## KBase Performance Metric for FY17: Develop improved open access platforms for computational analysis of large genomic datasets.

**Q4 Metric:** Report on new capabilities and approaches for analyzing metagenomics datasets.

# 1. Background

KBase, the DOE Systems Biology Knowledgebase (http://kbase.us/) [1] is an open access software and data platform that enables scientists to create, execute and collaborate on workflows that combine experimental evidence and computational analyses to model plant and microbial physiology and community dynamics.These workflows can be saved as KBase *Narratives*—shareable, interactive documents that include all the data, analysis steps, results, visualizations, scripts, commentary and conclusions of an experiment. Some example Narratives, which can be copied and run on the included data or user-uploaded data, can be found in the Narrative Library (http://kbase.us/narrative-library).

KBase was designed to enable systems biology analysis of communities of microbes and/or plants. A new suite of tools released in summer 2017 has streamlined the process of performing metagenomic analyses in KBase. A user can predict species interactions from metagenomic data by assembling raw reads, binning assembled contigs by species, annotating genomes, aligning RNA-seq reads, and reconstructing and analyzing individual and community metabolic models. Users have applied these tools to study: (i) interactions between plants and microbes in soil; (ii) why some microbes form stable communities; (iii) how a microbial community cooperates to produce a specific product; and (iv) how a community of heterotrophic species can feed on byproducts from an autotroph to grow autotrophically.

# 2. Overview of metagenomic analysis capabilities in KBase

Since its first release, KBase has supported the analysis of isolate genomes, with numerous apps available for genome annotation, genome comparison, model reconstruction, and model analysis. More recently, this scientific functionality in KBase has been expanded to also provide support for the analysis of metagenomic data. This initial support for metagenomics was designed to maximize synergy with the existing capabilities in KBase by enabling raw metagenomic reads to be converted into assembled genome sequences that can be piped into KBase's isolate analysis pipelines.

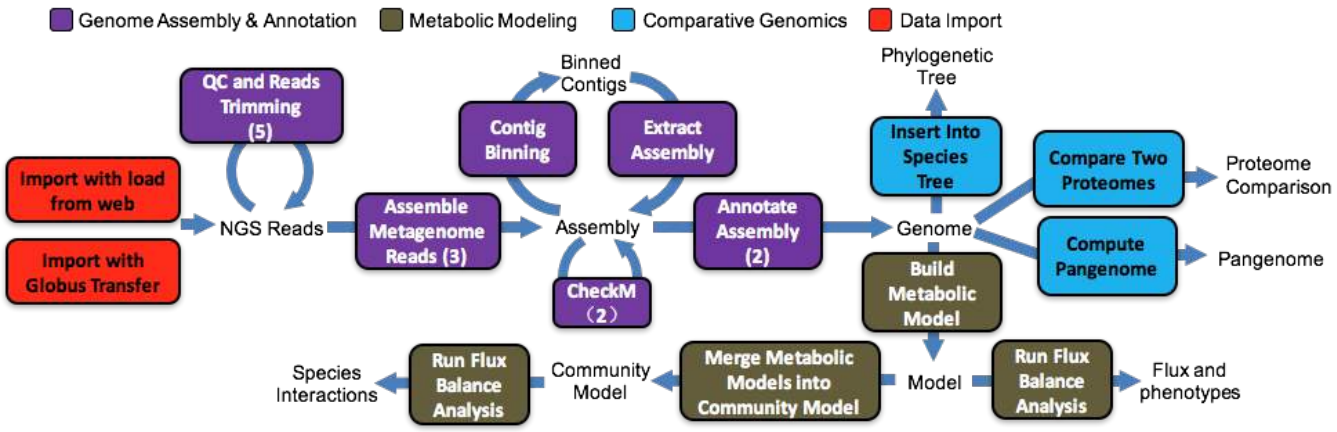


***Figure 1.*** *Workflow for metagenome analysis in KBase*

This new metagenome analysis pipeline (see Figure 1) begins with the upload of large reads files using either Globus Transfer or a new app that loads reads data from a public URL. Next, the user can apply one of three new apps for metagenome assembly (MEGAHIT [2], metaSPAdes [3], and IDBA [4]). The assembled contigs can then be binned (MaxBin [5]) and new genomes can be extracted from the bins. These extracted genomes can be analyzed for completeness using the genome quality analysis app (CheckM [6]), which ensures that genomes are complete before additional downstream analysis.

After metagenomic reads are assembled and binned into standard genomes, these can be piped into a wide range of downstream analysis apps in KBase, including genome annotation, genome comparison, metabolic modeling, and RNA-seq alignment. Some of these downstream analysis pipelines have also been updated to improve support for metagenome analysis. Specifically, new batch versions of the genome annotation [7] and model reconstruction [8] apps were added to facilitate the rapid generation of models for the potentially large number of genomes that might be produced from the assembly and binning of a metagenome.

Once a user has generated metabolic models from the genomes assembled from a metagenome, these models can then be combined into a community metabolic model, which can be applied with the Flux Balance Analysis app to predict trophic interactions between species. Model predictions can also be validated and even fit to available transcriptome data (including data processed by the RNA-seq pipeline in KBase).

