“Biosciences” at the Petascale

Martin Keller

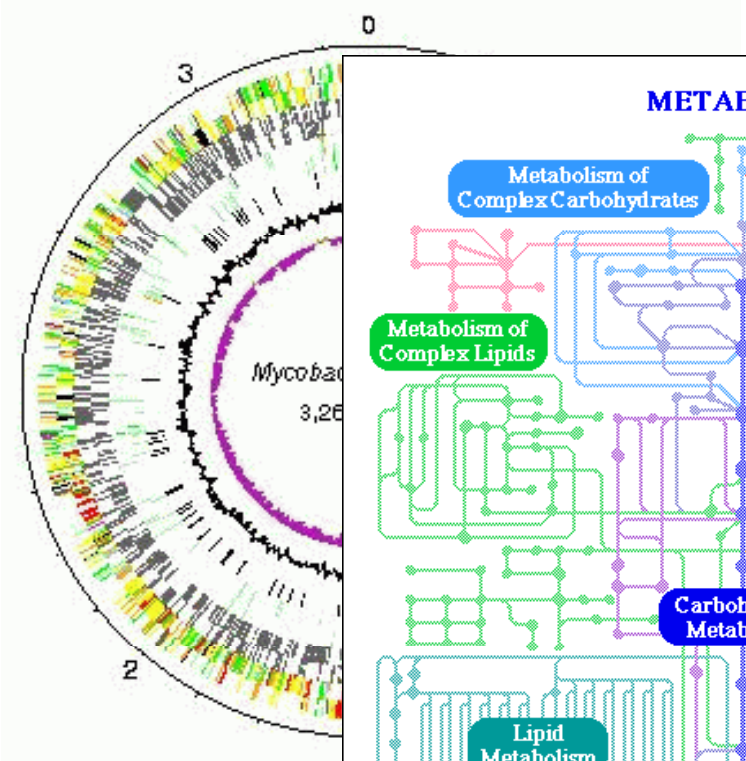
Oak Ridge National Laboratory

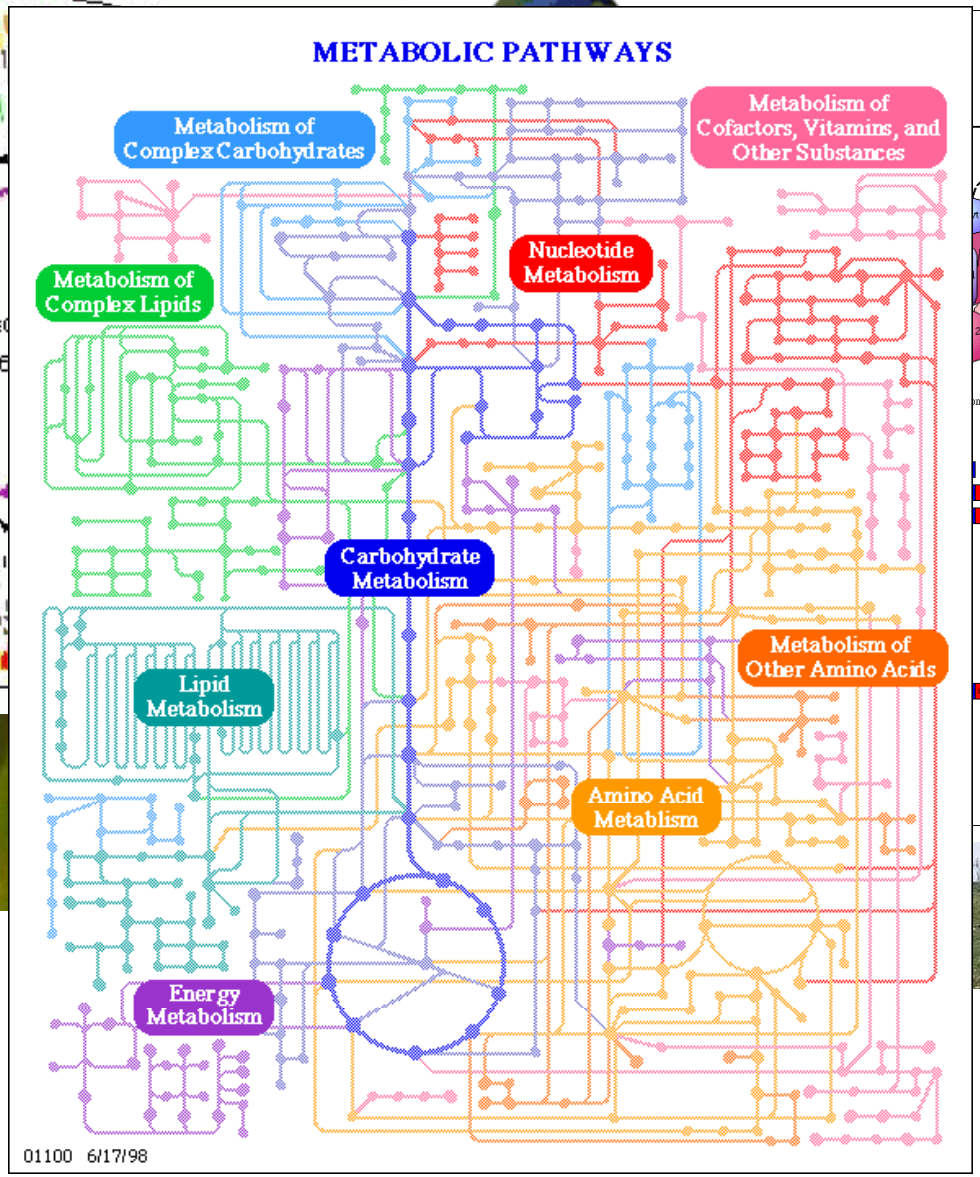
What is Biology?

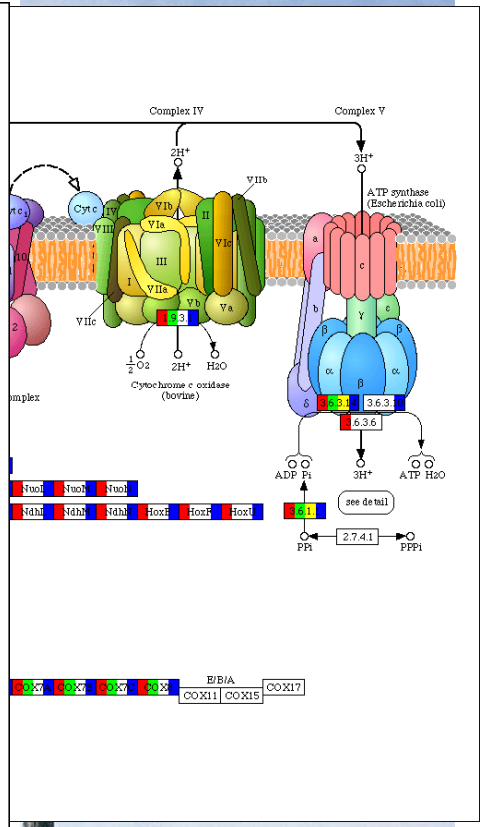
Study of living things and their vital processes. An extremely broad subject, biology is divided into branches. The current approach is based on the levels of biological organization involved (e.g., molecules, cells, individuals, populations) and on the specific topic under investigation (e.g., structure and function, growth and development). According to this scheme, biology's main subdivisions include morphology, physiology, taxonomy, embryology, genetics, and ecology, each of which can be further subdivided. Alternatively, biology can be divided into fields especially concerned with one type of living thing; for example, botany (plants), zoology (animals), ornithology (birds), entomology (insects), mycology (fungi), microbiology (microorganisms), and bacteriology (bacteria). biochemistry; molecular biology.

Copyright 1994-2008 Britannica Concise Encyclopedia.

Change in Biology Research





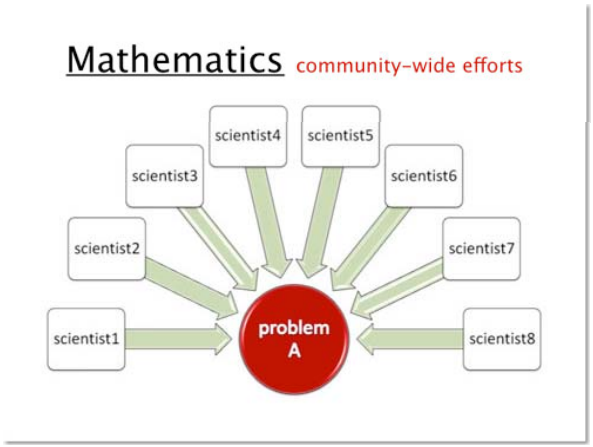


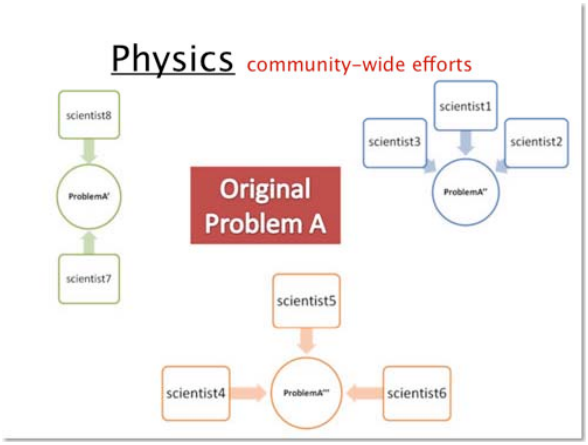
Mathematics vs. Physics vs. Biology

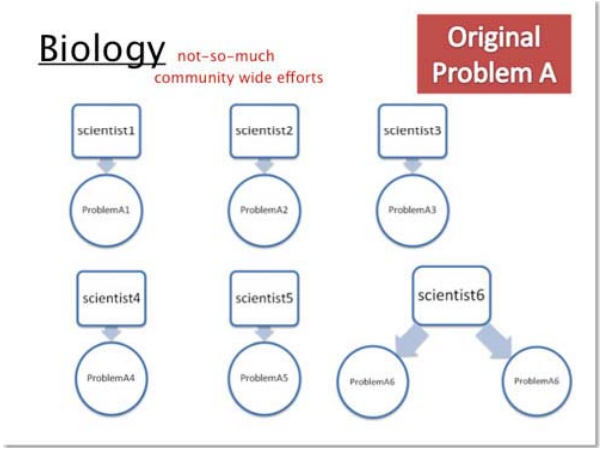
« O'Reilly Book: Programming Collective Intelligence Supercomputing 2007 »

Mathematics vs. Physics vs. Biology
Published by *Adam* on September 21, 2007 in Collective Intelligence
Following up on my previous post about community efforts and collective intelligence. I recreated some figures from my notes on a presentation by Jooyoung Lee. He talked about how we are approaching problems in the sciences.

In mathematics, scientists work together on some fundamental problem A. So there's community-wide efforts.







In physics, scientists work together on problems that are closely related to original problem A. So there's still community-wide efforts

Biology is somehow different. In biology, it seems that every scientist is working on their own problem, and some even have more than one!

The BESC Team

Joint Institute for Biological Sciences (JIBS)

Alternative Fuels User Facility

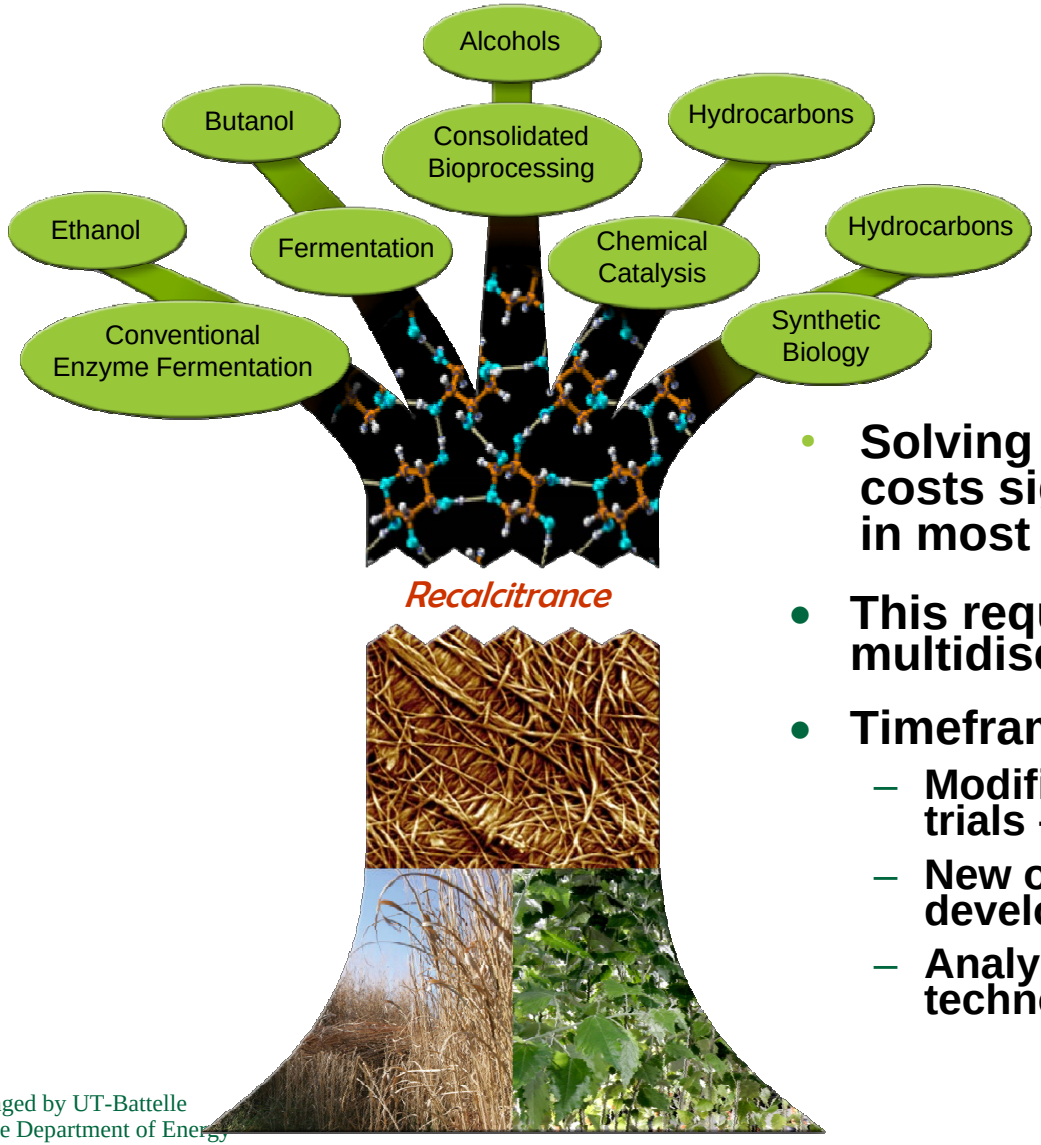
Complex Carbohydrate Research Center

- Oak Ridge National Laboratory
- University of Georgia
- University of Tennessee
- National Renewable Energy Laboratory
- Georgia Tech
- Samuel Roberts Noble Foundation
- Dartmouth
- ArborGen
- Verenium
- Mascoma
- Individuals from U California-Riverside, Cornell, Washington State, U Minnesota, NCSU, Brookhaven National Laboratory, Virginia Tech

