QTG-Finder: A Machine-Learning Algorithm to Prioritize Causal Genes of Quantitative Trait Loci in Plants

Fan Lin, Jue Fan, and **Seung Y. Rhee***

Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305, USA

Project Goals: This project (www.foxmillet.org) aims to leverage *Setaria viridis* as a model system to develop novel technologies and methodologies to redesign the bioenergy feedstock *Sorghum bicolor* to enhance water use and photosynthetic efficiencies. In this study we developed a computational pipeline is to accelerate the discovery of causal genes in QTLs by reducing the number of candidates to be tested experimentally.

Linkage mapping is one of the most commonly used methods to identify genetic loci that determine a trait. However, the loci identified by linkage mapping may contain hundreds to thousands of candidate genes and require a time-consuming and labor-intensive fine mapping process to find the causal gene controlling the trait. With the availability of a rich assortment of genomic and functional genomic data, it is possible to develop a computational method to facilitate faster identification of causal genes. We developed QTG-Finder, a machine-learning algorithm to prioritize causal genes by ranking genes within a quantitative trait locus (QTL). Two predictive models were trained separately based on known causal genes in Arabidopsis and rice. With an independent validation analysis, we demonstrate the models can correctly prioritize about 80% and 55% of Arabidopsis and rice causal genes when the top 20% ranked genes were considered. The models can prioritize different types of traits though at different efficiency. We also identified several important features of causal genes including non-synonymous SNPs at conserved protein sequences, paralog copy number, and being a transporter. This work lays the foundation for systematically understanding characteristics of causal genes and establishes a pipeline to predict causal genes based on public data. Currently, we are expanding this algorithm to other species such as Setaria and Sorghum by using the othologs of known causal genes as a training set.

Funding statement:

This project is funded by grant DE-SC0018277 from The DOE Department of Biological and Environmental Research.