolic behavior through FBA just like individual GEM members.

We propose a new probabilistic approach to develop GEMs and simulate FBA that captures the uncertainty of MAGs, ASVs, or clade-level assessments of microbiomes, which we collectively call Probabilistic Flux Balance Analysis (prFBA). Several methods have applied probability in reconstructing GEMs [39], predicting microbiome structure from chemical data [40], or merging regulatory networks into GEMs [41]; however, there is no GEM method that captures the functional uncertainty from MAG or ASV sequence approximations or the functional diversity of a phylogenetic clade, nor moreover simulates metabolism according to these uncertainties. prFBA favors the most conserved functions among the organisms or reference genomes according to the functional frequencies, or probabilities, of each function. prFBA is available in the ModelSEED Python API and is being integrated into KBase Applications for more reproducible and accessible modeling pipelines. The probabilistic modeling framework presented here – both the construction of probabilistic models and the probabilistic simulation of GEMs – will accelerate microbial ecology and offers a unique method for accommodating uncertainty in MAGs and ASV, and uniquely facilitate clade-level microbial ecology, for broad applications in microbiome research.

2. Methods

2.1. Mapping genome or MAG/ASV in phylogenetic space

Our method begins by localizing an organismal genome or MAG/ASV sequence, which is localized within phylogenetic space of a reference database or MAG collection. We used the Genome Taxonomical Database (GTDB) [42] because it covers vast phylogenetic space and has a convenient tool-kit API [23] that is integrated into KBase for rapid and reproducible computations [43]. The closest reference genomes for MAG and ASV sequences are determined as all of those that are within a threshold of 90% according to the average nucleotide identity (ANI) similarity.

2.2. Constructing a probabilistic model

The creation of probabilistic models from a collection of organisms is performed by determining the frequency of a given function from ideally the annotations or genes because avoids reconstructing metabolic models for all members and only reconstructs a metabolic model from the union of functions among the members, although examining the reactions could also be used as proxies for functions. It is pivotal to conserve core metabolism The mapping of reactions, by contrast, requires each organism to be reconstructed and then the probabilistic model is created from the union of organism reactions with their respective frequencies. The probability for a given function object

$$p_{\text{func}} = \frac{\text{num}_{\text{organisms,func}}}{\text{num}_{\text{organisms}}} \tag{1}$$

is determined as its frequency of the among the set of reference organisms, and is stored as an attribute of each respective reaction. The probability defined in **Equation 1** can be further refined for ASV and MAG systems by applying a secondary weighting α that favor organisms in the pool of MAG or ASV organisms who are closer to the reference organism. The α_i of organism i

$$\alpha_i = \frac{(\text{num}_{\text{point,func}} \cap \text{num}_{\text{i,func}}) * \text{completeness}}{\text{num}_{\text{i,func}}} \tag{2}$$

is the fraction of functions that are shared between the MAG or ASV point and the phylogenetically close reference organism i, according to ANI distance, multiplied by the completeness of the MAG or ASV. The $0 < \alpha < 1$ is then used to compute the $0 < p_{\text{func}} < 1$ function probability

$$p_{\text{func}} = \frac{\sum_i^I \left(b_{\text{func,i}} * \alpha_i\right)}{\sum_i^I (\alpha_i)} \tag{3}$$

for all genomes i that are close to the examined org, where $b_{\text{func,i}}$ is a binary variable that indicates whether the function is (1) or is not (0) in the examined genome i.

2.3. Simulating probabilistic FBA (prFBA)

The prFBA method is defined by several constraints, simulations, and a unique objective function. One of the constraints ("ElementUptake") limits carbon consumption to a specified $\text{ele}_{\text{limit}}$ number of atoms

$$\sum_{\text{ex}}^{\text{EX}} (\text{ele}_{\text{ex}} * (\text{ex}_{\text{forwardVar}} \oplus \text{ex}_{\text{reverseVar}})) = \text{ele}_{\text{limit}} \tag{4}$$

which imposes efficient utilization of resources from a given media, where complete media is the default environment. This constraint requires the atom count for each of the exchanged metabolites ele_{ex}, which

is then multipled by either the forward or reverse flux variables $\mathrm{ex}_{\mathrm{forwardVar}} \oplus \mathrm{ex}_{\mathrm{ReverseVar}}$ when calculating the impact of the variable on the total element balance. A second constraint ("CommKinetics"), limits the total reaction flux of each model to a kinCoef multiple of its biomass growth