Biocomputing needs in bioenergy

- **Simulations -** Molecular dynamics to understand complex assemblies of enzymes, solvents, and lignocellulose in regions which are difficult to access experimentally
- **Data-rich analysis of networks and pathways** for both plant cell wall biosynthesis and for deconstruction
- **Knowledge-base and improved annotations** for community exchange of data, insights and knowledge

Access to the Sugars in Lignocellulosic Biomass is the Current Critical Barrier



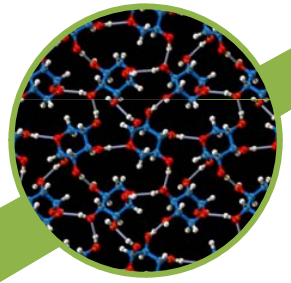
- **Solving this will cut processing costs significantly and be used in most conversion processes**
- **This requires an integrated multidisciplinary approach**
- **Timeframe**
  - **Modified plants to field trials – Year 5**
  - **New or improved microbes to development – Years 4-5**
  - **Analysis and screening technologies – Year 3 on**

The Fundamental Science of Biomass Recalcitrance is Poorly Understood

Overcoming recalcitrance is the single coherent overarching theme for the BESC

Recalcitrance: Resistance to breakdown into sugars

Cellulosic biomass

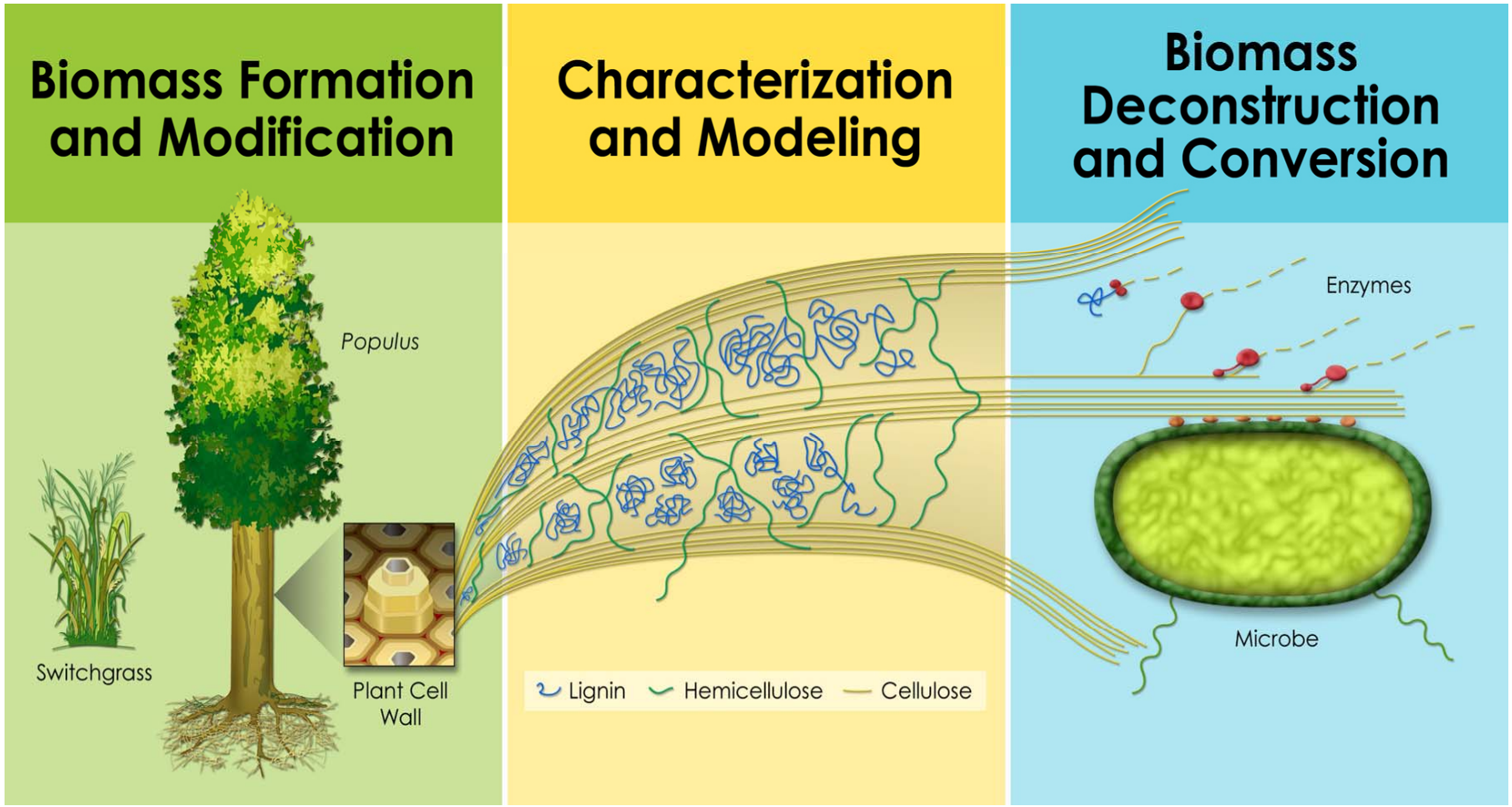


Sugars

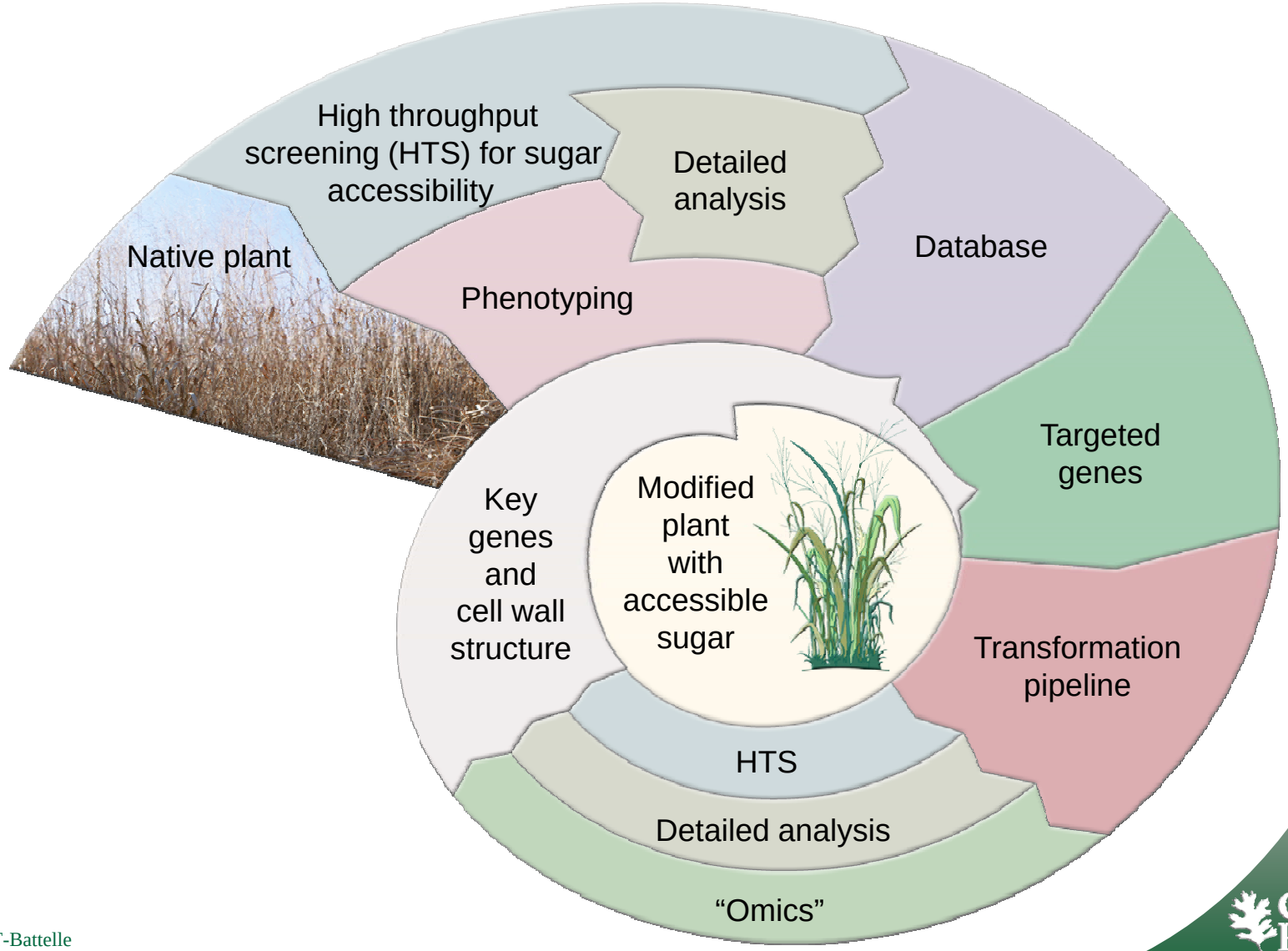
Fuel(s)

- A large-scale, integrated, interdisciplinary approach is needed to overcome this problem
  - Current research efforts are limited in scope
  - BESC will launch a broad and comprehensive attack on a scale well beyond any efforts to date
- Without advances, a cellulosic biofuels industry is unlikely to emerge
- Knowledge gained will benefit other biofuels and biofeedstocks

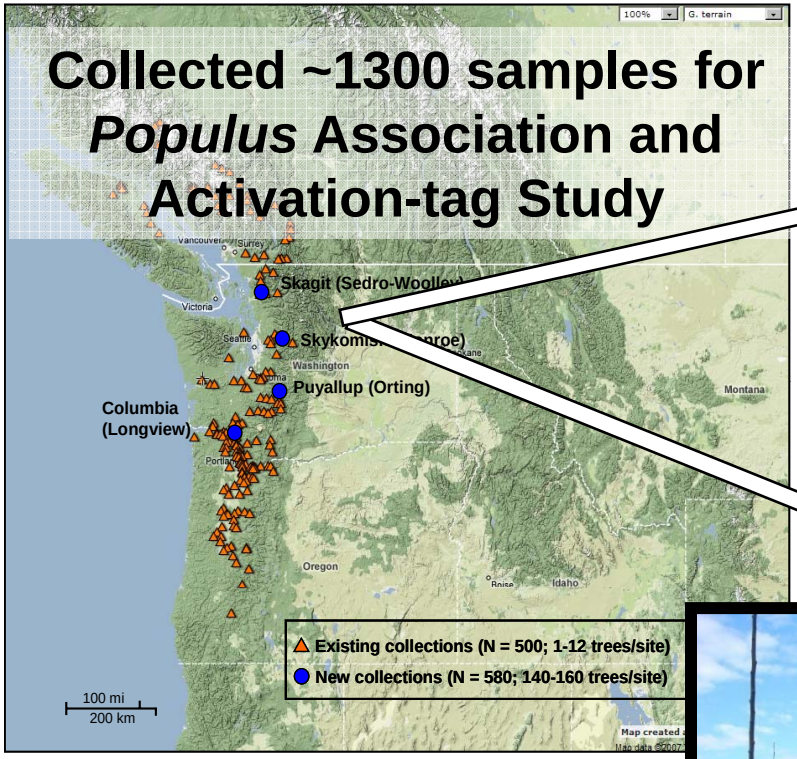
Three linked scientific focus areas will enable BESC to understand and overcome biomass recalcitrance



What Genes Control Cell Wall Synthesis (and Access to the Sugars)?

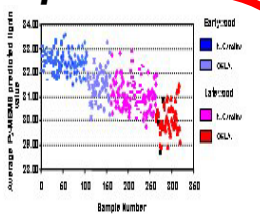


Mining Variation to Identify Key Genes in Biomass Composition and Sugar Release



HTS Pipeline

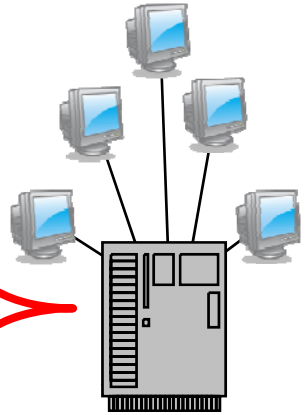
Sugar Release Assay



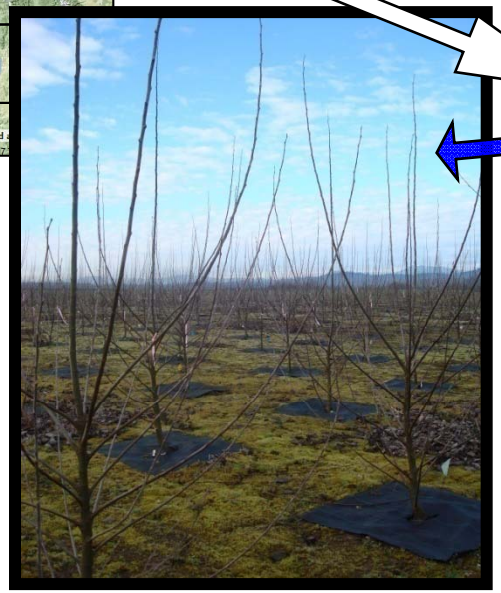
Analytical Pyrolysis

Create Genetic Marker Map to identify allelic variation

Identify Marker Trait Association



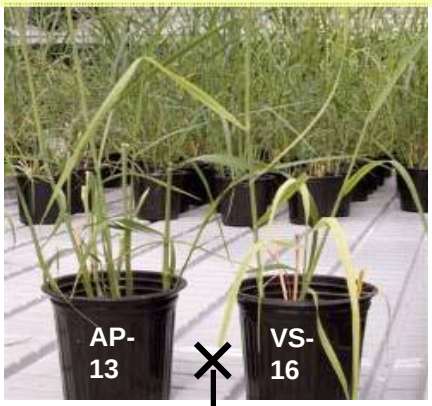
Cell Wall Biosynthesis Database



Establish common gardens for association and activation tag populations with 1000s of plants

Mining Genetic Variation in Switchgrass

Create diverse population by cross "lowland" SG AP-13 and "upland" SG VS-16 into 385 pseudo F1 clones

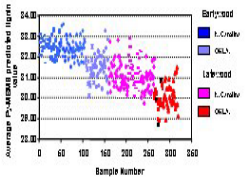


Pseudo F_1 population of 385 genotypes

Clones ready for field planting

HTS Pipeline

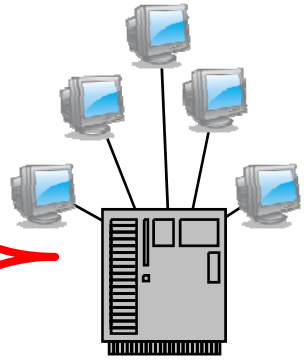
Sugar Release Assay



Analytical Pyrolysis

Create Genetic Marker Map to identify allelic variation

Identify Marker Trait Association



Cell Wall Biosynthesis Database

THE SAMUEL ROBERTS NOBLE FOUNDATION

Targeted Plant Genes and Transformation Pipeline

- **Gene transformation pipeline established and running**
  - **70 *Populus* genes per set**
  - **4 Switchgrass for stable transformation**
  - **30 Switchgrass by VIGS**
  - **First set totaling 104 genes in pipeline**
  - **Second set under review now**
- ***Populus***
  - **Transformation: 200 genes per year**
  - **Activation Tagging: 1000 genes per year**
- **Switchgrass**
  - **Transformation: 20 genes Year 1; 40-60 Year 2**
  - **VIGS (viral induced gene silencing): 200 genes per year, RNAi**
- **Higher perennial plants have fewer genetic tools and so targets must be selected carefully**

