

KBase

PREDICTIVE BIOLOGY



DOE Systems Biology Knowledgebase

**DOE BSSD Performance Management Metrics Report FY 2025 Q2:
Strategies developed within KBase to allow users to collaborate in team-oriented science**

Authors: Elisha-Wood Charlson¹ (elishawc@lbl.gov), Chris Henry² (chenry@mcs.anl.gov), Gazi Mahmud¹ (GaziMahmud@lbl.gov), Paramvir Dehal¹ (psdehal@lbl.gov), Roy Kamimura¹ (royk@lbl.gov), Adam Arkin¹ (aparkin@lbl.gov)

¹Lawrence Berkeley National Laboratory, Berkeley, CA 94720 and ²Argonne National Laboratory, Argonne IL 60439

KBBase's Mission: Data Interoperability and Collaborative Team-Oriented Science

When KBBase was conceived, it came with the fundamental goal of accelerating sophisticated analysis and modeling of biological systems within the context of their environment(s). Our premise was to provide access to 1) high-performance computing resources, 2) advanced suites of analytical tools that could be effectively and easily chained together, and 3) systems that make data and analyses FAIR (Findable, Accessible, Interoperable, and Reusable, [\(1\)](#)) for our users. It was also important that mechanisms were in place to ensure data could be “integrated” and made comparable among programs, and organized such that the relationships among the data could be identified. This well-structured and labeled biological information, alongside provenanced and reproducible tool chains, forms the basis for what is becoming principled, automated AI-supported workflows, which will be described in our Q4 report. One of the KBBase's core strengths is our ability to enable data interoperability across diverse environmental and technological biosciences research domains. Seamless integration and standardization of data from multiple sources enables collaboration among researchers across various DOE facilities and beyond. Data interoperability can accelerate insights and discoveries and reduce data siloing. By connecting data interoperability with a platform that supports a culture of sharing data, KBBase accelerates scientific progress.

In the FY2025 Q1 PMM, we focused on KBBase's current and future plans to support data interoperability. We also discuss how KBBase goes beyond data interoperability, loosely defined as data discovery and access using common standards, towards data *integration*, where data can be made Comparable and Organized for Predictive science by an Engaged community (COPE). Finally, we laid out ongoing plans to modernize our KBBase Central Data Model (CDM) and the KBBase architecture to support COPE and new ML/AI technologies. **For FY2025 Q2 PMM, we will focus on technical components specific to supporting collaborative research, as well as strategies to address sociological/cultural barriers to data interoperability, data integration, and team-oriented science in KBBase.** KBBase addresses these challenges by providing easy means for sharing data and analyses at multiple levels of granularity (single individuals, an organization, or publicly), while maintaining formal lines of credit for ownership and contributions. We also work closely with key scientific teams and the wider community to develop social and technical standards for effective collaborative science within and among teams and people. We describe these approaches here. Finally, we highlight several science use cases showcasing how scientists have used KBBase for collaborative, integrative biology.

KBBase's Current Approach to Team-Oriented Science

Within our remit to serve the biological data science community, the KBBase platform supports a wide array of programs to address research questions. Our focus is to enable team-oriented science by empowering research teams to explore, analyze, and share complex scientific samples, data, and tools in a collaborative environment that tracks provenance and supports reproducibility.

Upon publication, both data and analysis workflows are made public and given a Digital Object Identifier (DOI) to cite in primary scientific articles. Creating a citable data/workflow enables KBBase to rapidly and transparently track data reuse, which gives credit to teams that share their data/analyses, demonstrates data management best practices, and builds trust in open science.