$$\sum_{r}^{R}(r_{\mathrm{forwardVar}} + r_{\mathrm{reverseVar}}) = \mathrm{kinCoef} * \mathrm{bio}_{\mathrm{forwardVar}} \tag{5}$$

which effectively prevents organisms from being metabolically exploited for efficient reactions or pathways without themselves growing. The final constraint ("min_biomass") simulates the model with the default objective function of maximizing biomass growth and mandates that the biomass flux is at least 95% of its maximum value

$$\mathrm{cons} : \mathrm{bio1}_{\mathrm{flux}} > 0.95 * \mathrm{bio1}_{\mathrm{flux,\ max}} \tag{6}$$

to maintain optimum cellular growth despite later replacing the objective function.

The prFBA objective function

$$\begin{aligned} \min\{r, \mathrm{ex}\} : \sum_{r}^{R_{\mathrm{internal}}} \left(\left(1 - p^{\mathrm{prob\_exp}}\right) * r_{\mathrm{flux}}\right) + \mathrm{min\_prob}) + \\ \sum_{\mathrm{rd}}^{\mathrm{RD}}(-\mathrm{rd}_{\mathrm{expression}} * \mathrm{rd}_{\mathrm{flux}})) + \sum_{\mathrm{ex}}^{\mathrm{EX}}(100 * \mathrm{ex}_{\mathrm{flux}}) \end{aligned} \tag{7}$$

finds the prototypical metabolic activity of a MAG, clade, or ASV through three sums. The first sum penalizes intracellular reactions, for which data does not exist, inversely with their probability by scaling their flux r_{flux}. The min_prob parameter establishes the smallest probability that is assigned to a reaction, which prevents some reactions from becoming penalized into oblivion. The prob_exp parameter tailors the significance p_{object} in determining the reaction weightings. The second sum preferentially rewards internal reactions for which expression data exists (rd) proportionally with the relative expression of each respective reaction ($\mathrm{rd}_{\mathrm{expression}}$). The third sum penalizes exchanges, which maximizes transport and thus interactions among community members. This objective further mitigates degenerate solutions because alternative reaction pathways will nearly always have different weightings and thus the minimization will just use the most probable pathway.

2.4. Data integration

Meta-transcriptomics data can be invaluable for this method by yielding a probability of 1 for genes that are expressed in the data, because the uncertainty in gene expression that is captured by p_{object} disappears.

3. Results

3.1. New Probabilistic Flux Balance Analysis Approach to GEM Reconstruction and Analysis

Our workflow illustrated in **Figure 1** and analyzes GEMs for individual strains with an imcomplete sequence (MAG or ASV) or taxa with exemplar genomes, as a compliment to ModelSEED2 [44] or CarveME [45] and COBRApy [46] that build and analyze GEMs from complete high-quality genome sequences, respectively. Our approach reconstructs strain or clade probabilistic models by mapping a MAG, ASV, or taxonomical class into phylogenetic space (see **Section 2.1**). The phylogenetically closest reference

genomes or other MAGs are then identified, preferably being with genomes of >0.95 ANI score but ultimately including the closest representative genomes because MAGs and ASVs may not have very close reference genomes. Fortunately, our proposed method is somewhat flexible to the distance of the reference genomes, where the uncertainty is likely greater for each of the models.

All nearby genomes are then annotated with a model-friendly functional ontology such as RAST, GO, and KEGGKO. We employ RAST because it is compatible with our ModelSEED2 reconstruction pipeline and because we did not focus on annotation uncertainty in this work, but we strongly recommend aggregating the genome annotations from through several methods since functional annotation significantly contributes to uncertainty in these models. The annotated genomes of the selected genomes are then merged into a consolidated probabilitic annotation (CPA, see **Section 2.2**) that includes all unique functions among the selected genomes and is each assigned a probability based on two factors: (1) how prevalent the function is amongst the selected genomes included in our set; and (2) how distant each genome containing the function is from the MAGs and ASV sequence. Our approach also applies to existing metabolic models of reference genomes by merging their biochemical reactions as a proxy of functions instead of the original annotations, which is demonstrated by representing a gut microbiome with the AGORA2 model collection.