Functions of initial targets

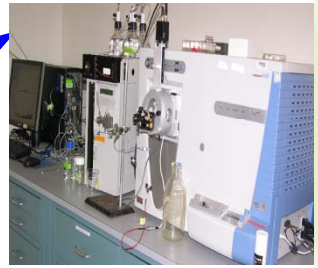
Functional category	# genes
Cell wall biosynthesis	50
Cell division and expansion	46
Signal transduction	26
Stress response	20
Metabolism	19
Intracellular traffic	9
Protein fate	9
Transcription	9
Plant defence	4
Nucleic acid or nucleotide binding	2
Transporters	2
Total	196

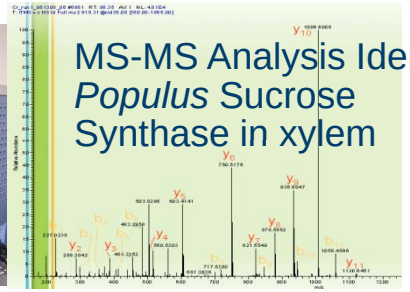
Systems Biology and "Omics" Provide Deeper Understanding

U.S. DEPARTMENT OF ENERGY

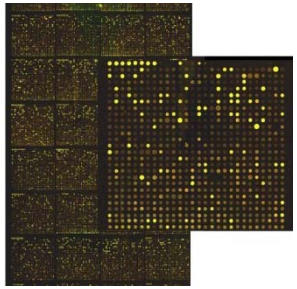
Transformation Pipeline

Proteomics

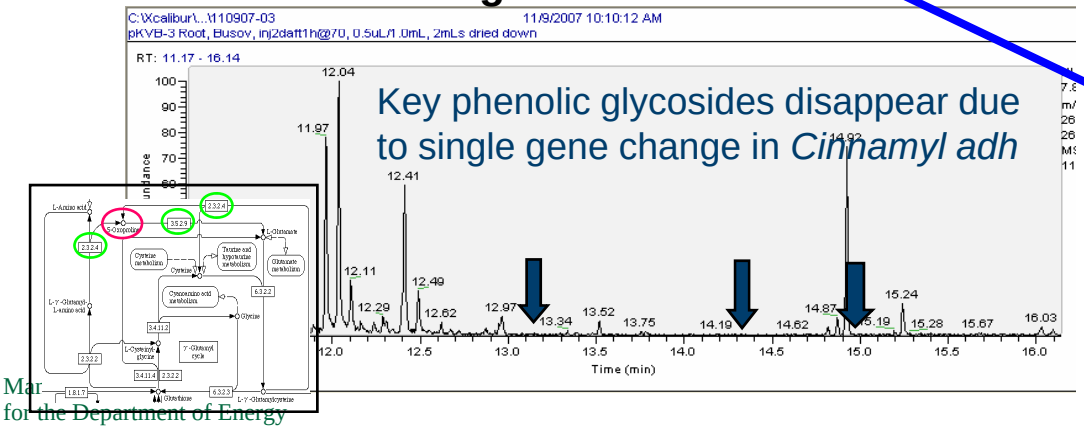


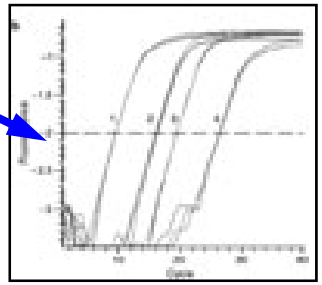


Transcriptomics: Microarray Profiling

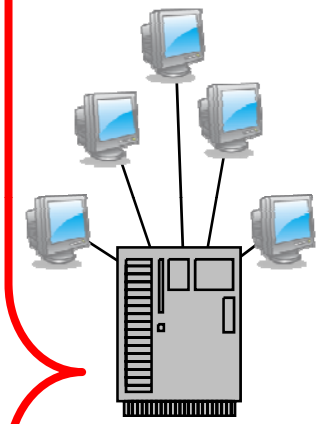


Metabolic Profiling





RT-PCR Gene Expression



Cell Wall Biosynthesis Database

OAK RIDGE National Laboratory

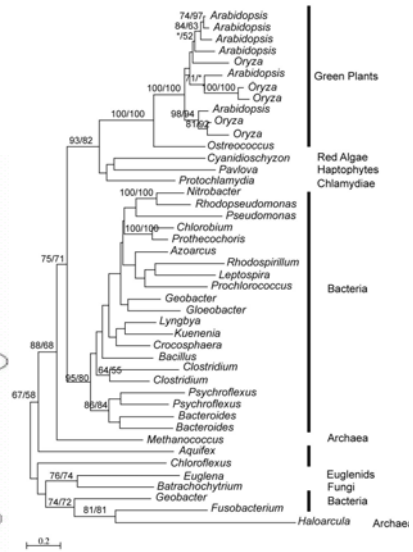
Plant synthesis pathways and networks

- **Goals: functional identification of genes and elucidation of pathways/networks involved in plant cell wall synthesis and remodeling,**
- **Example: Identification of genes involved in plant cell wall synthesis through protein-protein interaction prediction and phylogenomic analyses (Dam, Xu – UGA)**
  - **using a seed set of "known" cell wall synthesis genes**
  - **recruiting additional genes through prediction of protein-protein interactions**
  - **computational validation of through prediction of co-evolutionary and co-localization information**

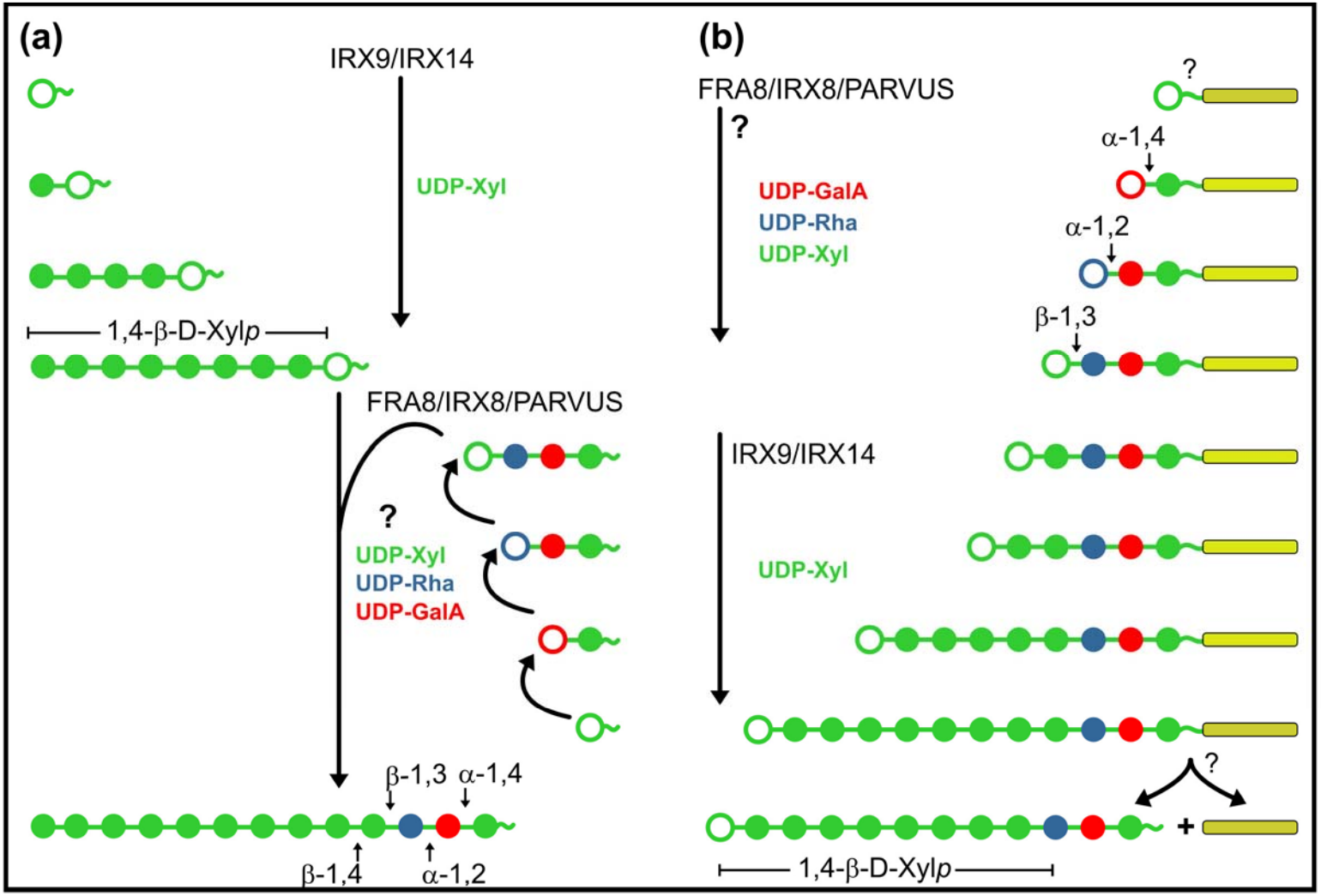
-Predicted ~400 genes with multiple interactions with the Purdue dataset

-~10% only have interactions with the Purdue set

-Doing computational validation

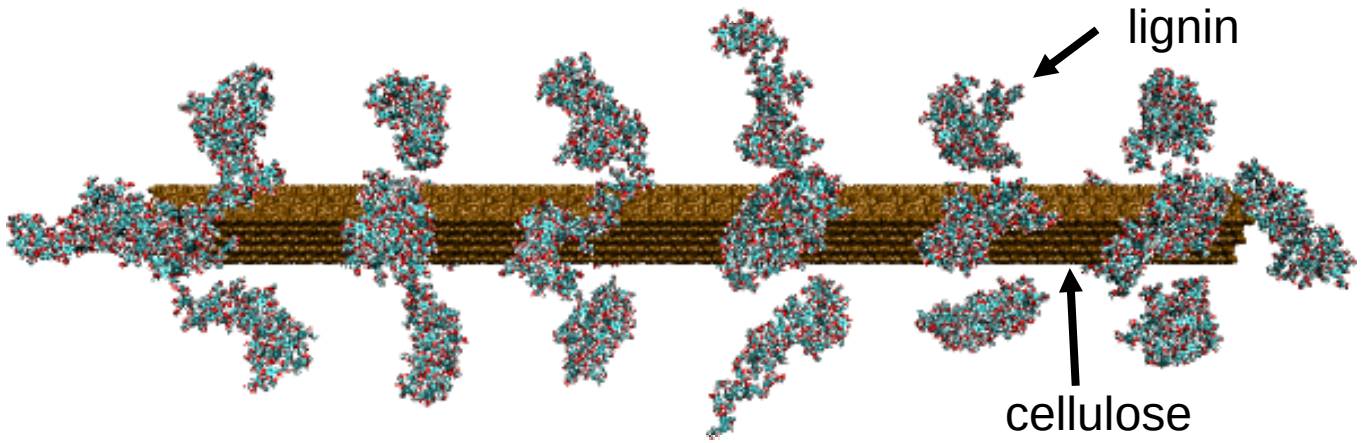


Working models of xylan biosynthesis



Computer Simulation of Lignocellulosic Biomass

Loukas Petridis, Benjamin Lindner and Jeremy C. Smith
ORNL Center for Molecular Biophysics



- **Recalcitrance of Biomass to Hydrolysis Limits Cellulosic Ethanol Production.**
- **Understanding Biomass Structure is Key to Overcoming Recalcitrance.**
- **Large-Scale Molecular Dynamics Simulation (1-3M atoms) using 2008 DOE INCITE award on ORNL Cray XT4.**
- **Multiscale Systems-Level Methodology under development.**
- **Results to be used to Interpret Biophysical Experiments *e.g.* Diffraction, Spectroscopy.**

HTP Characterization Pipeline for the Recalcitrance Phenotype

- **Screening of 1000's of samples**

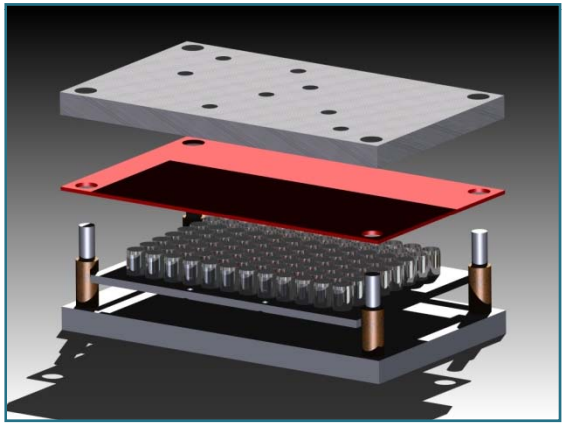
Composition
Analytical Pyrolysis, IR, confirmed by wet chemistry

→

Pre-Treatment
New method with Dilute acid and steam

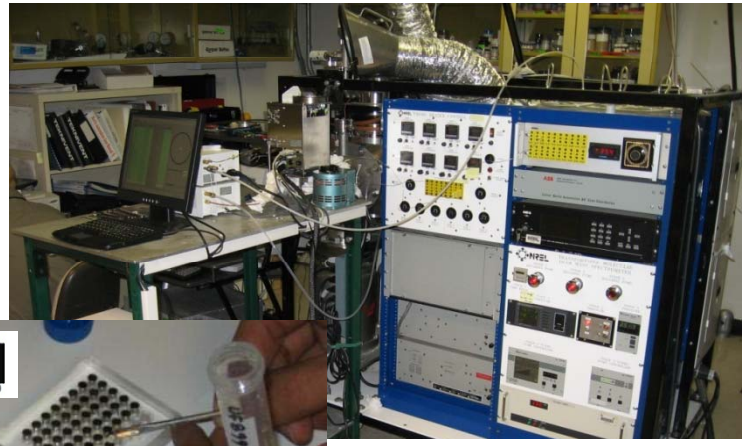
→

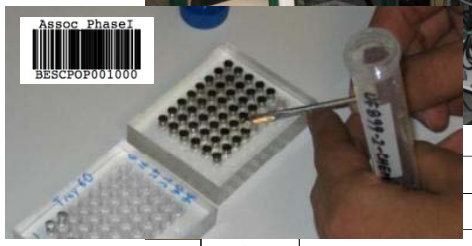
Enzyme Digestibility
Sugar release with enzyme cocktail

