

AN ABSTRACT OF THE DISSERTATION OF

Valerie N. Fraser for the degree of Doctor of Philosophy in Molecular and Cellular Biology presented on August 2, 2021.

Title: Investigating Transcriptional Control of Specialized Gene Expression in Plants

Abstract approved:

Molly Megraw

Specialized or secondary metabolism is a collection of pathways and small molecules that, while beneficial to an organism, are not strictly necessary for survival. Plants use secondary metabolites to, among other things, attract pollinators, defend against biotic and abiotic stressors, and form symbioses. Natural products from plants have seen an increase in scientific interest as many of these compounds have implications for human use. One of the main limitations in natural product research is the inability to produce relevant compounds in heterologous hosts or cell tissue culture, as some intermediate steps in the biosynthetic pathways are limited to a specific tissue or cell type.

In this work, my collaborators and I work to address some of the larger challenges that come from this limitation, largely through the lens of transcriptional control of gene expression. My colleagues and I began by using machine learning in the model plant *Arabidopsis thaliana* to explore the primary determinants of tissue specific gene expression. A pair of predictive models were built using precise transcriptomic and chromatin accessibility data; models of this type—L1-regularized logistic regression models—allow the user to extract the factors determined to be important for making the predictions. We found that most of the highly weighted factors for both models are sequence motifs, with actively transcribed promoters having a general openness of the chromatin.

Next, we investigated what factors affect biosynthesis of the chemotherapeutic compound vinblastine and its key intermediates in the medicinal plant *Catharanthus roseus*. Using a combination of metabolomics and gene expression analysis, we found that both plant variety and hormonal treatment are critical components in determining metabolite production levels. Additionally, we found that these factors have an impact on expression levels of master regulators and key biosynthetic pathway genes.

©Copyright by Valerie N. Fraser
August 2, 2021
All Rights Reserved

Investigating Transcriptional Control of Specialized Gene Expression in Plants

by
Valerie N. Fraser

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 2, 2021
Commencement June 2022

Doctor of Philosophy dissertation of Valerie N. Fraser presented on August 2, 2021.

APPROVED:

Major Professor, representing Molecular and Cellular Biology

Director of the Molecular and Cellular Biology Program

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Valerie N. Fraser, Author

ACKNOWLEDGEMENTS

It truly takes a village to do the research for and to write a dissertation. I would like to begin by thanking my major advisor, Dr. Molly Megraw. Thank you for taking a chance on me as a wet-lab student in your dry-lab world and for your patience and support. I am also immensely grateful to you for paving the way for me to make fabulous connections in the greater plant biology world and for trusting me to not run your lab into the ground when all our senior personnel retired.

To each of my committee members—Dr. John Fowler, Dr. Jeffrey Anderson, Dr. Michael Freitag, and Dr. Kerry McPhail—thank you for your personal and professional mentorship and support, as well as for your critical insights on everything from protocols to publications to this dissertation.

To Dr. BJ Philmus: thank you for introducing me to the wonderful world of natural products and for your kind mentorship throughout our collaborations. To Dr. Sergei Filichkin and to the late Dr. Maria Ivanchenko for the invaluable lessons about life in a lab.

To my labmates—Dr. Mitra Ansariola, Russ Gould, Zach Bright, and Olivia Ozguc: thank you for your hard work, perseverance in the face of cranky protocols and frustrating code, and thoughtful discussions about our research. But mostly thanks for your camaraderie.

To all the undergraduate techs we have had in the lab during my time here: thank you for keeping our plants alive and healthy until they were ready to die for science.

To the wonderful people in the Department of Botany and Plant Pathology: thank you for considering me one of your own and picking up where MCB left off.

Finally, I would like to thank my friends and my family. Without your unwavering support and encouragement, the last six years could have looked so very different. I love you all.

CONTRIBUTION OF AUTHORS

Chapter 1: VF wrote the chapter.

Chapter 2: MM, BP, and VF designed the study. VF and BP performed the experiments and generated the data. VF analyzed the data. MM, BP, and VF wrote the manuscript.

Chapter 3: MM designed the study. MI and SF carried out laboratory experiments for TSS-Seq, OC-Seq, and RNA-Seq dataset generation. MA performed algorithm implementation. MA and VF contributed to the design of data analysis experiments and to the evaluation of results. MA and VF worked together to select EMSA assay sites for examination, and VF carried out EMSA assays aided by OO. VF, RG, ZB, and SO contributed portions of the data analysis. OO and ZB contributed to data preparation for public distribution. MM, MA, and VF wrote the manuscript.

Chapter 4: VF wrote the chapter.

TABLE OF CONTENTS

	<u>Page</u>
CHAPTER 1: General Introduction.....	1
Gene expression begins with transcription	2
Understanding tissue specific gene expression through machine learning.....	5
Specialized metabolism and vinca alkaloid production.....	6
Central Questions.....	8
CHAPTER 2: Accurate Transcription Start Sites Enable Mining for the <i>cis</i> - Regulatory Determinants of Tissue Specific Gene Expression	9
ABSTRACT.....	10
INTRODUCTION	11
MATERIALS AND METHODS.....	16
Plant materials and sample preparation	16
Dataset generation.....	17
Sequencing.....	17
Read preprocessing and alignment	17
nanoCAGE-XL TSS-Seq data processing	18
DNase I SIM data processing	18
RNA-Seq data processing and differential expression analysis	19
TSS-Seq data quality analysis.....	19
Sequence depth analysis	19
3PEAT-style models	19

TABLE OF CONTENTS (Continued)

