DOE Systems Biology Knowledgebase

KBASE

Data and modeling for predictive biology

Present and Future Computing Requirements

Tom Brettin and Shane Canon

Oak Ridge National Laboratory and

Lawrence Berkeley National Laboratory

NERSC BER Requirements for 2017

September 11-12, 2012

Rockville, MD

U.S. DEPARTMENT OF ENERGY | Office of Science

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Key Personnel

Adam Arkin

Lawrence Berkeley National Lab

PI and Science

Rick Stevens

Argonne National Lab

Infrastructure (Hardware and Software)

Bob Cottingham

Oak Ridge National Lab

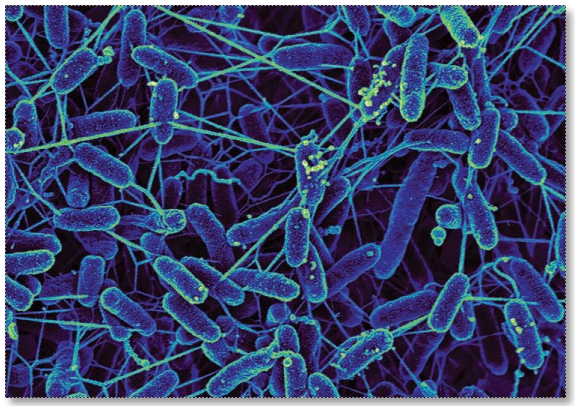
Operations

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Systems Biology Knowledge Base

Knowledgebase enabling ***predictive*** systems biology.

- Powerful modeling framework.
- **Community-driven**, extensible and scalable **open-source** software and application system.
- Infrastructure for integration and reconciliation of algorithms and data sources.
- Framework for standardization, search, and association of data.
- Enable model based **experimental design** and **interpretation** of results.