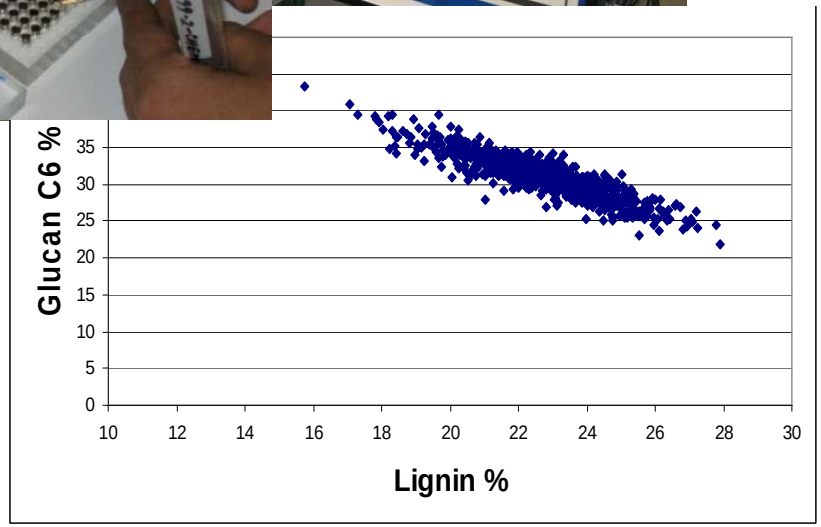


Detailed Chemical and Structural Analyses of Specific Samples

Composition Data from Analytic Pyrolysis (MBMS) for High-throughput Screening of Transgenic Populations



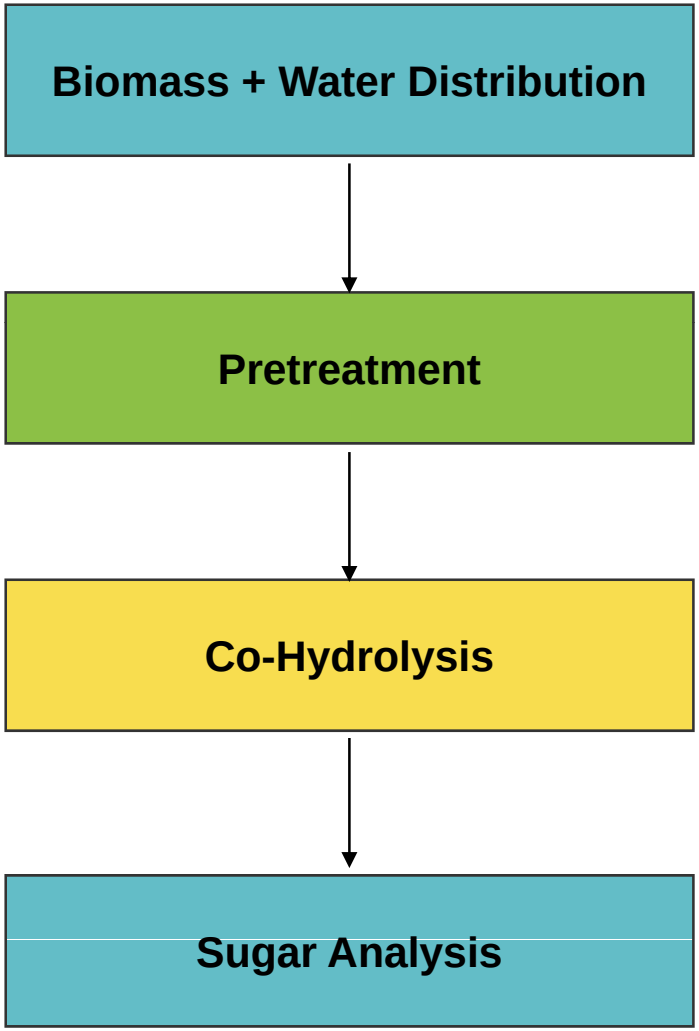


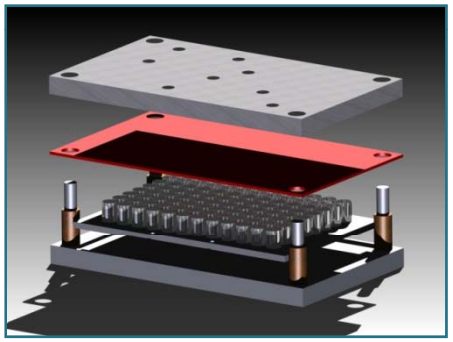


Composition data from *Populus* association study (798 samples) represents full range of known *Populus* variation. Sykes, Davis NREL

- **Rapid (50/h w/ 4mg)**
- **Reliable**
- **Gives values for glucan, xylan, lignin, and details on monomers – e.g., *S/G***
- **Complements time-consuming and more variable wet chemistry, molecular and biochemical analyses**

Enabling Technology: An HTP Pretreatment for 1000s of Small Samples



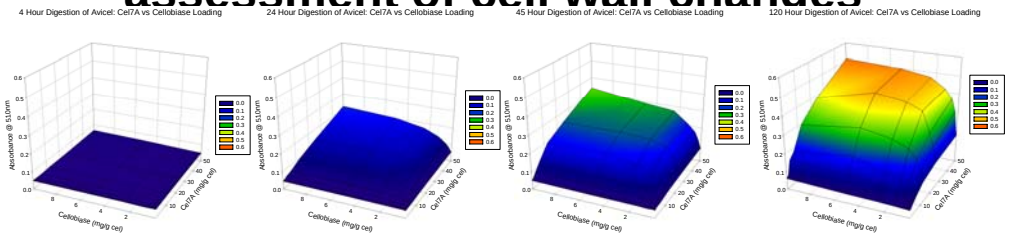


- **Unique and Important**
  - ***Steam*: efficient uniform heating**
  - ***No separation*: saves time and increases accuracy**
  - ***2-4mg sample size*: reduces material costs**

Studer et al., presented at the 30th *Symposium on Biotechnology for Fuels and Chemicals*, May 2008

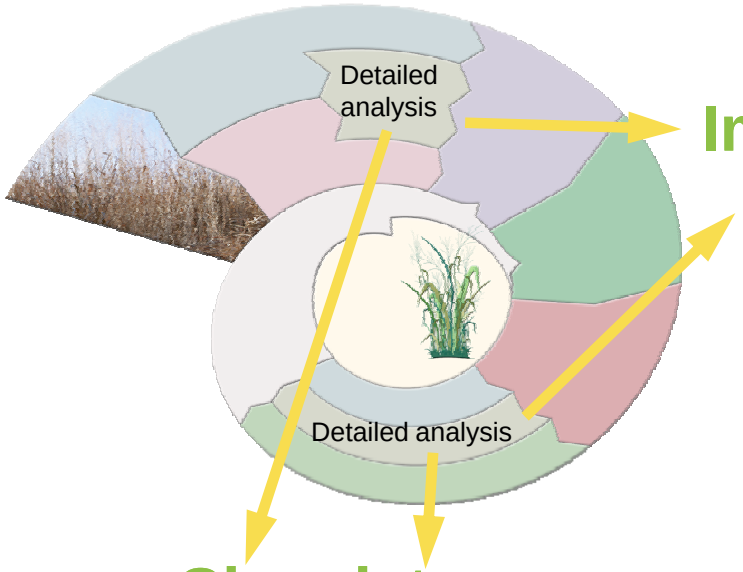
HTP Enzymatic Digestion Assays

- **Recalcitrance is ultimately determined by enzyme access to carbohydrates and sugar release.**
- **HTP assays are needed to assess recalcitrant phenotypes and to screen for more effective enzymes.**
- **1st tier assays: ~1000-5000 samples/day**
  - **Evaluate base-line susceptibility of pretreated biomass as well as enzymes from natural diversity**
- **2nd tier assays: ~200 samples/day**
  - **Hits from primary screen subjected to multi-dimensional assays using engineered enzyme cocktails for precise assessment of cell wall changes**



Response surface output for multi-dimensional digestibility assays

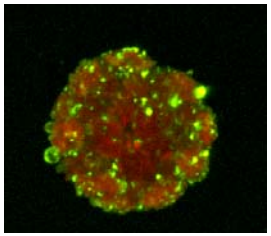
Detailed Analysis of Specific Samples Inform Cell-wall Chemistry and Structure

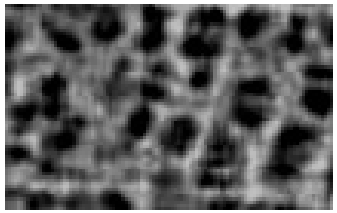


Imaging

AFM of switchgrass showing cellulose microfibrils

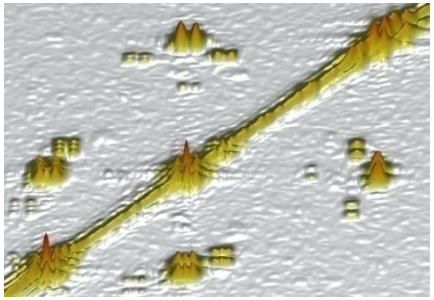
Immuno-localization using wall antibodies on Populus protoplasts



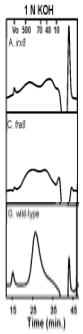


Bio-ultraCAT for 3-D density of Populus cell walls

Chemistry

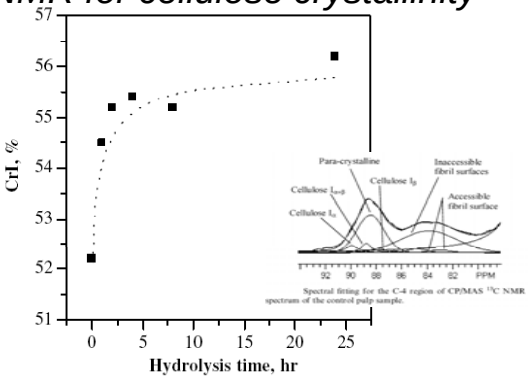


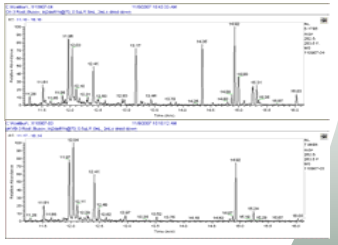
2D ^{1}H-NMR sees altered bonds in polysaccharides and lignin in biomass



Fractionation and chromatography

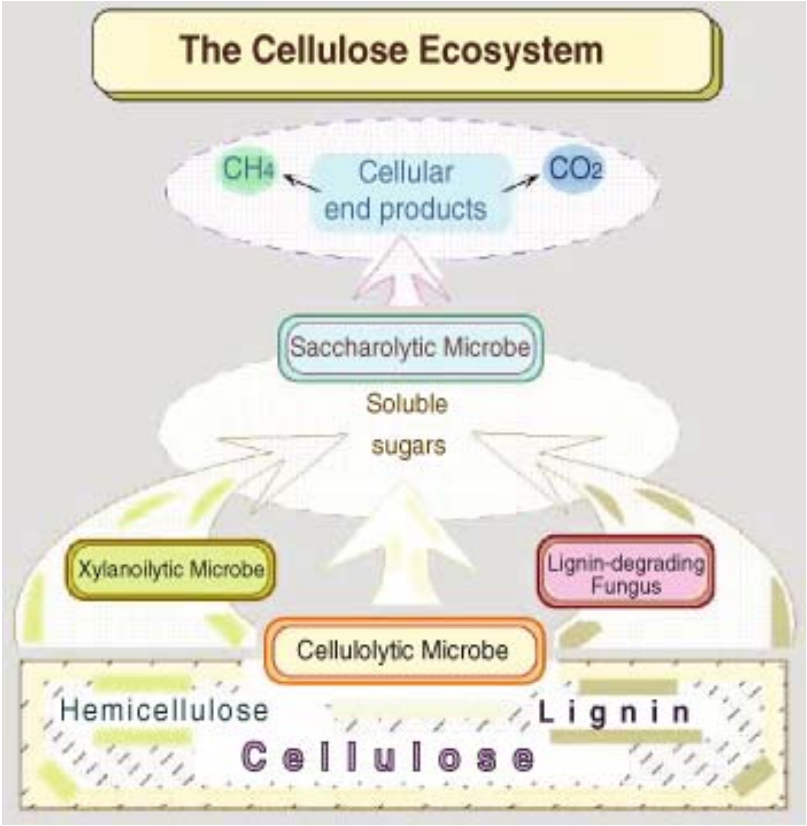
NMR for cellulose crystallinity





Mass Spectrometry for key metabolites

The Cellulose Ecosystem

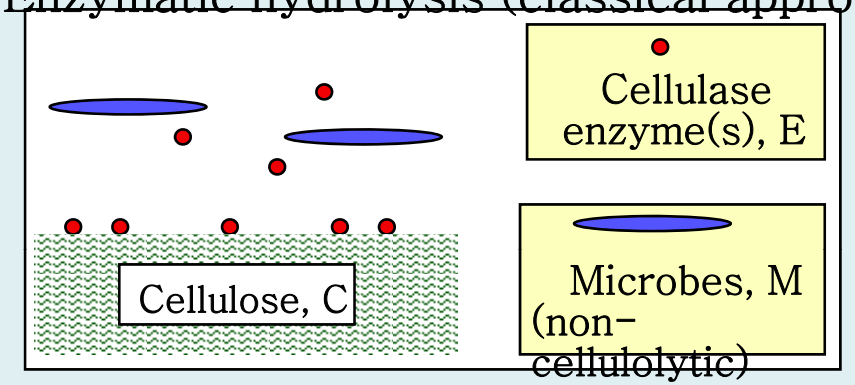


Simplified schematic description of a typical ecosystem comprising degrading plant matter. Cellulolytic, xylanolytic and ligninolytic microbes combine to decompose the major polysaccharide components to soluble sugars. “Satellite” microorganisms assimilate the excess sugars and other cellular end products, which are ultimately converted to methane and carbon dioxide.

