BESC Knowledgebase, Tools, and Laboratory Information Management System

Mustafa Syed[1]*(syedmh@ornl.gov), Susan Holladay,[1] Tatiana Karpinets,[1] Morey Parang,[2] Guru Kora,[1] Doug Hyatt,[2] Sheryl Martin,[1] Leslie Galloway,[1] Byung Park,[1] Ying Xu,[3] **Ed Uberbacher**,[1] Brian Davison,[1] and **Paul Gilna (BESC PI)**[1]

[1]BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of Tennessee, Knoxville; and [3]University of Georgia, Athens

http://besckb.ornl.gov
http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is a multi-institutional partnership undertaking large experimental campaigns to understand and reduce recalcitrance of biomass for degradation by enzymes and organisms, and converting plant biomass into biofuels in a single step. The BESC Knowledgebase and associated tools is a discovery platform for bioenergy research. It consists of a collection of metadata, data, and computational tools for data analysis, integration, comparison and visualization for plants and microbes in the center. BESC Knowledgebase (KB) and BESC Laboratory Information Management System (LIMS) enable bioenergy researchers to perform systemic research.

A unified way of sharing experimental workflows, standardized protocols, tracking samples, and capturing metadata/data across more than a dozen BESC institutions is critical to the proper functioning of BESC both from the standpoint of ensuring meaningful and reproducible results, as well as tracking the generation of intellectual property. The BESC LIMS has been used to track metadata from data collection to analysis for over 80,000 samples from more than sixty campaigns undertaken by BESC. For each of these campaigns details such as sample generation, sample splitting and shipping, sample processing, protocols, controls, replicates, and results have been captured. The LIMS web portal is available for BESC members that provides reports for material transfer, shipping, sample history and provenance, and analytical results. The LIMS is also linked to the BESC Knowledgebase, which contains the results of experiments interpreted in the context of biological systems.

The BESC KB infrastructure consists of separate public and private platforms cohesively integrated through leveraging a loosely coupled middleware layer which facilitates information exchange and access control. While the majority of the data and tools are available for public access via its public portal, a limited set of information and tools is only accessible by the BESC researchers, via its private portal through user authentication and access control. Some omics data, tiling array data, biochemical characterization data for biomass samples, and candidate genes for analysis through the transformation and biomass characterization are private. These results are integrated with reference genomic, metabolic and other omics data obtained from public resources which form core KB.

The core KB for plants consists of 21 plant genomes including six algal genomes along with a rich set of annotated data and computed information for: (a) gene structures; (b) protein products; (c) homology-based functional prediction; (d) domain structures; (e) ortholog and paralog prediction; (f) gene ontology; and (g) metabolic and enzymatic pathways. Currently, the KB's reference plant data consists of over 500,000 coding genes from which nearly 400,000 protein-coding genes with function prediction have been identified. The Plant KB maintains available gene model variations, alternative gene models (including alternative splicing), and historical versioning. For example, our *Populus* database alone contains over 500,000 versions of gene models each of which are associated with one or more annotation data including functional predication, domains, GO terms, KEGG pathways, ortholog and in-paralog data. We have also collected rich set of omics data from external resources like NCBI, GEO and EMBL

ArrayExpress related to assembly of the cell wall pectic matrix, cellulose synthesis and cell elongation, cellulose synthase mutants, cell wall stress, primary to secondary stem development and many other experiments for *Arabidopsis*, rice and *Populus*.

The microbial core KB for microbes consists of 37 microbes, including biomass degraders, fuel producers, endophytes and model organisms. Data collected on these organisms include genome annotation, biochemical data including enzymes, ligands and pathway annotation generated using the Pathway Tools software and provided in KEGG, carbohydrate active enzyme data generated by CAT and provided by CAZy, operon predictions from BeoCyc and DOOR, protein functional domain predictions from CDD, Pfam, COGS, TigrFam, SMART. Pair-wise predictions of orthologs and inparalogs for the genomes generated using bi-directional BLAST hits and Inparanoid software, BLAST hits against NCBI NR database. Trans membrane protein predictions using TMHMM and signaling proteins predictions using SignalP. We also have omics data on effect of alcohol on cell, fermentation time course study, growth on model substrates found in lignocelluloses, growth on simple and complex sugars, biomass deconstruction, and many other experiments from several *Clostridium* and other bioenergy relevant microbes. Recent addition to microbial core includes a microbial resequencing database (ReSeqDB); we have made available resequencing data for an ethanol-adapted *Clostridium thermocellum* strain. We provide not only high confidence differences, but also processed results from several different tools (Genome Resequencing Toolkit) and comprehensive annotations of those sequence changes, position of sequence change with respect to gene and intergenic region, changed protein sequence, and probable change in gene function or regulation. SNPs indels are also visualized and made available from the genome browser. The goal of the resequencing projects is to discover genomic modifications underlying the specific phenotype of the mutant. The Genome Resequencing Toolkit reveals mutational changes in the mutant strains to understand the biological effects of each change on the mutant phenotype.

The KB is tightly integrated with computational tools for genome annotation, comparative genome analysis, data integration, data mining, analysis and visualization. Genome annotations tools such as CAT for annotating protein sequences with CAZy family, DOOR for predicting operons, BLAST, ClustalW, Muscle, and tools to build phylogenetic trees. Comparative analysis tools to compare multiple genomes in terms of pathway, enzymes, domain architecture, or sequence similarity. Browse interfaces, such as organism card, and gene card, allow the user to easily retrieve detailed information about a specific gene or organism. Search interfaces provide mechanisms to search data objects using keyword, protein, enzyme and pathways with custom filters. Integrated views for genes and organisms (gene cards/organism cards) which provide a rich set of information, omics data, annotations, and tool results. Visualization tools such as Gbrowse – genome browser, CMAP (Comparative Map Viewer) can be used to view and compare maps (genetic, physical) across multiple genomes. Layers of heterogeneous data and annotations can be overlaid on GBrowse or CMAP browsers for browsing and comparing across maps. Tools to search omics data in resources such as GEO, and ArrayExpress integrate experimental data into the knowledgebase OnDemand, making it possible to draw interactive heat maps, scatterplot, overlay omics data on pathways, build networks and identify regulatory motifs.

[The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.]