## 2.1 Uploading large reads datasets

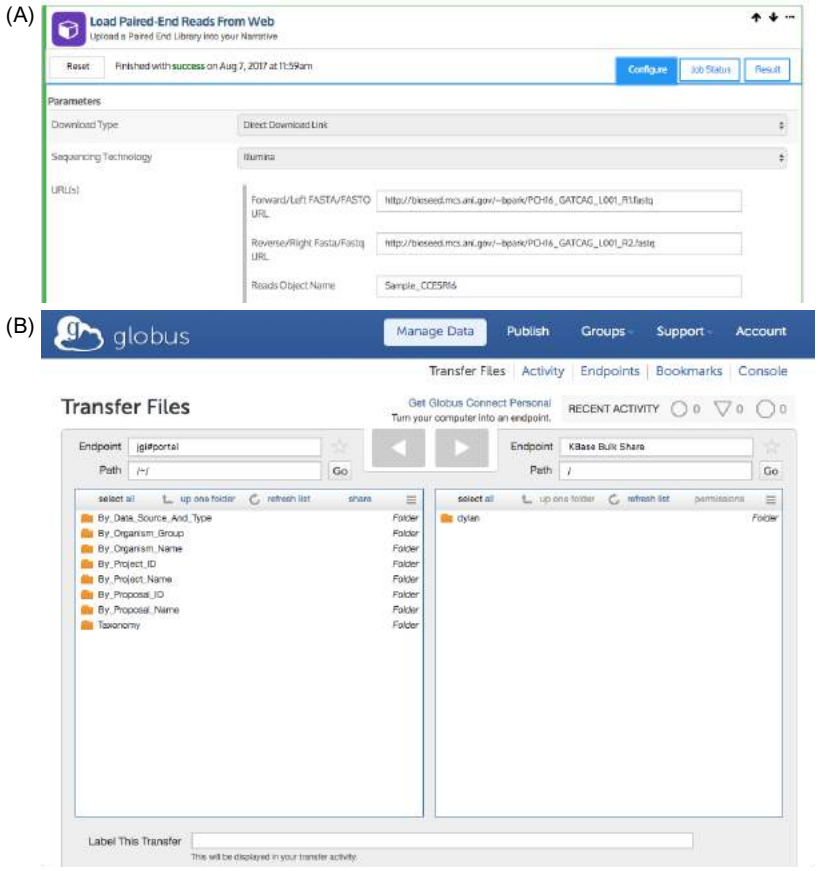


***Figure 2.*** *Example Load From Web (a) and Globus Transfer (b) Interfaces in KBase*

The microbiome pipeline in KBase requires shotgun metagenome reads as its primary initial input. These reads files can be quite large (ranging from 20GB to over 50GB), making it a challenge simply to load the reads into an online platform like KBase. To address this, KBase offers two primary mechanisms for loading large files into the platform.

The easiest approach is to apply one of four *Load from Web* apps in the KBase interface (Figure 2a). The file to be loaded must be available in a publically accessible web location (although this can be a private link)--for example, a public FTP, a Dropbox share link, a Google Drive share link, or an HTTP link. The *Load Paired-End Reads From Web* and *Load Single-End Reads From Web* apps can be used to load FASTQ (or zipped FASTQ) files for paired-end and single-end datasets respectively. The *Upload Web File* app loads any file into the user's staging area, requiring the user to transform the file into a KBase object

in a separate step. Finally, the *Import SRA File as Reads From Web* app supports the upload and processing of SRA-formatted reads, which may also be zipped.

In some cases, reads are not available from a publically accessible location. For these scenarios, KBase supports Globus Data Transfer (Figure 2b). To use this approach, data files must be placed on a Globus Data Transfer endpoint, which can include a user's own computer. The user then opens the Globus Data Transfer large file upload interface in KBase, moving the files from their Globus endpoint into the KBase Globus endpoint. The files are then immediately available in the user's private staging area in KBase, from which they can be converted into any KBase object.

## 2.2 Processing reads

After uploading reads into KBase, the next step is to process them. Raw sequencing reads contain base calls for the adapters and often other signatures from multiplexing or other barcoding technologies. Additionally, individual base calls are variable in quality, typically suffering in fidelity as the read gets longer. We have added the industry-standard tool FastQC [9] as an app for assessing the quality of read libraries. Additionally, for removing adapter/barcode sequences and regions of low quality from the reads, we have added apps that wrap the popular read-processing tools Cutadapt [10] and Trimmomatic [11]. These apps are typically run within Narratives that start with sequence read library data and allow for far superior genome/ metagenome assembly and other analyses where limiting the reads to the high-fidelity base calls is essential.

## 2.3 Assembling metagenomes into contigs

Shotgun sequencing of direct DNA extraction from environmental samples, so-called "metagenomic" sequencing, has in recent years been amenable to genome assembly into contiguous fragments, called "contigs". Leaps in algorithms and compute resources have enabled these approaches, and while many biologists can now fairly easily and inexpensively obtain environmental shotgun sequencing, the metagenome assembly step remains difficult for them to access. KBase recently added apps that run several of the most commonly used metagenome assemblers (MEGAHIT [2], metaSPAdes [3], and IDBA [4]). These metagenome assemblers produce "Assembly" objects that are groups of contigs that come from the mixture of species present in the samples. These algorithms are improving rapidly, but the best assembly for a given member of the community and the community as a whole may be algorithm and parameterization-dependent, so our expectation is that a user may decide to run several assemblies with different methods and parameterizations and compare the results. Toward this end, KBase provides the ability to compare the distributions of the contigs produced by the assemblers with the QUAST [12] tool, as well as an internally-developed variant of QUAST (the *Compare Assembled Contig Distributions* app) that displays histograms of the contig length distributions. Other metagenome assemblers and approaches to comparing assemblies to determine which contigs or methods of deriving consensus contigs are in the near-term plans for KBase to build out a more complete suite of applications in this area.

