Collaboratively Assembling a Toolkit in KBase to Leverage Probabilistic Annotation and Multi-omics Data to Enable Metabolic Modeling of Microbial Community Ecology

José P. Faria[1](jplfaria@anl.gov), Filipe Liu[1], Andrew P. Freiburger[1], Mikayla Borton[9], Kelly Wrighton[9], Patrik D'haeseleer[2], Jeff Kimbrel[2], Jeremy Jacobson[3], Bill Nelson[3], Jason McDermott[3], Aimee K. Kessell[4], Hugh C. McCullough[4], Hyun-Seob Song[4], Janaka N. Edirisinghe[1], Nidhi Gupta[1], Samuel M.D. Seaver[1], Qizhi Zhang[1], Pamela Weisenhorn[1], Neal Conrad[1], Raphy Zarecki[5], Matthew DeJongh[5], Aaron A. Best[5], KBase Team[1,6,7,8], Robert W. Cottingham[6], Adam P. Arkin[7], Rhona Stuart[2], Kirsten Hofmockel[3], and Christopher S. Henry[1]

[1]Argonne National Laboratory, Lemont, IL; [2]Lawrence Livermore National Laboratory, Livermore, CA; [3]Pacific Northwest National Laboratory, Richland WA; [4]University of Nebraska–Lincoln, Lincoln, NE; [5]Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel; [5]Hope College, Holland, MI; [6]Oak Ridge National Laboratory, Oak Ridge, TN; [7]Lawrence Berkeley National Laboratory, Berkeley, CA; [8]Brookhaven National Laboratory, Upton, NY; [9]Colorado State University, Fort Collins, CO

KBase PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

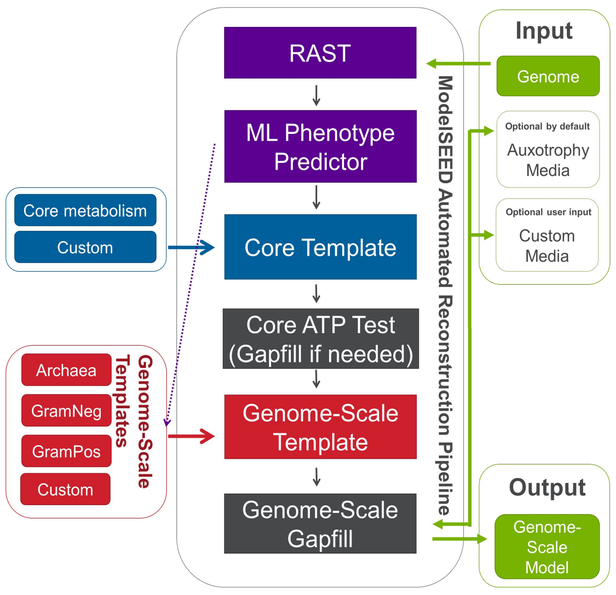
kbase.us – engage@kbase.us – @DOEKBase

Abstract:

Mechanistic understanding of biological systems relies on accurate protein annotations, which are often uncertain and error-prone. Genome-scale metabolic models (GEMs) enable evaluation of these annotations within their biological context, offering a means to refine them by considering experimental observations. KBase has developed a set of tools for this purpose, including protein sequence annotation and support for external annotations. The novel ModelSEED2 (MS2) tool enhances GEM construction with improved energy metabolism representation and pathway curation, leading to more comprehensive models. From probabilistic protein annotations, ensemble modeling approaches generate multiple GEM drafts which are evaluated for ATP biosynthesis, necessary gap-filling, and omics data congruence. The best models are further analyzed, with gap-filling algorithms like OMEGGA selecting annotations that align with experimental data. This collaborative effort across KBase, µBiospheres SFA, and PNNL Soil SFA demonstrates improved GEM pathway completeness and annotation accuracy through applications to diverse species and datasets, showcasing the system's ability to refine our understanding of metabolic functions across organisms.

Improvements to the GEM reconstruction pipeline, templates, and KBase apps:

MS2 genome-scale metabolic reconstruction pipeline enabling quantitative prediction of ATP production



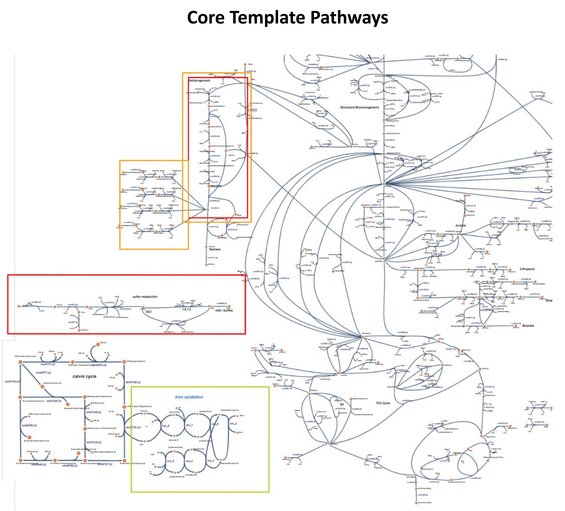
A genome annotated with RAST is inputted. Users may choose a reconstruction template, or ML classifiers can select one. ATP production is tested in 54 media, representing various energy biosynthesis strategies, with gap-filling as needed. The core metabolism model is then expanded to genome-scale.