Samples to Support Collaborative Science in KBBase


Biological systems are studied by making observations, taking measurements, and building models. Every study begins with a “sample” as an expression of what is being observed, measured, or modeled (from molecules to ecosystems). This makes “samples” difficult to represent, so KBBase has partnered with other DOE Office of Science data platforms in the Biological and Environmental Research (BER) program to support how scientists define “samples”. Our current Samples framework enables users to upload samples with a variety of templates, including the International General Sample Number (IGSN) standard template, created by the System for Earth and Extraterrestrial Sample Registration (SESAR²), which enables samples to get a DOI from DataCite for citation and tracking ([2](#)). Once samples are uploaded into KBBase, each has a landing page displaying who, when, and where the sample was collected, and any linked data derived from the sample within the KBBase platform (Figure 1).

To improve sample linking and discoverability, KBBase is part of a broader effort to integrate samples and data across BER data platforms, the BER Bio-Eco Data Integration project, which includes KBBase, Joint Genome Institute (JGI), National Microbiome Data Collaborative (NMDC), Environmental Molecular Sciences Laboratory (EMSL), and Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE). While the BER Bio-Eco Data Integration effort is broader than just standardizing sample metadata and linking, KBBase is working closely with that effort to establish the NMDC Sample Submission Portal as the main entry point for all BER sample metadata. KBBase connects directly to the NMDC Application Programming Interface (API) via the Data Transfer Service (DTS), and can receive sample metadata for any requested NMDC data objects. This feature ensures sample metadata can be automatically linked to the NMDC data objects once they are imported into a KBBase Narrative. We are currently conducting user research to streamline the user experience. The long-term goal of this collective effort is to enable researchers to create a set of samples that address a

research question, quickly discover and access any linked data generated from those samples across all BER data platforms, and bring that data into KBase for rapid, reproducible analysis and sharing.

KBase Sample View for "NASQAN2010_172"

Name NASQAN2010_172
ID NASQAN2010_172

Owner  GROWdb data account
growdb
Last Saved Feb 15
Versions 1

Sample Geolocation Linked Data Access

Description	Template	SESAR
	IGSN	IEGRW002I
	Material (SESAR)	Liquid>aqueous
	Field name (informal classification)	surface water
	Other name(s)	MississippiThebes_USGS_stn_07022000_06Dec2010_N
	Collection method	Depth-integrated
	Collection method description	Water samples were collected with isokinetic, depth-integrated sampling method of the U.S. Geological Survey (DOI: 10.3133/twri09A4). DNA samples were collected on 0.2 micron pore-size Sterivex filters (Millipore, Billerica, MA, United States), and preserved and extracted following Fortunato and Crump (2011; DOI 10.1007/s00248-011-9805-z).
	Purpose	microbial sampling of rivers for ROMEO
Geolocation	Latitude	37.216 degrees
	Longitude	-89.468 degrees
	Primary physiographic feature	stream
	Name of physiographic feature	Mississippi River
	Country	United States
Collection	Field program/cruise	ROMEO
	Collector/Chief Scientist name	Byron C. Crump
	Collection date	2010-12-06 00:00:00

Figure 1. Public view of sample information available in KBase.

KBase Brings Together Diverse Public Data to Enable Collaborative Research

Biological systems require a lot of data to understand. Reference data provides a consistent, standardized baseline that can be built upon quickly. As such, KBase provides a centralized location for many key reference data sources, including microbial, fungal, and plant genomes from the National Center for Biotechnology Information (NCBI) and several JGI data portals, references for various culture media formulations, and pre-loaded community ontologies that

provide standardized labels for making data Comparable and Organized (the C & O in COPE). Overall, KBase provides access to ~25TB of public data, with reference data and shared user data combined. The > 27,000 KBase users have shared more than 17,000 KBase Narratives, our reproducible data analysis notebooks, with collaborators or with the entire KBase community.

One important component of data science is being able to organize data (the O in COPE). KBase Organizations help projects with one critical aspect of that, by giving team members a central location to share KBase Narratives within a team/collaboration (Figure 2). KBase currently supports ~260 public Organizations, ranging from BER Science Focus Areas (SFAs) like ENIGMA to university courses that use KBase for professional development. Organizations can also be kept private, so they are not discoverable by other KBase users. These are typically preferred by industry partners that use KBase for data analysis.