## 2.4 Binning contigs and evaluating genome quality

Genome recovery from metagenomic assembly is often likened to the problem of trying to solve many jigsaw puzzles simultaneously when the pieces have all been mixed together and most of the puzzles are missing a significant fraction of their pieces. The analogy is apt, but advances in sequencing technology that permit affordable and deeper shotgun sequencing with longer reads have allowed at least the most abundant members of microbial communities to be sequenced at sufficient depth-of-coverage to not only assemble reasonable length contigs, but also sometimes achieve near-complete genome coverage by those contigs. However, identifying which of these contigs belong together in the same species or other lineage grouping, called a genome "bin", is non-trivial. Concomitant with metagenome contig assembly, groups have been working on various strategies for taking advantage

of information within the contigs and read libraries to group the contigs into putative genome bins. The terms frequently used by the various algorithms include nucleotide k-mer frequencies, phylogenetically informative protein-coding marker genes within the contigs, and depth-of-coverage of reads to the contigs. KBase incorporated the MaxBin2 [5] method developed at the DOE Joint BioEnergy Institute (JBEI, www.jbei.org) as it does a good job with the three principle terms. As with assembly, the intent is to add additional binning tools to KBase that allow for comparison and perhaps improved performance by combination, beginning with MetaBAT [13] from the DOE Joint Genome Institute (JGI, jgi.doe.gov).

It is not sufficient to generate putative draft genomes without assessing their quality. In collaboration with the original tool developers KBase chose to wrap the CheckM [6] tool (Figure 3) for assessment of genome completeness and contamination because it goes a step further in that it derives clade-specific marker genes that may significantly improve the reliability of marker gene assessment compared to the smaller set of universal markers that would otherwise be used. It is at this point in the process, considering the overall quality of genomes extracted from the metagenome data, that a user can iterate the assembly/binning/contig assignment steps to obtain the best possible genomes from their samples. Once satisfied with the completeness and that the bin doesn't contain contigs from other organisms, the user moves forward with extracting the contigs for each bin as an assembly for genome annotation.

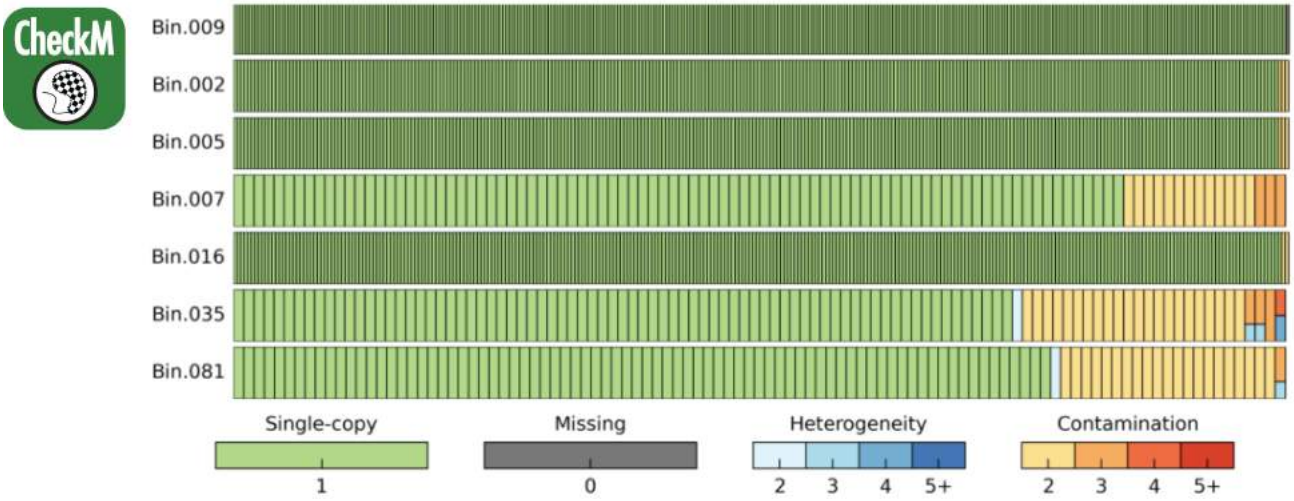


***Figure 3.*** *The CheckM app shows a visual overview of genome bin quality.*

## 2.5 Genome annotation, metabolic model reconstruction, and gapfilling

The binning step of the KBase microbiome analysis pipeline produces a series of KBase assembly objects, each representing the contigs of an individual strain within the microbiome being analyzed. These individual strain assemblies may now be input into the isolate analysis pipeline of KBase, which includes steps for genome annotation, metabolic model reconstruction, and model gapfilling. This pipeline was described in detail in our Q1 report, so we will not discuss it in detail here. The genome objects produced by this pipeline can be used as inputs to the comparative genomics apps in KBase, which are described in section 2.7. The metabolic model objects produced by this pipeline serve as inputs to the microbiome analysis pipeline in KBase, which is described next in section 2.6.