Improvements in energy biosynthesis pathway reconstruction based on community-driven collaborative curation

ANME and SRB, **Victoria Orphan**

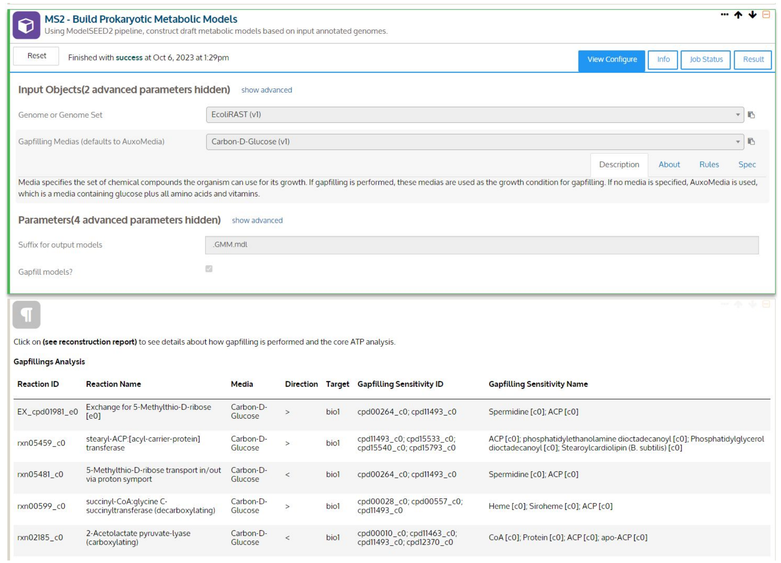
Methanogenesis and Methyltrophy, JGI **Susannah Tringe**

Iron Oxidation, University of Delaware **Clara Chan**



Many pathways of interest for DOE researchers are still poorly represented in public databases. Working with experts, we expanded our templates to model metabolisms of anaerobic methanotrophic archaea (ANME) and sulfate reducing bacteria (SRB) as well as pathways in methanogenesis, methyltrophy, and iron oxidation.

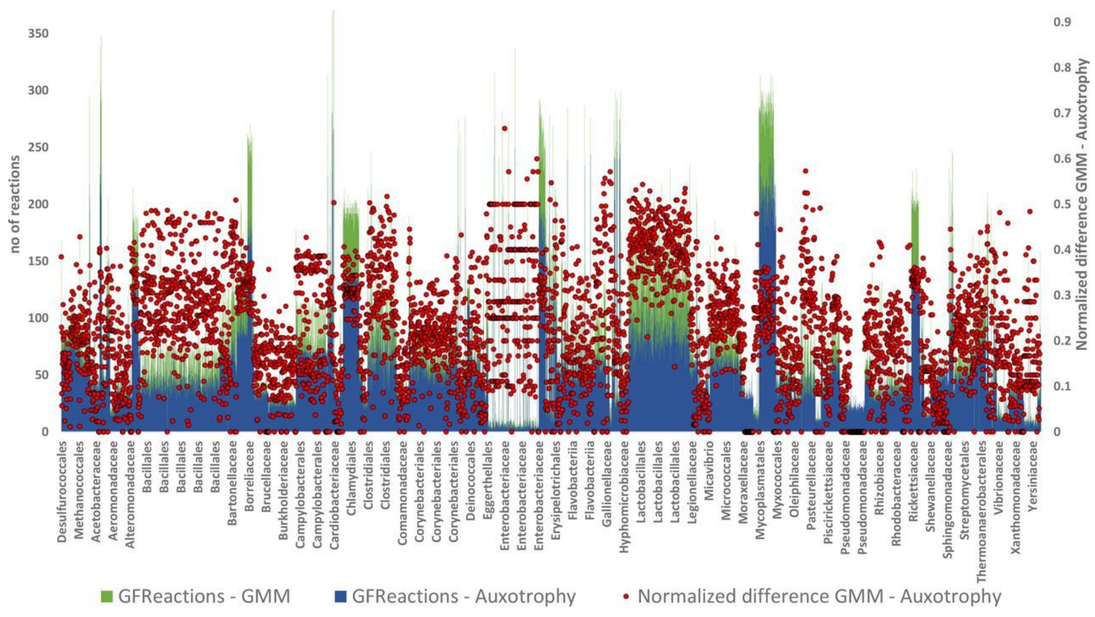
New modeling apps in KBase



A new reconstruction app implements the MS2 pipeline and uses the latest modeling templates. In addition, detailed reports provide insights into the gap-filling results and ATP production.

Insights from building models for a large set of phylogenetic diverse organisms:

Finding annotation gaps, classification errors, and exploring the level of auxotrophic dependencies by comparing model gap-filled reactions in glucose minimal media (GMM) and auxotrophy media.



Comparison of total gap-filled reactions (left axis) for two sets of models representing 5,420 genomes. Models gap-filled in GMM are shown in green. Models gap-filled in auxotrophy media are shown in dark blue. Red points (right axis) show the difference between the GMM and auxotrophy gap-filling counts normalized by the GMM gap-filling counts. If this normalized gap-filling difference is close to one, then the organism is more likely to be highly auxotrophic; if the number is close to zero, then the organism is likely to grow in near-minimal media

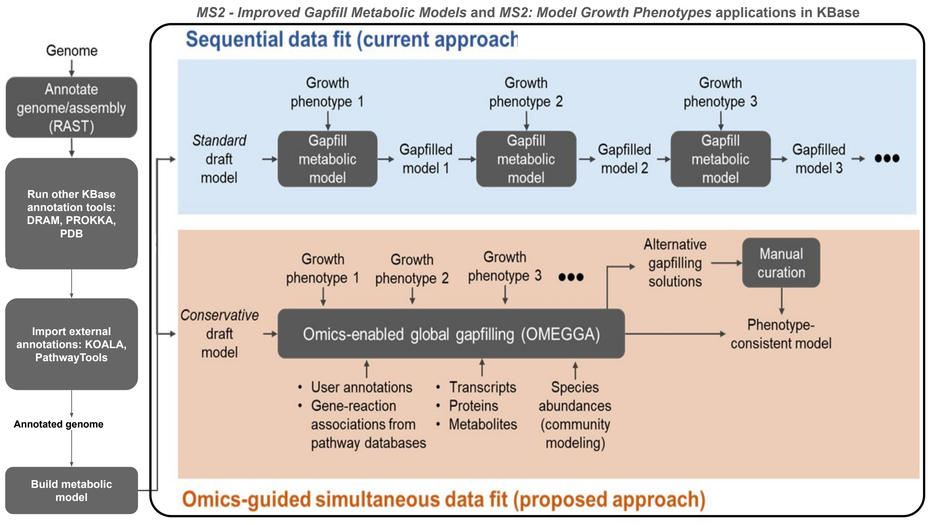
Gap-filling analysis for energy biosynthesis across diverse genomes.

