# Collaboratively Assembling a Toolkit in KBase to Leverage Probabilistic Annotation and Multi-omics Data to Improve Mechanistic Modeling of Metabolic Phenotypes


**José P. Faria[1](jplfaria@anl.gov), Filipe Liu[1], 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], Andrew P. Freiburger[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.


KBase PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

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

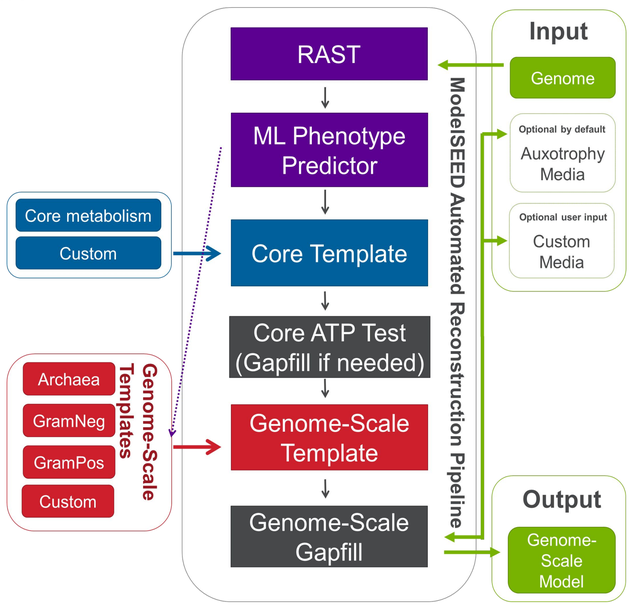
Argonne NATIONAL LABORATORY · BERKELEY LAB Bringing Science Solutions to the World · Brookhaven National Laboratory · OAK RIDGE National Laboratory · Lawrence Livermore National Laboratory · Pacific Northwest NATIONAL LABORATORY


## 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*



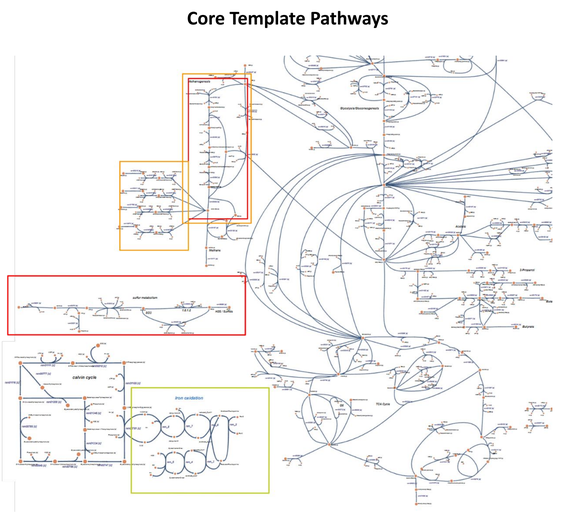

A genome annotated with RAST is inputted. Users may choose a reconstruction template, or ML classifiers can select them. ATP production is tested in 54 media, representing various energy biosynthesis strategies, with gap-filling as needed. The core metabolism model is then expanded to genome-scale.

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

ANME and SRB, **Victoria Orphan**

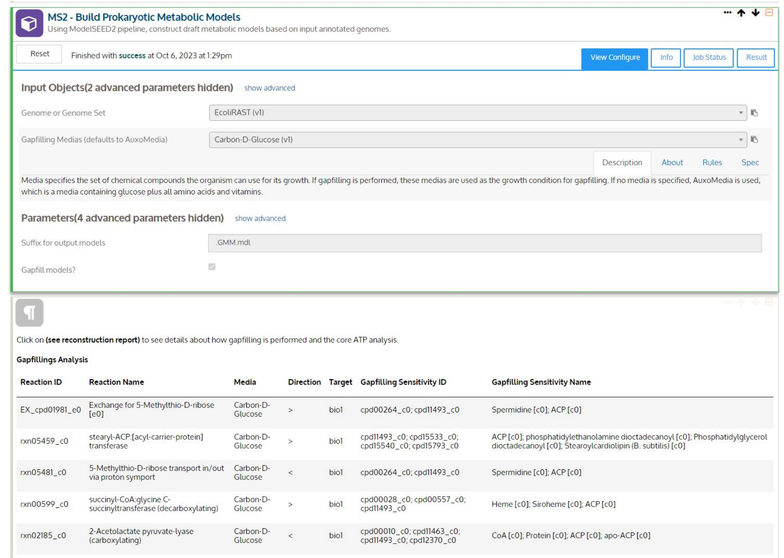
Methanogenesis and Methyltrophy, JGI **Susannah Tringe**

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 I have expanded our templates to properly model Anaerobic methanotrophic archaea (ANME),sulfate reducing bacteria (SRB), Methanogenesis, Methyltrophy and Iron Oxidation.