A reconstruction method is next applied based on the CPA, or the CPA becomes the GEM itself when merging existing metabolic models. We used the ModelSEED2 method to build our CPA model for the MAG and clade systems, and ascribed our function-level probabilities to all GEM reactions. CPA-constructed GEMS can be simulated individually or aggregated into a compartmentalized community metabolic model to probabilistically study a microbiome system where each probabilistic model is constrained based on species abundance of the microbiome sample. Our example here all study microbiomes through CPA-constructed community metabolic models as the more complex use-case of probabilistic modeling.

Finally, flux balance analysis with numerous potential constraints and modeling modalities is employed to predict metabolic activity and fluxes that maximally use the highest probability model reactions, which captures as much top-down knowledge of the system behavior such as gene/protein expression, species abundance, global flux measures, metabolomics profiles, QSIP profiles, and TN-seq data. Low probability reactions will generally be avoided, but could still be used if they are essential for explaining systems-level experimental observations, which allows CPA-constructed models, compartmentalized community models of these individual models, are able to resolve uncertainties based on systems level observations.

4. Conclusion

Probabilistic GEM modeling, introduced herein, should be powerful tool for microbial ecology. A MAG from an unstudied soil sample, for example, could map to the Psueodmonad genus but an insufficient amount of genetic sequence is available to map to a species. Our method would be able to assemble a genome-scale metabolic model that pulls reactions from the closest reference genomes and would exercise the most conserved behavior of the phylogenetic space while implementing possible niche behavior as environmental constraints necessitate.

Our method further uniquely permits an analysis of prototypical clade behavior and specifically how multiple clades interact with each other. This introduces a new paradigm where the general roles of microbes can be contextualized in otherwise excessively complex ecological systems.