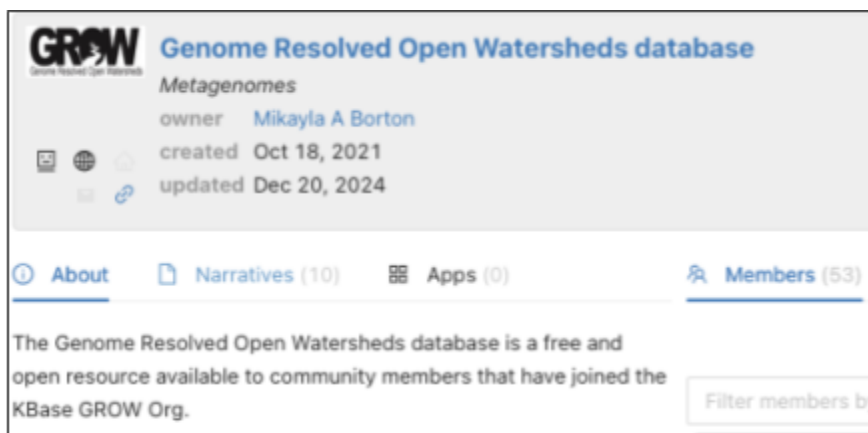


Figure 2. Example of a KBase Organization, with a summary of the group, a tab for shared Narratives, and list of members.

KBase also hosts useful public datasets that are either uniquely integrated within KBase or that lack a home elsewhere. Examples of this include: (1) a series of narratives containing [Web of Microbes](#) exometabolomic (3) data integrated with associated genomes, annotations and metabolic models; (2) a large [compedia of assembled Biolog phenotype data](#) with associated genomes; and (3) a collection of [published metabolic models](#) of central carbon metabolism (4).

KBase Supports Development and Sharing of Open-Source Analysis Tools Built by Teams

One of the pillars of KBase is that the platform's graphical user interface (GUI) makes it possible for anyone to run complex bioinformatics analyses without needing to learn computer programming. The KBase Software Development Kit (SDK) enables software developers on the KBase team and our broader KBase developer community to wrap analysis tools for deployment in the KBase Narrative interface. With guidance on how to define inputs/outputs, the KBase SDK ensures all tools on the platform can detect data types appropriate for analysis, and generate data products that are interoperable with downstream analysis tools.

Many of KBase’s open-source tools are the product of team science. In partnership with our User Working Groups (<https://www.kbase.us/research/user-working-groups>), KBase has enabled more than a dozen projects to co-develop analysis tools for the broader KBase community. Functionality includes tools to process long-read sequence data, assign taxonomic labels to microbial genomes, such as the Genome Taxonomy Database (GTDB, [5](#)), and tools for an array of functional annotation such as Distilled and Refined Annotation of Metabolism (DRAM, [6](#)), omics-enabled global gapfilling (OMEGGA) for predicting metabolic networks, and Snekmer for targeted annotation or generating novel protein families (Table 1).

Table 1. Top UWG tools integrated into KBase as part of a team-oriented science effort.

Tool / Package	Collaborator	UWG site links	Number of times run since release
Long-read sequence analysis	ENIGMA SFA	https://www.kbase.us/research/adams-sfa	Released: 2024 Unique users: 427 App runs: 2452
Genome Taxonomy Database (GTDB) (5)	University of Queensland, Australia	https://narrative.kbase.us/collections/GTDB (login required)	Released: 2020 Unique users: 6147 App runs: 66326
Distilled and Refined Annotation of Metabolism (DRAM) (6)	Microbial ecoSystems Laboratory	https://www.kbase.us/research/wrighton	Released: 2020 Unique users: 2686 App runs: 34444
Omics-enabled global gapfilling (OMEGGA)	Soil Microbiome SFA	https://www.kbase.us/research/hofmockel-sfa	Released: 2024 Unique users: 569 App runs: 10091
Snekmer (7)	Persistence Control SFA	https://www.kbase.us/research/egbert-sfa	Released: 2024 Unique users: 47 App runs: 292