The Prokaryotes
Volume 2: Ecophysiology and Biochemistry19. Cellulose-decomposing Bacteria and Their Enzyme Systems
Edward A. Bayer, Yuval Shoham and Raphael Lamed

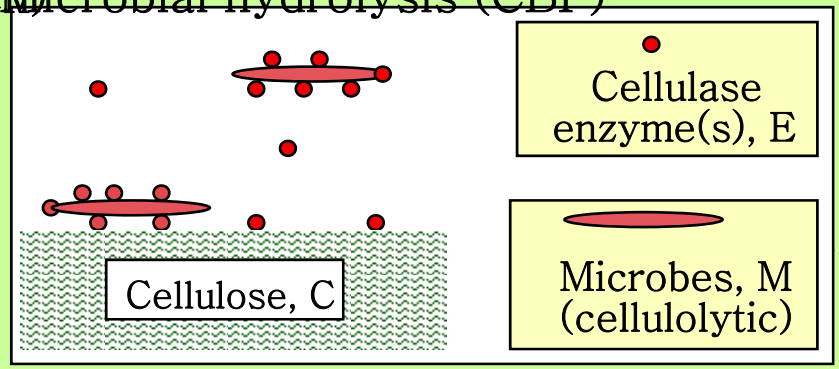
Microbial Hydrolysis and Enzymatic Hydrolysis: A Fundamentally Different Relationship Between Microbes and Cellulose

Enzymatic hydrolysis (classical approach)



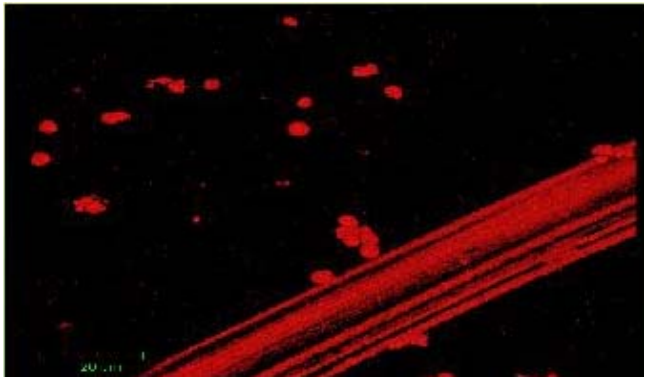
- Hydrolysis mediated by CE complexes
- Enzymes (several) both bound and free
- Cells may or may not be present

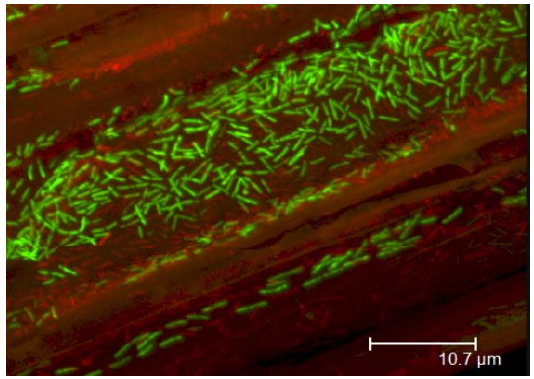
Microbial hydrolysis (CBP)



- Hydrolysis mediated mainly by CEM complexe
- Enzymes both bound and free
- Cells both bound and free

Yeast, enzymes with biomass, Dumitrache and Wolfaardt





C. thermocellum on poplar, Morrell-Falvey and Raman, ORNL

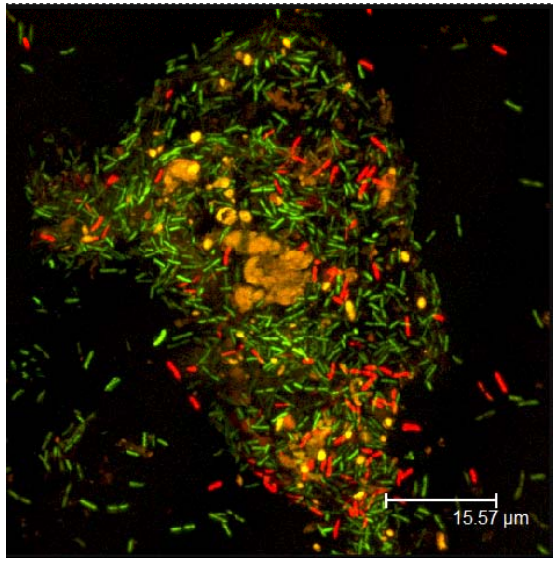
Biodiversity Access for New Biocatalysts

- **State-of-the-art cultivation techniques to isolate novel high-temperature microbes with powerful lignocellulolytic enzymes**
  - **Collect samples from thermal biotopes**
  - **Establish primary enrichment cultures at relevant temperatures and conditions**

Sampling at Yellowstone National Park, October 2007 and July 2008

High-throughput Isolation of Cellulolytic Extreme Thermophiles Using Flow Cytometry

Complex Enrichment

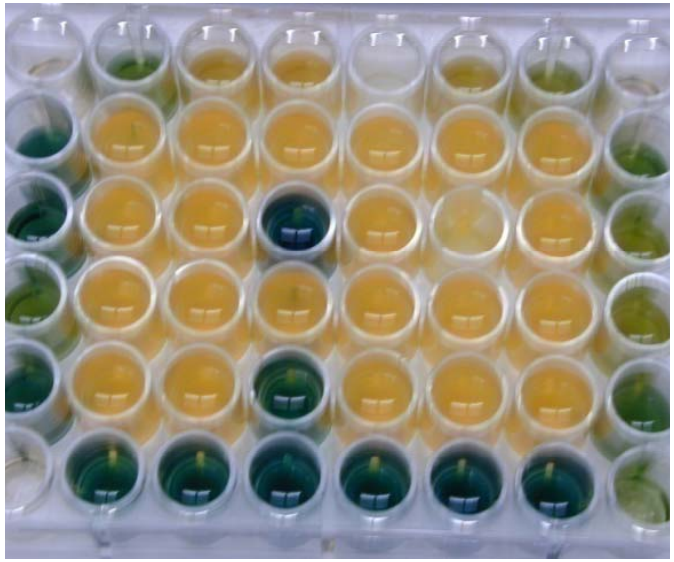


1. Establish primary enrichments from environmental samples on biomass.
2. Screen for growth and hydrolysis of pretreated biomass.

Single Cell Isolation

3. A single cell is deposited by Flow Cytometry in a culture well containing pretreated biomass.
4. Multi-well plates are incubated at 70-80 ^{o}C in the absence of oxygen.

High-throughput Screening

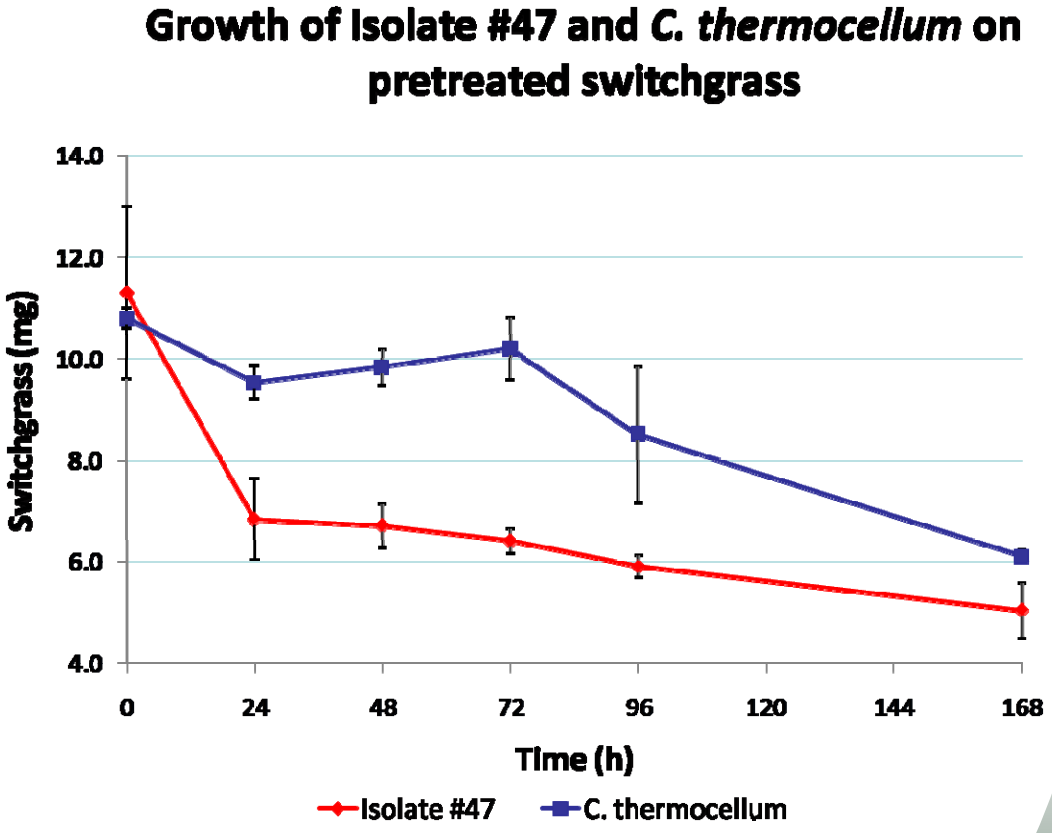


5. Plates are screened for growth and biomass hydrolysis.
6. High-throughput screening allows thousands of isolates to be evaluated with natural substrates.

New Isolates Show Enhanced Biomass Hydrolysis Rates

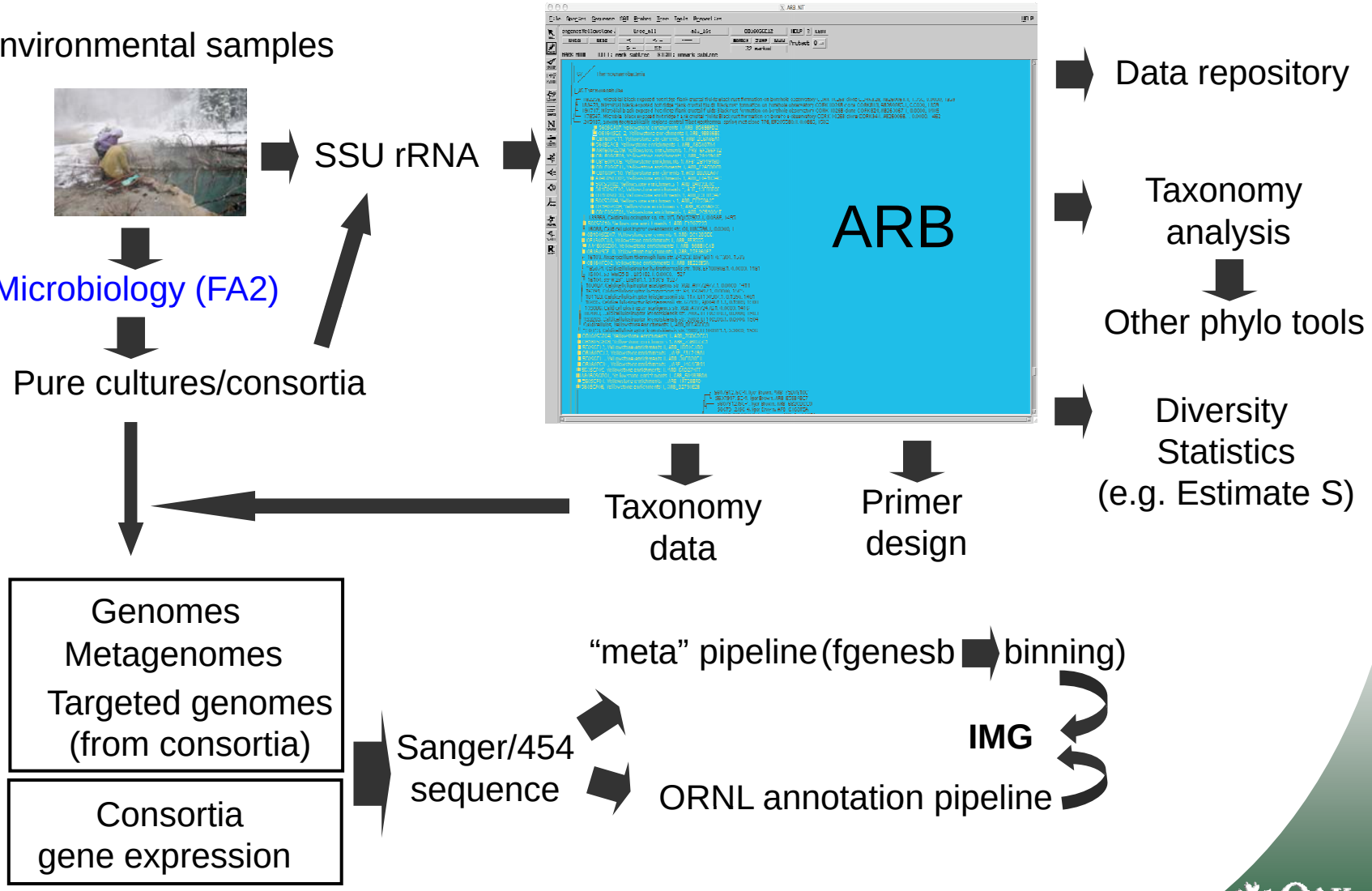
Isolate #47

Control

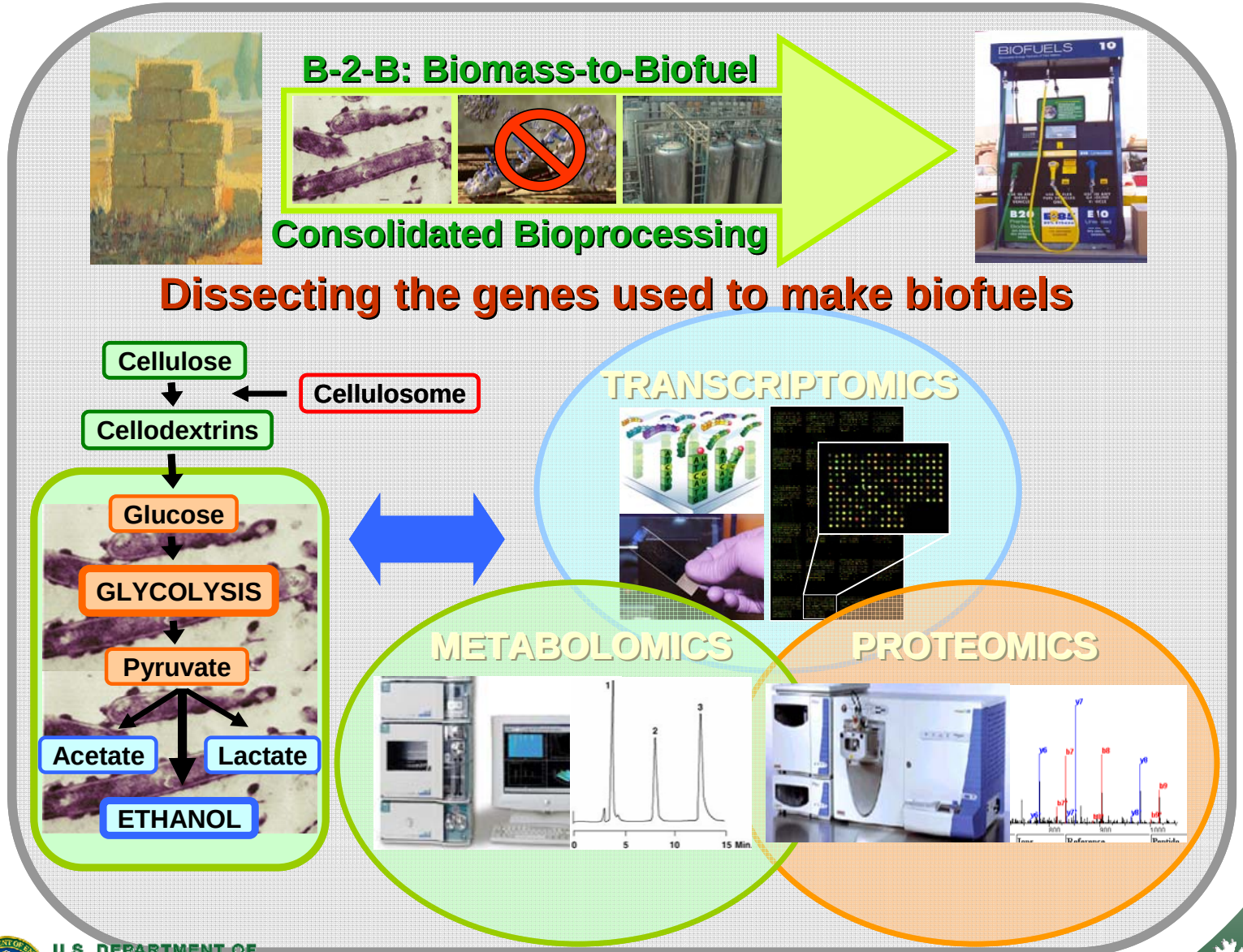


Preliminary results show visual disappearance of pretreated switchgrass solids during growth at 78˚C relative to a benchmark organism