## 2.6 Merging models into a community model and predicting species interactions

Although individual-strain metabolic models can be used with flux balance analysis to predict the potential activity of the strain within a microbiome, it can still be challenging to use individual models

to predict potential interspecies interactions. For this reason, KBase offers the *Merge Metabolic Models into Community Model* app [14, 15], which combines multiple individual strain models into a single compartmentalized model of the entire microbiome. In this process, each strain model occupies a different compartment within the larger microbiome model, with all strain compartments occupying a single shared extracellular compartment. Based on this approach, models of different strains can exchange metabolites by exporting and importing from the shared extracellular compartment. Community models can also be gapfilled [16, 17] with the *Gapfill Metabolic Model* app. However, in this case, the gapfilling algorithm will attempt to add the ideal content to each individual strain model such that the entire community can grow in a specific growth condition. Once a community model is gapfilled in the appropriate condition, the model can be analyzed with the *Run Flux Balance Analysis* app to predict overall microbiome behavior, while also predicting the flux through every reaction in every strain comprising the microbiome. This will include predictions of the specific metabolites consumed, excreted, and exchanged by every strain comprising the microbiome, providing a detailed prediction of potential species interactions. This approach has already been applied extensively within KBase to predict potential interactions in a variety of microbiome systems [14, 15, 18].

## 2.7 Comparative analysis of genomes to identify enriched functions

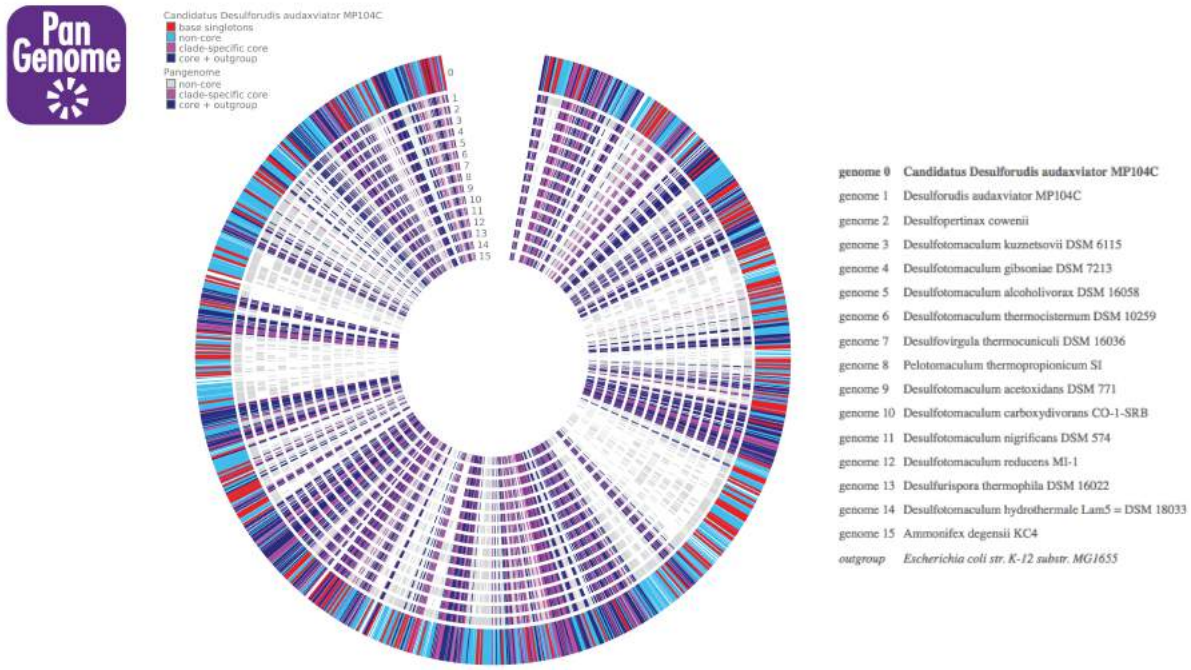


***Figure 4.*** *The Pangenome Circle Plot app provides a way to visualize the sets of genes in a pangenome.*

In addition to constructing metabolic models of individuals strains and entire communities, the microbiome pipeline in KBase also includes some comparative genomics tools, which can be used to assign functional roles to members of a microbiome and compare that genetic complement against the rest of the microbiome and against other sequenced genomes, including from the RefSeq database. At the single-gene level, KBase includes the BLAST+ suite [19] and HMMER [20] apps to search genomes for specific protein or gene sequences. KBase also includes the MUSCLE [21] app for multiple-sequence alignment (MSA), gBlocks [22] for trimming MSAs, and FastTree-2 [23] for phylogenetic reconstruction of an MSA to an evolutionary tree model. At the level of entire genomes, KBase includes the app *Compare Two Proteomes*, which computes bidirectional best hits between two genomes. To find orthologs among many genomes at once, KBase includes the OrthoMCL algorithm [24] which produces a pangenome that can be visualized and divided into sets of genes in

several ways, including a Pangenome Circle Plot (Figure 4).

Functional gene and domain family profiling is offered in several namespaces including The SEED [25], COG [26], Pfam [27], and TIGRFAMs [28]. Domains are summarized as a heatmap or numerically with the domain-based *View Function Profile* app (Figure 5). Finally, to determine the species of assembled genomes and identify nearby reference genomes for comparison, KBase includes an app called *Insert Genome into Species Tree* [23]. Code cells can be used in combination with these apps to identify functions or gene families that have been enriched between two metagenomes. This functionality will continue to improve as the microbiome pipeline in KBase undergoes further development.


***Figure 5.*** *The domain-based View Function Profile app displays functional gene and domain family profiles as a heatmap.*

# 3. Examples of scientific use cases for KBase metagenome analysis

The microbiome pipeline in KBase, though still quite new, has been applied to analyze microbiome systems at varying levels of complexity ranging from two species to 75 species. Here we explore three of these studies in detail: (1) predicting trophic interactions in a simple multi-species soil isolate to explore why the isolate forms a stable community; (2) modeling a 13-species electrosynthetic microbiome [29]; and (3) modeling interactions between heterotrophs and autotrophs in a hypersaline lake microbiome [15, 30, 31]. These applications demonstrate most steps of KBase's microbiome analysis pipeline, showing the insights exposed by the pipeline in each microbiome system.