The power of these tool integrations is in the synergy offered by their interactions. For example, in collaboration with the uBiosphere’s SFA, we added the capability to load multiple independent function annotations into a KBase genome, and to import annotations obtained outside of KBase (e.g., Blast Koala, NetGo, or DEEPEC) into the KBase platform. At the same time, the Persistence Control SFA and KBase collaborators (Miller Lab at UC Denver and Wrighton Lab at CSU) added new annotation algorithms to the KBase platforms (Snekmer, DRAM, GLM4EC). We also collaborated with the Protein Data Bank (PDB) team to integrate a tool that annotated genomes with functions and structures from PDB. Together, all of these tools produce numerous alternative potential hypotheses of the possible functions for the genes in a genome. These hypotheses then feed into algorithms like OMEGGA (Soil Microbiome SFA) and the

Exometabolomics Fitting App (Northen Lab) by offering candidate genes for gap-filled reactions based on the competing annotations from these other algorithms. These candidate genes greatly improve the quality of solutions proposed by gapfilling, as well as offering gene targets for experimental validation of model predictions. Now that this integrated synergistic set of tools is in place, contributed by teams from ten different labs across the country, most projects and teams are now applying this tool set to study their distinctive systems, including ENIGMA SFA, PMI SFA, Soil Microbiome SFA, uBiospheres SFA, and Persistence Control SFA. In all cases, teams are uploading new isolate microbial and fungal genomes, applying multiple annotation pipelines to those genomes, uploading their own multi-omics data, and using OMEGGA to gapfill models to fit growth phenotype data, ultimately improving genome annotations and models generated.

KBBase is Paving the Way for More Collaborative Team-Oriented Science

The scientific enterprise has benefited from both large collaborative, team-oriented science and small-scale, single-investigator science. However, many scientists and institutions built on single-investigator science are beginning to hit the limit of what they can accomplish solo. As new research fields move into more collaborative, team-oriented science, there has to be intentional support behind the culture change to ensure transparency and trust are built into the new paradigm. KBBase is leading the way by demonstrating how provenanced, reproducible, data science supports transparency, and good data management practices build trust by enabling credit, recognition, and tracking of impact.

As a collaborative science platform, it is imperative that our 45,000+ users have access to training and support, and feel like valued members of our community. The KBBase platform's outreach/engagement and devops teams provide:

- Regular training via workshops, webinars, and video recordings on our KBBase YouTube Channel (<https://www.youtube.com/DOEKBBase>).
- Up-to-date news (<https://www.kbase.us/news/>) regarding new feature releases, bug fixes, and highlights demonstrating how various community members are using KBBase.
- KBBase Community Developers (<https://www.kbase.us/develop>), responsible for Table 1 .
- KBBase Educators program (<https://www.kbase.us/engage/educators>) and the associated NSF-funded Microbiomes In Computational Research Opportunities Network (MICROnet; <https://www.kbase.us/engage/microbiome-training>) as a way for educators to help each other.
- KBBase Help Board (<https://www.kbase.us/support>) to ask questions.
- KBBase Users Slack group (invite by request, mostly for developers, educators) for users to interact with other KBBase users.

Our outreach/engagement teams work closely with invested users to build out best practice examples of how to leverage KBBase for team-oriented science, including creating project Organizations in KBBase, uploading and linking samples to data, sharing and collaborating in KBBase, and finally publishing KBBase Narratives in alignment with the FAIR data principles.