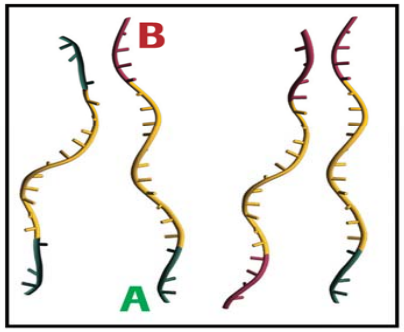
Microbial degradation pathways and networks: Diversity analysis and phylotyping



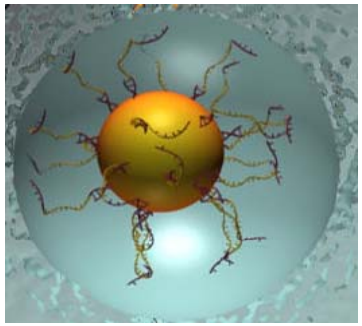
Systems Biology



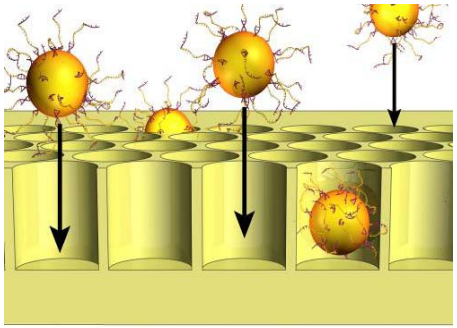
454 Process Overview



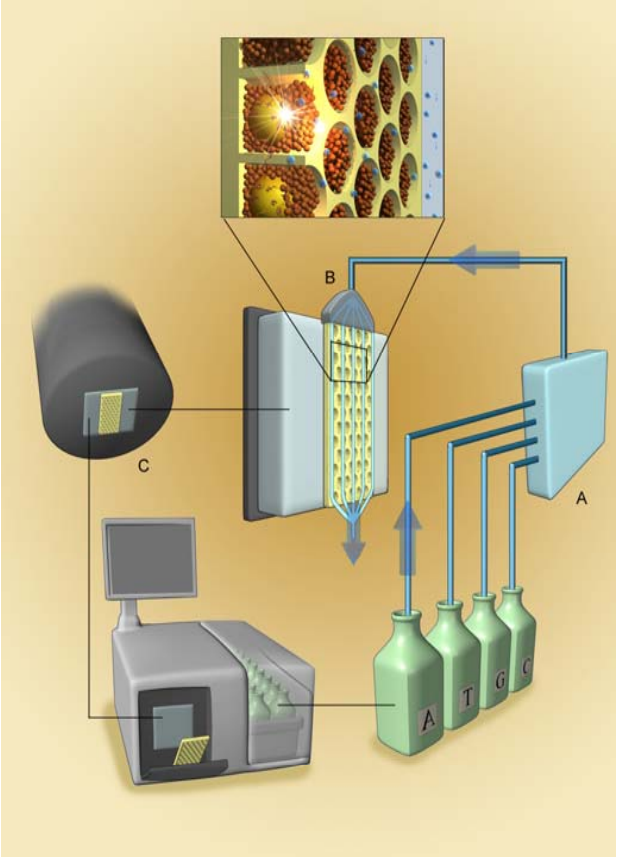
1) Prepare Adapter Ligated ssDNA Library (A-[insert]-B)



2) Clonal Amplification on 28 µ beads

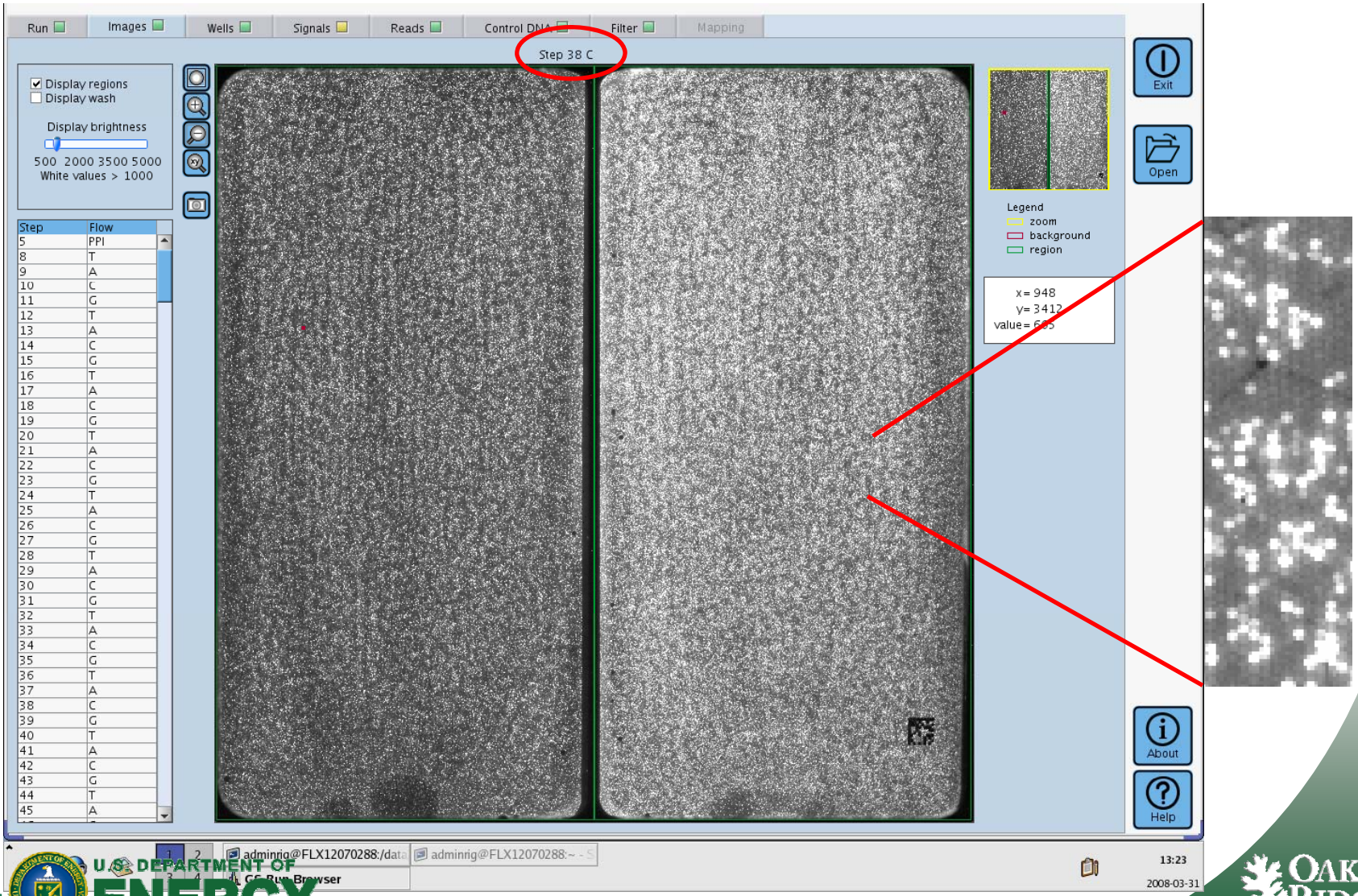


3) Load beads and enzymes in PicoTiterPlate™

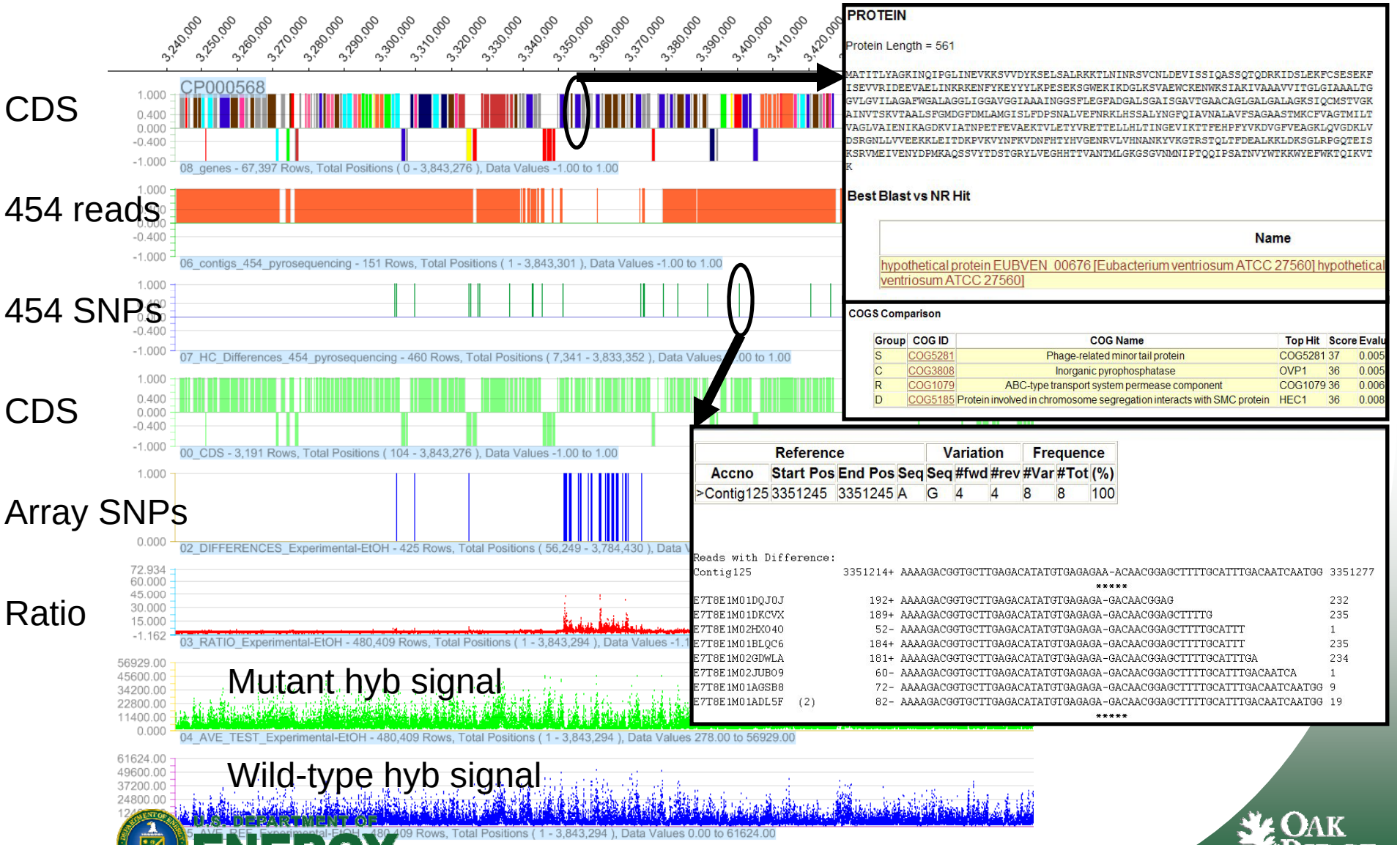


4) Perform Sequencing by synthesis on the 454 Instrument

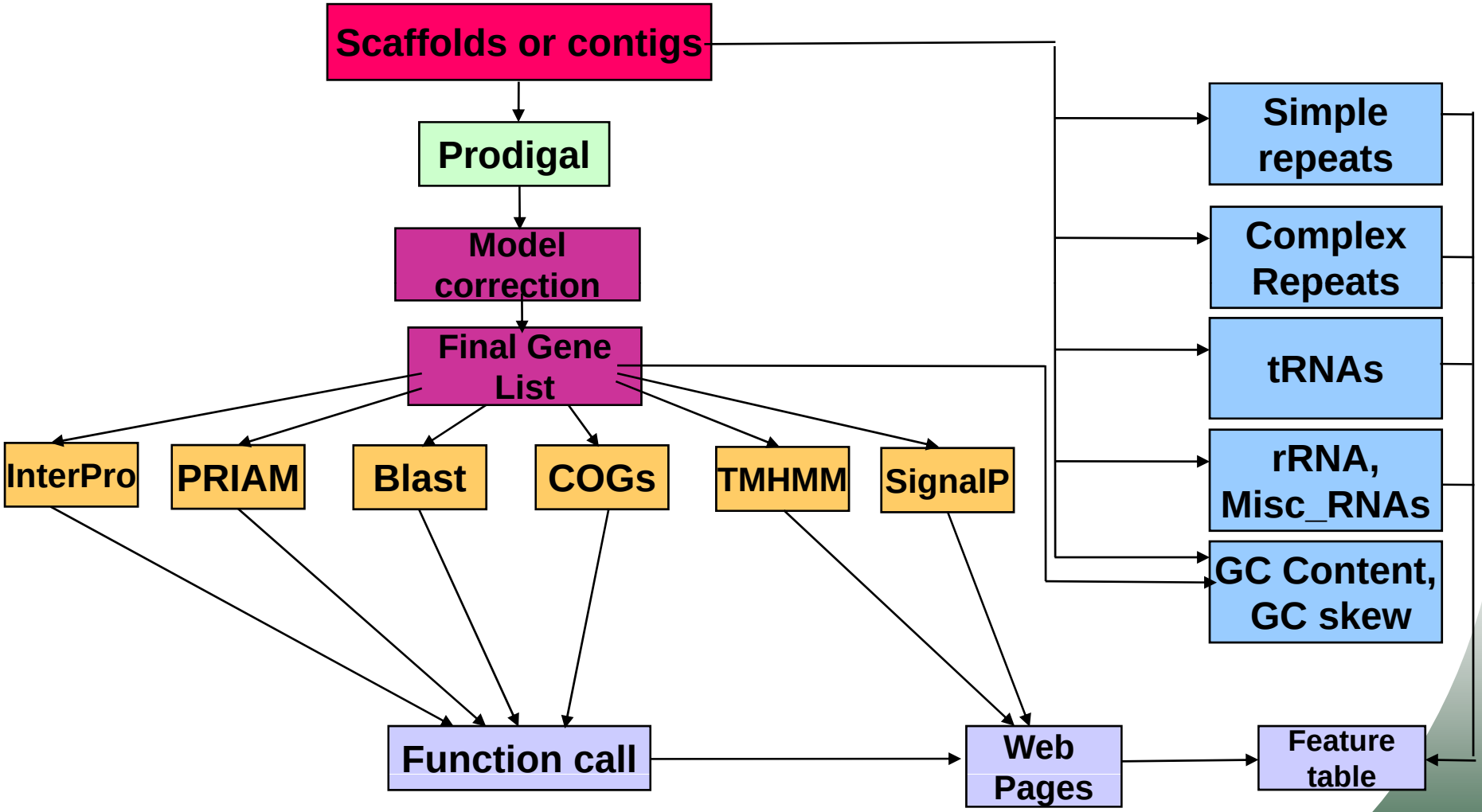
454 Run- data acquisition snapshot



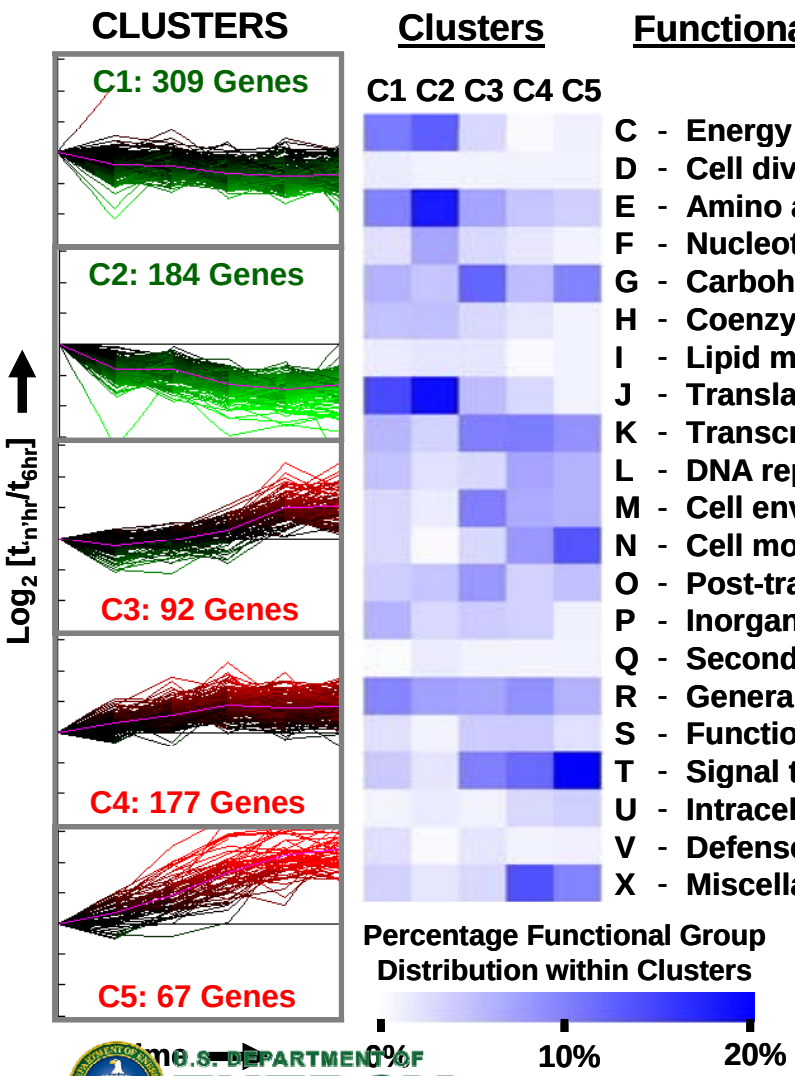
454 and Microarray Resequencing of *Clostridium thermocellum* mutant



Annotation Pipeline



Gene Expression Changes During *Cellulose* Fermentation



Clusters

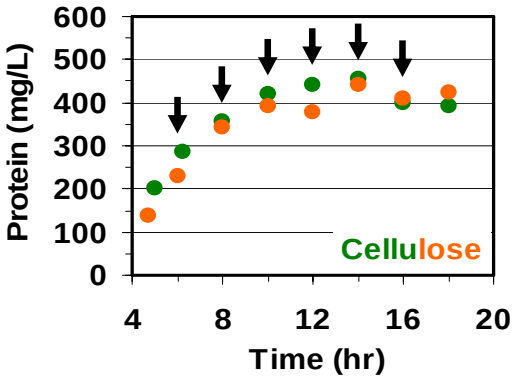
C1 C2 C3 C4 C5

Functional Group

- C - Energy production, conversion
- D - Cell division, chromosome partitioning
- E - Amino acid transport, metabolism
- F - Nucleotide transport, metabolism
- G - Carbohydrate transport, metabolism
- H - Coenzyme metabolism
- I - Lipid metabolism
- J - Translation, ribosomal structure, biogenesis
- K - Transcription
- L - DNA replication, recombination, repair
- M - Cell envelope biogenesis, outer membrane
- N - Cell motility, secretion
- O - Post-translational modification, protein turnover, chaperones
- P - Inorganic ion transport, metabolism
- Q - Secondary metabolites biosynthesis, transport, catabolism
- R - General function prediction only
- S - Function unknown
- T - Signal transduction mechanisms
- U - Intracellular trafficking, secretion, vesicular transport
- V - Defense mechanisms
- X - Miscellaneous

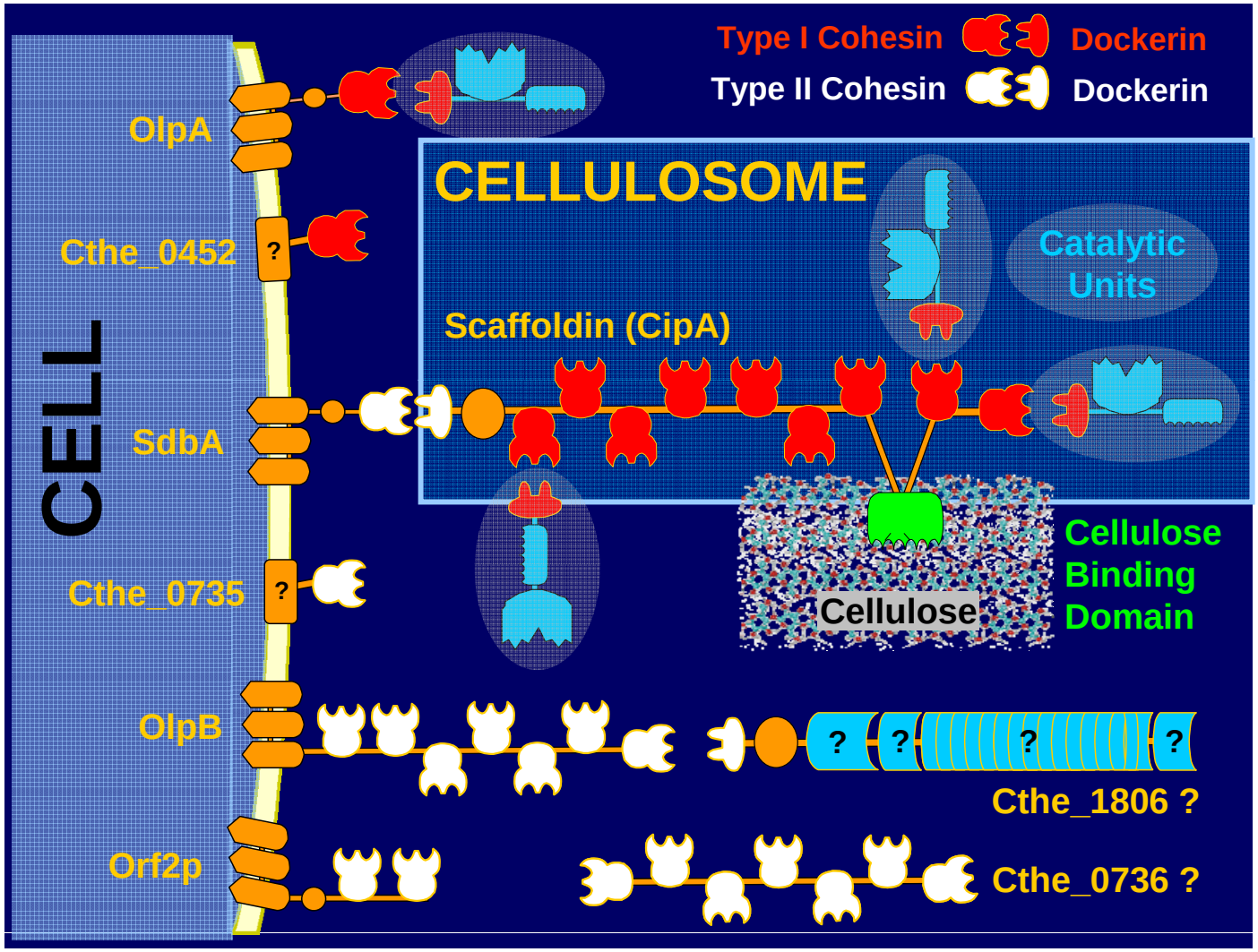
Percentage Functional Group Distribution within Clusters

0% 10% 20%



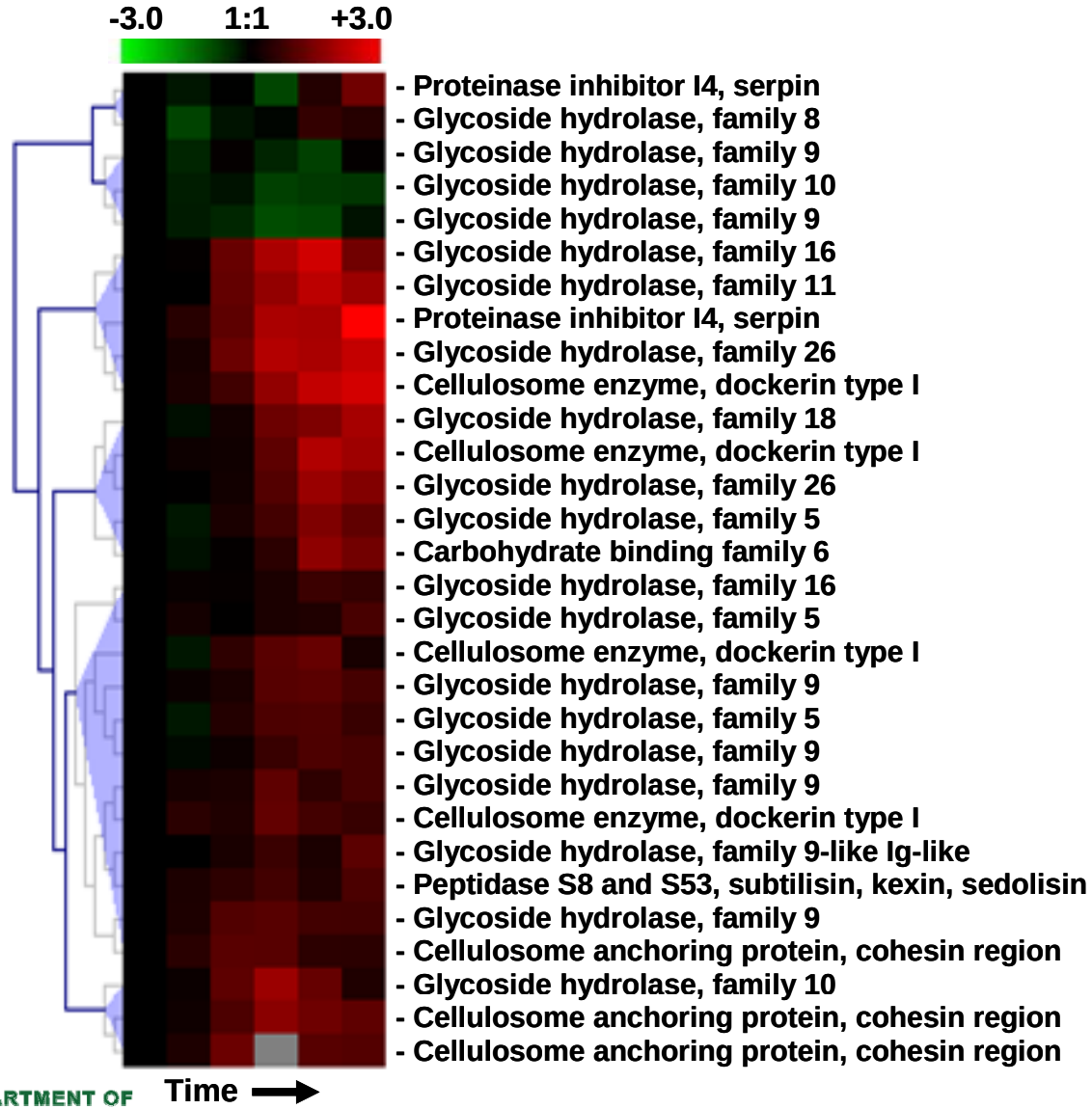
Paper in draft
Do not distribute

Schematic of Cellulosome



(adapted from Carlos Fontes, 2007 Gordon Research Conference on 'Cellulases and Cellulosomes')

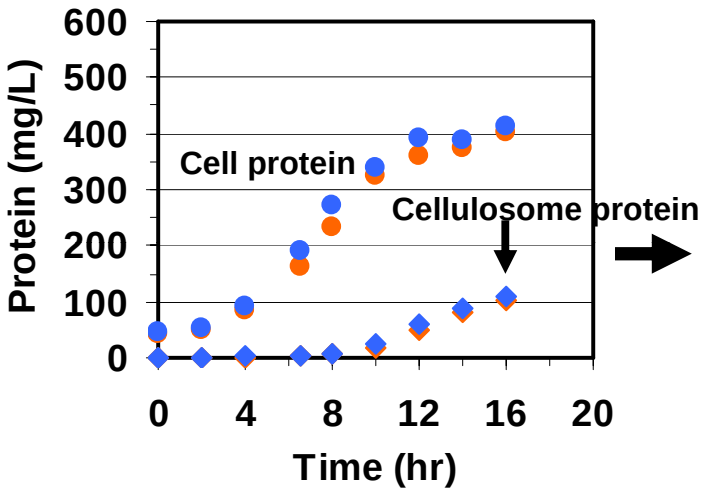
Transcript Changes in Cellulosomal Genes

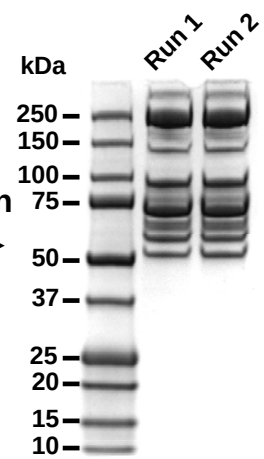


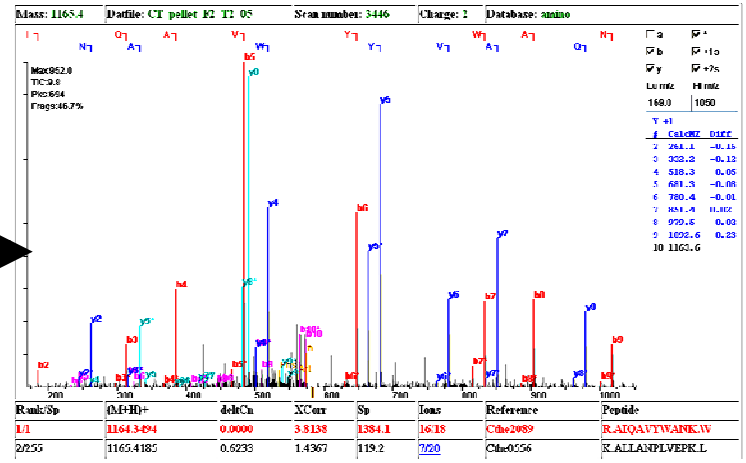
Paper in draft
Do not distribute

C. thermocellum Cellulosome

Metabolic Labeling:
Shot-gun LC-MS/MS





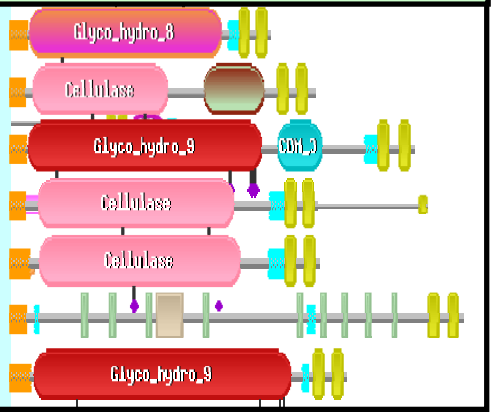


- Cellulosome is released when growth begins to slow down
- Cellulosome isolation via affinity digestion method
- In-solution trypsin digestion, following by shot-gun proteomics

Qualitative Analysis of Cellulosome Composition

Protein	Description	MW (Da)	Sum of Spectrum Count	Maximum Sequence Coverage	Relative to CipA	PFAM Domain Organization
CipA	Cellulosome anchoring protein, cohesin region	196831	1512	33.2	1.000	
	Cellulosome anchoring protein, cohesin region	248166	1427	15.8	0.944	
	Cellulosome anchoring protein, cohesin region	74971	101	33	0.067	
	Cellulosome anchoring protein, cohesin region	140561	83	35.1	0.055	
	Cellulosome anchoring protein, cohesin region	68619	57	27.9	0.038	
	Cellulosome anchoring protein, cohesin region	28469	13	16.7	0.009	
CelS	Glycoside hydrolase, family 48	83558	1118	55.2	0.739	
CelK	Glycoside hydrolase, family 9	100622	491	48.9	0.325	
XynA/U	Glycoside hydrolase, family 11	74471	470	43.9	0.311	
CbhA	Glycoside hydrolase, family 9	137116	324	45	0.214	
	Glycoside hydrolase, family 9	89488	19	8	0.013	
	Carbohydrate binding family 6	103126	19	8.8	0.013	
	Glycoside hydrolase, family 9	80223	18	14.9	0.012	
CelH	Carbohydrate binding family 11	102416	13	8.6	0.009	
	Carbohydrate binding family 6	70385	13	17.9	0.009	
	Cellulosome enzyme, dockerin type I	64556	12	11	0.008	
	Glycoside hydrolase, family 5	59921	10	12	0.007	
	Glycosyl hydrolase 53	47067	10	8.9	0.007	
	Cellulosome enzyme, dockerin type I	51385	9	12.6	0.006	
	Ricin B lectin	63863	9	7.5	0.006	
	Cellulosome enzyme, dockerin type I	46800	9	18	0.006	
	Glycoside hydrolase, family 9	109001	8	8.3	0.005	
	Cellulosome enzyme, dockerin type I	55169	8	9.8	0.005	
LicB	Glycoside hydrolase, family 16	37897	6	17.4	0.004	
	Alpha-L-arabinofuranosidase B	110674	4	3.1	0.003	

Protein	Description	MW (Da)	Sum of Spectrum Count	Maximum Sequence Coverage	Relative to CipA	PFAM Domain Organization
CelA	Glycoside hydrolase, family 8	52562	243	48.8	0.161	
	Coagulation factor 5/8 type-like	63023	213	41.4	0.141	
CelF	Glycoside hydrolase, family 9	82071	194	48.3	0.128	
CelB	Glycoside hydrolase, family 5	63929	167	43.7	0.110	
CelG	Glycoside hydrolase, family 5	63199	165	26.3	0.109	
XghA	Cellulosome enzyme, dockerin type I	92365	162	52.6	0.107	
CelT	Glycoside hydrolase, family 9	68511	139	46.8	0.092	



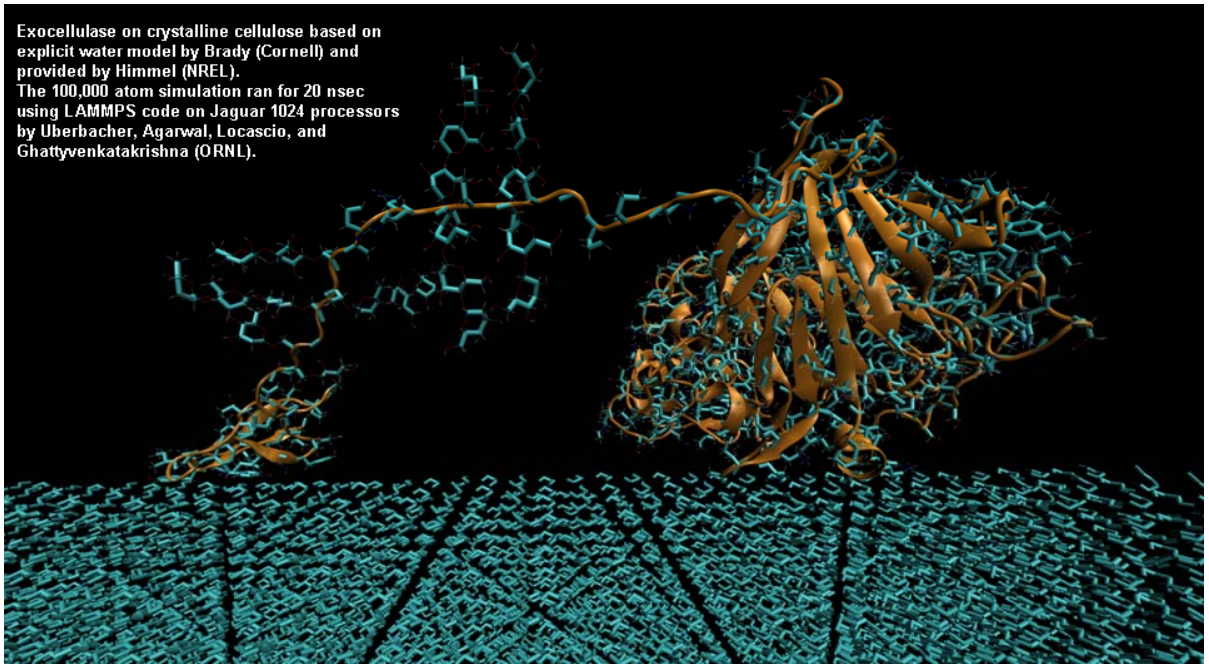
Paper in draft
Do not distribute

U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

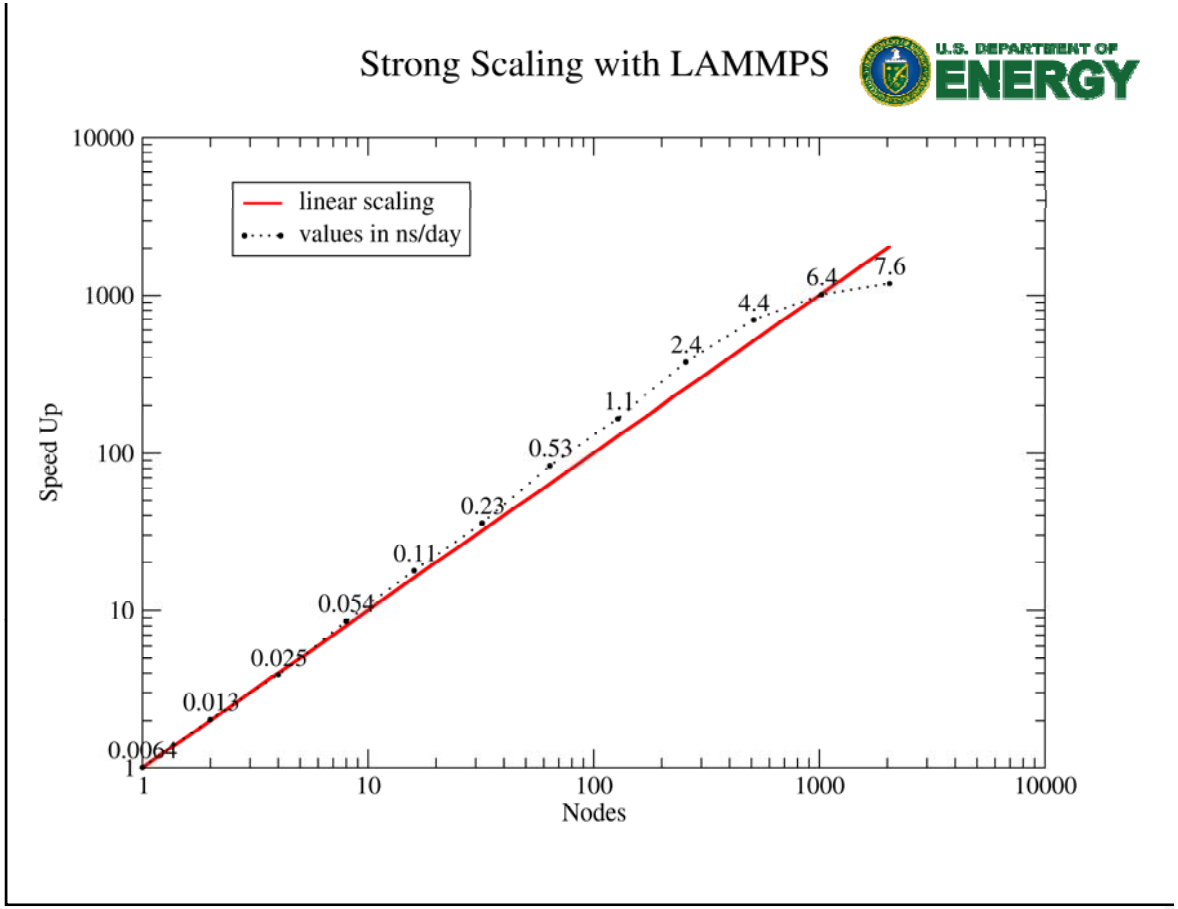
Modeling CBH1 cellulase on cellulose

- **100,000 atom model of enzyme cellulase on crystalline cellulose in water**



- **Computation for 50 nS on Jaguar by Uberbacher, Agarwal et al.**
- **Based on explicit model of Brady (Cornell) and Himmel (NREL) – leads of SciDac project.**
- **Shows cellulose binding domain and active site interaction with surface**

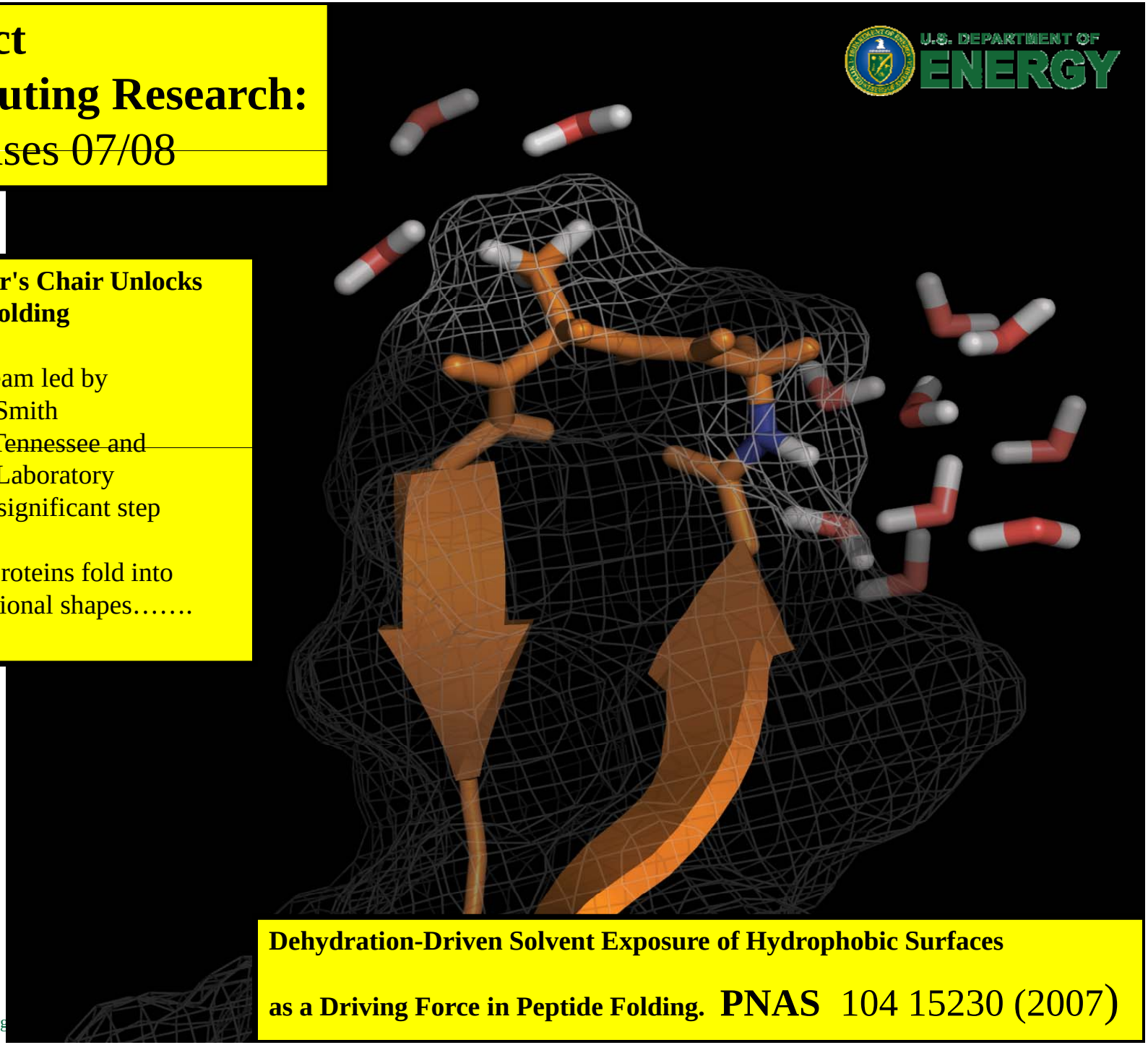
One Million Atoms – Molecular Dynamics



News flash: Biosystems now scale to 30,000 processors (Aug 08)

High-Impact Supercomputing Research:
3 press releases 07/08

UT-ORNL Governor's Chair Unlocks Secrets of Protein Folding
September 17, 2007
KNOXVILLE -- A team led by biophysicist Jeremy Smith of the University of Tennessee and Oak Ridge National Laboratory (ORNL) has taken a significant step toward unraveling the mystery of how proteins fold into unique, three-dimensional shapes…….



Dehydration-Driven Solvent Exposure of Hydrophobic Surfaces as a Driving Force in Peptide Folding. **PNAS** 104 15230 (2007)

Biology and Peta-scale Computing: The Future

- **How do we handle data?**
- **How do we compare data?**
- **Do we need to create data standards for biology?**
- **We need to create tools to compare and analyze data across multiple types and systems**

→ Need for central data storage and "data rich" computing

Thank you

BESC 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