Energy strategies from 1,250 Bacteria and Archaea GEMs. Row background colors indicate different media conditions for gap-filling: light blue = oxic, light green = anoxic, dark blue = anoxic nitrate, orange = anoxic sulfate. Abbreviations in the legend represent media substrates like glucose (Glc), acetate (Ac), etc. For each genome (column), the colored symbol indicates the number of gap-filled reactions needed for the GEM to produce ATP: green squares = no extra reactions needed, green diamond = one reaction, blue diamond = two reactions, yellow diamond = three reactions, pink diamond = four reactions, no shape = five or more reactions needed for ATP. Dashed boxes highlight phylogenetic groups of interest: **A** - Desulfovibrionales and Desulfobacterales; **B** - *Thioalkalivibrio paradoxus*; **C** - Pseudomonadaceae; **D** - Erwiniaceae; **E** - Rickettsiaceae; **F** - Synechococcales; **G** - *Rhodococcus opacus*; **H** - Clostridium and Fusobacterium; **I** - Mycoplasmataceae; **J** - Bacillales; **K** - Chlamydiales

Energy pathways define the amount of ATP an organism can derive from the environment given the availability of required nutrients. By combining this energy pathway knowledge with measured abundances of species within a sample, we gain understanding of resource richness of the environment, environment parameters like redox availability, and how organisms within the environment might work together. Some energy strategies are more synergistic than others. With predictions of energy pathways from genome sequences, we gain causal insights into metabolic drivers that govern microbiome structure.

Collaborating with PNNL Soil Microbiome SFA to integrate phenotype and multi-omics data with OMEGGA:

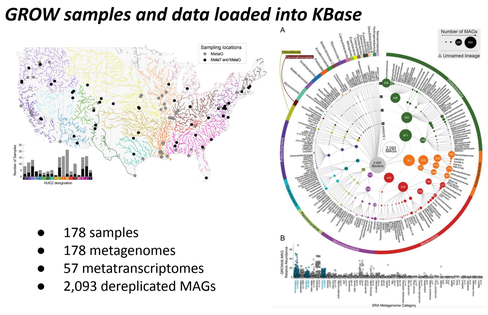
The PNNL Soil Microbiome SFA team incorporated the omics-enabled global gap-filling algorithm (OMEGGA) into two KBase apps: MS2 - Improved Gap-fill Metabolic Models and MS2 - Model Growth Phenotypes. OMEGGA utilizes growth phenotype and multi-omics data to fill gaps in an organism's metabolic pathways and annotations, optimizing the metabolic model to simultaneously match multiple observed growth conditions and produce experimentally observed metabolites. The algorithm selects gene candidates from KBase annotation tools (right), weighted by probabilities, with the highest probability gene chosen for each gap-filled reaction. OMEGGA refines these probabilities using transcriptomic, proteomic, or gene fitness data, prioritizing gene candidates with omics-based evidence for expression. The OMEGGA pipeline was applied in KBase to enhance MS2-built models for 7 PNNL Soil Microbiome SFA strains in the Model Soil Consortium (MSC)-2 across 11 experimentally tested growth conditions. The table tallies the number of gap-filled reactions for each media / the number of those reactions for which OMEGGA found a gene candidate.



Strain	Glucose	NAG	Serine	Alanine	Maltose	Xylose	Glutamate	Fructose	Arabinose	Sucrose	Glycine
Streptomyces (G1)	80/42	80/42	80/42	82/43			80/43	80/43	83/43	80/43	
Neorhizobium (G5)	66/43	68/46			66/46	67/47	67/46	70/47	69/46	69/46	
Dyadobacter (G7)	90/51	92/51	92/52	92/52	92/52	91/53	92/52	92/52	90/52	94/53	
Sphingopyxis (G8)	83/37	83/37	85/37	85/37	84/38	85/37	83/37	84/38	86/37		
Ensifer (G11)	77/46	78/46	76/47		75/46	76/47	76/47	77/46	79/50	78/47	76/47
Variovax (G12)	70/41	71/40		72/41	70/41	71/42	70/41	70/41	70/41		
Rhodococcus (G16)	80/50	81/49	80/50	81/49		82/49	78/48		84/48	81/50	

Applying KBase tools to analyze and model Genome Resolved Open Wetlands (GROW) samples:

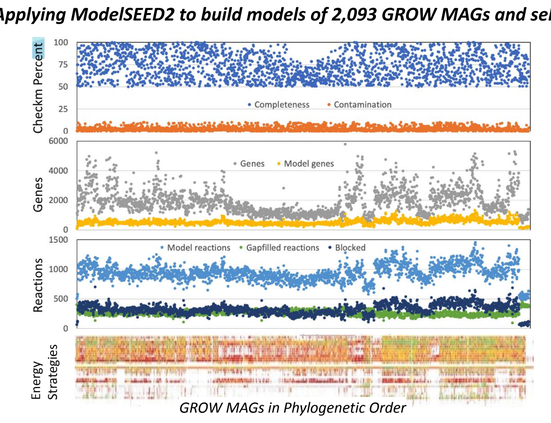
GROW samples and data loaded into KBase

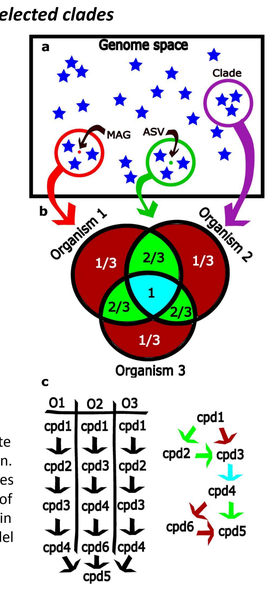


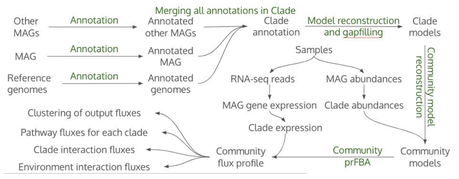
- 178 samples
- 178 metagenomes
- 57 metatranscriptomes
- 2,093 dereplicated MAGs