	<u>Page</u>
Model construction	19
Functional binding site selection	20
Electrophoretic mobility shift assays (EMSA)	20
Evaluation of peak classification in same-tissue and other-tissue data-sets.....	21
Construction of tissue of expression prediction (TEP) models	22
Feature generation.....	22
Positional weight matrix set.....	23
Regions of enrichment and TF selection	23
Promoter tiling	24
Feature scaling	24
Model training and testing using nanoCAGE-XL TSSs.....	24
Tissue of expression modeling analyses	25
Model stability assessment	25
Comparison to model using TAIR10 annotated TSSs	26
Hard-coded promoter analyses	26
<i>In silico</i> knockouts	27
RESULTS	28
TSS-Seq, RNA-Seq, and OC-Seq dataset in <i>Arabidopsis</i> roots and shoots captures chromatin state together with promoter utilization in different plant organs.....	28

TABLE OF CONTENTS (Continued)

	<u>Page</u>
Highly expressed TSSs can be accurately modeled in each tissue type using only DNA sequence	31
TSSs enable meaningful TF binding-site-based feature set construction ..	33
TFBS locations and their chromatin state accurately predict tissue of expression for differentially expressed genes	36
Promoter ‘tiling’ model offers complementary view of important feature locations	40
TEP models suggest some promoters may express almost solely based on patterns of functionally bound sites	44
TF site presence and location are predominant explainers of tissue of expression	47
DISCUSSION	49
Model success in predicting tissue of expression fundamentally derives from precise TSS information.....	49
Accurate TSSs implicate proximal cis-regulatory regions as primary determinants of tissue-specific gene expression	52
Implications for synthetic biology: systematic design of tissue-specific promoters	53
DATA AND MODEL AVAILABILITY	54
ACCESSION NUMBERS	55
ACKNOWLEDGEMENTS	55
CHAPTER 3: Metabolomics Analysis Reveals Both Plant Variety and Choice of Hormone Treatment Modulate Vinca Alkaloid Production in <i>Catharanthus roseus</i>	56

TABLE OF CONTENTS (Continued)

	<u>Page</u>
ABSTRACT.....	57
INTRODUCTION	58
MATERIALS AND METHODS.....	63
Plant materials and growth.....	63
Extraction protocol validation.....	64
Phytohormone treatments and sample collection	64
Alkaloid extraction.....	65
LCMS quantitation of <i>C. roseus</i> alkaloids.....	67
RNA extraction and qRT-PCR	68
Data analyses	69
RESULTS	69
Alkaloid levels substantially differ between varieties	70
Induction of alkaloid levels differs markedly based on which phytohormone is used	71
Master regulators are upregulated by hormonal induction	74
Evidence supports transcriptional regulation of key pathway steps	75
DISCUSSION	77
Questions generated by our study	78
Examining varietal differences in the roles of master regulators	79
Linking transcriptional changes to metabolite production.....	80

TABLE OF CONTENTS (Continued)

	<u>Page</u>
Conclusions.....	82
ACKNOWLEDGEMENTS.....	82
CHAPTER 4: Conclusions	83
Future directions for investigating transcriptional regulation of specialized metabolism.....	86
Literature Cited.....	89
Appendix A: Supplementary Materials for Chapter 2.....	102
Appendix B: Supplementary Materials for Chapter 3	144

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1.1 Schematic of the architecture of an example promoter	3
2.1 Conceptual diagram of the datasets generated for the tissue specific gene expression modeling experiment	29
2.2 Summary of data outcomes from differential expression analyses	30
2.3 Selection and testing of putative functional binding sites	35
2.4 TEP model concept and TEP-ROE feature generation schematic and composition.....	39
2.5 TEP-Tiled feature generation schematic and composition	42
2.6 Comparison of top PWMs between TEP models	44
2.7 TEP model variations performance summary.....	48
3.1 Vinblastine biosynthetic pathway diagram from MVA and MEP to TIA	60
3.2 Experimental design of the <i>C. roseus</i> metabolomics study	66
3.3 Comparison of alkaloid concentrations between varieties in control group.....	71
3.4 Comparison of alkaloid concentrations between varieties in treatment groups.....	73
3.5 qRT-PCR analyses of master regulator genes	75
3.6 qRT-PCR analyses of key biosynthetic genes	77

LIST OF APPENDIX FIGURES

<u>Figure</u>	<u>Page</u>
Figure A1 TSS-Seq root and shoot sampled read depth saturation analysis.....	103
Figure A2 Chromatin accessibility surrounding shared TSS mode in root and shoot.....	104
Figure A3 Proportion of top features shared by 3PEAT root and shoot models.....	105
Figure A4 False positive and false negative rates for 3PEAT root and shoot models.....	106
Figure A5 EMSA evaluations of putatively functional binding sites.....	107
Figure A6 TEP-ROE model performance.....	108
Figure A7 TEP-Tiled model performance.....	109
Figure A8 Cross-validation ROC curves for ROE and Tiled models.....	110
Figure A9 TEP model performance comparison.....	111
Figure A10 Feature rank variability TEP-ROE model.....	112
Figure A11 Feature rank variability TEP-Tiled model.....	113
Figure A12 Feature removal performance plot for TEP-ROE model.....	114
Figure A13 Feature removal performance plot for TEP-Tiled model.....	115
Figure A14 Top-weighted feature comparison between TEP-Tiled and enhancer model.....	116
Figure A15 TEP-ROE rank correlation plot.....	117
Figure A16 TEP-Tiled rank correlation plot.....	118
Figure A17 ROE feature products vs openness.....	119
Figure A18 Tiled feature products vs openness.....	120