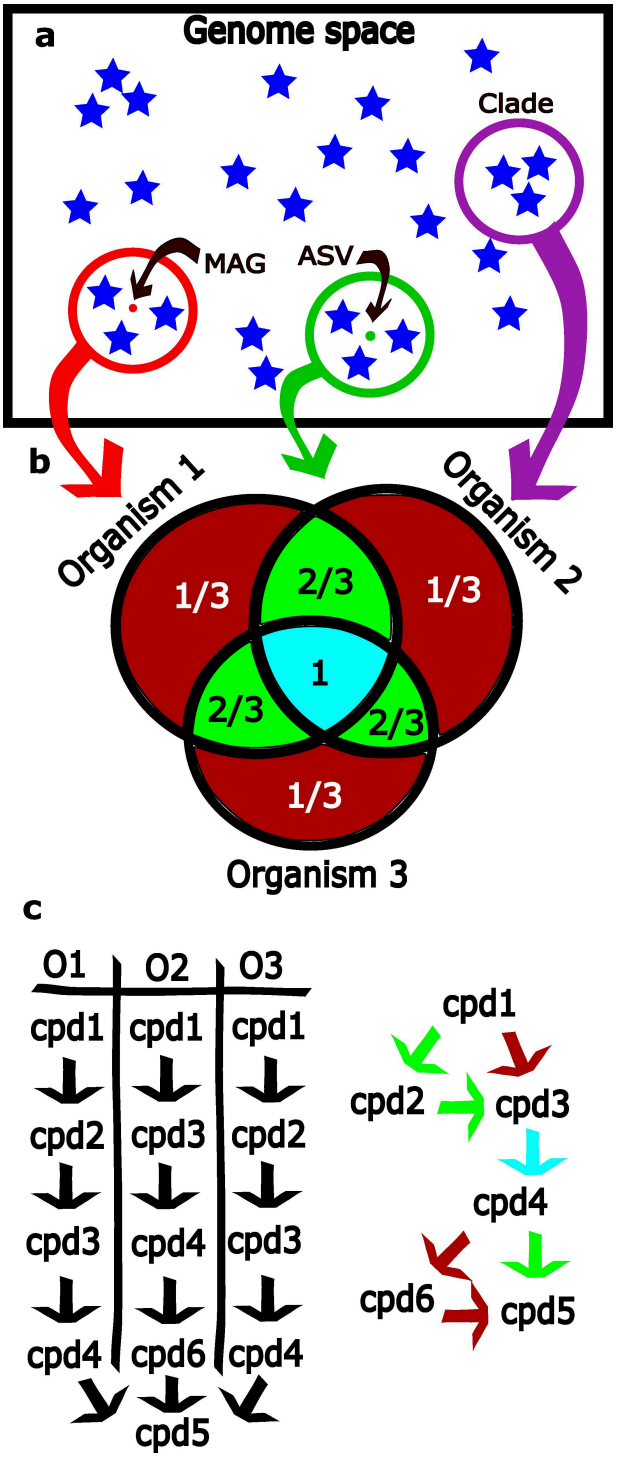

Figure 1: Panel a denotes the genome space defined by reference genomes and either a MAG, ASV, or clade of interest. The MAG and ASV sequences are mapped into this space based on average nucleotide identity (ANI). The phylogenetically close reference genomes and (other) MAGs can then be used for gapfilling and model reconstruction by merging the annotations. The development of a clade model simply assigns a cut-off at a taxonomical level and involves merging the reactions per se from the set of all models.

Panel b depicts three reference organisms that are either phylogenetically close to a MAG/ASV or are in a common clade, and are merged into a single model that captures metabolic uncertainty and diversity from the closest reference organisms by assigning a probability to each reaction according to its frequency among the reference organisms. Reactions that are already annotated to the MAG or are empirically detected are overwritten with a probability of 1. The red sections of the model Venn diagram are the reactions that are present in only one of the three models; the green sections are the reactions that are present in two of the three models; and the teal center are the reactions present in all of the models.

Panel c illustrates a sample PrFBA simulation of a probabilistic model. Example alternative pathways from the 3 reference models that convert compound 1 (cpd1) into compound 5 (cpd5) in the left columns are examined probabilistically in the right diagram, where the reaction arrows are painted with the respective probabilities of each reaction according to the color schema from panel b. PrFBA would preferentially follow the pathway cpd1 $\Rightarrow$ cpd2 $\Rightarrow$ cpd3 $\Rightarrow$ cpd4 $\Rightarrow$ cpd5 because this is the route that employs the most probable reactions, unless additional constraints incentivize the use of less probable reactions. This probabilistic framework allows a broad assessment of MAG, ASV, or clade behavior in a single-strain GEM. The probabilistic model can also model sample-level systems, where all possible hypotheses of functional content in a MAG are evaluated to discern how biological activity is delegated across a microbial community.

Bibliography

[1] S. Nayfach *et al.*, "A genomic catalog of Earth's microbiomes," *Nature Biotechnology*, vol. 39, no. 4, pp. 499–509, 2021, doi: 10.1038/s41587-020-0718-6.

[2] S. Shoaie *et al.*, "Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome," *Cell Metabolism*, vol. 22, no. 2, pp. 320–331, 2015, doi: 10.1016/j.cmet.2015.07.001.

[3] M. Kumar, B. Ji, K. Zengler, and J. Nielsen, "Modelling approaches for studying the microbiome," *Nature Microbiology*, vol. 4, no. 8, pp. 1253–1267, 2019, doi: 10.1038/s41564-019-0491-9.

[4] S. Shoaie, F. Karlsson, A. Mardinoglu, I. Nookaew, S. Bordel, and J. Nielsen, "Understanding the interactions between bacteria in the human gut through metabolic modeling," *Scientific Reports*, vol. 3, pp. 1–10, 2013, doi: 10.1038/srep02532.

[5] L. Mcguigan and M. Callaghan, "The evolving dynamics of the microbial community in the cystic fibrosis lung," *Environmental Microbiology*, vol. 17, no. 1, pp. 16–28, 2015, doi: 10.1111/1462-2920.12504.

[6] H. F. Jenkinson and R. J. Lamont, "Oral microbial communities in sickness and in health," *Trends in Microbiology*, vol. 13, no. 12, pp. 589–595, 2005, doi: 10.1016/j.tim.2005.09.006.

[7] E. Tzamali, P. Poirazi, I. G. Tollis, and M. Reczko, "A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities," *BMC Systems Biology*, vol. 5, 2011, doi: 10.1186/1752-0509-5-167.

[8] R. K. Pettit, "Mixed fermentation for natural product drug discovery," *Applied Microbiology and Biotechnology*, vol. 83, no. 1, pp. 19–25, 2009, doi: 10.1007/s00253-009-1916-9.

[9] E. T. Hillman, L. E. Caceres-Martinez, G. Kilaz, and K. V. Solomon, "Top-down enrichment of oil field microbiomes to limit souring and control oil composition during extraction operations," *AIChE Journal*, no. September, pp. 1–13, 2022, doi: 10.1002/aic.17927.