Many of our users' first interaction with KBBase staff is via our online training webinars, either live or via videos on our YouTube channel. Approximately half of the publications that use KBBase for data analysis have at least one author who has attended an outreach event. In speaking with users, there is often a "KBBase person" on a team or in a lab group. These individuals are tasked with using KBBase for the project's data management, data sharing, and/or data analysis, freeing up other team members to complete other parts of the project, while making it easy for them to rapidly access, explore, and analyze project data. These champions often spread KBBase usage to their new labs as they progress throughout their career. Examples include a graduate student from University of São Paulo who got a postdoctoral position at Princeton University ([8](#), [9](#)), and a new faculty member at SUNY College of Environmental Science and Forestry that originally published using KBBase while a postdoc at University of Georgia, Athens ([10](#), [11](#), KBBase Highlight: <https://www.kbase.us/news/jennifer-goff>)

The KBBase Educators program ([12](#)) was born out of the COVID-19 pandemic, which required educators to suddenly teach science remotely. Now, with representation in almost all 50 states (see [Q1 2025 PMM](#)), the KBBase Educators program has evolved into a NSF-funded Research Coordination Network, MICRONet, that supports educator training for the next generation of microbiome researchers in the full scientific method, from question and hypothesis generation, through sample collection and processing, which culminates in analysis and publishing using KBBase. The goal of MICRONet is to build sustainable, regional networks of educators that integrate KBBase into their curriculum. These networks will provide peer support and training, resource sharing, and collaborative opportunities for students to work on projects larger than a single course can offer.

Finally, the KBBase Help Board is a public forum for asking questions, reporting bugs, or requesting new features (Figure 3A). For many KBBase users, this might be the main support mechanism they seek out, with ~1000 users submitting tickets to the Help Board so far, and about half of those returning to submit additional tickets. In July 2019, we started gathering feedback from users who submitted tickets to the Help Board, with >80% of responses indicating that we are providing quality service, and that users were very likely to recommend KBBase to their colleagues (Figure 3B, C). One user in particular submitted 64 tickets, starting Dec 2016! The questions and recommendations were so valuable that we ended up hiring them on as full-time staff in 2023.

