Using Iterative Random Forest to Predict the Progeny of Crosses in *Populus trichocarpa*

Jonathon Romero*,[1,2] (romerojc@ornl.gov), David Kainer,[1], Ashley Cliff,[1,2], Daniel Jacobson,[1,2] and **Gerald A. Tuskan**[1]

[1]Center for Bioenergy Innovation, Oak Ridge National Laboratory, TN; [2]Bredesen Center, University of Tennessee, Knoxville, TN

cbi.ornl.gov

The Center for Bioenergy Innovation (CBI) vision is *to accelerate domestication of bioenergy-relevant, non-model plants and microbes to enable high-impact innovations at multiple points in the bioenergy supply chain.* CBI addresses strategic barriers to the current bioeconomy in the areas of 1) high-yielding, robust feedstocks, 2) lower capital and processing costs via consolidated bioprocessing (CBP) to specialty biofuels, and 3) methods to create valuable byproducts from the lignin. CBI will identify and utilize key plant genes for growth, composition and sustainability phenotypes as a means of achieving lower feedstock costs, focusing on poplar and switchgrass. We will convert these feedstocks to specialty biofuels (C4 alcohols, C6 esters and hydrocarbons) using CBP at high rates, titers and yield in combination with cotreatment, pretreatment or catalytic upgrading. CBI will maximize product value by *in planta* modifications and biological funneling of lignin to value-added chemicals.

The goal of this task was to develop a method that uses Explainable Artificial Intelligence (X-AI) to assist in the Genomic Selection of *Populus trichocarpa* thus improving desired positive phenotypes such as yield and pathogen resistance in a manner that takes fewer generations than traditional Genomic Selection techniques.

Iterative Random Forest (iRF) [1,2] is a machine learning model with multiple uses in the field of Genomics. We have shown that iRF, when used in conjunction with *in silico* breeding software (SBVB) [3], can accurately predict the real outcome of crossing two parents together by training the model on a GWAS population that does not contain the parents. This method takes advantage of the data collected from a GWAS population, but by using iRF, the method is able to incorporate non-additive effects that it detects in the population to aid in model accuracy. Once the iRF model has been trained for the phenotype of interest, the sequenced genomes of potential parents are synthetically crossed to produce virtual full-sibling families. The virtual genomes of each progeny are then used as input to the iRF model to predict the distribution of the phenotypes expected from the progeny of each parental cross. The rank order of these families proved to be a highly accurate representation of the rankings of the actual families they represent when grown in a common garden. This method selected crosses from a large diversity of parents instead of from only the top ranked males and females that a traditional Genomic Selection technique would use. This achieves an important goal in genomic selection, which is to keep as much genetic diversity as possible

while improving a selected phenotype. This method has been validated in *Populus trichocarpa* by accurately predicting the ranking of height of *in silico* progeny and comparing to real progeny crossed from the same parents.

References:

1. Sumanta Basu, Karl Kumbier, James B. Brown, Bin Yu, Iterative random forests, *Proceedings of the National Academy of Sciences* Feb 2018, 115 (8) 1943-1948; DOI: 10.1073/pnas.1711236115
2. Cliff A, Romero J, Kainer D, Walker A, Furches A, Jacobson D. A High-Performance Computing Implementation of Iterative Random Forest for the Creation of Predictive Expression Networks. *Genes* (Basel). 2019 Dec 2;10(12):996. doi:10.3390/genes10120996. PMID: 31810264; PMCID: PMC6947651.
3. Pérez-Enciso M, Forneris N, de Los Campos G, Legarra A "Evaluating Sequence-Based Genomic Prediction with an Efficient New Simulator." *Genetics* 2017; 205(2):939-953 https://doi.org/10.1534/genetics.116.194878

The Center for Bioenergy Innovation 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.