LIST OF APPENDIX FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
Figure A19 TEP-ROE model construction	121
Figure A20 TEP-Tiled model construction.....	122
Figure A21 Feature product sums for hard-codedness evaluation.....	123
Figure B1 Boxplots comparing ajmalicine/tetrahydroalstonine concentrations between varieties and treatments	145
Figure B2 qRT-PCR analyses for genes upstream of the TIA pathway ..	146
Figure B3 qRT-PCR analyses for addition biosynthetic genes	147
Figure B4 Summary of changes in metabolite levels in response to treatment	148

LIST OF APPENDIX TABLES

<u>Table</u>	<u>Page</u>
Table A1 Number of TSS peaks and their mapped locations.....	124
Table A2 Number of TSS peaks and covered number of transcripts	125
Table A3 Basic statistics on TSS-seq, RNA-seq and OC data	126
Table A4 Basic RNA-seq expression statistics.....	127
Table A5 Top weighted features for 3PEAT root and shoot models.....	129
Table A6 Cross-tissue model performance.....	130
Table A7 Cross-tissue 3PEAT-style GO-analysis for misclassified TSSs	131
Table A8 TEP-Tiled vs TEP-ROE models top feature comparison	132
Table A9 Top 30 features of enhancer-covering Tiled model.....	133
Table A10 Putatively “hardcoded” promoter examples from TEP- ROE.....	134
Table A11 Putatively “hardcoded” promoter examples from TEP- Tiled	136
Table A12 GO enrichment analyses for “hardcoded” promoter examples	140
Table A13 “In-silico knockout” results for TEP-ROE model	141
Table A14 “In-silico knockout” results for TEP-Tiled model.....	142
Table B1 The ’omics data available for selected varieties of <i>C. roseus</i> ..	150
Table B2 Mean alkaloid concentrations	151
Table B3 p-values for absolute alkaloid concentrations from ANOVA analysis.....	152

LIST OF APPENDIX TABLES (Continued)

<u>Table</u>	<u>Page</u>
Table B4 p-values for absolute alkaloid concentrations from pairwise post-hoc analysis.....	153
Table B 5 p-values for peak intensities relative to internal standard from pairwise post-hoc analysis	154
Table B6 p-values for normalized qRT-PCR from ANOVA analysis	155
Table B7 p-values for normalized qRT-PCR in shoots from pairwise post-hoc analysis.....	156
Table B8 p-values for normalized qRT-PCR in roots from pairwise post-hoc analysis.....	157
Table B9 qPCR primers	158

LIST OF APPENDIX DATASETS

<u>Table</u>	<u>Page</u>
Dataset A1 3PEAT logistic regression coefficients for 3PEAT root and shoot TSS location prediction models	143
Dataset A2 TEP logistic regression coefficients for ROE and Tiled Tissue-of-Expression-Prediction models.....	143

DEDICATION

This work is dedicated to my parents, who taught me the value of hard work and perseverance; without them I would never have gotten this far.

Chapter One

General Introduction

Valerie N. Fraser

Gene expression has many steps and, thus, has many points where control can be exerted. Thinking about the central dogma of biology—for protein coding genes, DNA is transcribed to RNA which is translated into proteins that then perform specific functions—many of those regulatory points are quite obvious, occurring at transition points from one macromolecule to the next. Other regulatory mechanisms occur by modification or degradation of these molecules. Therefore, how gene expression is controlled is likely a complex combination of these physical patterns and biochemical mechanisms. For my dissertation work, however, I chose to focus on regulation at that very first transition point: the process of transcribing DNA into RNA.

Gene expression begins with transcription

The first layer of information transmission during gene expression occurs as the cell converts DNA to RNA in a process called transcription. RNA polymerase II is the enzyme responsible for transcribing messenger RNA (mRNA)—single-stranded oligonucleotides that act as the direct templates for protein production. This enzyme is recruited to a particular point in the genome, called the Transcription Start Site (TSS), and proceeds along the genic sequence to synthesize a new mRNA. Immediately upstream of the TSS is a regulatory region spanning between 0.5kb and 10kb in size, depending on species (International Rice Genome Sequencing, 2005; Yamamoto et al., 2011), known as the promoter. This stretch of sequence is comprised of a series of cis-regulatory elements (CREs), short DNA sequences that impact expression of a gene typically located on the same strand (Figure 1.1A). binding sites for proteins called transcription factors (TFs). These TFs interact with their binding sites (TFBSs) and with each other to either positively or negatively affect transcription of the associated gene located downstream of the promoter sequence (i.e., in the 3' direction).