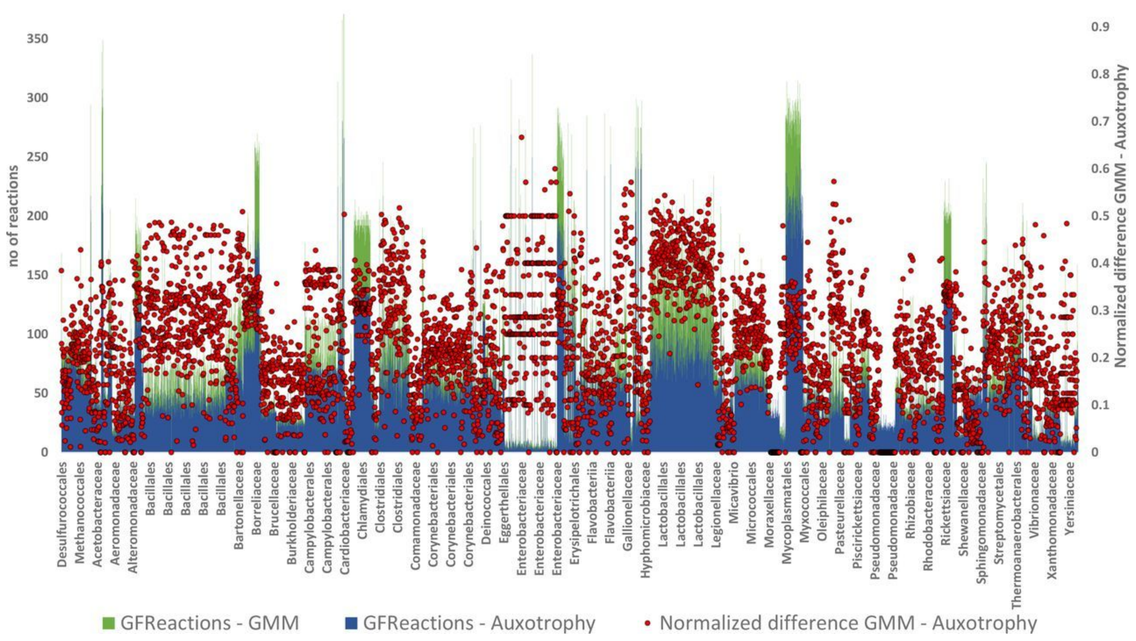
### *New modeling apps in KBase*




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.

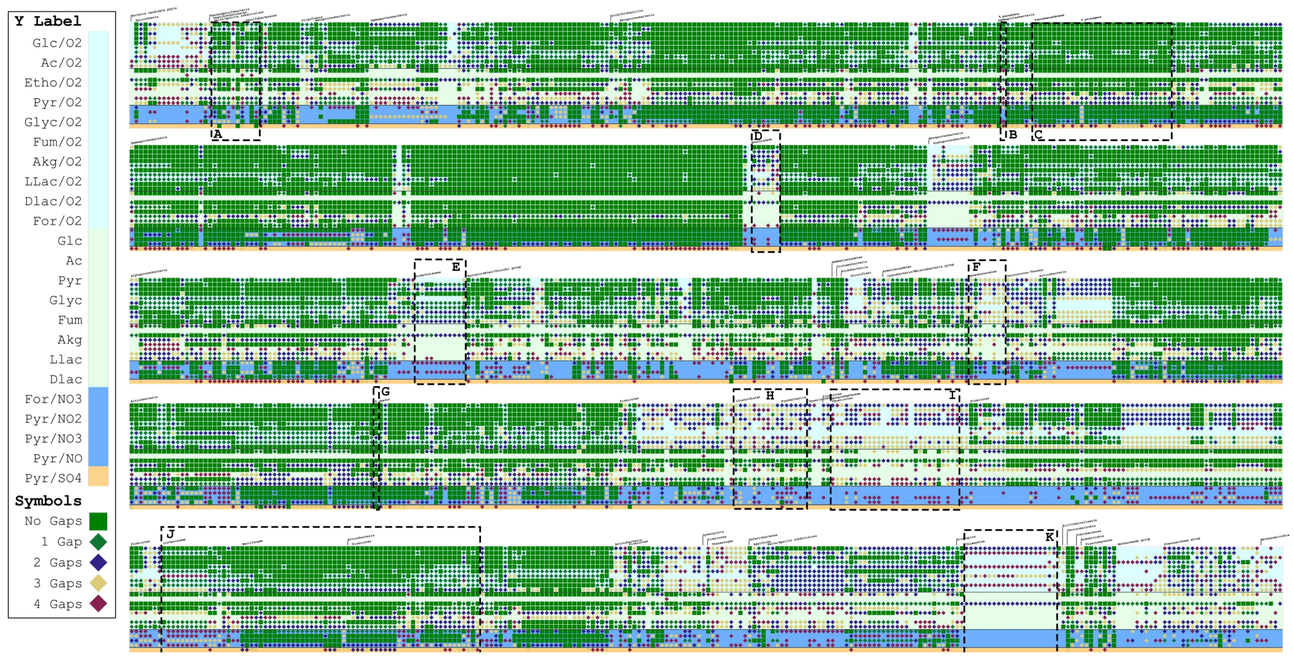
## Insights from building models for a large set of phylogenetic diverse organisms:

### *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

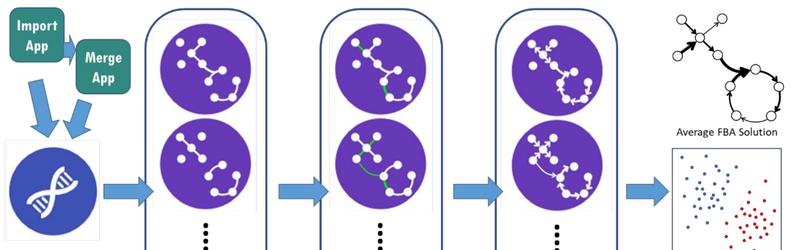
### *Gap-filling analysis for energy biosynthesis across diverse genomes.*




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 the LLNL µBiospheres SFA to Build Probabilistic Annotation and Ensemble Modeling in KBase

The LLNL µBiospheres SFA introduced a pipeline in KBase for probabilistic annotation and ensemble modeling. Functional annotations, from various algorithms inside or outside KBase, are integrated into a genome object with associated probabilities. Sampling from these annotations generates an ensemble of potential models. Probabilistic annotations aid in suggesting gene candidates for gap-filling using the OMEGGA tool (bottom right panel). Applying model ensembles predicts pathway flux, producing a set of flux solutions for statistical analysis.




| Domain | Algorithm | Total proteins | Reaction match | EC match | No mapping | Not EC match | No prediction |
|---|---|---|---|---|---|---|---|
| Bacteria | RAST | 5203 | 27.21% | 57.03% | 27.28% | 15.69% | 20.39% |
| Fungi | RAST | 1725 | 18.79% | 49.83% | 37.12% | 13.05% | 48.46% |
| Other | RAST | 468 | 37.29% | 63.84% | 21.47% | 14.69% | 62.18% |
| Viridiplantae | RAST | 2741 | 25.39% | 51.02% | 22.45% | 26.53% | 55.31% |
| Metazoa | RAST | 5242 | 23.58% | 51.78% | 28.99% | 19.23% | 65.78% |
| Archaea | RAST | 671 | 34.45% | 69.46% | 18.06% | 12.48% | 19.97% |
| Bacteria | DRAM_KO | 5203 | 41.75% | 53.76% | 38.88% | 7.36% | 12.80% |
| Fungi | DRAM_KO | 1725 | 34.33% | 47.57% | 45.69% | 6.74% | 26.03% |
| Other | DRAM_KO | 468 | 38.80% | 45.60% | 49.60% | 4.80% | 46.58% |
| Viridiplantae | DRAM_KO | 2741 | 36.41% | 52.74% | 30.10% | 17.16% | 21.34% |
| Metazoa | DRAM_KO | 5242 | 25.19% | 42.67% | 52.87% | 4.45% | 4.06% |
| Archaea | DRAM_KO | 671 | 48.79% | 61.42% | 30.97% | 7.61% | 13.86% |
| Bacteria | Prokka | 5203 | 65.03% | 92.66% | 1.14% | 6.20% | 22.24% |
| Fungi | Prokka | 1725 | 28.97% | 45.59% | 18.53% | 35.88% | 60.58% |
| Other | Prokka | 468 | 43.30% | 67.01% | 5.67% | 27.32% | 58.55% |
| Viridiplantae | Prokka | 2741 | 32.79% | 53.12% | 3.37% | 43.51% | 64.28% |
| Metazoa | Prokka | 5242 | 30.00% | 62.56% | 5.50% | 31.94% | 69.48% |
| Archaea | Prokka | 671 | 42.86% | 54.29% | 1.04% | 44.68% | 42.62% |