The GROW Community Sequencing Project sequenced 178 metagenome samples with 57 metatranscriptomes. From this data, 2,093 dereplicated MAGs were created, loaded into KBase, and used to construct metabolic models (stats on right). Unfortunately, the MAG models have many gaps.

Applying ModelSEED2 to build models of 2,093 GROW MAGs and selected clades



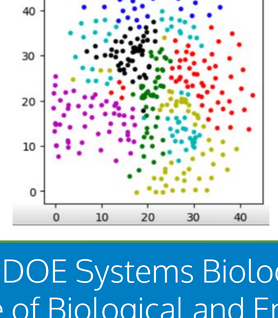




To mitigate gaps in MAG models, KBase developed a pangenome framework to aggregate annotations from many phylogenetically close genomes/MAGs into a probabilistic annotation. These annotations can be built from: (1) genomes/MAGs close to an input MAG; (2) genomes with 16S similar to an input ASV; or (3) all genome falling into a particular taxonomic group of interest. We applied this approach to build probabilistic annotations for 6 clades of interest in the GROW data. We then constructed a probabilistic model for each clade, applying the model to predict clade interactions using the community simulation method displayed to the left.

Ordination plot of intracellular fluxes within each clade of the 6-clade GROW model

We applied our 6-clade community model to simulate growth across 57 GROW river samples with metatranscriptome data. These samples displayed different community growth rates, environmental interactions, and species interactions. By plotting the flux through each species in a self-organized-map (SOM), we see that each clade occupies a distinct metabolic niche, with some clades jumping amongst multiple niche clusters. This plot demonstrates how the modeling approach captures the distinct metabolic role of each of our 6 GROW clades.



Yellow: UBA3064 (f__Burkholderiaceae)
Black: UBA954 (f__Burkholderiaceae_A)

Cyan: Polynucleobacter
Purple: Pirellula_B

Blue: Planktophila
Green: Methylopumilus

Other

References

Henry, Christopher S., et al. "High-throughput generation, optimization and analysis of genome-scale metabolic models." Nature biotechnology 28.9 (2010): 977-982.

Faria, José P., et al. "ModelSEED v2: High-throughput genome-scale metabolic model reconstruction with enhanced energy biosynthesis pathway prediction." bioRxiv (2023): 2023-10.

Borton, Kayla, et al. "A functional microbiome catalog crowdsourced from North American rivers." bioRxiv, 2023.07. 22.550117

Funding

This work is supported as part of the BER Genomic Science Program. The DOE Systems Biology Knowledgebase (KBase), SFA-KBase supplements, and GROW GSP program are funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, DE-AC02-98CH10886, DEAC0576RL01830, and DE-AC52-07NA27344.

U.S. DEPARTMENT OF ENERGY Office of Science

Funding: This work is supported as part of the Genomic Sciences Program DOE Systems Biology Knowledgebase (KBase) funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.

kbase.us YouTube

Collaboratively Assembling a Toolkit in KBase to Leverage Probabilistic Annotation and Multi-omics Data to Enable Metabolic Modeling of Microbial Community Ecology

José P. Faria[1](jplfaria@anl.gov), Filipe Liu[1], Andrew P. Freiburger[1], Mikayla Borton[9], Kelly Wrighton[9], Patrik D'haeseleer[2], Jeff Kimbrel[2], Jeremy Jacobson[3], Bill Nelson[3], Jason McDermott[3], Aimee K. Kessell[4], Hugh C. McCullough[4], Hyun-Seob Song[4], Janaka N. Edirisinghe[1], Nidhi Gupta[1], Samuel M.D. Seaver[1], Qizhi Zhang[1], Pamela Weisenhorn[1], Neal Conrad[1], Raphy Zarecki[5], Matthew DeJongh[5], Aaron A. Best[5], KBase Team[1,6,7,8], Robert W. Cottingham[6], Adam P. Arkin[7], Rhona Stuart[2], Kirsten Hofmockel[3], and Christopher S. Henry[1]

[1]Argonne National Laboratory, Lemont, IL; [2]Lawrence Livermore National Laboratory, Livermore, CA; [3]Pacific Northwest National Laboratory, Richland WA; [4]University of Nebraska–Lincoln, Lincoln, NE; [5]Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel; [5]Hope College, Holland, MI; [6]Oak Ridge National Laboratory, Oak Ridge, TN; [7]Lawrence Berkeley National Laboratory, Berkeley, CA; [8]Brookhaven National Laboratory, Upton, NY; [9]Colorado State University, Fort Collins, CO