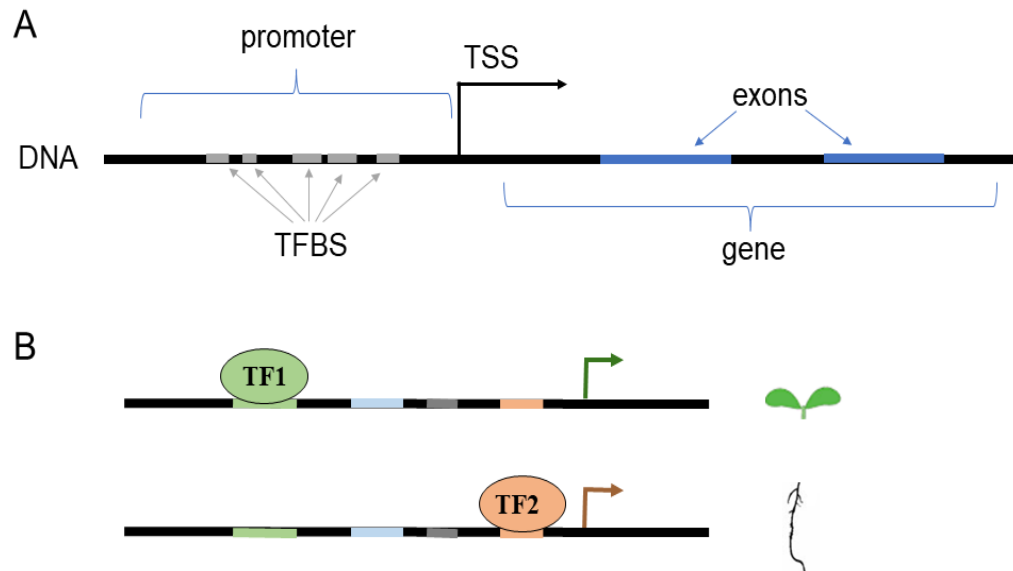


Figure 1.1 A) Schematic highlighting the structure of a gene and its promoter. The Transcription Start Site (TSS) is located upstream of the genic sequence and is the point at which transcription will begin. Upstream of the TSS is the promoter, which contains regulatory sequences that act as binding sites for transcription factor proteins (TFBSs). B) Diagram illustrating one possible way in which genes are differentially transcribed for tissue specific expression. These two example promoters are identical and are upstream of the same gene. The transcription factor bound in each, however, is different. In the top example, TF1 is bound to its binding site in the 5' end of the proximal promoter, causing the gene to express in shoots. In the bottom example, TF2 is bound to *its* binding site—which is closer to the TSS than TF1's—and this causes the gene to express in root.

General transcription factors (GTFs) can work alone or in complexes but either way are essential components of basal gene expression (Orphanides et al., 1996; Holstege et al., 1998). The GTFs bind to CREs and interact with a larger protein complex known as Mediator (Malik and Roeder, 2000; Myers and Kornberg, 2000; Backstrom et al., 2007). Chromatin looping then occurs and, in coordination with DNA-sequence specific TFs bound to their binding sites in the promoter region, Mediator forms the preinitiation complex (PIC) by recruiting RNA Pol II to the TSS (Holstege et al., 1998; Soutourina et al., 2011). Once the PIC is established, transcription can occur.

In addition to the availability of TFs for RNA pol recruitment, chromatin accessibility is another factor that impacts gene expression. Nucleosomes are structures formed when short stretches of DNA wrap around a collection of histone proteins; this is the first level of packaging that allows an entire genome to fit into the nucleus of a cell. Histone proteins have tails that can be decorated with a variety of biochemical modifications at specific amino acid residues that change the conformation of the proteins and affect chromatin accessibility. Dimethylation and trimethylation of histone H3 lysine 9 (H3K9me2 and H3K9me3), for example have long been associated with heterochromatin, which is defined by closely packed nucleosomes that make the associated DNA inaccessible or ‘closed’ (Rea et al., 2000; Nakayama et al., 2001). Interestingly, promoters with H3K27me3 marks can still be bound by general transcription factors and even (Dellino et al., 2004) RNA polymerase (Breiling et al., 2001), though gene expression is still repressed (Hawkins et al., 2010; Margueron and Reinberg, 2011), suggesting that its mechanism of repression may be different than that of H3K9me2/3. Meanwhile, acetylation of H3K27 (H3K27ac) is associated with euchromatin, a looser arrangement of histones that leave DNA accessible or ‘open’, and active transcription (Charron et al., 2009; Creighton et al., 2010). Repressive histone modifications can be added and removed by methyltransferases and demethylases in the nucleus (Rea et al., 2000; Loh et al., 2007), but the scientific community has yet to determine whether the chromatin is made accessible prior to the binding of transcription factors or if the process of transcription initiation triggers the change in chromatin state (Huminiecki and Horbańczuk, 2017). Previous studies have produced conflicting results—some finding that TFs require chromatin to be accessible before they can bind (Robertson et al., 2008; Guertin and Lis, 2010; John et al., 2011), while others demonstrate that binding of certain TFs (often ligand-dependent TFs) is necessary for the formation and maintenance of accessible chromatin (Cirillo et al., 2002; Wang et al., 2007;

Biddie et al., 2011). These contradictory findings have made coming to a consensus on the subject challenging.

Understanding tissue specific gene expression through machine learning

Although all the cells of an organism contain the same genome sequence, not all genes are expressed at the same time and in the same place. Some genes are expressed only at certain times of day or in a specific tissue or cell type.

Photosynthesis-related genes, for example, would not do the plant much good if they are expressed in roots, which are rarely exposed to the light needed for that process to occur; as a result, these genes are leaf-specific (Huang et al., 2018; Hoopes et al., 2019). Unsurprisingly, genes specific to seeds tend to be involved in nutrient storage activity, as they endosperm layer must feed the developing embryo until leaves emerge and photosynthesis can occur (Huang et al., 2018; Hoopes et al., 2019). Accurate spatial and temporal gene expression is crucial for the proper development and physiological functioning of an organism and, as a result, investigating the primary determinants of tissue specific gene expression using machine learning methods has become increasingly common.