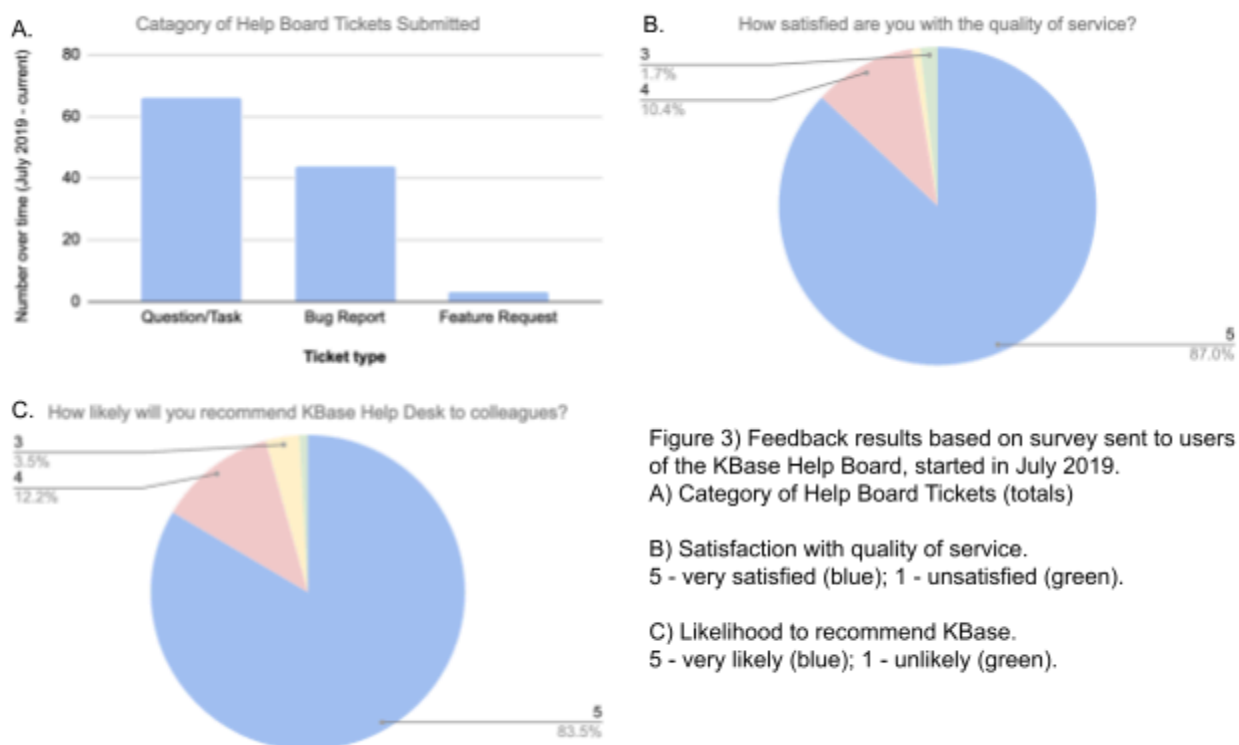


Figure 3) Feedback results based on survey sent to users of the KBase Help Board, started in July 2019.

A) Category of Help Board Tickets (totals)

B) Satisfaction with quality of service.
5 - very satisfied (blue); 1 - unsatisfied (green).

C) Likelihood to recommend KBase.
5 - very likely (blue); 1 - unlikely (green).

Collaborative Science Stories Supported by KBase

Many of the large collaborative projects that leverage KBase for sample and data analysis, sharing, and publishing are Department of Energy and academic partnerships. Mentioned previously as KBase Collections, we'll now highlight the team-oriented science by projects, such as GROW, ENIGMA, and PMI.

Genome Resolved Open Watershed database (GROWdb)

Projects working to understand ecosystem-level processes and how they respond to perturbations (fires, floods, etc) have illuminated how a lack of genome-resolved functionality outside of the core, culture-based model organisms (e.g., *Escherichia coli*) hinders the ability to model such systems. To address this problem, a collective of microbiologists that study river systems built the GROWdb, which “profiles the identity, distribution, function and expression of microbial genomes across river surface waters covering 90% of United States watersheds.” (13) With >100 teams collecting 163 samples across 106 sites in US rivers, the resulting database contains 3.8 terabases (Tb) of metagenomic and metatranscriptomic

sequencing data with extensive geochemical and geospatial measurements at each site. Many of the collecting partners are field scientists or laboratory researchers. They are not experienced data scientists. Hence, KBBase provides a needed sample and data sharing platform, coupled with easy-to-access analysis capabilities able to support terabytes of data and complex system biology research questions. In turn, this has enabled researchers in countries outside the US to better use their collected global rivers' data.

Long-read Isolate Sequencing and Assembly (LISA) workshop with ENIGMA SFA

In addition to the ENIGMA Collection, the ENIGMA SFA also partnered with KBBase to deliver a 3-day Long-Read Isolate Sequencing and Assembly (LISA) workshop with the goals of introducing the benefits of long-read sequencing to the BER research community, sharing and training protocols for data interoperability from different labs, and to establish a network of researchers using this technology for future collaborative conversations ([14](#)). Coupled with a 2-day wet-lab training, the LISA workshop has a full 3rd day dedicated to data analysis and publishing of isolate genomes using KBBase. ENIGMA contributed 4 long-read tools (see Table 1), including 2 genome assembly tools that can do *de novo* assembly of long reads or can perform a hybrid assembly using long and short sequence reads, along with 2 quality control tools to filter out short sequences and repair assembly errors. Beyond enabling members of their own SFA to accelerate and share their metagenomic research pipelines, these form a critical capability for the community. The inaugural workshop was held at LBNL in late 2023 and trained ~20 participants, mostly early-career, from 6 National labs (ANL, LBNL, LLNL, PNNL, ORNL, Sandia) and 8 universities.