[10] B. S. Griffiths and L. Philippot, "Insights into the resistance and resilience of the soil microbial community," *FEMS Microbiology Reviews*, vol. 37, no. 2, pp. 112–129, 2013, doi: 10.1111/j.1574-6976.2012.00343.x.

[11] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, "Prokaryotes: The unseen majority," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 12, pp. 6578–6583, 1998, doi: 10.1073/pnas.95.12.6578.

[12] A. E. Dukas, R. S. Poretsky, and V. J. Orphan, "Deep-Sea Archaea Fix and Share Nitrogen in Methane-Consuming Microbial Consortia," *Science*, vol. 326, no. October, pp. 422–427, 2009.

[13] R. Cavicchioli *et al.*, "Scientists' warning to humanity: microorganisms and climate change," *Nature Reviews Microbiology*, vol. 17, no. 9, pp. 569–586, 2019, doi: 10.1038/s41579-019-0222-5.

[14] V. J. Orphan, C. H. House, K. U. Hinrichs, K. D. McKeegan, and E. F. DeLong, "Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis," *Science*, vol. 293, no. 5529, pp. 484–487, 2001, doi: 10.1126/science.1061338.

[15] J. B. Glass and V. J. Orphan, "Trace metal requirements for microbial enzymes involved in the production and consumption of methane and nitrous oxide," *Frontiers in Microbiology*, vol. 3, no. FEB, pp. 1–20, 2012, doi: 10.3389/fmicb.2012.00061.

[16] A. R. Zomorrodi and D. Segrè, "Synthetic Ecology of Microbes: Mathematical Models and Applications," vol. 428. Academic Press, pp. 837–861, 2016. doi: 10.1016/j.jmb.2015.10.019.

[17] A. F. White *et al.*, "A unified initiative to harness Earth's microbiomes," *Science*, vol. 315, no. 6260, pp. 2015–2017, 2015.

[18] J. I. Prosser *et al.*, "The role of ecological theory in microbial ecology," *Nature Reviews Microbiology*, vol. 5, no. May, pp. 384–392, 2007.

[19] A. Varma and B. O. Palsson, "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110," *Applied and Environmental Microbiology*, vol. 60, no. 10, pp. 3724–3731, 1994, doi: 10.1128/aem.60.10.3724-3731.1994.

[20] L. X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, and J. F. Banfield, "Accurate and complete genomes from metagenomes," vol. 30. Cold Spring Harbor Laboratory Press, pp. 315–333, 2020. doi: 10.1101/gr.258640.119.

[21] D. Chivian *et al.*, "Metagenome-assembled genome extraction and analysis from microbiomes using KBase," *Nature Protocols*, vol. 18, no. 1, pp. 208–238, 2023, doi: 10.1038/s41596-022-00747-x.

[22] B. J. Callahan, D. Grinevich, S. Thakur, M. A. Balamotis, and T. B. Yehezkel, "Ultra-accurate microbial amplicon sequencing with synthetic long reads," *Microbiome*, vol. 9, no. 1, 2021, doi: 10.1186/s40168-021-01072-3.

[23] P. A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk v2: memory friendly classification with the genome taxonomy database," *Bioinformatics (Oxford, England)*, vol. 38, no. 23, pp. 5315–5316, 2022, doi: 10.1093/bioinformatics/btac672.

[24] R. Eisenhofer, I. Odriozola, and A. Alberdi, "Impact of microbial genome completeness on functional metagenomics," *ResearchSquare*, 2022, doi: 10.21203/rs.3.rs-1955526/v1.

[25] J. Liu, X. Wang, H. Xie, Q. Zhong, and Y. Xia, "Analysis and evaluation of different sequencing depths from 5 to 20 million reads in shotgun metagenomic sequencing, with optimal minimum depth being recommended," *Genome*, vol. 65, no. 9, pp. 491–504, 2022, doi: 10.1139/gen-2021-0120.

[26] D. K. K. Baskaran, S. Umale, Z. Zhou, K. Raman, and K. Anantharaman, "Metagenome-based metabolic modelling predicts unique microbial interactions in deep-sea hydrothermal plume microbiomes," *ISME Communications*, vol. 3, no. 1, p. 42–43, 2023, doi: 10.1038/s43705-023-00242-8.

