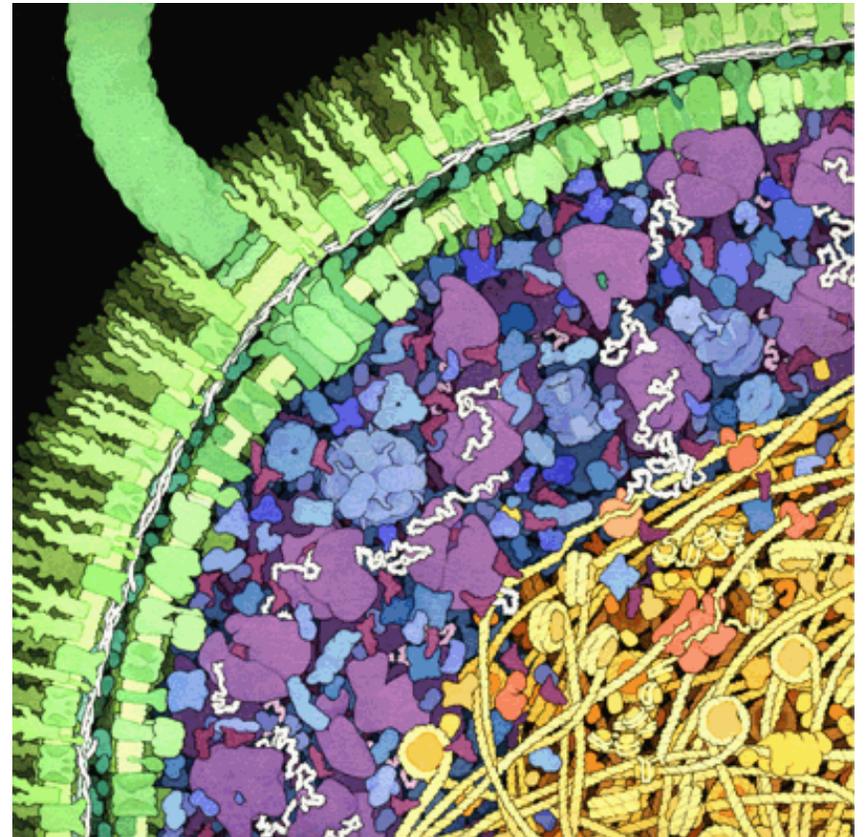
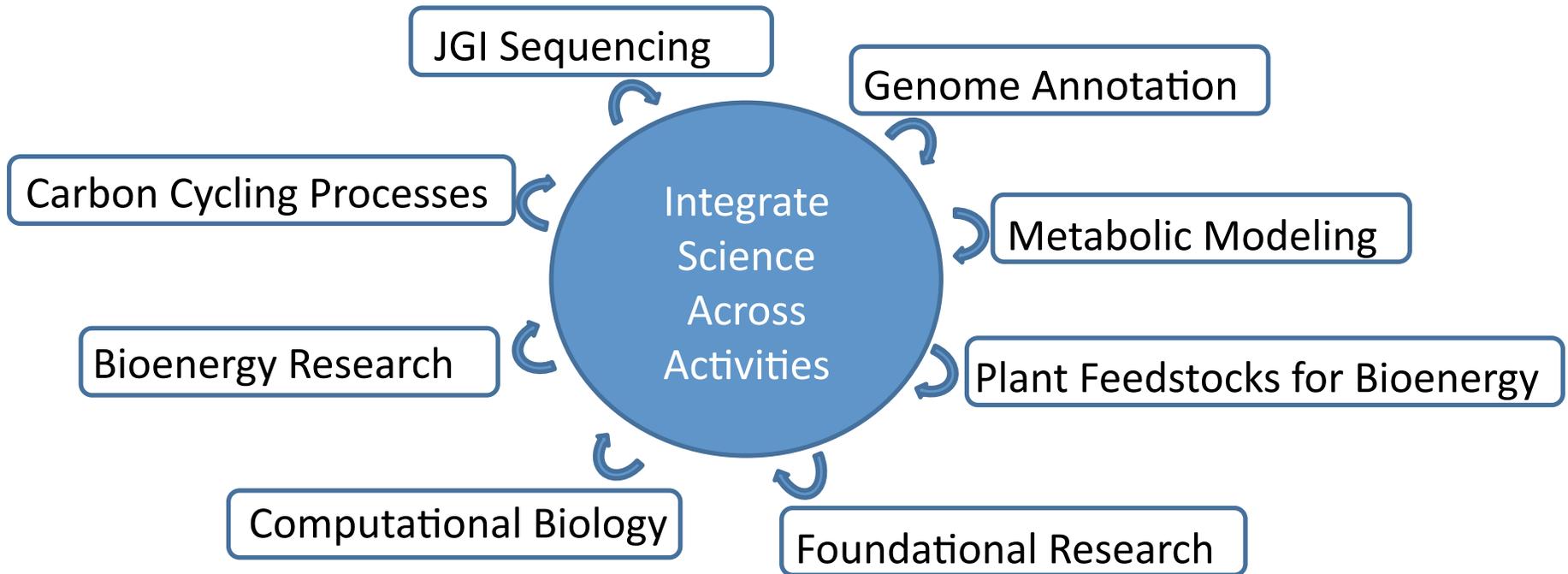