Isolate Phenotype Collection, with the ENIGMA and Plant Microbe Interfaces (PMI) SFAs

The Genomic Science Program SFAs explore a range of scientific questions that pertain to the BER mission, including plant-microbe interactions, biodesign and biosecurity, and soil ecosystem health and resilience (<https://www.genomicscience.energy.gov/sfas>). Within those programs, research priorities may generate related resources, like microbial isolates, for example. In a collaborative effort to explore resource sharing and evaluate scientific reproducibility, the ENIGMA and PMI SFAs reached out to KBBase to support the development of an SFA microbial isolate genome database, and then take it a step further to associate genome predictions of function with observed phenotypes, as in growth on various carbon substrates. The goal was to share data with all SFAs, and build a set of protocols that could result in distributed phenotyping of isolate collections that could be brought into KBBase for correlation analyses and refinement of genome-based metabolic models. The teams have established joint laboratory growth assays on 24 known carbon media sources, and have established growth curve profiles for isolates that typically show measurable growth within 24 hours. The next step is to expand phenotype growth characteristics across the 24 known carbon sources for isolates

representing the breadth of the phylogenetic tree where isolates exist, and compare the empirical growth data with the predicted growth data based on the genomic sequence of the microbe alone. As with many cross-validation experiments, we have learned that small variations in laboratory protocols introduce discrepancies, and microbial diversity, as represented by isolates, is sparse and not evenly distributed. Future developments in automated laboratory practices will greatly improve our reproducibility, and the generation of AI models validated on these growth data will enable us to grow more microbes in culture, as we get better at decoding their genomic potential.

Impact of Open, Collaborative, Team-Oriented Science

The impact of the science stories described above is only beginning to emerge. With the release of the GROWdb, KBbase has been contacted by groups wanting to create similar “genome reference databases” for other targeted environments, including:

- National Energy Technology Laboratory’s Produced Water Project “to help determine options for treatment, reuse, and recovery of valuable resources”, from 38 studies across North America of Shale, Coal Bed Methane, and Crude Oil resources (<https://narrative.kbase.us/narrative/156785> requires login, full release coming soon).
- A university collaborative focused on creating a global atlas of marine and terrestrial subsurface environment microbiomes ([15](#)).
- A NSF Critical Zone project looking at defining organism interactions from bedrock to treetop (<https://criticalzone.org>).
- A start-up company interested in creating an open fermented foods database for rapid bioactive compound discovery and exploration.

For these groups, KBbase provides a dynamic environment where secondary users can quickly and efficiently begin to explore, analyse, and compare their data to these resources, without having to download 100s of files and TBs of data. As these projects mature and the range and diversity of open, team-oriented projects evolve, KBbase will need to provide access to diverse data resources and modeling capabilities, including those beyond our current platform. Those strategies will be discussed in the following Q3 report.

References

- 1) Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).
<https://doi.org/10.1038/sdata.2016.18>
- 2) Buys, M., & Lehnert, K. (2021). Partnership between IGSN and DataCite. DataCite.
<https://doi.org/10.5438/7Z70-1155>

