166. Design of KBase Infrastructure

Thomas Brettin*[1] (brettin@cels.anl.gov), Daniel Olson[1], Jason Baumohl[2], Aaron Best[3], Jared Bischof[1], Ben Bowen[2], Tom Brown[1], Shane Canon[1], Stephen Chan[2], John-Marc Chandonia[2], Dylan Chivian[2], Ric Colasanti[1], Neal Conrad[1], Brian Davison[4], Matt DeJongh[3], Paramvir Dehal[2], Narayan Desai[1], Scott Devoid[1], Terry Disz[1], Meghan Drake[4], Janaka Edirisinghe[1], Gang Fang[7], José Pedro Lopes Faria[1], Mark Gerstein[7], Elizabeth M. Glass[1], Annette Greiner[2], Dan Gunter[2], James Gurtowski[6], Nomi Harris[2], Travis Harrison[1], Fei He[5], Matt Henderson[2], Chris Henry[1], Adina Howe[1], Marcin Joachimiak[2], Kevin Keegan[1], Keith Keller[2], Guruprasad Kora[4], Sunita Kumari[6], Miriam Land[4], Folker Meyer[1], Steve Moulton[4], Pavel Novichkov[2], Taeyun Oh[8], Gary Olsen[9], Bob Olson[1], Dan Olson[1], Ross Overbeek[1], Tobias Paczian[1], Bruce Parrello[1], Shiran Pasternak[6], Sarah Poon[2], Gavin Price[2], Srivdya Ramakrishnan[6], Priya Ranjan[4], Bill Riehl[2], Pamela Ronald[8], Michael Schatz[6], Lynn Schriml[10], Sam Seaver[1], Michael W. Sneddon[2], Roman Sutormin[2], Mustafa Syed[4], James Thomason[6], Nathan Tintle[3], Will Trimble[1], Daifeng Wang[7], Doreen Ware[5,6], David Weston[4], Andreas Wilke[1], Fangfang Xia[1], Shinjae Yoo[5], Dantong Yu[5], **Robert Cottingham[4], Sergei Maslov[5], Rick Stevens[1], Adam P. Arkin[2]**

[1]Argonne National Laboratory, Argonne, IL, [2]Lawrence Berkeley National Laboratory, Berkeley, CA, [3]Hope College, Holland, MI, [4]Oak Ridge National Laboratory, Oak Ridge, TN, [5]Brookhaven National Laboratory, Upton, NY, [6]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [7]Yale University, New Haven, CT, [8]University of California, Davis, CA, [9]University of Illinois at Champaign-Urbana, Champaign, IL, [10]University of Maryland, College Park, MD

http://kbase.us

Project Goals: The KBase project aims to provide the capabilities needed to address the grand challenge of systems biology: to predict and ultimately design biological function. KBase enables users to collaboratively integrate the array of heterogeneous datasets, analysis tools and workflows needed to achieve a predictive understanding of biological systems. It incorporates functional genomic and metagenomic data for thousands of organisms, and diverse tools for (meta)genomic assembly, annotation, network inference and modeling, allowing researchers to combine diverse lines of evidence to create increasingly accurate models of the physiology and community dynamics of microbes and plants. KBase will soon allow models to be compared to observations and dynamically revised. A new prototype Narrative interface lets users create a reproducible record of the data, computational steps and thought process leading from hypothesis to result in the form of interactive publications.

At the core of the KBase architecture is a set of rich data models and stores, scalable computing, and workflow management. Our KBase physical infrastructure is built on the successes of DOE investment in our national scientific cyber-infrastructure and therefore leverages enormous intellectual resources present in the DOE community. Building on ESNet allows us to construct a wide area network between the partner labs that enables a virtual hardware infrastructure. Our use of cloud-computing supports development of new tools and provides compute resources for production services. The acceptance of virtualization technology is growing, and the use of machine images produced by others is already visible in our core services. Additionally, machine images are now provided which contain multiple components of the KBase infrastructure and services. Cluster Computing has long been a critical part of biological data

analysis. In collaboration with computing centers created by the Office of Advanced Computing Research such as NERSC, our underlying cluster services can leverage these resources and scale to meet needs.

KBase aims to power the next wave of biological research in DOE and beyond. Enabling these capabilities requires a software and hardware infrastructure that is integrated, extensible, and scalable. The architecture is designed to meet these needs and support user functionality to visualize data, create models or design experiments based on KBase- generated suggestions.

KBase is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research.