Building the Systems Biology Knowledgebase

Rick Stevens
Argonne National Laboratory
The University of Chicago
Stevens@anl.gov
Stevens@ci.uchicago.edu

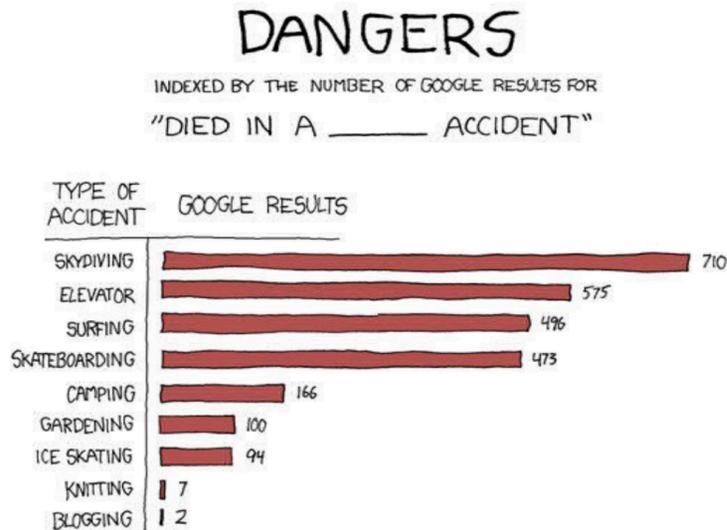


Integrate science and the science community



There is a tremendous wealth of data and information in the Genomic Sciences program. The [Knowledgebase \(Kbase\)](#) is an opportunity to integrate this data and information both within individual activities as well as to integrate together different activities.

Data is extracted and displayed



This is the database model

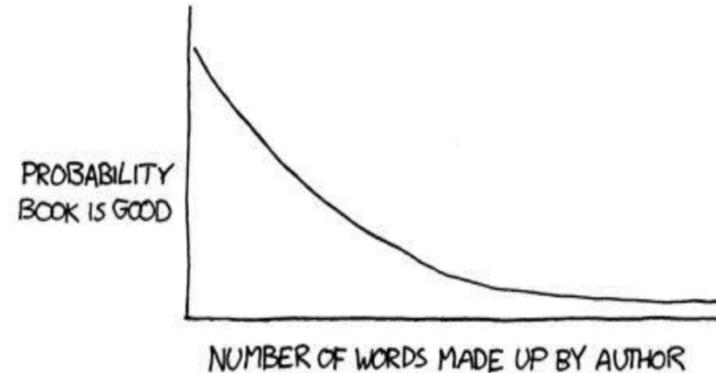
Knowledgebases should *learn* a "model" of the data to provide "conclusions" (hypotheses)

M-C Jenkins (<http://www.scienceforseo.com>)

