Improvement of Rice Genome Annotation through Comparative Analyses

Wei Zhu, Shu Ouyang, Haining Lin, John Hamilton, Matthew A Campbell, Kevin Childs, Françoise Thibaud-Nissen, Joshua Orvis, Brian Haas, Jennifer Wortman and C. Robin Buell

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

A critical component of any genome project is generating high quality annotation at the structural and functional level. For the grasses, genome projects are complete, in progress, or planned for rice, maize, sorghum, and Brachypodium. With a finished genome, rice provides an important resource for accurate annotation of other genomes. To determine the extent to which we can annotate across grass genomes, we performed genome level comparisons between rice and all plant species in which large genomic or transcriptomic datasets are available. Using four plant genome sequence datasets (maize, sorghum, Arabidopsis, and poplar) and transcript assemblies from 185 plant species representing 2.6 Gb of sequence, we were able to identify support for 38,109 (89.3%) of the total 42,653 non transposable element-related genes in the rice genome in at least one plant species. Through these alignments, we were able to support not only known, putative, and expressed genes but also genes that are currently lacking expression or database similarity and are currently annotated as hypothetical genes. With respect to the grasses, a majority of the rice genes could be aligned to a Poaceae sequence with only a very small number supported solely by non-Poaceae sequence data. Through comparative alignments, we could identify missed genes in our rice genome annotation, identify non-coding sequences, and confirm gene nesting in rice. The comparative alignments were also valuable in a reciprocal manner and the rice genome annotation will clearly be an important resource for annotation of maize and sorghum.