Machine learning is an ideal tool for predicting gene expression, as it applies pattern recognition to large datasets. Without computation, identifying which biological information drives expression and determining the importance of each datatype's contribution on a genome-wide scale would be a near impossible task requiring enormous amounts of resources. In the last ten years, genome-scale studies have used a variety of machine learning methods and biochemical markers to examine control of general gene expression (Ong and Corces, 2011; Spitz and Furlong, 2012; Singh et al., 2016; Li et al., 2019). The earliest of these studies utilized linear regression and histone marks to gene expression in supervised learning models known as Support Vector Machines (SVMs) to predict expression (Karlic et al., 2010; Cheng et al., 2011). Subsequent studies used other forms of machine learning models, from

Random Forest Classifier (Dong et al., 2012) to Deep Learning Models (Singh et al., 2016). All of these early studies focus on histone modifications as the main data type. Control of cell type specific and tissue specific gene expression is likely even more complex than control of general expression, and likely depends on many different factors—including biochemical processes such as changes in chromatin state through histone and DNA modification, and availability of TFs in the nucleus (Ernst et al., 2011; Huang et al., 2018)—though machine learning is still an appropriate tool for investigating how this occurs. Currently, the most common biochemical features considered to be determinants of tissue specificity are DNA sequence elements and chromatin state (as assessed through enzyme-based degradation studies), though some studies also include ChIP-seq data for histone tail and DNA modifications (Vera et al., 2014b; Huminiecki and Horbańczuk, 2017; Lee et al., 2019). These studies use models ranging from SVMs to Deep Learning with various degrees of success (Cheng et al., 2011; Ernst et al., 2011; Leung et al., 2014; Sonawane et al., 2017). The presence, order, and occupation of specific regulatory sequences can play a role in determining a gene's tissue of expression (Figure 1.1B); work done on human data, however, has demonstrated that tissue specific gene expression is not necessarily dependent on tissue specific TFs, but rather on unique TF targeting patterns in the promoter regions (Sonawane et al., 2017).

Specialized metabolism and vinca alkaloid production

Secondary or specialized metabolism refers to a collection of pathways and compounds that are not strictly needed for survival, but generally convey a benefit to the organism. Examples of secondary metabolites in plants range from pigments for attracting pollinators to bitter defense molecules that deter herbivory. Of particular interest in the scientific world are plants that produce these small molecules that also have some benefit for humans as medications (Balunas and Kinghorn, 2005).

Catharanthus roseus is one such plant, a perennial shrub native to the island of

Madagascar that produces the monoterpene indole alkaloids vinblastine and vincristine. These compounds are used as chemotherapeutics to fight diseases such as breast cancer and lymphoma (Noble et al., 1958; Johnson et al., 1963).

While natural products scientists aim to increase production of vinblastine and vincristine in heterologous organisms such as yeast or in cell culture, they have run into numerous bottlenecks and roadblocks along the way (Tyler, 1988). *In planta*, secondary metabolite biosynthesis pathways often have one or more steps that occur in a specific tissue or cell type, and the monoterpene indole alkaloid pathway is no different (Nascimento and Fett-Neto, 2010). Translating these pathways to heterologous hosts or single-tissue culture, however, can be notoriously difficult due to these tissue specific steps create bottlenecks and other similar limitations (Nascimento and Fett-Neto, 2010; Isah et al., 2018); if these steps were not so troublesome and rate limiting, these strategies would otherwise be a solution to the low levels of secondary metabolite accumulation *in planta*.

For the monoterpene indole alkaloid pathway in *Catharanthus roseus*, the steps resulting in the formation of vindoline from tabersonine are the basis of the bottleneck. *In planta*, these steps occur in the laticifer cells—specialized, elongated parenchymal cells that are only found in leaf and modified leaf tissues (St-Pierre et al., 1999). Meanwhile, researchers using hairy root culture have discovered that the intermediate alkaloid vindoline cannot be produced at acceptable levels in their system (Van der Heijden et al., 1989; Besseau et al., 2013), as its biosynthetic genes are specifically expressed in the aerial parts of the plant (Murata and De Luca, 2005).

For most medicinally relevant secondary metabolites, control of their biosynthesis is thought to be transcriptionally regulated (Colinas and Goossens, 2018). In *C. roseus*, much of the regulation of the expression of biosynthetic genes has been worked out (Pan et al., 2016; Liu et al., 2019), though there are still some gaps left to be filled. Additionally, to date, there has been little investigation into the tissue-specificity of

the transcription factors associated with expression of biosynthetic genes in the monoterpene indole alkaloid pathway.

Central questions

This body of work revolves around the following three central research questions. First, can we use a machine learning model to understand the primary determinants of the transcriptional control of tissue specific gene expression in general? Second, how does *Catharanthus roseus* regulate the production of its bioactive alkaloids in a tissue specific manner and is there a transcriptional component to the regulatory mechanism? And finally, what directions can we go with the knowledge gained from the answers to the first two research questions?

Chapter Two

**Accurate transcription start sites enable mining for the cis-regulatory
determinants of tissue specific gene expression**

Valerie N. Fraser*, Mitra Ansariola*, Sergei A. Filichkin, Maria G. Ivanchenko,
Zachary A. Bright, Russell A. Gould, Olivia R. Ozguc, Shawn T. O'Neil, & Molly
Megraw

bioRxiv

September 10, 2020

<https://doi.org/10.1101/2020.09.01.278424>

ABSTRACT

Across tissues, gene expression is regulated by a combination of determinants, including the binding of transcription factors (TFs), along with other aspects of cellular state. Recent studies emphasize the importance of both genetic and epigenetic states – TF binding sites and binding site chromatin accessibility have emerged as potentially causal determinants of tissue specificity. To investigate the relative contributions of these determinants, we constructed three genome-scale datasets for both root and shoot tissues of the same *Arabidopsis thaliana* plants: TSS-seq data to identify Transcription Start Sites, OC-seq data to identify regions of Open Chromatin, and RNA-seq data to assess gene expression levels. For genes that are differentially expressed between root and shoot, we constructed a machine learning model predicting tissue of expression from chromatin accessibility and TF binding information upstream of TSS locations. The resulting model was highly accurate (over 90% auROC and auPRC), and our analysis of model contributions (feature weights) strongly suggests that patterns of TF binding sites within ~500 nt TSS-proximal regions are predominant explainers of tissue of expression in most cases. Thus, in plants, cis-regulatory control of tissue-specific gene expression appears to be primarily determined by TSS-proximal sequences, and rarely by distal enhancer-like accessible chromatin regions. This study highlights the exciting future possibility of a native TF site-based design process for the tissue-specific targeting of plant gene promoters.