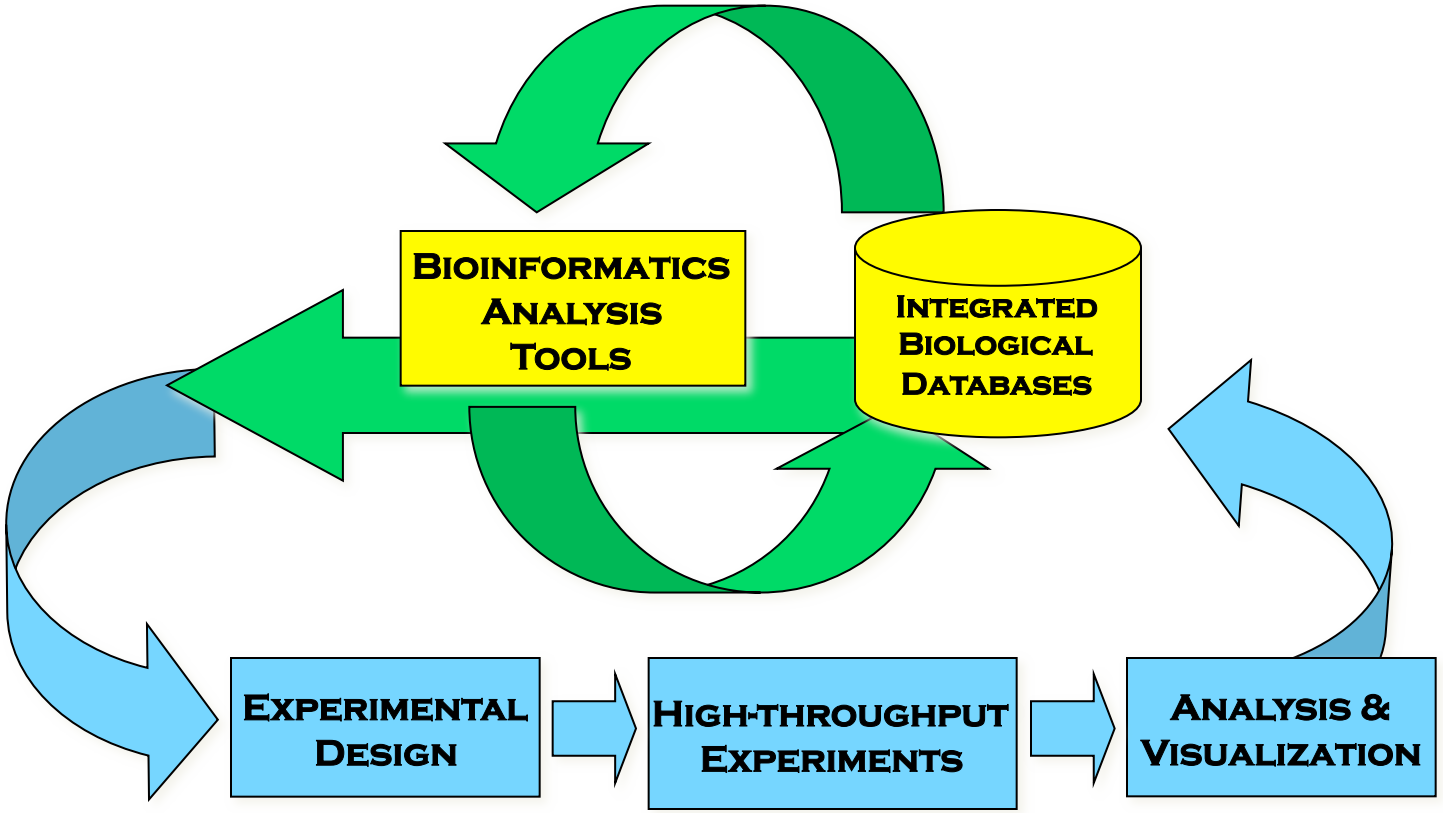
Microbes

Communities

Plants

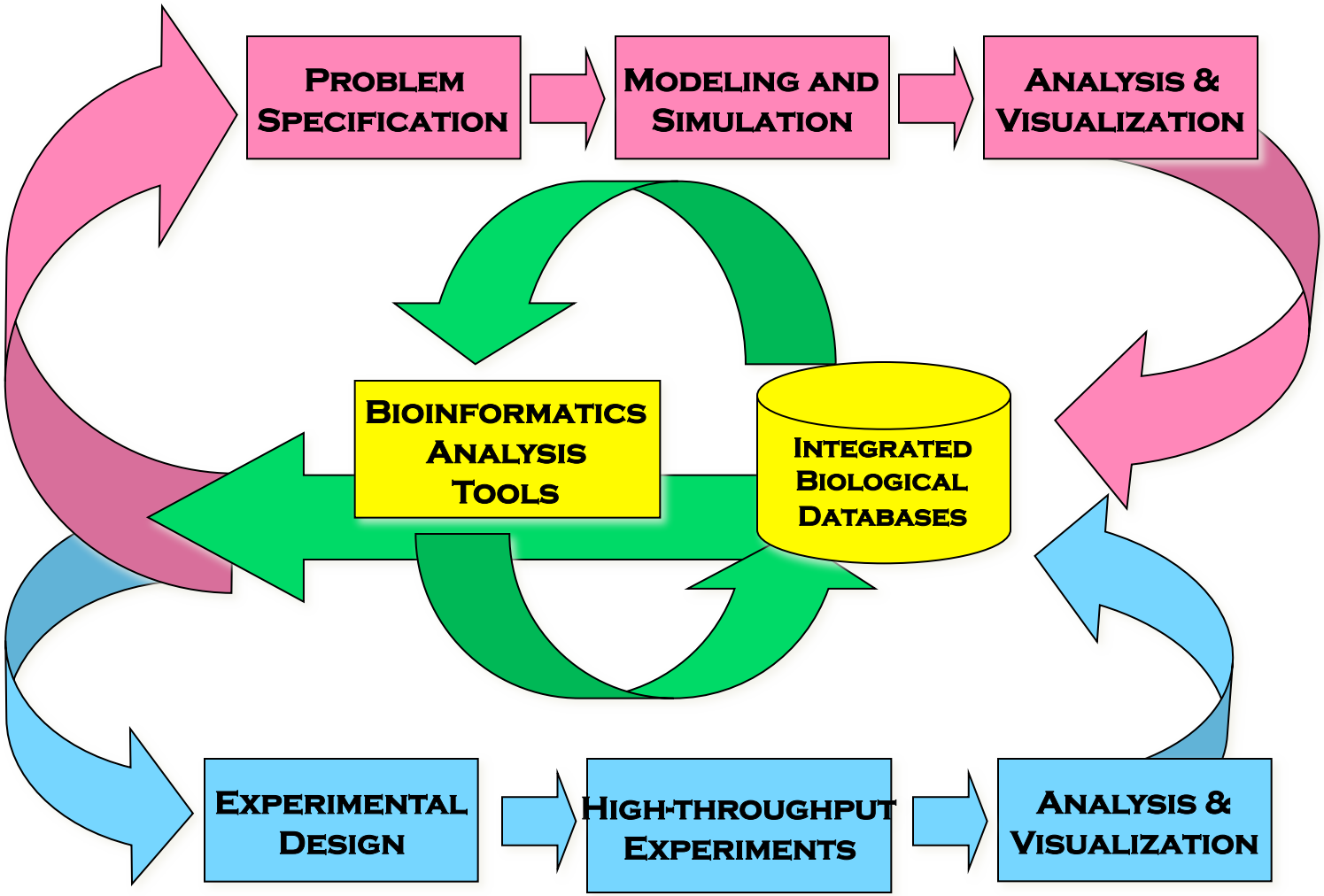
KBASE
predictive biology
DOE Systems Biology Knowledgebase

Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



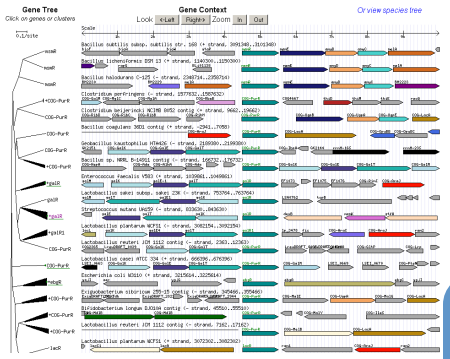
KBASE
predictive biology
DOE Systems Biology Knowledgebase

Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



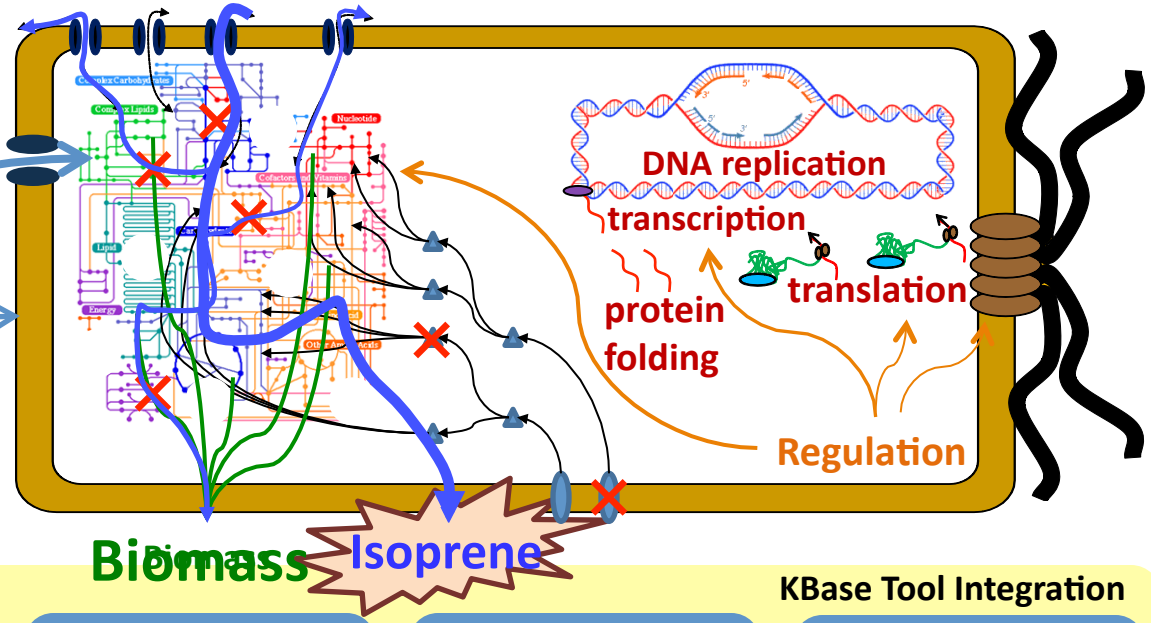
Engineering a Microbe for Biofuel Production

