

IOWA STATE UNIVERSITY

Department of Genetics, Development and Cell Biology

Jing Li, Ph.D. Defense

Wednesday, June 15, 2022

10 a.m.

Zoom link: <https://iastate.zoom.us/j/91822373063?pwd=TC8ydmlUVitQbGVmWExCbDFqMIUydz09>

Major: Genetics and Genomics

Major professor: GDCB Professor Eve Wurtele

Annotating and characterizing orphan gene in *Zea mays* via diverse RNA-Seq data

In the past, many studies have dismissed the pervasive transcribed but unannotated transcripts as transcriptional “noise”, which I referred as “dark transcriptome”. Some functional genes have been identified from this dark transcriptome. Most of the dark transcriptome are orphan genes. Orphan genes are the recently emerged young genes, which shares no sequence similarity with proteins in any other species. In recent 20 years, thousands of orphan genes have been experimentally shown to play important roles in diverse species. However, there remain significant limitation in our knowledge about orphan genes and their function.

The traditional gene model prediction pipeline based on *ab initio* method and homologous sequence, which is lack in the orphan genes. It is hard to predict orphan genes by traditional method. Moreover, orphan genes usually only expressed in some specific conditions. Even though many *de novo* genes have been evaluated in several studies from multiple RNA-Seq evidence, the limited library conditions may restrict the identification of orphan genes. Currently, we have no idea about the amount of orphan gene in a genome.

The gene functional prediction is majority based on sequence and domain similarity, only a small set of gene function inferred directly from experimental evidence. Orphan gene cannot be inferred functional annotation via the traditional method. Even though some orphan genes have been experimentally characterized, most of their functions are not integrated in the public database.

This dissertation presents methods and tools to evaluate potential orphan genes, predict potential orphan genes and their function efficiently. First, I comprehensively evaluated all potential ORFs in yeast using over 3,000 RNA-Seq and Ribo-Seq samples for transcription and translation evidence. Next, I developed light weight, flexible, reproducible, and well-documented pipeline, BIND, to improve orphan gene prediction. Finally, I provide improved gene model predictions using BIND, and comprehensive functional annotations using co-expression analysis from over 1,000 RNA-Seq samples for 26 inbred lines in *Zea mays* subsp. *mays*. The functional annotation was validated by enrichment analysis with differential expression analysis. Thousands of orphan genes showed specific expression in at least one stress condition and tissue. The annotation of pan-orphan genes, especially for the line-specific genes, hold potential to help agronomists and geneticists to use as molecular markers for marker-assisted selection and to develop desired varieties for maize.