Functional annotation mapping is crucial for our ensemble modeling. Annotation pipelines employ various ontologies, sometimes none, for protein functional annotations. Evaluation requires comparison to a gold standard, here, all SwissProt experimental evidence-associated annotations. Applying RAST, DRAM, and PROKKA to SwissProt proteins, using EC numbers and biochemistry, reveals discrepancies. Cross-ontology comparisons prove error-prone, yet they underpin our annotation assessment and ensemble modeling. BioBERT is now employed to enhance the curation and refinement of our annotation mappings, recognizing the challenges in ensuring accuracy and reliability in functional annotations.

| Name1 | Name2 | Term1 | Term2 | similarity |
|---|---|---|---|---|
| (+)-3-carene synthase 1, chloroplastic | (+)-alpha-pinene synthase. (EC:4.2.3.121) | RHEA:25500; | EC:4.2.3.121 | 0.900726 |
| (+)-3-carene synthase 2, chloroplastic | (+)-alpha-pinene synthase. (EC:4.2.3.121) | RHEA:32539 | EC:4.2.3.121 | 0.90106 |
| (+)-T-muurolol synthase ((2E,6E)-farnesyl diphosphate cyclizing) | Name not found (4.2.3.-) | RHEA:32011 | EC:4.2.3.- | 0.797332 |
| (+)-T-muurolol synthase. (EC:4.2.3.98) | (+)-T-muurolol synthase ((2E,6E)-farnesyl diphosphate cyclizing) | EC:4.2.3.98 | RHEA:32011 | 0.918895 |
| (+)-T-muurolol synthase. (EC:4.2.3.98) | 5-epi-alpha-selinene synthase [EC:4.2.3.90] (4.2.3.90) | EC:4.2.3.98 | KO:K18109 | 0.959187 |

From Uniprot:(2E,6E)-farnesyl diphosphate + H2O = +-T-muurolol + diphosphate EC:4.2.3.9 **(79 % similarity)**

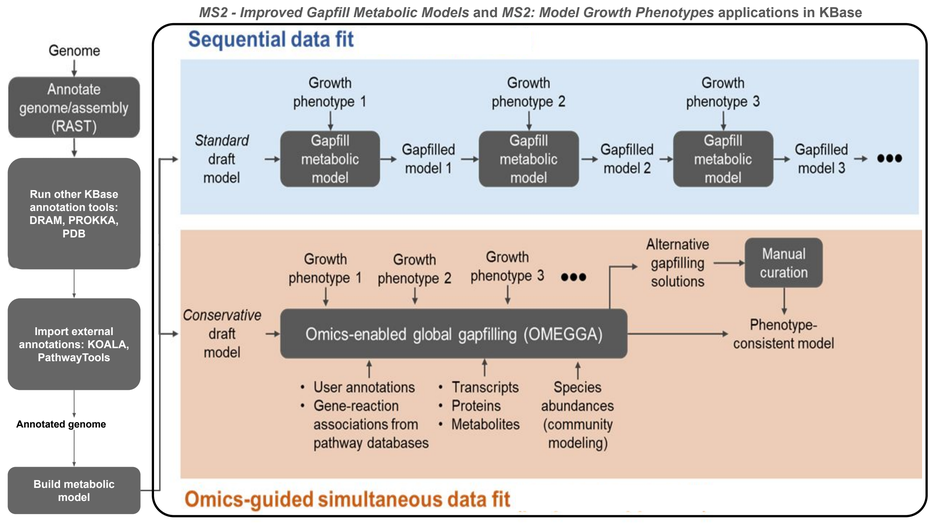
- Even with "name not found" BioBert find high similarity based on incomplete EC number

Both enzymes catalyze the conversion of geranyl pyrophosphate (GPP) to different monoterpenes: **(90% similarity)**

- These have separate reactions associated and are not a "good" mapping.
- Additional information (alternative names and ECs) from Uniprot can help provide further context for comparison when EC is missing

## 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 LLNL pipeline annotations, 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 | |
| **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 | |

### *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.


### *Funding*


This work is supported as part of the BER Genomic Science Program. The DOE Systems Biology Knowledgebase (KBase) and the SFA-KBase supplements 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