## 3.1 Predicting trophic interactions linking tightly coupled species in a soil isolate

In a recent study, the KBase microbiome platform was applied to study a large number of soil isolates collected from the Cedar Creek Ecosystem Science Reserve (https://www.cedarcreek.umn.edu/). These isolates were selected from soil beneath specific plant species to explore potential variations in species type and function.

While there were significant ecological goals driving the design of this study and the selection of specific consortia for analysis, here we will focus on the analysis and modeling of a single isolate that turned out to be comprised of two distinct, tightly coupled species. We started with the assembly of our selected isolate using all three of the metagenome assembly algorithms in KBase. A QUAST comparison of these assemblies revealed that the IDBA assembler worked best on this specific data. CheckM was then run on the contigs produced by IDBA, revealing one complete genome, but also

exposing a very large fraction of contamination in the sample. This output demonstrated that this sample actually contained multiple species (a surprise considering these cells were harvested from a single colony on our soil extract plate). Next, the MaxBin tool was applied, which reported the presence of two distinct bins, *Flavobacterium* and *Burkholderiales*. These bins were extracted into separate assemblies in KBase, and then the KBase genome annotation app was applied. Finally, both bins were individually inserted into a species tree using the *Insert Genome Into Species Tree* app, which identified the species of bin 1 and bin 2 to be *Flavobacterium reichenbachii* and *Variovorax paradoxus* respectively.

At this stage, we wanted to understand why these two genomes were so tightly coupled that they would survive as a consortium through multiple rounds of plating and culture. To accomplish this, metabolic models of each bin were constructed, and then merged into a community model. This community model was then gapfilled with glucose minimal media, which resembled the media used to culture this consortium in our batch reactor system. The gapfilling added a variety of reactions to each species in the consortium, distributing functions based on the presence of pathways in each species model. Finally, we ran flux balance analysis, which revealed a fascinating series of trophic interactions. First, one species, *V. Paradoxus*, consumed all the glucose, producing malate and succinate as byproducts, which *F. reichenbachii* then consumed. *V. Paradoxus* also produced just enough aspartate and nicotinamide to meet the need for these compounds by *F. reichenbachii*. This reveals a mutual trophic interaction between *F. reichenbachii* and *V. Paradoxus*, with *F. reichenbachii* consuming a potentially inhibiting acidic byproduct of *V. Paradoxus* while *V. Paradoxus* produces two essential metabolites for *F. reichenbachii* (see Figure 6). All steps of this analysis are shown in the KBase Narrative https://narrative.kbase.us/narrative/ws.24499.obj.88. Overall, this study demonstrates how the KBase microbiome analysis pipeline can predict potential species interactions from raw sequence data to provide a mechanistic explanation of an observed ecological behavior of a very simple two-member microbial consortium. The next example case study demonstrates the application of KBase to a more complex 13-member consortium.

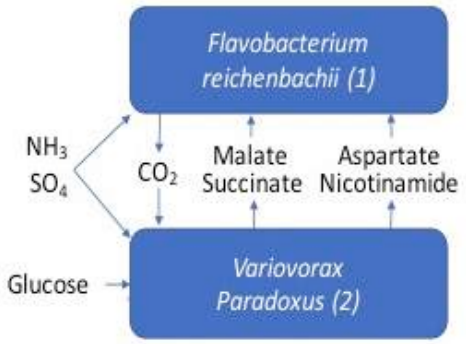


***Figure 6.*** *A map of trophic interactions predicted between species in a simple soil consortium.*

## 3.2 Modeling species interactions in an electrosynthetic microbiome

In a recent published study [29], researchers applied KBase annotation and metabolic modeling pipelines to analyze a 13-species electrosynthetic community that captures electrons from a cathode and fixes carbon dioxide. Metabolic models of the predominant community members belonging to *Acetobacterium*, *Sulfurospirillum*, and *Desulfovibrio* revealed that *Acetobacterium* is the primary carbon fixer for the community, excreting large amounts of acetate which serves as the main carbon source for the rest of the community. Using model predictions produced by flux balance analysis coupled with differential gene expression data, the researchers inferred interactions between the diverse members of this community.

To perform this anslysis, the researchers selected the three most abundant members from this 13-species community based on the relative abundance data (*Acetobacterium*, *Sulfurospirillum*, and *Desulfovibrio)* and imported their genomes into KBase followed by constructing draft metabolic models with the *Build Metabolic Model* app. These models were extensively curated in order to accurately represent the key metabolic pathways. Next, the *Acetobacterium* model was gapfilled in the presence of carbon dioxide and hydrogen (or electrons), where carbon dioxide serves as the sole

carbon source while hydrogen/electrons serve as the electron donor. *Sulfurospirillum* and *Desulfovibrio* were gapfilled in acetate minimal media where these organisms use acetate as the sole carbon source. Metabolic flux analysis was run for each model; the researchers then used the *Compare Flux with Expression* app to evaluate the flux distributions, gapfilled reactions and the differential expression data in order to determine the agreement of the model predictions with reactions based on differentially expressed genes. Trophic interactions predicted between these three species based on metabolic model analyses are shown in Figure 7. The analysis workflow for this study can be found in https://narrative.kbase.us/narrative/ws.15248.obj.1.