Images from xkcd comics

Databases enable the rapid organization and **search** of data

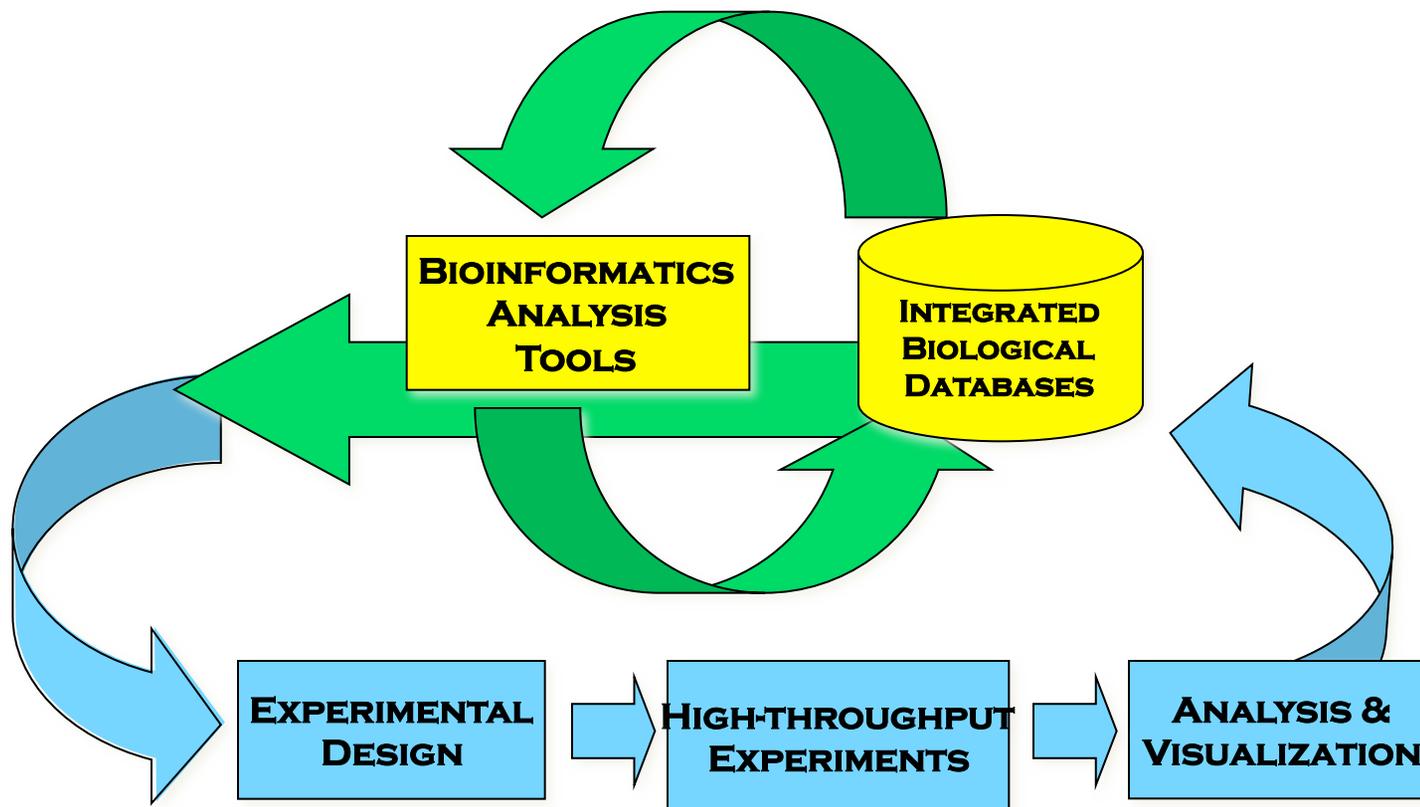
Knowledge is learning & answering



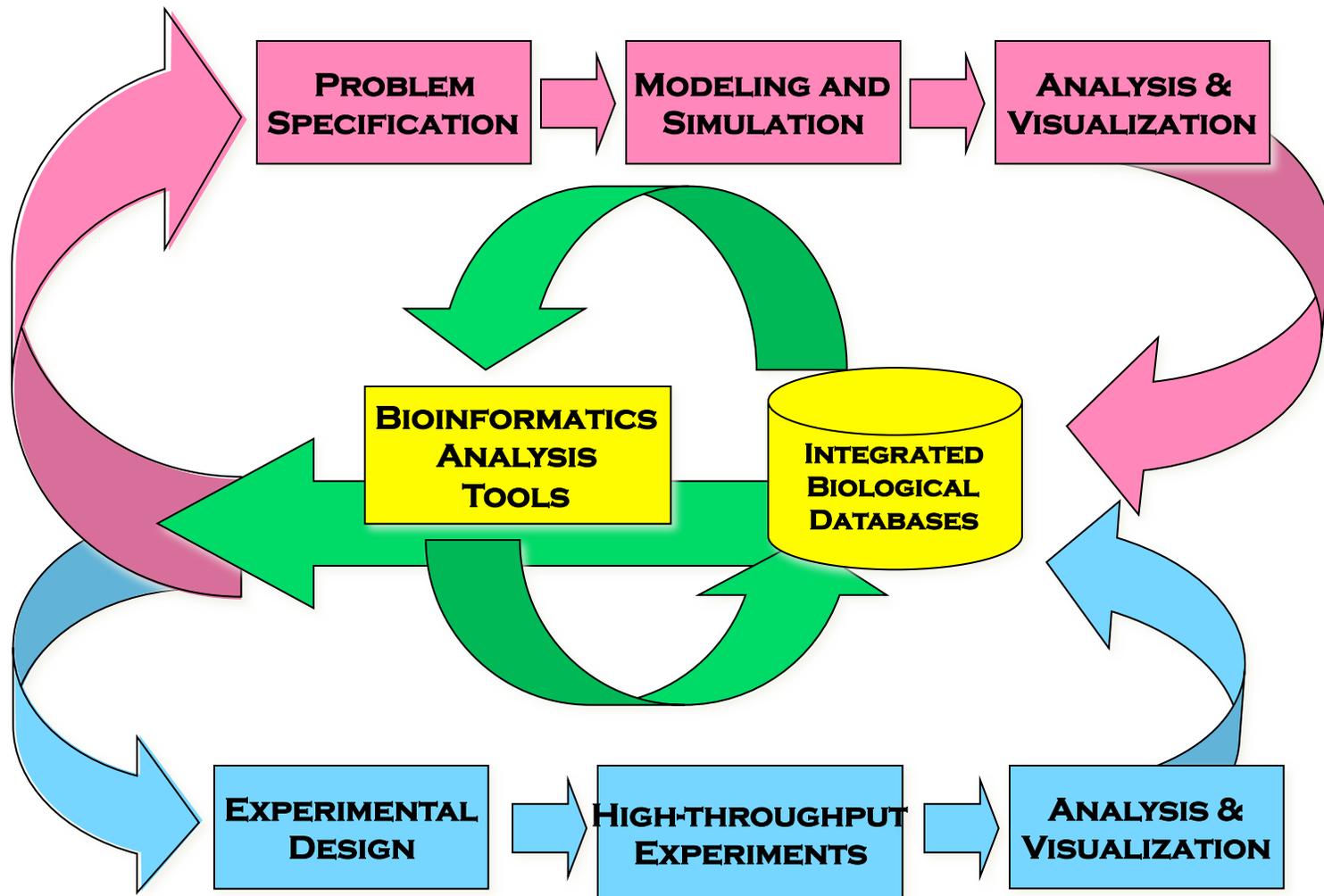
"THE ELDERS, OR FRAÄS, GUARDED THE FARMLINGS (CHILDREN) WITH THEIR KRYTOSES, WHICH ARE LIKE SWORDS BUT AWESOMER..."

