

Genome-wide, Organ-delimited gene regulatory networks (OD-GRNs) provide high accuracy in candidate TF selection across diverse processes

Rajeev Ranjan^{1,2*}, Sonali Srijan¹, Somaiah Balekuttira¹, Tina Agarwal^{1,2}, Melissa Ramey¹, Madison Dobbins¹, Xiaojin Wang^{1,2}, Karen Hudson³, Ying Li^{1,2}, Kranthi Varala^{1,2#},

¹Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN, 47907, USA; ²Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA; ³USDA-ARS Crop Production and Pest Control Research Unit, West Lafayette, IN 47907

Construction of organ-specific gene expression datasets that include hundreds to thousands of experiments would greatly aid reconstruction of gene regulatory networks with organ-level spatial resolution. However, creating such datasets is greatly hampered by the requirements of extensive and tedious manual curation. Here we trained a supervised classification model that can accurately classify the organ-of-origin for a plant transcriptome. This K-Nearest Neighbor based multiclass classifier was used to create organ-specific gene expression datasets for the leaf, root, shoot, flower, seed, seedling, silique, and stem in the model plant *Arabidopsis thaliana*. In the leaf, root, flower, seed and, a gene regulatory network (GRN) inference approach was used to determine: i. influential transcription factors (TFs) in that organ and, ii. The most influential TFs for specific biological processes in the organ. These genome-wide, organ-delimited GRNs (OD-GRNs), identified *de novo* many known regulators of organ development and processes operating in those organs. Moreover, many previously unknown TF regulators were highly ranked as potential master regulators of organ development or organspecific processes. As a proof-of-concept, we focused on experimentally validating the predicted TF regulators of lipid biosynthesis in seeds, with relevance to food and biofuel production. Of the top twenty candidate TFs, eight (e.g., WRI1, LEC1, and FUS3) are known regulators of seed oil content. Importantly, we validated that seven more candidate TFs, whose role was previously unknown in seed lipid biosynthesis, indeed affect this process by genetics and physiological approaches, thus yielding a net accuracy rate of >75% for the *de novo* TF predictions. The general approach developed here could be extended to any species with sufficiently large gene expression datasets to speed up hypothesis generation and testing for constructing gene regulatory networks at a high spatial resolution.