Annotated Genome



Feed Stock

Stresses
Hydrolysate,
pH, Salt,
End product,
intermediates



KBase Tool Integration

Annotation algorithms

Metabolic model generation

Regulatory network inference

Fitting kinetic model parameters

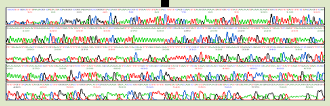
Proposing strain optimizations

Metabolic reconstruction

Model optimization algorithms

Other functional modeling

Predicting pathway fluxes



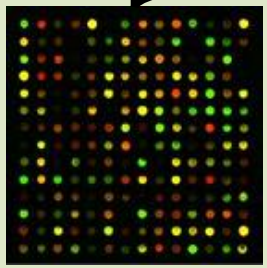
Genome Sequence

Comparative Genomics

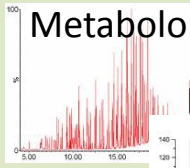
KEGG
Brenda
BioCyc
Published models

BIOLOG

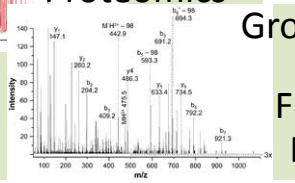
Gene KO Phenotypes



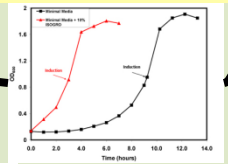
Transcriptomics



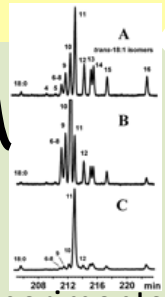
Metabolomics



Proteomics



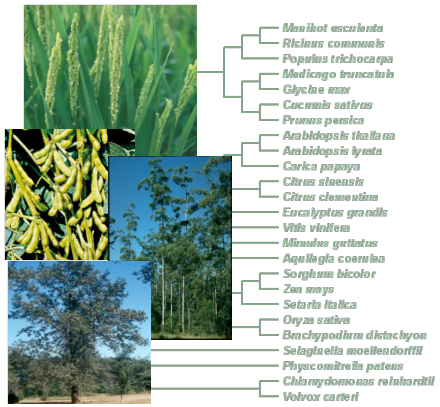
Growth curves

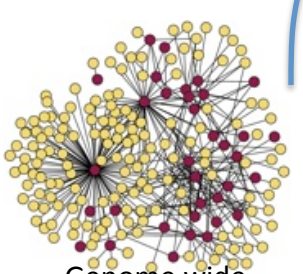


Flux tracing experiments

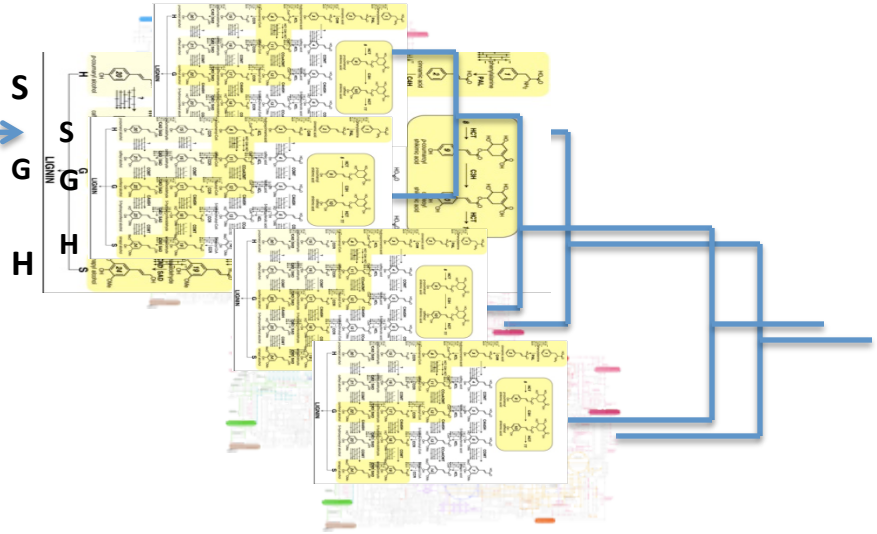
KBase Data Integration

Modifying Lignin Biosynthesis





Genome wide
Correlative analysis



SNPs3D

PolyPhen-2

SNP influenced changes in protein structure and function

Pathway predictions

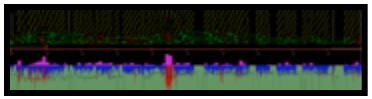
- Model optimization
- validation

Plant systems modification

- Genome annotation algorithms
- Comparative genomics

- Network inference
- Pathway reconstruction
- Omics & SNP overlay

Phylogenomics
Modeling phase I



phytozome

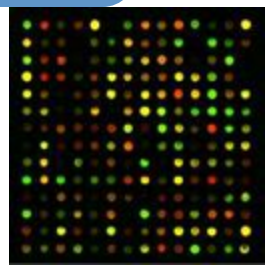
NCBI

Phenotype
Mutant population
Resequencing data

iPlant Collaborative

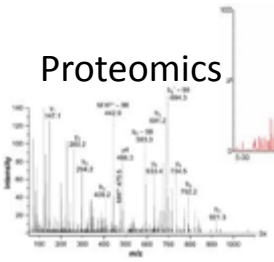
BESC BioEnergy Science Center

jbei Joint BioEnergy Institute

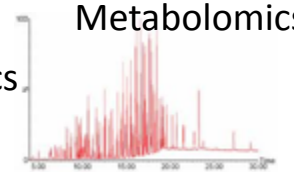


Transcriptomics

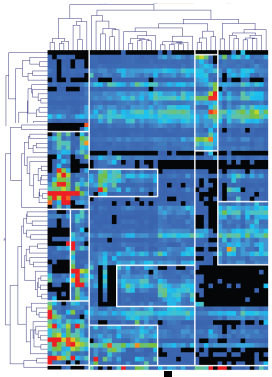
Proteomics



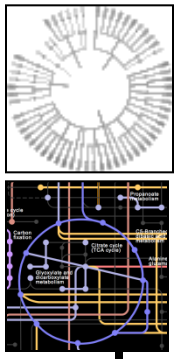
Metabolomics



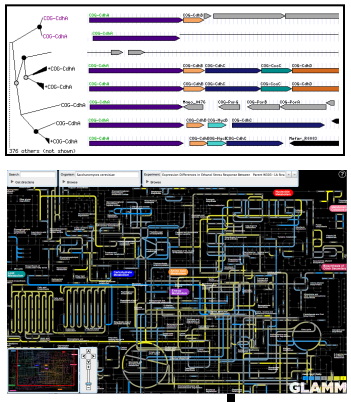
Culturing Recalcitrant Microbes from Communities



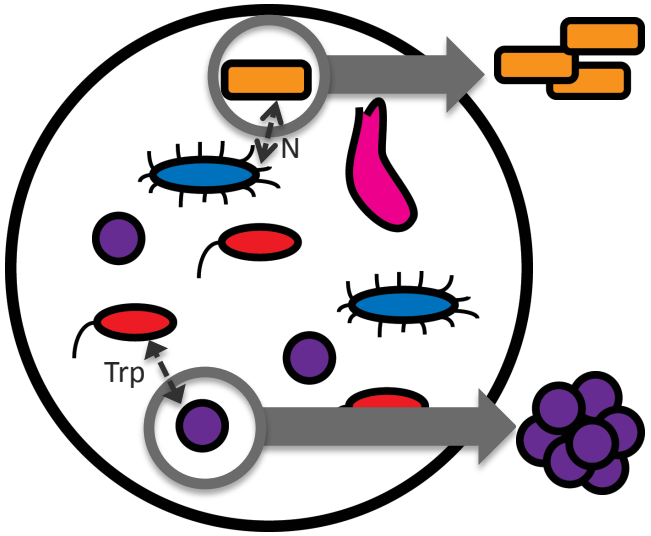
Covariation Analysis, Phylogenetically and Functionally Interesting Keystone Species