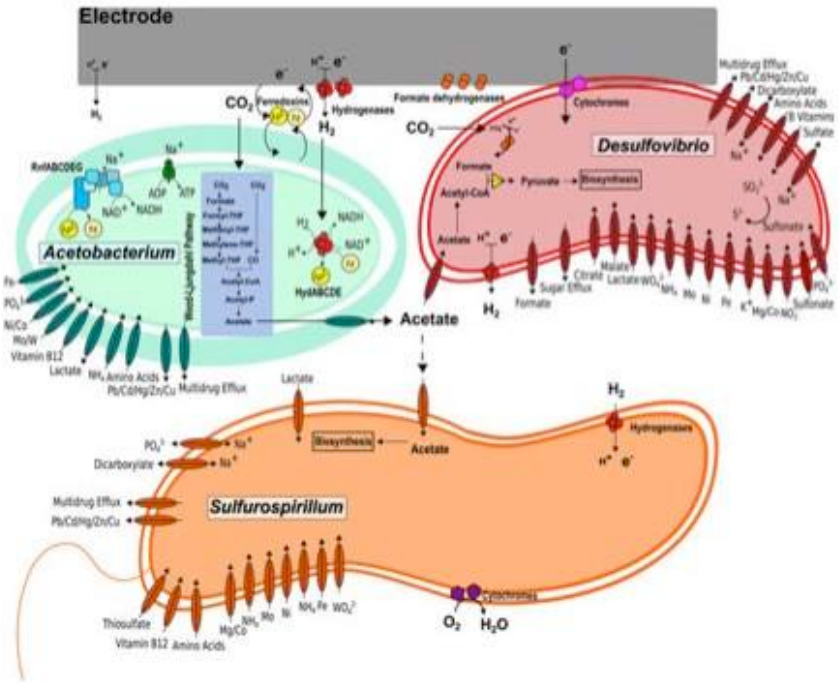


***Figure 7.*** *Interactions predicted between three key species in an electrosynthetic microbiome.*

When this electrosynthesis analysis was originally performed, KBase lacked any tools for metagenome assembly and binning, so those steps had to be performed outside of the system. Now that these tools are available in KBase, we applied them to the same reads used in the original study (see https://narrative.kbase.us/narrative/ws.23992.obj.1). The analysis process started with uploading metagenomics sequencing data into KBase followed by de novo assembly of the community using the *Assemble with metaSPAdes* app. Next, we applied the *MaxBin2 Contig Binning* app, which reported 18 bins. We ran *Assess Genome Quality with CheckM* on each bin to determine the genome completeness and the degree of contamination. Due to high contamination and poor coverage, five of bins were not considered for further analysis. The thirteen bins that were selected for downstream analysis were extracted and annotated with the KBase annotation pipeline by utilizing a code cell to iteratively annotate each of the assemblies (bins). We then inserted the 13 genomes into a species tree that consists of KBase reference genomes in order to identify the closest taxonomic lineage. Based on the tree analysis data, the three newly binned genomes closely resemble the three genomes that were originally used in the study (Table 1).

| KBase species identifier | Closest species name based on the species tree | Completeness | Features | Closest initial electrosynthetic genome KBase identifier | Matching features | Matching features as percentage |
|---|---|---|---|---|---|---|
| Electrosynthesis_Bin.003 | Acetobacterium dehalogenans | 99.20 | 4537 | Acetobacterium_MES1 | 4291 | 94.58 |
| Electrosynthesis_Bin.001 | Sulfurosprillium multivorans DSM 12446 | 99.30 | 2775 | Sulfurosprillium_MES13 | 2679 | 96.54 |
| Electrosynthesis_Bin.002 | Desulfovibrio desulfuricans subsp. desulfurcans ATCC 27774 | 99.70 | 3916 | Desulfovibrio_high-cov_MES4 | 3515 | 89.76 |

*Table 1. Comparison of KBase-produced genomes with genomes produced by an external pipeline.*

This demonstrates that KBase could be used today to repeat the electrosynthesis analysis starting with raw reads. Now, with the metagenome assembly and binning apps in KBase, the entire workflow can be performed on a single integrated platform. This integration brings significant benefits: since all work can be performed within a single system, there is no need to transfer data between multiple

systems, provenance is now tracked throughout the study, and if there are any updates to the data or tools the analyses can easily be rerun, thanks to KBase's Narratives which make it easy to reproduce an analysis workflow.

## 3.3 Modeling species interactions in an autotrophic microbiome

The KBase microbiome analysis pipeline has also been applied by a team at Pacific Northwest National Laboratory to investigate a cyanobacterial mat community that forms seasonally in the mixolimnion of an epsomitic, hypersaline lake (Hot Lake in Oroville, WA). Despite great variation in the environmental parameters (e.g., temperature and salinity), the composition of the Hot Lake mat changes little over the course of a year [30]. These mats also display limited taxonomic diversity overall (500 and 1000 operational taxonomic units throughout the seasonal cycle), making these systems ideal candidates for analysis by the microbiome analysis pipeline in KBase. In this study, two unicyanobacterial consortia (UCC-A and UCC-O) were developed by physical isolation of filaments from two morphologically distinct regions of a mat. These two physical samples were subsequently subjected to shotgun metagenome sequencing, assembly, and contig binning outside of KBase using a highly customized approach [31]. The results of this analysis revealed two consortia, each dominated by a single autotrophic cyanobacteria strain and 14-15 heterotrophic species. While the cyanobacteria in these consortia were very distinct, the two consortia overlapped almost completely in their heterotrophic membership. The genomes assembled from this analysis were subsequently loaded into KBase, which was then used to construct metabolic models for each species. A Narrative containing these genomes and linking to these models can be found here: https://narrative.kbase.us/narrative/ws.24526.obj.46. This team also applied the KBase modeling pipeline to predict interactions between species in a simplified two-member microbiome system comprised of one heterotrophic species (*Meiothermus ruber*) and one autotrophic species (*Thermosynechococcus elongatus* BP-1)—see http://www.kbase.us/predict-interspecies-interactions/ for more information about this study [15].