KBase PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

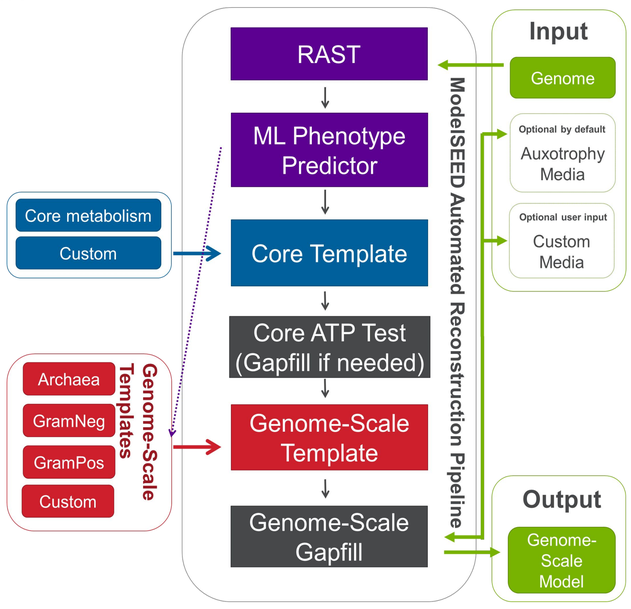
kbase.us – engage@kbase.us – @DOEKBase

Abstract:

Mechanistic understanding of biological systems relies on accurate protein annotations, which are often uncertain and error-prone. Genome-scale metabolic models (GEMs) evaluate these annotations within their biological context, offering a means to refine them by considering experimental observations. KBase has developed an ecosystem of tools for this purpose, starting with protein sequence annotation using various tools, and supporting external annotations. The novel ModelSEED2 (MS2) tool enhances GEM construction with improved energy metabolism representation and pathway curation, leading to more comprehensive models. Ensemble modeling approaches then generate multiple GEM drafts from probabilistic protein annotations, evaluated against ATP biosynthesis, necessary gap-filling, and omics data congruence. The best models are further analyzed, with gap-filling algorithms like OMEGGA selecting annotations that align with experimental data. This collaborative effort across KBase, µBiospheres SFA, and PNNL Soil SFA demonstrates improved GEM pathway completeness and annotation accuracy through applications to diverse species and datasets, showcasing the system's ability to refine our understanding of metabolic functions across organisms.

Improvements to the reconstructions pipeline, templates and KBase Apps:

MS2 genome-scale metabolic reconstruction pipeline enabling quantitative prediction of ATP production



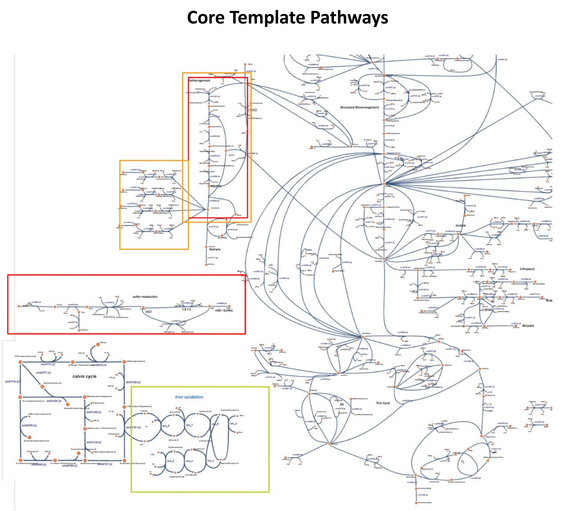
A genome annotated with RAST is inputted. Users may choose a reconstruction template, or ML classifiers can select them. ATP production is tested in 54 media, representing various energy biosynthesis strategies, with gap-filling as needed. The core metabolism model is then expanded to genome-scale.

Improvements in Energy Biosynthesis Pathway Reconstruction Based on Community-Driven Collaborative Curation

ANME and SRB, **Victoria Orphan**

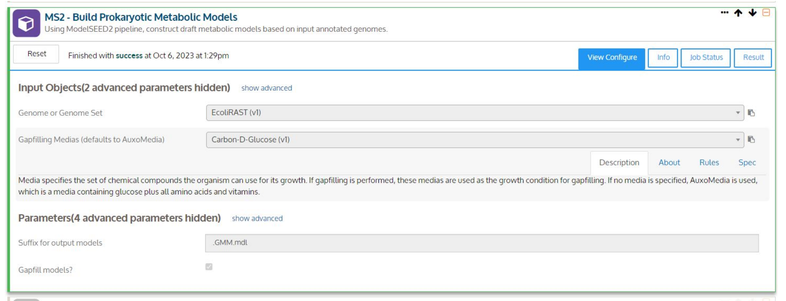
Methanogenesis and Methyltrophy, JGI **Susannah Tringe**

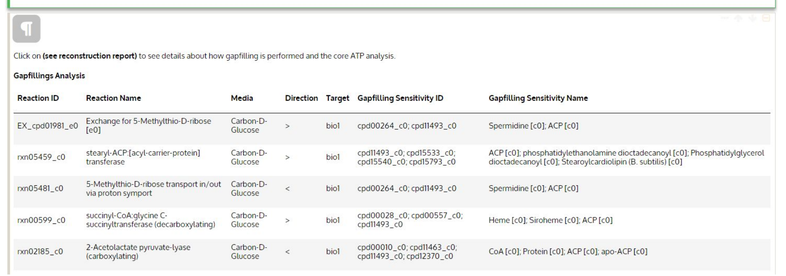
Iron Oxidation, University of Delaware **Clara Chan**



Many pathways of interest for researcher in the DOE space are still poorly represented in public databases. Working with experts we have expanded our templates to properly model Anaerobic methanotrophic archaea (ANME), sulfate reducing bacteria (SRB), Methanogenesis, Methyltrophy and Iron Oxidation.

New modeling apps in KBase





Click on **(see reconstruction report)** to see details about how gapfilling is performed and the core ATP analysis.

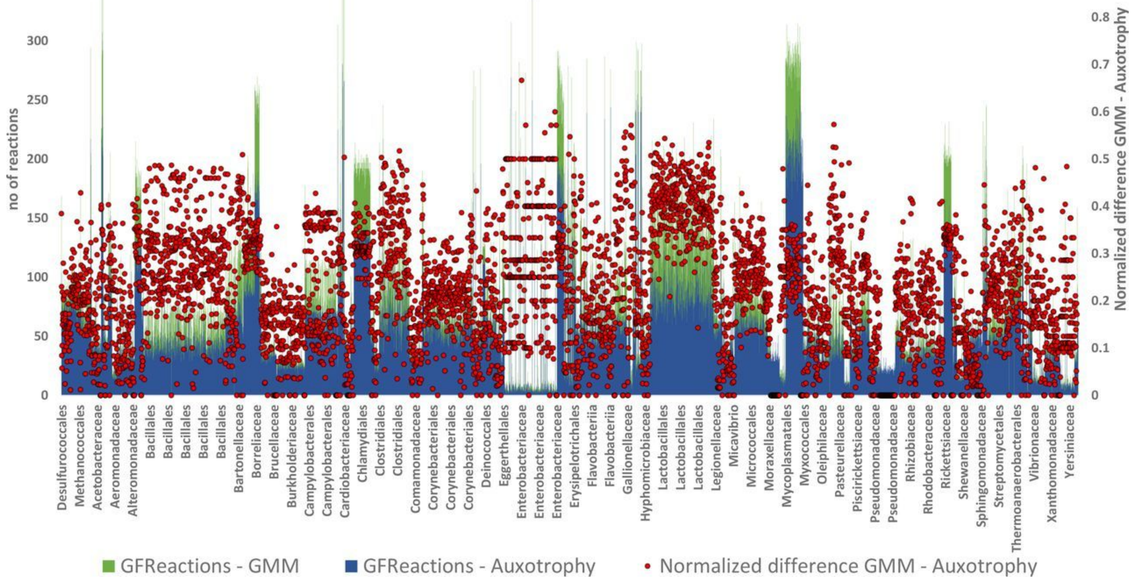
Gapfillings Analysis

Reaction ID	Reaction Name	Media	Direction	Target	Gapfilling Sensitivity ID	Gapfilling Sensitivity Name
EX_cpd01981_e0	Exchange for 5-Methylthio-D-ribose [e0]	Carbon-D-Glucose	>	bio1	cpd00264_c0; cpd11493_c0	Spermidine [c0]; ACP [c0]
rxn05459_c0	stearyl-ACP [acyl-carrier-protein] transferase	Carbon-D-Glucose	>	bio1	cpd11493_c0; cpd15533_c0; cpd15540_c0; cpd15793_c0	ACP [c0]; phosphatidylethanolamine dioctadecanoyl [c0]; Phosphatidylglycerol dioctadecanoyl [c0]; Stearoylcardiolipin (B. subtilis) [c0]
rxn05481_c0	5-Methylthio-D-ribose transport in/out via proton symport	Carbon-D-Glucose	<	bio1	cpd00264_c0; cpd11493_c0	Spermidine [c0]; ACP [c0]
rxn00599_c0	succinyl-CoA:glycine C-succinyltransferase (decarboxylating)	Carbon-D-Glucose	>	bio1	cpd00028_c0; cpd00557_c0; cpd11493_c0	Heme [c0]; Siroheme [c0]; ACP [c0]
rxn02185_c0	2-Acetolactate pyruvate-lyase (carboxylating)	Carbon-D-Glucose	<	bio1	cpd00010_c0; cpd11463_c0; cpd11493_c0; cpd12370_c0	CoA [c0]; Protein [c0]; ACP [c0]; apo-ACP [c0]

New reconstruction app implements the MS2 pipeline and uses the latest modeling templates. In addition, detailed reports provide insights into the gap filling results and ATP production.

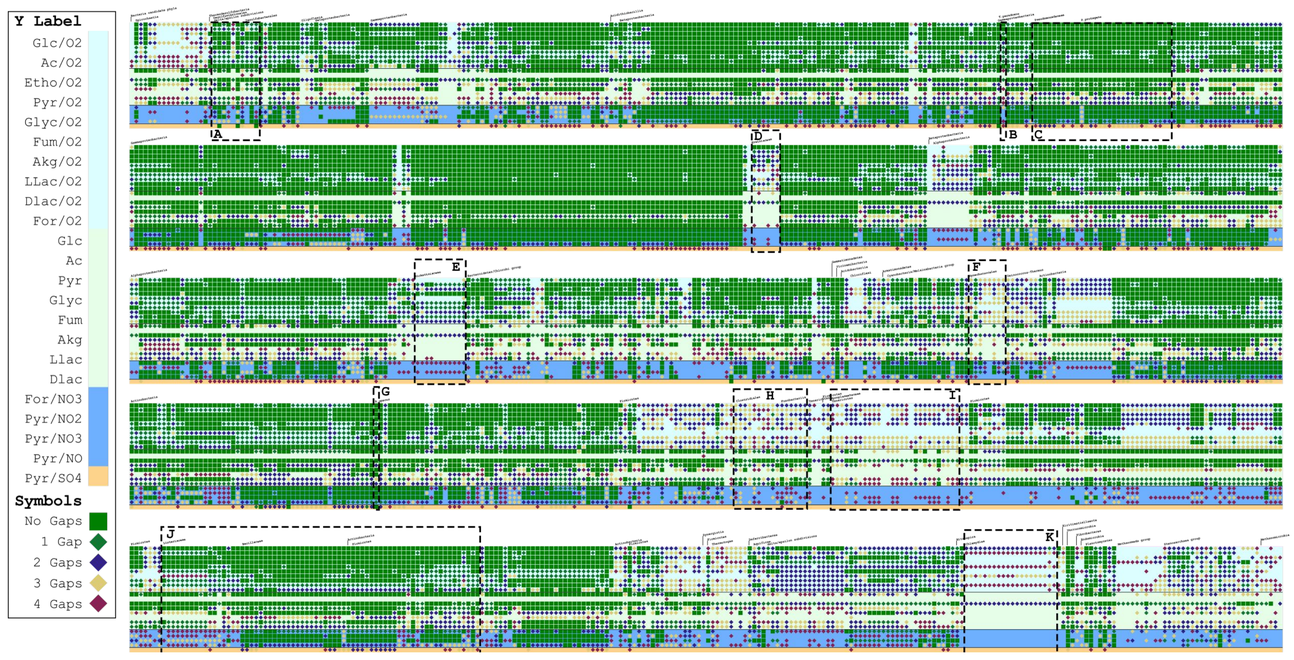
Insights from building models for a large set of phylogenetic diverse organisms:

Model comparison of total model reactions and gap-filled reactions in glucose minimal media (GMM) and auxotrophy media.



Comparison of total gap-filled reactions for two sets of models representing 5420 genomes. Models gap-filled in GMM are shown in green. Models gap-filled in auxotrophy media are shown in dark blue. The difference between the GMM and auxotrophy gapfilling counts normalized by the GMM gapfilling counts as a third data element (red points, second axis). If this normalized gap-filling difference is close to 1, then the organism is more likely to be highly auxotrophic; if the number is close to zero, then the organism is likely to grow in near-minimal media

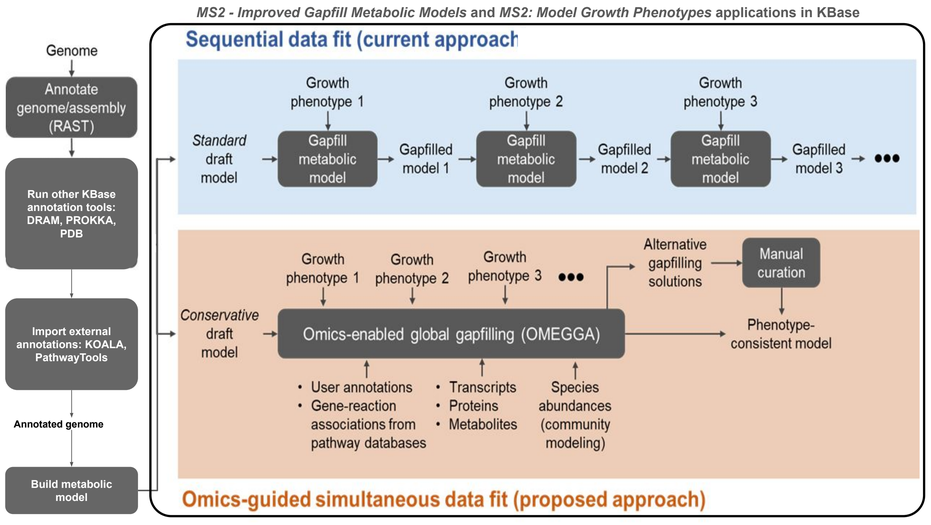
Gap-filling analysis for energy biosynthesis across diverse genomes.



Green squares: No extra reactions needed for ATP in specific media. Diamonds: Extra reactions needed for ATP; green for one, dark blue for two, yellow for three, dark pink for four. No shape: Five or more reactions needed. Light blue/green backgrounds: oxic/anoxic conditions. Dark blue: anoxic nitrate media; orange: anoxic sulfate media. Dashed boxes: Phylogenetic groups, labeled A-K. Data from 1,250 Bacteria and Archaea genomes. Abbreviations represent compounds like glucose (Glc), acetate (Ac), etc.Dashed line boxes represent phylogenetic groups of interest: **A** - Desulfovibrionales and Desulfobacterales; **B** - *Thioalkalivibrio paradoxus*; **C** - Pseudomonadaceae; **D** - Erwiniaceae; **E** - Rickettsiaceae; **F** - Synechococcales; **G** - *Rhodococcus opacus*; **H** - Clostridium and Fusobacterium; **I** - Mycoplasmataceae; **J** - Bacillales; **K** - Chlamydiales

Collaborating with PNNL Soil Microbiome SFA to Integrate Phenotype and Multi-omics Data in KBase with OMEGGA Tool

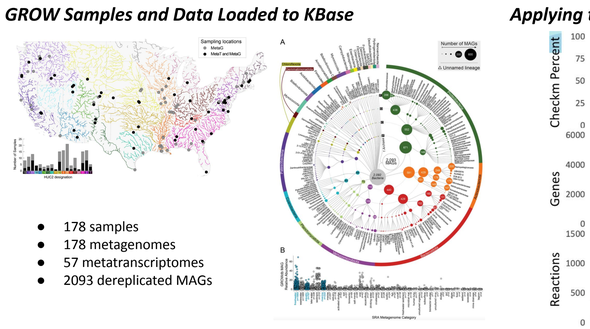
The PNNL Soil Microbiome SFA team incorporated the Omics-enabled global gap-filling (OMEGGA) algorithm into the MS2 - Improved Gap-fill Metabolic Models and MS2 - Model Growth Phenotypes apps on the KBase platform. OMEGGA utilizes growth phenotype and multi-omics data to fill gaps in an organism's metabolic pathways and annotations, optimizing the metabolic model to simultaneously match multiple observed growth conditions and produce observed metabolites. The algorithm selects gene candidates from the KBase annotation pipeline and tools (right), weighted by probabilities, with the highest probability gene chosen for each gap-filled reaction. OMEGGA refines these probabilities using transcriptomic, proteomic, or gene fitness data, prioritizing gene candidates with omics-based evidence for expression. The OMEGGA pipeline was applied in KBase to enhance MS2-built models for 7 PNNL Soil Microbiome SFA strains in the Model Soil Consortium (MSC)-2 across 11 experimentally tested growth conditions. The table below illustrates gap-filled reactions/gene candidates added by OMEGGA in this analysis.