This is the knowledgebase model

An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



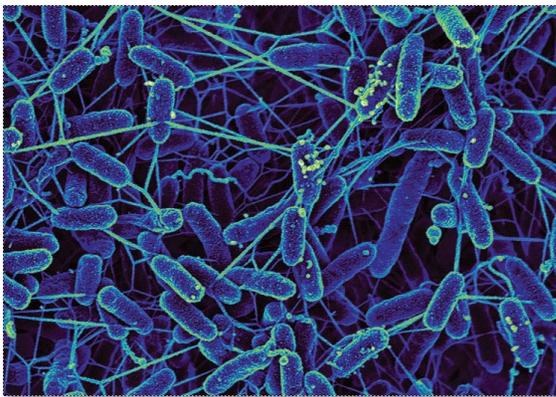
An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



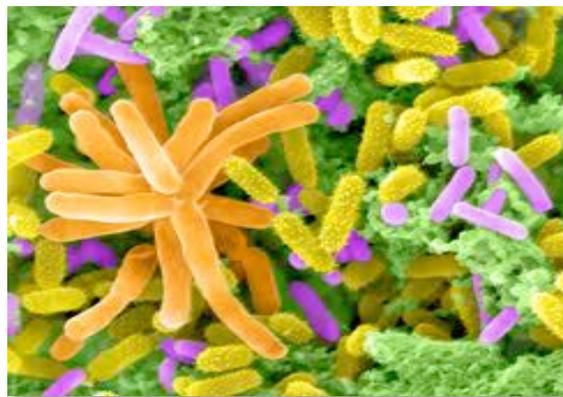
Systems Biology Knowledge

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.
- Resource to enable **experimental design** and **interpretation** of results.



Microbes



Communities

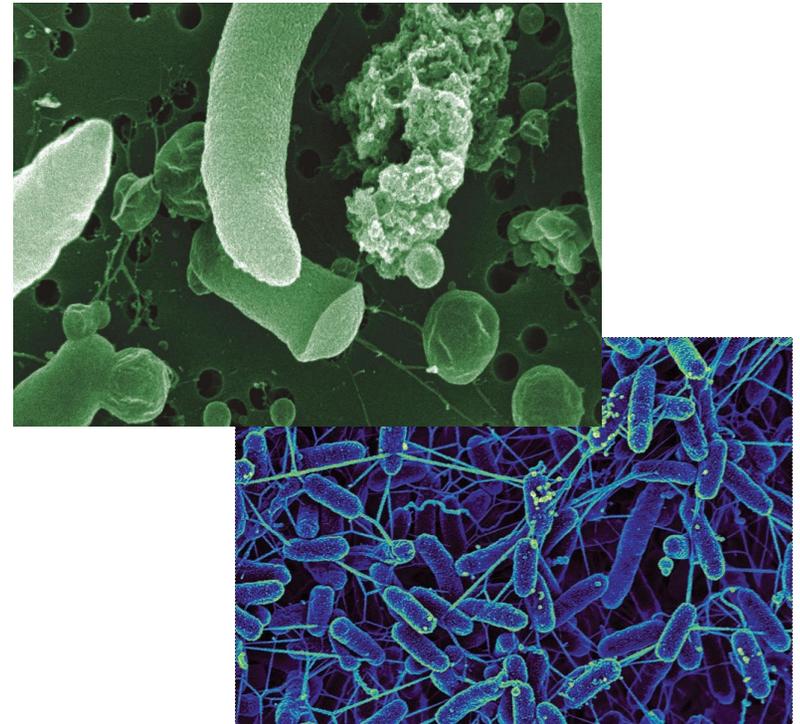


Plants



Our overall goals are to:

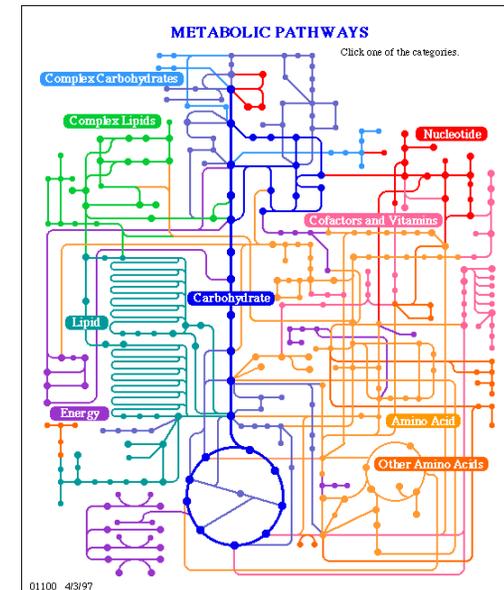
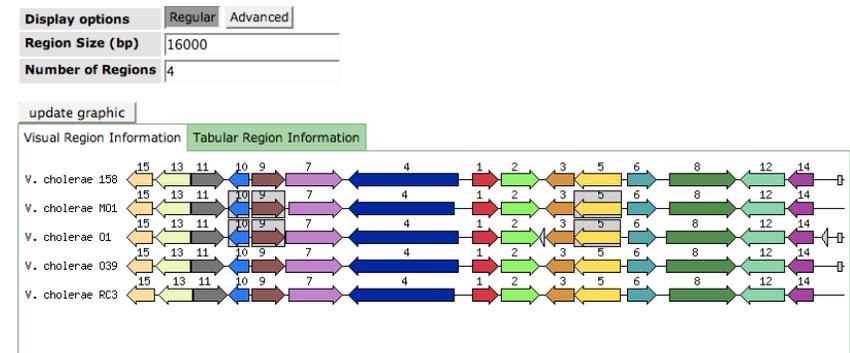
- Reconstruct and predict metabolic and gene expression regulatory networks to manipulate microbial function
- Vastly increase the capability of the scientific community to communicate and utilize their existing data
- Enable the planning of effective experiments and maximizing understanding of microbial system function





We propose to do this by:

- Annotating genomes and assigning confidence
- Reconstructing metabolism and optimizing for function
- Reconstructing regulation and assessing agreement with expression data
- Integrating and standardizing -omics data from multiple data sources
- Constructing models of microbial organisms and interlinking models with data





Within 13 months, we will be able to demonstrate use of the following:

- Data integration and data model
- Next generation organism pages
- Phylogenetic tree services
- Next generation gene pages
- Metabolic modeling
- Regulatory/Transcriptional networks

A microbiologist with a genome sequence and phenotypic growth data will be able to create a metabolic model fully reconciled with the data.



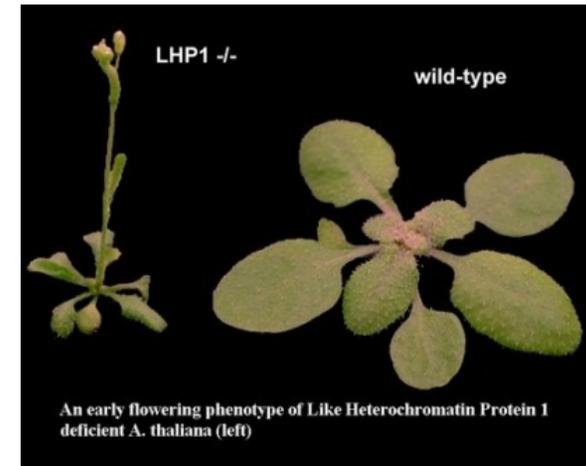


Our overall goals:

In order to extract knowledge from the wealth of high-throughput data in plant biology we need the ability to meaningfully integrate data.

Therefore we aim to:

- Deploy Kbase capability that will allow for interactive, data-driven analysis and exploration across multiple -omics experiments.
- Provide researchers in plant sciences access to comprehensive datasets from high-throughput experiments together with relevant analytical tools and resources.
- Provide a platform for researchers to analyze their own experimental data, and have these results incorporated into the data exploration framework.



