ABSTRACT

USING TRANSCRIPTOME AND DATA SCIENCE METHODS TO UNCOVER GENE REGULATORY AND FUNCTIONAL INFORMATION

By

Sahra Uygun

Even in the well-studied model organisms, there are still genomic regions with unknown function. These genomic regions include protein-coding genes and regulatory elements that are key components of transcriptional regulation. With technological advances, more biological data are being generated including spatial, temporal, developmental, and conditional gene expression data. Gene expression data, and specifically co-expression analyses have been widely used to predict gene function through guilt by association. However, it remains to be seen to what degree co-expression is informative, whether it can be applied to genes involved in different biological processes, and how the choice of gene expression dataset and clustering algorithms impact inferences about gene functions. To answer these questions, I used co-expression to identify novel genes that function in a biological process, and the impact of different clustering algorithms on the ability to identify genes that function in the same pathway. Apart from the functional associations, gene co-expression analyses can also be used to identify the putative cis-regulatory elements that are over-represented in co-expressed gene promoters. These elements can be used to build models of gene regulation under changing environments and genome-wide models of how different organ and cell type gene expression are regulated under changing environments have not yet been built in plants. I used Arabidopsis thaliana organ and cell type stress responsive gene expression data and co-expression clusters to identify putative cis-regulatory elements. Using these elements and machine learning models, I predicted high salinity responsive gene expression in shoots, roots and six root cell types. I found that plant organ and cell type transcriptional response to high salinity

are likely regulated by a core set of elements that we identified and built predictive models of plant

spatial transcriptional responses to environmental stress. Overall, this research contributes to

understanding the role of “big data” in biology, provides guidelines for effectively using gene coexpression in functional associations and shows how computational approaches help in identifying

gene regulatory information.