Phylogenetic Inference



Gene Functional Annotation



Differential Gene Expression

Population Statistics

Isolate Genomes and Models

Annotation and Metabolic Reconstruction

Predict Syntrophic Interactions

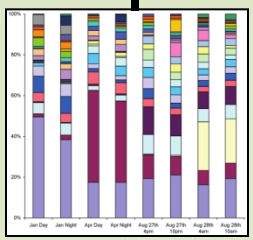
Isolate vs. Community Phenotype

Comparative Metagenomics

Genome Assembly from Metagenomics

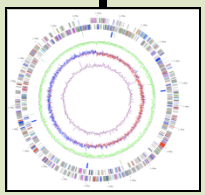
Regulation and Functional Modeling

Predict Culturing Conditions



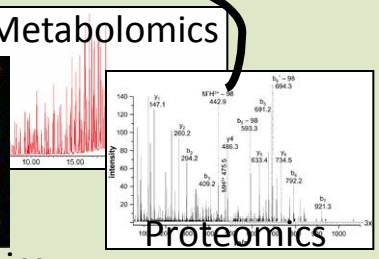
Species Abundance

Functional Gene Abundance



Phylo binning and scaffolding

Transcriptomics



Temp
pH
Salinity
Amino Acids
Cofactors
Syntrophies

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Scientific Goals 2017

- Analyze understudied microbial phyla
- Interpret metagenomics data to identify conditions required for growth
- Construct, simulate and validate plant life models
- Integrate Descriptions and Annotations of Microbial Genomic Features
- Integrate Descriptions and Annotations of Microbial Genomic Features
- Improve Plant Genome Annotation Datasets and Make Them More Accessible

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Computational Strategies

Most KBase scientific targets are data driven and for the foreseeable future, would allow their computing requirements to be met by modest-scale commodity clusters.

HOWEVER, several sub problems will require substantially more computing

- Combinatorial Analysis and Optimization of Biological Networks
- Metagenome Indexing, Assembly, and Analysis
- Computing Sequence Similarities and Indexing Kbase Reference Databases

These problems are characterized by these features:

- Matrix operations
- Large aggregate memory and tightly coupled processors
- Hardware acceleration for local alignments

These problems could benefit from:

- Low level support for many-task parallelism at OS/hardware level
- Low level support for fast string comparisons and associative memory type operations
- Hardware acceleration for local alignments, K-mer indexing and associative arrays

KBASE predictive biology
DOE Systems Biology Knowledgebase

Current NERSC HPC Usage

- To date NERSC has only been used for development and early testing
- System will be exposed via KBase Cluster Service
- Both Hopper and Carver
- Applications ported to date:
  - Sequence - aligners BLAST, BLAT
  - Assemblers - Kiki
  - Phylogenetic Clustering - RaiPHY
- In Progress
  - Metabolic Modeling - Gap Analysis
  - Assembler - Bowtie

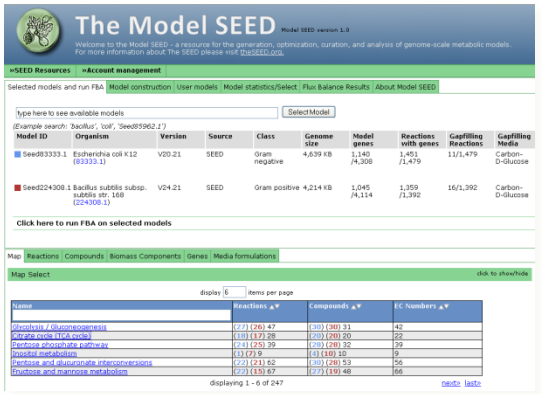
KBASE predictive biology
DOE Systems Biology Knowledgebase

HPC Requirements for 2017

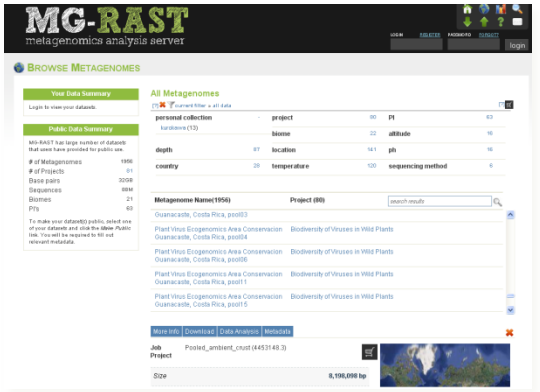
- Very difficult to predict today
  - KBase not in general release yet
  - Needs will heavily depend on demand from community
  - Dependent on how KBase infrastructure grows
  - Dependent on level of contribution from ALCF and OLCF
- Expect similar demands to JGI once KBase is generally available?

Microbes Online

Model SEED



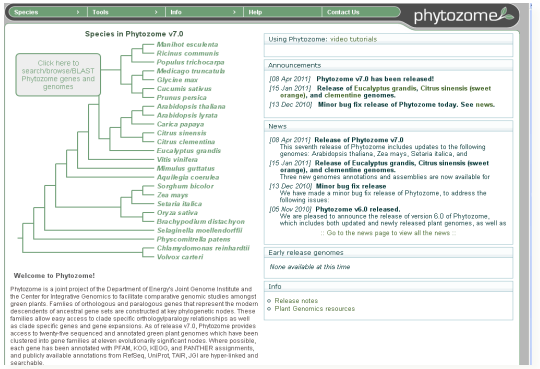
MG-RAST



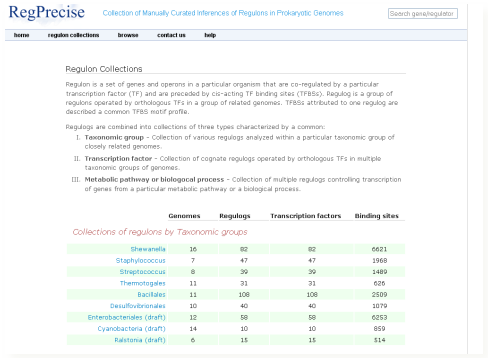
Meta Microbes Online

20,000+ users

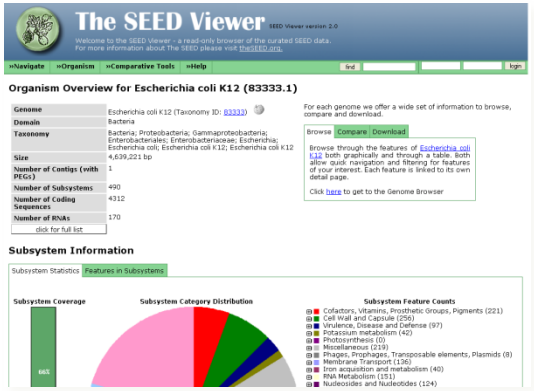
Phyotozome



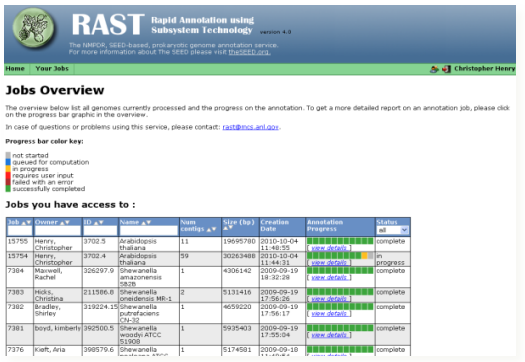
RegFam



The SEED



RAST



KBASE predictive biology
DOE Systems Biology Knowledgebase

Strategies for New Architectures

- No project plans to develop or port applications to GPU or Many Core
- Leverage community developed codes for new architectures
  - Several initiatives for GPU (especially nVidia)
  - Intel is starting to engage community
- Recent workshop on next generation sequence analysis libraries
  - Attended by vendors
  - KBase and the JGI staff represented
  - Others from the community with sequence analysis libraries and HPC experience

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Summary

What new science results might be afforded by improvements in NERSC computing hardware, software and services?

- **Engineering a Microbe for Biofuel Production**
  - Ability to grow novel microorganisms in the laboratory
  - Directed engineering
- **Modifying Lignin Biosynthesis**
  - Understanding of the genetic determinants of complex phenotypic traits in eukaryotic organisms and modification for industrial and other applications
- **Culturing Recalcitrant Microbes from Communities**
  - Understanding of population structure and interactions in complex mixtures of microorganisms and replicating these populations in laboratory settings.

Thank You

Additional Reading:

DOE Systems Biology Knowledge Base Implementation Plan