Strain	Glucose	NAG	Serine	Alanine	Maltose	Xylose	Glutamate	Fructose	Arabinose	Sucrose	Glycine
Streptomyces (G1)	80/42	80/42	80/42	82/43			80/43	80/43	83/43	80/43	
Neorhizombium (G5)	66/43	68/46			66/46	67/47	67/46	70/47	69/46	69/46	
Dyadobacter (G7)	90/51	92/51	92/52	92/52	92/52	91/53	92/52	92/52	90/52	94/53	
Sphingopyxis (G8)	83/37	83/37	85/37	85/37	84/38	85/37	83/37	84/38	86/37		
Ensifer (G11)	77/46	78/46	76/47		75/46	76/47	76/47	77/46	79/50	78/47	76/47
Variovax (G12)	70/41	71/40		72/41	70/41	71/42	70/41	70/41	70/41		
Rhodococcus (G16)	80/50	81/49	80/50	81/49		82/49	78/48		84/48	81/50	

Applying KBase Tools to Analyzing and Modeling Genome Resolved Open Wetlands (GROW) Samples

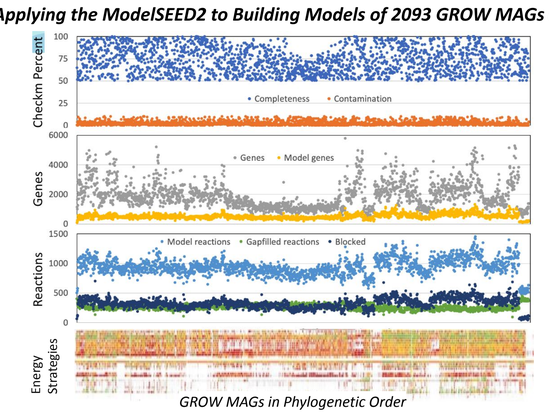
GROW Samples and Data Loaded to KBase



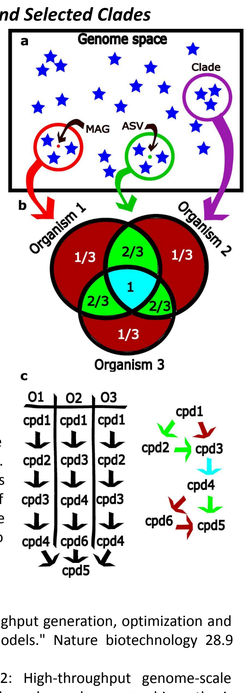
- 178 samples
- 178 metagenomes
- 57 metatranscriptomes
- 2093 dereplicated MAGs

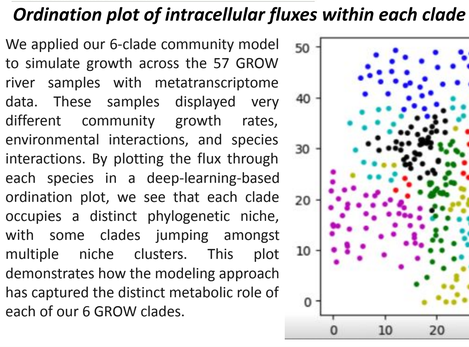
GROW Community Sequencing Project sequenced 178 metagenome samples with 57 metatranscriptomes. From this data, 2093 dereplicated MAGs were created, loaded into KBase, and applied to construct metabolic model (model stats on right). Unfortunately, MAG models has many gaps.

Applying the ModelSEED2 to Building Models of 2093 GROW MAGs and Selected Clades



GROW MAGs in Phylogenetic Order





To mitigate gaps in MAG models, KBase developed a pangenome framework to aggregate annotations from many phylogenetically close genomes/MAGs into a probabilistic annotation. These annotations can be built from: (1) genomes/MAGs close to an input MAG; (2) genomes with 16s similar to an input ASV; or (3) all genome falling into a particular taxonomic group of interest. We applied this approach to build probabilistic annotations for 6 clades of interest in the GROW dataset. We then constructed a probabilistic model for each clade, applying the model to predict clade interactions using the community simulation method displayed to the left.

Ordination plot of intracellular fluxes within each clade of the 6-clade GROW model

We applied our 6-clade community model to simulate growth across the 57 GROW river samples with metatranscriptome data. These samples displayed very different community growth rates, environmental interactions, and species interactions. By plotting the flux through each species in a deep-learning-based ordination plot, we see that each clade occupies a distinct phylogenetic niche, with some clades jumping amongst multiple niche clusters. This plot demonstrates how the modeling approach has captured the distinct metabolic role of each of our 6 GROW clades.

Yellow: UBA3064 (f__Burkholderiaceae)
Black: UBA954 (f__Burkholderiaceae_A)

Cyan: Polynucleobacter
Purple: Pirellula_B

Blue: Planktophila
Green: Methylopumilus

Other

References

Henry, Christopher S., et al. "High-throughput generation, optimization and analysis of genome-scale metabolic models." Nature biotechnology 28.9 (2010): 977-982.

Faria, José P., et al. "ModelSEED v2: High-throughput genome-scale metabolic model reconstruction with enhanced energy biosynthesis pathway prediction." bioRxiv (2023): 2023-10.

Borton, Kayla, et al. "A functional microbiome catalog crowdsourced from North American rivers." bioRxiv, 2023.07. 22.550117

Funding

This work is supported as part of the BER Genomic Science Program. The DOE Systems Biology Knowledgebase (KBase), SFA-KBase supplements, and GROW GSP program are funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, DE-AC02-98CH10886, DEAC0576RL01830, and DE-AC52-07NA27344.

U.S. DEPARTMENT OF ENERGY Office of Science

Funding: This work is supported as part of the Genomic Sciences Program DOE Systems Biology Knowledgebase (KBase) funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.

kbase.us YouTube