INTRODUCTION

With the advent of genome-scale technologies, data has become increasingly available to address the intriguing question of when and where a gene will express in multi-cellular organisms. Since the entire genome of DNA sequence is identical in all of an organism's cells, what information does RNA Polymerase II (pol-II) use to drive the transcription of many copies of a coding gene's mRNA in one tissue or cell type, and very few copies in another? We know that "cause" is ultimately connected to a complex series of interrelated events that determine cellular state at the moments leading up to transcription initiation, including concentrations of various transcription factors (TFs) and nucleosomes, DNA methylation states, and histone modification states. Nonetheless, it is possible that DNA sequence alone contains all or most of the information necessary to determine the tissues in which the gene will strongly express.

Previous studies have largely focused on chromatin state and available TF binding sites as candidates for the primary determinants of tissue-specific gene expression, based on present mechanistic understanding of pol-II transcription initiation. The depth of understanding of pol-II promoter structure differs across multi-cellular eukaryotes (Smale and Kadonaga, 2003; Kadonaga, 2004; Thomas and Chiang, 2006; Sandelin et al., 2007; Juven-Gershon and Kadonaga, 2010; Kadonaga, 2012; Kumari and Ware, 2013) due to the timing of extensive genome-scale data availability across species. Foundational studies in *Drosophila* have strongly influenced concepts of 'core' promoter elements that reside at or immediately adjacent to the transcription start site (TSS) and regulate basal transcription, 'proximal' regions that extend beyond the core promoter but are also fundamentally important for transcription, and more distal 'enhancer' regions that are thought to regulate spatial and temporal control of transcription (Kadonaga, 2004; Ong and Corces, 2011; Spitz and Furlong, 2012; Kumari and Ware, 2013). While additional studies have broadened

consideration of this paradigm over time particularly in vertebrates (Andersson, 2015; Feuerborn and Cook, 2015; Kim and Shiekhattar, 2015) and plants (Morton et al., 2014), still relatively little is known in many species about how ‘core’, ‘proximal’, and ‘enhancer’ regions are precisely defined genomically; the literature continues to focus on TF binding sites in more TSS-distal enhancer-like regions as candidate master-regulators of tissue specific gene expression (Ko et al., 2017).

The concept that ‘accessible chromatin regions’ or ‘chromatin footprints’ seem likely to pinpoint to specific regions of functionally bound TF sites, particularly in TSS-distal regions, has been presented since the advent of genome-scale open chromatin studies (Heintzman et al., 2007; Xi et al., 2007); this idea provides an attractive hypothesis that perhaps chromatin differences between tissues or cell types ‘modulate’ the patterns of TF binding sites that are available, thereby explaining gene expression differences between tissues. Many bioinformatic analyses support various forms of correlation between patterns of open chromatin and gene expression in a given tissue (Dong et al., 2012; Sheffield et al., 2013; Vera et al., 2014a; Wilken et al., 2015; Rodgers-Melnick et al., 2016; Snyder et al., 2016). A recent study in plants (Ricci et al., 2019) specifically supports the idea that distal regions of open chromatin are statistically correlated with tissue-specific gene expression, and that some of these regions are likely to be enriched for relevant TF binding sites. Strikingly, however, there has been little evidence across the literature that distal accessible chromatin regions are primary drivers of tissue-specific gene expression in the case of most differentially expressed genes, or even that chromatin accessibility itself is largely determining which TFs are able to bind and functionally interact with pol-II. In fact, a recent study that includes mouse cells concludes in the title that “Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation” (Chereji et al., 2019). In plants, studies comparing open chromatin landscapes across tissues and cell types (Zhang et al., 2012b; Zhang et al., 2012a; Pajoro et al., 2014; Sullivan et al., 2014; Maher et al., 2018; Lu et al., 2019) observe a

surprising degree of qualitative similarity in general, including chromatin patterns surrounding differentially expressed genes, and it is not clear whether more refined, quantitative chromatin state differences may explain transcriptional program differences. The primary determinants of gene expression in a given cell type or tissue thus remain a provocative open question.

Machine learning models can potentially speak to this question by integrating genome-scale datasets of different types— including both DNA sequence and chromatin state information— in order to test whether enough information is present in these data types to predict tissue-of-expression related outcomes. Several studies have specifically contributed to this line of inquiry in the literature. An early study (Vandenbon and Nakai, 2010) used TF binding site information to test whether DNA sequences alone could predict the cell or tissue type in which a gene would be specifically expressed in 26 human and 34 mouse tissue and cell types. The study did this by training a series of models that would effectively predict whether a gene was more likely to tissue-specifically express in one tissue vs another in the same species, with the highest inter-tissue prediction success coming in at 73% auROC for human Kidney vs Fetal Liver (auROC is a performance measure of sensitivity and specificity, with a perfect model having auROC of 100% and a random classifier 50%). This study did use TSS information to define promoter regions, but high-precision genome scale TSS-sequencing data was not widely available at that time. A study from several years later incorporated genome-scale OC data (Natarajan et al., 2012) to examine 19 human cell lines, generating classifiers that predicted whether a gene would be strongly upregulated in different cell types. Median performance was reported for a variety of feature-generation techniques, the most successful technique achieved a median auROC of 73% by incorporating open chromatin information, with several of the top-performing models (out of 19 models) achieving an auROC of nearly 90%. Performance is not directly comparable with the (Vandenbon and Nakai, 2010) study, because the goal was not to predict the tissue of expression in a pair-

wise setting but rather to distinguish genes that express very differently in a certain cell type than in other cell types. Features were interpreted for well-performing models, and several examples of tissue-specificity-associated TF binding locations within open chromatin were observed to be important to the models in these cases. It was clear in this study that use of chromatin information overall boosted performance, but it was very difficult to explain why just a few models performed very well while others did poorly. This study used the single annotated TSS location per gene, likely as genome-scale TSS-sequencing information was not available in all cell types of interest at that time.