We will accomplish this by using:

- Advance storage and indexing strategies for fast but persistent retrieval of large-scale genomics data
- Massively parallel processing of raw data using cloud compute resources
- Interactive charting libraries
- Using controlled vocabularies and ontologies for describing and storing genomic data





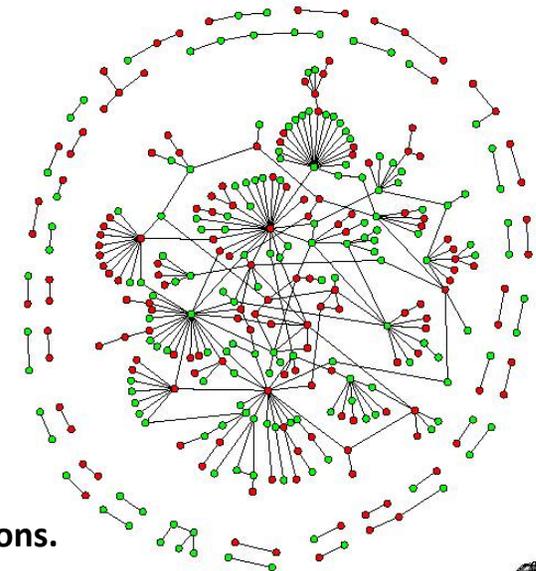
Within 13 months, we will have the following capabilities:

Genotyping Workflow

- Create a workflow for rapidly converting sequencing reads into genotypes
- Demonstrate more than 100-fold speedup over serial version by leveraging capabilities of KBase cloud
- Workflows will be developed as part of the KBase CyberInfrastructure

Data exploration: Linking of gene targets from phenotype and genotype studies with co-expression, protein-protein interaction, and metabolic models

- Allow users to narrow candidate gene lists based on these integrated data types
- Recommender system based on “guilt-by-association” principle: identify other genes, expression datasets, etc. associated with the user-selected group of genes
- Project genetic variations and network edges onto metabolic pathways
- Visualize co-expression network and node Interactions among subnetworks correlated to phenotype of interest



Plants work is cross-cutting with microbial and communities data and predictions.



KBASE
predictive biology

DOE Systems Biology Knowledgebase

Microbial Communities



There has been an explosion of metagenomics data:

- Systems biology is driven by the ever-increasing wealth of data
- Metagenomics is >90% of the data
- Computation needs to be smarter

Our overall goal is to build a Kbase metagenomic platform that provides:

- Scalable, flexible analyses, link physiological and metadata sets to metagenomic sequences
- Data QC and GSC compliant data and standards for data collection
- Enable modeling of metabolic processes within a community
- Predict microbial growth in isolation and in a community



Within 13 months, we will have the following capabilities:

Metagenomic Experimental Design Wizard

The foundation laid will enable researchers to perform *in silico* experimentation and hypothesis testing

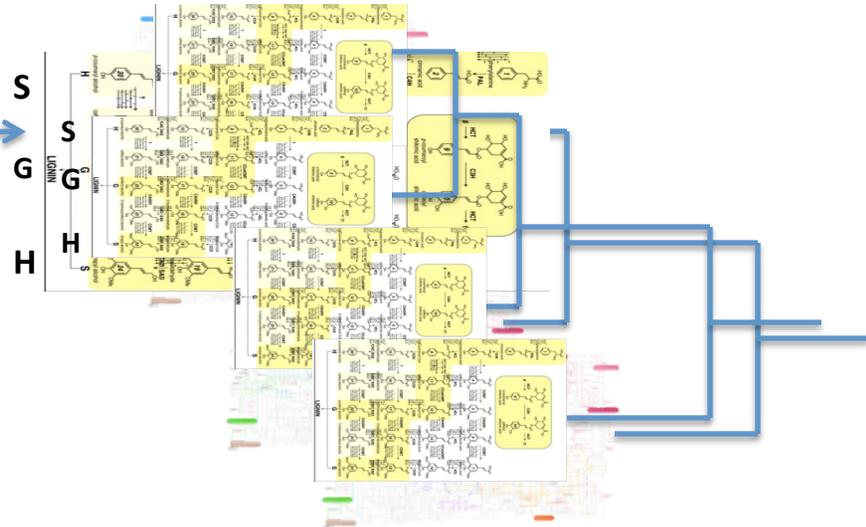
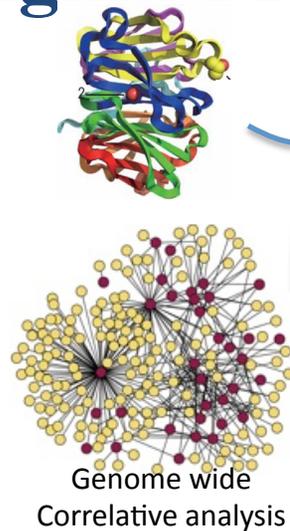
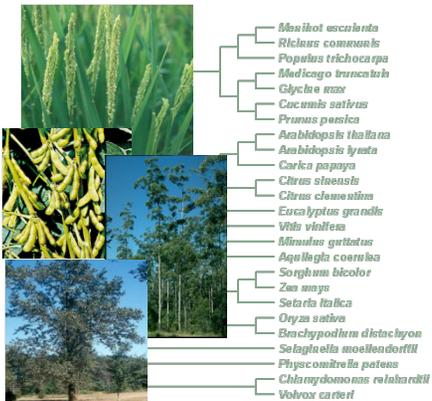
Bioprospecting