[27] M. H. Meek and W. A. Larson, "The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era," *Molecular Ecology Resources*, 2019, doi: 10.1111/1755-0998.12998.

[28] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," *Nature Methods*, vol. 13, no. 7, pp. 581–583, 2016, doi: 10.1038/nmeth.3869.

[29] P. Yilmaz, P. Yarza, J. Z. Rapp, and F. O. Glöckner, "Expanding the world of marine bacterial and archaeal clades," *Frontiers in Microbiology*, vol. 6, no. JAN, 2016, doi: 10.3389/fmicb.2015.01524.

[30] E. A. Mcdaniel *et al.*, "Metabolic Differentiation of Co-occurring Accumulibacter Clades Revealed through Genome-Resolved Metatranscriptomics." [Online]. Available: https://journals.asm.org/journal/msystems

[31] S. Fodelianakis *et al.*, "Microdiversity characterizes prevalent phylogenetic clades in the glacier-fed stream microbiome," *ISME Journal*, vol. 16, no. 3, pp. 666–675, 2022, doi: 10.1038/s41396-021-01106-6.

[32] J. S. Edwards and B. O. Palsson, "The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 10, pp. 5528–5533, 2000, doi: 10.1073/pnas.97.10.5528.

[33] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from large-scale biology," *Science*, vol. 300, no. 5617, pp. 286–290, 2003, doi: 10.1126/science.1084564.

[34] M. W. Covert *et al.*, "Metabolic modeling of microbial strains in silico," *Trends in Biochemical Sciences*, vol. 26, no. 3, pp. 179–186, 2001, doi: 10.1016/S0968-0004(00)01754-0.

[35] J. D. Orth, I. Thiele, and B. O. Palsson, "What is flux balance analysis?," *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, 2010, doi: 10.1038/nbt.1614.

[36] J. M. Lee, E. P. Gianchandani, and J. A. Papin, "Flux balance analysis in the era of metabolomics," *Briefings in Bioinformatics*, vol. 7, no. 2, pp. 140–150, 2006, doi: 10.1093/bib/bbl007.

[37] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data," *Nature Biotechnology*, vol. 19, no. 2, pp. 125–130, 2001, doi: 10.1038/84379.

[38] S. H. J. Chan, M. N. Simons, and C. D. Maranas, "SteadyCom: Predicting microbial abundances while ensuring community stability," *PLoS Computational Biology*, vol. 13, no. 5, pp. 1–25, 2017, doi: 10.1371/journal.pcbi.1005539.

[39] Q. Qi, J. Li, and J. Cheng, "Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods," *BMC Proceedings*, vol. 8, 2014, doi: 10.1186/1753-6561-8-S6-S5.

[40] D. Eveillard, N. J. Bouskill, D. Vintache, J. Gras, B. B. Ward, and J. Bourdon, "Probabilistic modeling of microbial metabolic networks for integrating partial quantitative knowledge within the nitrogen cycle," *Frontiers in Microbiology*, vol. 10, no. JAN, 2019, doi: 10.3389/fmicb.2018.03298.

[41] H. Yu and R. H. Blair, "Integration of probabilistic regulatory networks into constraint-based models of metabolism with applications to Alzheimer's disease," *BMC Bioinformatics*, vol. 20, no. 1, 2019, doi: 10.1186/s12859-019-2872-8.

[42] D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P. A. Chaumeil, and P. Hugenholtz, "GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy," *Nucleic Acids Research*, vol. 50, no. D1, p. D785–D794, 2022, doi: 10.1093/nar/gkab776.

[43] A. P. Arkin *et al.*, "KBase: The United States department of energy systems biology knowledgebase," *Nature Biotechnology*, vol. 36, no. 7, pp. 566–569, 2018, doi: 10.1038/nbt.4163.

[44] J. P. Faria *et al.*, "ModelSEED v2: High-throughput genome-scale metabolic model reconstruction with enhanced energy biosynthesis pathway prediction," *BioRxiv*, 2023, doi: 10.1101/2023.10.04.556561.

[45] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities," *Nucleic Acids Research*, vol. 46, no. 15, pp. 7542–7553, 2018, doi: 10.1093/nar/gky537.

[46] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "COBRApy: COnstraints-Based Reconstruction and Analysis for Python," *BMC Systems Biology*, vol. 7, 2013, doi: 10.1186/1752-0509-7-74.