Recently, the study (Agarwal and Shendure, 2020) trained a deep-learning model with the benefit of human ENCODE data that includes accurate transcription start site (TSS) locations in each cell type, as well as mRNA stability data. This study focused primarily on modeling transcript expression levels using TSS-sequencing data, but also trained a classifier to examine whether the cell type could be correctly predicted for cell-type-specifically expressing genes in human cell lines GM12878 and K562. The classifier achieved an auROC of 65% in predicting cell type from promoter sequence, but without use of TF binding site profiles of any kind. The promising model performance success of all of these outcomes supports the idea that both TF binding sites and chromatin accessibility carry considerable predictive power in classifying tissue of expression. Performance outcomes in the case of predicting tissue of expression in inter-tissue or inter-cell-line comparisons remain relatively low compared to the high 80%'s auROC that would be desirable for model interpretation. However, all of these past studies were necessarily limited to available data sets that either did not have precise TSS and chromatin information available in these same tissues, or—in cell line studies— cases where the material under examination did not come from the same individuals and did not come from normally functioning tissues. It was therefore not possible to inquire directly into the relative predictive success

contributions of primary sequence information and chromatin state with the benefit of precise TSS locations in tissues or cells from the same healthy individuals.

This limitation may well have hindered predictive success in these past efforts, as precise TSS locations are necessary to correctly define promoter sequences relative to the pol-II binding site. TF binding sites are short (6-12nt) and sometimes degenerate sequences that appear throughout the genome by statistical chance. Therefore, if one does not know the actual location of each gene's highly expressing TSS(s) in a tissue, and instead 'guesstimates' for each gene with a single annotated TSS (which is very likely to differ by at least 30-50 nt (nucleotides) from the actual TSS by as much as 500 nt in *Arabidopsis* (Morton et al., 2014)), then any 'promoter sequence' under consideration may be shifted many binding sites away from the actual TSS-proximal sequence. In this situation, not only is one unable to identify cases in which a different promoter sequence is being used to transcribe a gene in a different tissue, but one is completely unable to take advantage of accurate binding site patterns within each promoter—for example, on a very simple level one cannot even know whether a TATA site seen ~25-35 nt upstream of a TSS is likely to be functional. Without the ability to approximate TSS location within a few nucleotides, it simply isn't possible for a model to take advantage of precise patterns of relationships surrounding these binding sites over thousands of TSSs expressed on the genome in a given tissue sample. It will also be impossible to detect differences in these patterns for TSSs expressed in a different tissue sample, omitting an important source of information, as pol-II can utilize different transcription initiation locations for transcribing the same gene in different organs (Forrest et al., 2014; Mejía-Guerra et al., 2015). In essence, because binding sites are short and seen everywhere, important patterns in their relationships observed over thousands of TSSs (in either the same or different tissues) can simply be 'washed away in the noise' if each of those TSSs is randomly shifted away from its actual location by many binding sites in genomic distance.

In our study, we set out to construct a dataset that could help begin to quantitatively address the informative components in pol-II gene promoters that explain tissue of expression. The large and relatively complex yet well-studied genome of *Arabidopsis*, with many datasets from distinct tissues/organs during plant development, presented an ideal organism in which to undertake this task. We were able to construct a dataset from the same seedlings where each data component—while using relatively new technologies at the time for Transcription Start Site Sequencing (TSS-Seq) and Open Chromatin Sequencing (OC-Seq)—was able to be corroborated with other published datasets derived from similar material, indicating some stability in these data with regard to tissue/organ type, and allaying our concerns that our results might be particular to our sample material or particular to the technology/protocol that we used to produce the different dataset components. Our observations from machine learning model analysis of this dataset challenged our previous assumptions that distal chromatin-accessible TF site locations play a primary role in tissue of expression of most promoters, and suggest a possible paradigm shift in the way we generally assume plant promoters to operate. Specifically, our findings suggest that for the vast majority of differentially expressed genes in developing *Arabidopsis* organs, it is the pattern of cis-regulatory sites in the TSS-proximal DNA of these regions, regardless of chromatin state, that is most explanatory of the tissue of expression.

MATERIALS AND METHODS

Plant materials and sample preparation

Arabidopsis thaliana ecotype Columbia 0 seeds were sterilized in a solution of 50% (v/v) bleach solution with 0.1% Tween 20 for 10 minutes, then rinsed extensively with sterile deionized water. Sterilized 100 micron nylon mesh was placed on top of solidified medium (30 mM sucrose, 4.2 g Murashige and Skoog medium (PhytoTech Labs), and 0.8 % Phytagar, pH adjusted to 5.8 with KOH) in large petri plates (Genesee Scientific). Following a 4 day vernalization period in water, sterilized seeds

were suspended in a 0.75% agar solution and were transferred to each plate in two dense rows (~500 seeds per row) in a laminar flow hood under sterile conditions. Seedlings were grown vertically in a Conviron PGR15 growth chamber at 21°C under a 12:12 hour light:dark cycle (50% humidity, and 250 mol/m²/s light intensity). At seven days, the seedlings were harvested and divided into three batches. For each batch, seedlings were dissected using a surgical blade and the root tissue was separated from the shoot tissue. For our purposes here, shoot tissues include the hypocotyl, cotyledons, and any stems and true leaves that had developed by the time of collection. Each batch of seedlings was handled identically, and harvested tissues were flash-frozen in liquid nitrogen, then stored at -80°C until needed for the following protocols.