- 3) Kosina, S.M., Greiner, A.M., Lau, R.K. et al. Web of microbes (WoM): a curated microbial exometabolomics database for linking chemistry and microbes. *BMC Microbiol* 18, 115 (2018). <https://doi.org/10.1186/s12866-018-1256-y>
- 4) Edirisinghe, J.N., Weisenhorn, P., Conrad, N. et al. Modeling central metabolism and energy biosynthesis across microbial life. *BMC Genomics* 17, 568 (2016). <https://doi.org/10.1186/s12864-016-2887-8>
- 5) Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, Philip Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy, *Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D785–D794, <https://doi.org/10.1093/nar/gkab776>
- 6) Michael Shaffer, Mikayla A Borton, Ben Bolduc, José P Faria, Rory M Flynn, Parsa Ghadermazi, Janaka N Edirisinghe, Elisha M Wood-Charlson, Christopher S Miller, Siu Hung Joshua Chan, Matthew B Sullivan, Christopher S Henry, Kelly C Wrighton, kb_DRAM: annotation and metabolic profiling of genomes with DRAM in KBBase, *Bioinformatics*, Volume 39, Issue 4, April 2023, btad110, <https://doi.org/10.1093/bioinformatics/btad110>
- 7) Chang CH, Nelson WC, Jerger A, Wright AT, Egbert RG, McDermott JE. Snekmer: a scalable pipeline for protein sequence fingerprinting based on amino acid recoding. *Bioinform Adv.* 2023 Feb 2;3(1):vbad005. <https://doi.org/10.1093/bioadv/vbad005> . PMID: 36789294; PMCID: PMC9913046.
- 8) Venturini AM, Gontijo JB, Mandro JA, Paula FS, Yoshiura CA, da França AG, Tsai SM. Genome-resolved metagenomics reveals novel archaeal and bacterial genomes from Amazonian forest and pasture soils. *Microbial Genomics* 2022 8(7), <https://doi.org/10.1099/mgen.0.000853>
- 9) Gontijo, J.B.; Paula, F.S.; Venturini, A.M.; Mandro, J.A.; Bodelier, P.L.E.; Tsai, S.M. Insights into the Genomic Potential of a *Methylocystis* sp. from Amazonian Floodplain Sediments. *Microorganisms* 2022, 10, 1747. <https://doi.org/10.3390/microorganisms10091747>
- 10) Goff, J.L., Szink, E.G., Durrence, K.L. et al. Genomic and environmental controls on *Castellaniella* biogeography in an anthropogenically disturbed subsurface. *Environmental Microbiome* 19, 26 (2024). <https://doi.org/10.1186/s40793-024-00570-9>
- 11) Jennifer L Goff, Lauren M Lui, Torben N Nielsen, Farris L Poole, Heidi J Smith, Kathleen F Walker, Terry C Hazen, Matthew W Fields, Adam P Arkin, Michael W W Adams, Mixed waste contamination selects for a mobile genetic element population enriched in multiple heavy metal resistance genes, *ISME Communications*, Volume 4, Issue 1, January 2024, ycae064, <https://doi.org/10.1093/ismeco/ycae064>
- 12) Dow EG, Wood-Charlson EM, Biller SJ, Paustian T, Schirmer A, Sheik CS, Whitham JM, Krebs R, Goller CC, Allen B, Crockett Z and Arkin AP (2021) Bioinformatic Teaching Resources – For Educators, by Educators – Using KBBase, a Free, User-Friendly, Open Source Platform. *Front. Educ.* 6:711535. doi: 10.3389/educ.2021.711535

- 13) Borton, M.A., McGivern, B.B., Willi, K.R. et al. A functional microbiome catalogue crowdsourced from North American rivers. *Nature* 637, 103–112 (2025).
<https://doi.org/10.1038/s41586-024-08240-z>
- 14) <https://sites.google.com/lbl.gov/lisaworkshop/home>
- 15) Ruff SE, de Angelis IH, Mullis M, Payet JP, Magnabosco C, Lloyd KG, Sheik CS, Steen AD, Shipunova A, Morozov A, Reese BK, Bradley JA, Lemonnier C, Schrenk MO, Joye SB, Huber JA, Probst AJ, Morrison HG, Sogin ML, Ladau J, Colwell F. A global comparison of surface and subsurface microbiomes reveals large-scale biodiversity gradients, and a marine-terrestrial divide. *Sci Adv.* 2024 Dec 20;10(51):eadq0645. doi: <https://doi.org/10.1126/sciadv.adq0645> Epub 2024 Dec 18. PMID: 39693444; PMCID: PMC11654699.