| KBase species identifier | Closest species name based on the species tree | Metagenomic Sample* | Complete-ness | Features | Closest PNNL genome KBase identifier | Matching features | Matching features (%) |
|---|---|---|---|---|---|---|---|
| 2370.4.1902.CGATGT_Bin.007 | Synechococcus sp. PCC 7355 | 1 | 99.9 | 4705 | bin16_bin16.genome | 4074 | 86.59 |
| 2370.4.1902.GATCAG.reads_Bin.003 | Synechococcus sp. PCC 7355 | 2 | 97.8 | 5640 | bin11_bin11.genome | 4812 | 85.32 |
| 2370.4.1902.GATCAG.reads_Bin.015 | Algoriphagus marincola HL-49 | 2 | 99.6 | 4132 | bin10_HL-49_bin_10HL-49.genome | 3762 | 91.05 |
| 2370.4.1902.CGATGT_Bin.003 | Algoriphagus marincola HL-49 | 1 | 38.5 | 1350 | bin01.bin01.genome | 1191 | 88.22 |
| 2370.4.1902.CGATGT_Bin.004 | Algoriphagus marincola HL-49 | 1 | 57.5 | 1458 | bin01.bin01.genome | 1365 | 93.62 |
| 2370.4.1902.CGATGT_Bin.006 | Idiomarina sp. A28L | 1 | 99 | 2567 | bin02_HL-53_bin02_HL-53.genome | 2488 | 96.92 |
| 2370.4.1902.CGATGT_Bin.014 | Marinobacter sp. ES-1 | 1 | 96.4 | 7189 | bin14_HL-55_bin14_HL-55.genome | 5327 | 74.10 |
| 2370.4.1902.CGATGT_Bin.005 | Marinobacter sp. HL-58 | 1 | 99.5 | 3928 | bin03_HL-58_bin03_HL-58.genome | 3816 | 97.15 |
| 2370.4.1902.CGATGT_Bin.011 | Holomonas sp. HL-48 | 1 | 97.3 | 4070 | bin06_HL-93_bin06_HL-93.genome | 3734 | 91.74 |
| 2370.4.1902.GATCAG.reads_Bin.010 | Holomonas sp. HL-48 | 2 | 100 | 3527 | bin13_HL-48_bin13_HL-48.genome | 3317 | 94.05 |
| 2370.4.1902.CGATGT_Bin.015 | Porphyrobacter sp. HL-46 | 1 | 90 | 4200 | bin21_HL-46_bin21_HL-46.genome | 3216 | 76.57 |
| 2370.4.1902.CGATGT_Bin.009 | Erythrobacter litoralis | 1 | 98.6 | 2964 | bin15_HL-111_bin_15_HL-111.genome | 2860 | 96.49 |
| 2370.4.1902.CGATGT_Bin.001 | Salinarimonas rosea DSM 21201 | 1 | (contigs are very short) | 151 | bin17_HL-109_bin17_HL-109.genome | 117 | 77.48 |
| 2370.4.1902.CGATGT_Bin.002 | Salinarimonas rosea DSM 21201 | 1 | 95.9 | 3633 | bin17_HL-109_bin17_HL-109.genome | 3426 | 94.30 |
| 2370.4.1902.CGATGT_Bin.017 | Oceanicola granulosus HTCC2516 (far off - no relatively close genome) | 1 | 91.9 | 7058 | bin09_bin09.genome | 5000 | 70.84 |
| 2370.4.1902.CGATGT_Bin.016 | Octadecabacter antarcticus 238 | 1 | 90.4 | 3667 | bin08_bin08.genome | 3558 | 97.03 |
| 2370.4.1902.CGATGT_Bin.010 | Sediminimonos qiaohouensis DSM 21189 | 1 | 98.2 | 3641 | bin07_bin07.genome | 3327 | 91.38 |
| 2370.4.1902.CGATGT_Bin.013 | Roseibacterium elongatum DSM 19469 | 1 | 99.5 | 3912 | bin18_bin18.genome | 3647 | 93.23 |
| 2370.4.1902.GATCAG.reads_Bin.007 | Defluviimonas sp. 20V17 (far off - no relatively close genome | 2 | 98.9 | 3610 | bin12_bin12.genome | 3436 | 95.18 |
| 2370.4.1902.CGATGT_Bin.012 | Defluviimonas sp. 20V17 (far off - no relatively close genome | 1 | 100 | 3140 | bin05_HL-91_bin05_HL-91.genome | 2987 | 95.13 |
| 2370.4.1902.CGATGT_Bin.008 | Oceanicaulis sp. HL-87 | 1 | 99.5 | 2727 | bin04_bin04.genome | 2614 | 95.86 |

* Metagenomic samples are UCC-A (1) and UCC-O (2)

*Table 2. Comparison of KBase genomes with genomes produced by a custom external pipeline.*