- Find communities with similar alpha diversity
- Find communities in similar biomes
- Locate novel proteins (unknowns)
- Suggest functions (based on metadata) that might be encoded in “abundant unknowns”
- Identify optimal candidates for screening



Communities work is cross-cutting with microbial and plant data and predictions.

Modifying Lignin Biosynthesis



SNPs3D

PolyPhen-2

SNP influenced changes in protein structure and function

Pathway predictions

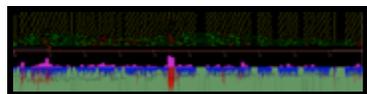
- Model optimization
- validation

Plant systems modification

- Genome annotation algorithms
- Comparative genomics

- Network inference
- Pathway reconstruction
- Omics & SNP overlay

Phylogenomics
Modeling phase I



phytozome

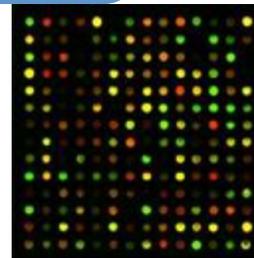


Phenotype
Mutant
population

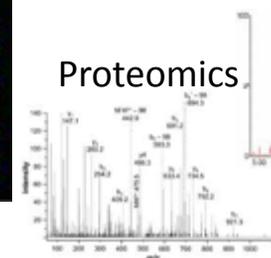
Resequencing data



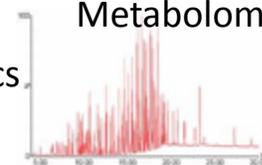
Transcriptomics



Proteomics



Metabolomics



What the KBase Needs To Provide?

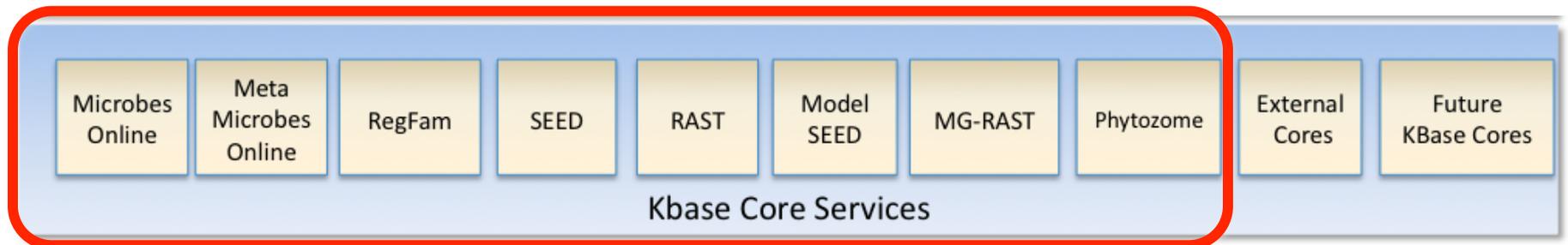
- Scalable compute and data capabilities beyond that available locally
- Distributed infrastructure available 24x7 worldwide
- Integration with local bioinfo systems for seamless computing and data management
- Enables leverage of remote systems administration and support via service providers
- Enables access to state of the art facilities at fraction of the cost (SPs just add more servers)
- Centralized support of tools and data
- Bottom line \Rightarrow enable biologists to focus on biology

Large-Number of Bioinformatics Tools and Environments **O(100)** core codes

- Raw sequencing clean up
- Assembly (prok, euk, meta), Gene Calling (prok)
- Comparative analysis (basic tools, e.g. BLAST, BLAT, etc.)
- Protein family maintenance and decision procedures
- MG Fragment id and function/taxa inference (e.g. MG-RAST)
- Annotation, reconstruction (e.g. RAST etc)
- Multiple sequence alignment (e.g. T-COFFEE, Clustalw)
- Statistical analysis (k-mer spectra, co-occurrences, etc.)
- Sequence data mining (e.g. motifs, RNAi)
- Evolutionary analysis (16S, phylogenetic trees)
- Large-scale data integrations (SEED, IMG, etc.)
- Systems level integrations (e.g. modeling environments)

Leverage Existing Investments

- We leverage the considerable investments in existing integrated databases and analysis environments
- Key challenge: How we build on these systems yet provide to the community an integrated view for future development

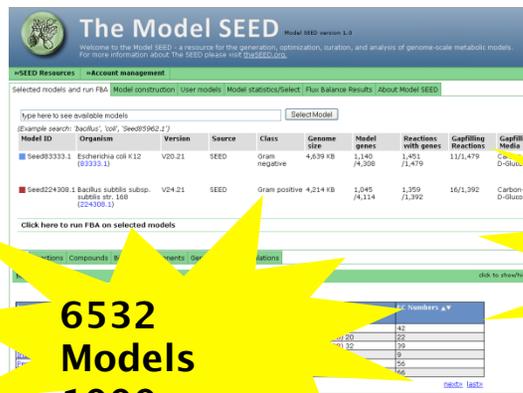


Microbes Online



1000s Data Sets
300+ Daily Users

Model SEED



6532 Models
1000+ Users

MG-RAST



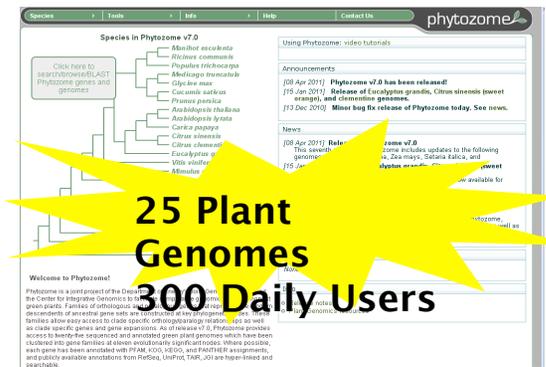
41,000 Metagenomes
500+ Daily Users

Meta Microbes Online



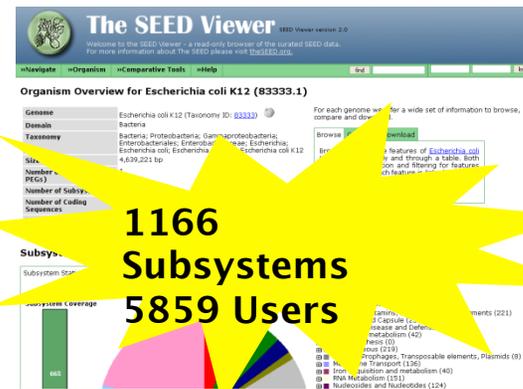
153 Metagenomes
100+ Daily Users

Phytozome



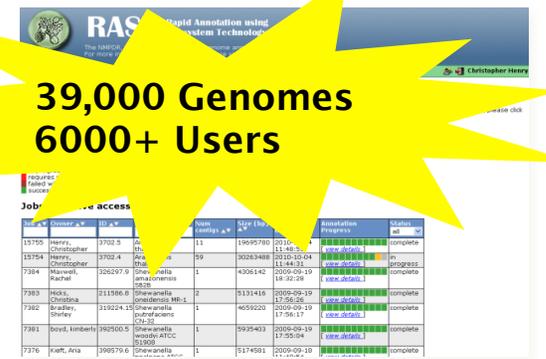
25 Plant Genomes
300 Daily Users

The SEED



1166 Subsystems
5859 Users

RAST

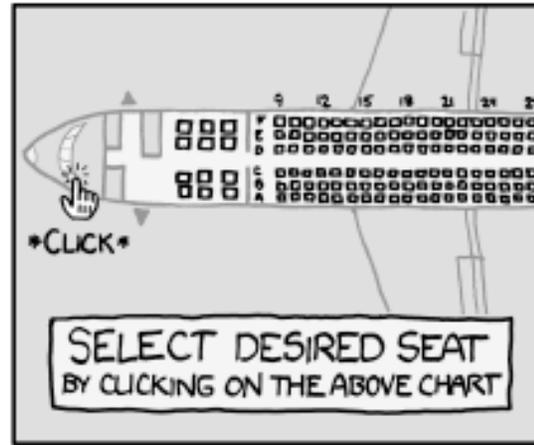
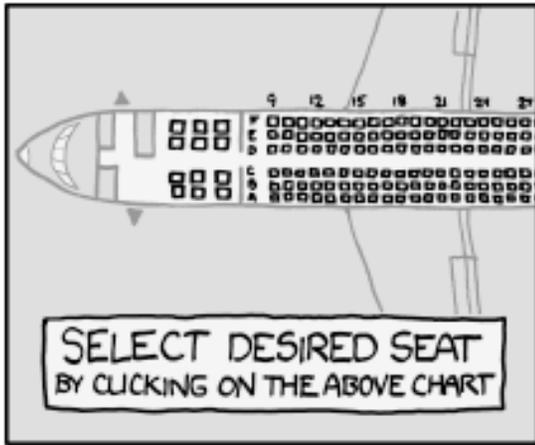


39,000 Genomes
6000+ Users

RegFam



1000s Papers
100+ Daily Users



Our vision is to put users in the drivers seat.