Dataset generation

Sequencing

TSS-Seq was performed for both root and shoot samples as described in (Cumbie et al., 2015b) using the nanoCAGE-XL protocol in conjunction with the HiSeq-2000 sequencing platform. For DNase-seq, chromatin from isolated nuclei was digested with DNase I and libraries were prepared for both root and shoot samples according to the DNase-I-SIM protocol (Filichkin and Megraw) as published in (Cumbie et al., 2015a). RNA was isolated from both root and shoot samples for RNA-Seq using the RNeasy kit (Qiagen). Samples were analyzed for quality on the Bioanalyzer 2100 (Agilent) and only RNA with a RIN > 9.0 was used. Single-end libraries were sequenced on the Illumina HiSeq-2000 sequencing platform in triplicate.

Read preprocessing and alignment

CapFilter software (Cumbie et al., 2015b) was used to pre-process all nanoCAGE-XL TSS sequence files prior to alignment, removing library artifacts such as extra

guanines at the beginning of the reads. For all three sequencing experiments, single-end reads were aligned to the TAIR10 reference genome (Lamesch et al., 2012), using Bowtie version 2.0 (Langmead and Salzberg, 2012) with the parameter settings '-v 0 -m 1 -a best strata' (uniquely mapped reads with only one mismatch allowed).

nanoCAGE-XL TSS-Seq data processing

After cap-filtering and aligning the TSS-Seq reads for the root and shoot samples, the JAMM peak finder (Ibrahim et al., 2015) was used to identify TSS read clusters. For TSS datasets in our study, the fragment size and bin size were both set to 10. The output of JAMM is a list of peaks along with their genomic coordinates. The coverage subcommand from the bedtools software suite (Quinlan and Hall, 2010) was used with the parameter settings -s (requiring same-strandedness) and -d (for reporting depth at each position) to retrieve the number of aligned reads in each peak region for peak annotation. An R script was developed to process the aligned reads within peak regions and generate peak information, such as the number of aligned reads in peak, TSS peak mode location, and mode read count. Each TSS peak was assigned to the closest TAIR10 annotated transcript, and peaks which fell within 250 bps upstream of the annotated translation start site and contained more than 50 read counts were selected for use.

DNase I SIM data processing

After alignment to the genome, the F-Seq peak-calling software (Boyle et al., 2008) was used to identify DNase-I hypersensitive sites (DHSs), as in (Cumbie et al., 2015a). We chose this OC-Seq peak caller because it provides compatible output with the original DNASE-I ENCODE data, and is therefore comparable with DNase-I peak usage by the machine learning modeling studies discussed in the Introduction and Discussion sections. Subsequent peak callers have more parameters that can be tuned, but all peak callers are smoothing algorithms that have limitations and tradeoffs in

signal processing parameter selection. F-Seq was run with a specified feature length of 300 and a minimum DHS length of 50 nt.

RNA-Seq data processing and differential expression analysis

Individual transcript abundance was determined using the RSEM software package (Li and Dewey, 2011) for each root and shoot RNA-seq sample. RSEM enables accurate transcript quantification using its built-in bowtie2 alignment by building the reference sequence from a user-provided genome annotation and calculating expression for each isoform. We used *rsem-prepare-reference* and *rsem-calculate-expression* for preparing reference sequence and computing transcript abundances, respectively. We then used EBseq (Leng et al., 2013), which is included in the RSEM package and is robust to outliers, in order to detect differentially expressed transcripts (*rsem-run-ebseq*).

TSS-Seq data quality analysis

Sequence depth analysis

To determine whether our sampling depth was sufficient to accurately represent gene expression in root and shoot samples, we performed saturation analysis on our nanoCAGE-XL data. This analysis was performed as described in (Morton et al., 2014). Supplementary Figure A1 contains the results of this analysis.

3PEAT-style models

Model construction

Using our nanoCAGE-XL peak data, we constructed models to predict whether a given genomic location is a TSS for each tissue type. Starting with annotated peaks, we removed those with less than 100 reads per peak or less than 30 reads at the peakmode and kept only those peaks labeled as being a TSS, within the 5'UTR region, or within 500 nucleotides of the TSS. We then generated the model features

from TAIR10 sequences surrounding these peak regions using the TFBSscanner (Morton and Megraw, 2014). The filtered peaks were then used to train and test 3PEAT models exactly as described in (Morton et al., 2014).

Functional binding site selection

The TSBS sequence features, their model-assigned weights, and the TSS probabilities generated from the root-trained model were then used to construct a table of putative functional binding sites and their associated metrics. This table was then filtered to retain only Narrow Peak promoters. Next, the data was sorted by model output probability, descending total feature score, descending model weight, descending false negative rate, ascending false positive rate, peak count read, and the absolute value of relative location. TFBS sites were selected from the top 500 rows of this sorted table for wet-lab validation of functional binding.

Electrophoretic mobility shift assays (EMSA)

Nuclear extracts were purified from *Arabidopsis thaliana* