Caldicellulosiruptor Pan-Genomics: Perspectives on Newly Sequenced Species, and Genus-Wide Diversity of Cellulose Binding Proteins (tāpirins)

Laura L. Lee[1,2]* (lllee@ncsu.edu), Sara E. Blumer-Schuette,[1,2] Javier A. Izquierdo,[1,2] Kasey E. Koballa,[1,2] Kevin T. Mullen,[1,2] Jonathan M. Conway,[1,2] Jeffrey V. Zurawski,[1,2] Piyum A. Khatibi,[1,2] Robert M. Kelly,[1,2] and **Paul Gilna**[2]

[1]North Carolina State University, Raleigh; [2]BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves (1) designing plant cell walls for rapid deconstruction and (2) developing multi-talented microbes or converting plant biomass into biofuels in a single step [consolidated bioprocessing (CBP)]. BESC research in biomass deconstruction and conversion targets CBP by studying thermophilic anaerobes to understand novel strategies and enzyme complexes for biomass deconstruction and manipulating these microorganisms for improved conversion, yields, and biofuel titer.

Extremely thermophilic organisms have a promising, but yet unrealized, role to play in the microbial production of lignocellulosic ethanol. Discovery of their potential has been pursued through genomics (metagenomics and pan-genomics), as well as the investigation of novel cellulose binding proteins (tāpirins). First, samples from Obsidian Pool in Yellowstone National Park were metagenomically sequenced to identify thermophilic cellulases and ethanol-forming enzymes. The 16S rRNA and Illumina DNA sequences revealed novel enzymes and organisms of interest, while PacBio sequencing has been used to obtain longer reads and potentially closed genomes. Next, we characterized the *Caldicellulosiruptor* core and pan-genomes with the GET_HOMOLOGUES software in order to determine the genetic diversity within the genus. Presently, the core genome contains 1284 genes, but the pan-genome is open, as the size is still increasing as the number of sequenced isolates grows. The three newest species' genomes (*C.* sp. str. Rt8.B8, *C.* sp. str. Wai35.B5, and *C.* sp. str. NA10) were examined for the presence of new or uncommon glycoside hydrolases (GH) and surface layer homology domain proteins. One of the most interesting finds was a multi-modular enzyme with a GH12 domain (along with a GH48, multiple CBMs, and either a GH5 or GH10); these are the first *Caldicellulosiruptor* species sequenced at this point to have a multi-modular CAZyme with a GH12 domain. Finally, novel proteins (tāpirins) were identified via transcriptomics and proteomics to be highly expressed in cellulose-bound fractions. Structural homology to other classes of proteins could not be assigned, indicating that this is truly a new class of biomolecules and establishing a new paradigm for how cellulolytic bacteria adhere to cellulose. Overall, these results bring a more comprehensive understanding of the *Caldicellulosiruptor* genus, as well as shed light on novel CAZymes and proteins in both characterized and novel species, which could have important roles in how these microbes degrade polysaccharides.

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.