Now that KBase includes a pipeline for metagenome assembly and contig binning, we revisited the Hot Lake study to apply this new pipeline to the UCC-A and UCC-O metagenomes:
https://narrative.kbase.us/narrative/ws.24490.obj.1 - UCC-A sample
https://narrative.kbase.us/narrative/ws.24509.obj.1 - UCC-O sample
https://narrative.kbase.us/narrative/ws.24535.obj.85 - proteome comparisons

Our analysis revealed 18 distinct genome bins found in the metagenome sample, which matched the number found in the original study done by the PNNL team. We matched each of our bins to the closest bin provided by the PNNL team and conducted a detailed comparison. Nearly all genomes were extremely similar, with >90% of genes overlapping (see Table 2). It is expected that the PNNL genomes will be better than those obtained directly from metagenome sequences, as many of these genomes were produced from subsequent species isolation and pure sequencing analysis. Thus, the genome similarity in this study provides another excellent validation of the metagenome assembly and binning capabilities in KBase.

# 4. Conclusions and Future Plans

In our previous reports, we explained how the KBase platform can take genomes as input, and through a process of annotation, model reconstruction, gapfilling, and community modeling, predict species interactions as an output. Previously, this pipeline had the limitation of being unable to start with raw metagenome reads. In the three case studies described above, we demonstrated how, with the addition of the pipeline for metagenome assembly and binning, KBase can now reliably predict genome sequences from metagenome reads, and these genomes can subsequently serve as input to the community modeling and comparative genomics pipelines. The uses cases described in this report also show how code cells and batch operations in the KBase platform facilitate the application of the genome sequence analysis pipelines to the large numbers of genomes often produced from metagenome assembly and binning. These enhancements represent significant advances in KBase's functionality, unlocking the potential of the platform for analysis of metagenomic data.

Of course, this analysis has its limitations, as it is best applied to metagenomes of limited complexity in terms of microbial diversity, and the pipeline presently can only be applied to shotgun metagenome data. For this reason, our roadmap for 2018 includes a plan to extend the scientific tools in KBase to also support analysis of 16S amplicon data, as well as enabling annotation and analysis of raw metagenome reads.

**Q4 Report Authors**: Chris Henry, Dylan Chivian, Nomi Harris, José Pedro Lopes Faria, Janaka Edirisinghe
**KBase PIs**: Adam Arkin, Robert Cottingham, Chris Henry

## References

1. Arkin, A.P., et al., *The DOE Systems Biology Knowledgebase (KBase).* bioRxiv, 2016. **preprint first posted online Dec. 22, 2016**.
2. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-6.
3. Nurk, S., et al., *Assembling single-cell genomes and mini-metagenomes from chimeric MDA products.* J Comput Biol, 2013. **20**(10): p. 714-37.
4. Peng, Y., et al., *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.* Bioinformatics, 2012. **28**(11): p. 1420-8.
5. Wu, Y.W., et al., *MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.* Microbiome, 2014. **2**: p. 26.
6. Parks, D.H., et al., *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.* Genome Res, 2015. **25**(7): p. 1043-55.

7. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.
8. Henry, C.S., et al., *High-throughput generation, optimization, and analysis of genome-scale metabolic models.* Nature Biotechnology, 2010. **Nbt.1672**: p. 1-6.
9. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2016: Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
10. Martin, M., *Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads.* EMBnet.journal, 2011. **17**(1): p. 10-12.
11. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
12. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-5.
13. Kang, D.D., et al., *MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities.* PeerJ, 2015. **3**: p. e1165.
14. Faria, J.P., et al., *Constructing and Analyzing Metabolic Flux Models of Microbial Communities*, in *Hydrocarbon and Lipid Microbiology Protocols. Springer Protocols Handbooks*, M. T., T. K., and N. B., Editors. 2016, Springer: Berlin, Heidelberg. p. 247-273.
15. Henry, C.S., et al., *Microbial community metabolic modeling: A community data-driven network reconstruction.* J Cell Physiol, 2016.
16. Latendresse, M., *Efficiently gap-filling reaction networks.* BMC Bioinformatics, 2014. **15**: 225.
17. Dreyfuss, J.M., et al., *Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM.* PLoS Comput Biol, 2013. **9**(7): p. e1003126.
18. Marshall, C., et al., *Electron transfer and carbon metabolism in an electrosynthetic microbial community.* mSystems, 2016. **Submitted**.
19. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: 421.
20. Eddy, S.R., *Accelerated Profile HMM Searches.* PLoS Comput Biol, 2011. **7**(10): p. e1002195.
21. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
22. Talavera, G. and J. Castresana, *Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.* Syst Biol, 2007. **56**(4):564-77.
23. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
24. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.
25. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.
26. Galperin, M.Y., et al., *Expanded microbial genome coverage and improved protein family annotation in the COG database.* Nucleic Acids Res, 2015. **43**(Database issue): p. D261-9.
27. Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Res, 2016. **44**(D1): p. D279-85.
28. Selengut, J.D., et al., *TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes.* Nucleic Acids Res, 2007. **35**(Database issue): p. D260-4.
29. Marshall, C.W., et al., *Metabolic Reconstruction and Modeling Microbial Electrosynthesis.* Sci Rep, 2017. **7**(1): p. 8391.
30. Lindemann, S.R., et al., *The epsomitic phototrophic microbial mat of Hot Lake, Washington: community structural responses to seasonal cycling.* Front Microbiol, 2013. **4**: p. 323.
31. Nelson, W.C., et al., *Identification and Resolution of Microdiversity through Metagenomic Sequencing of Parallel Consortia.* Appl Environ Microbiol, 2015. **82**(1): p. 255-67.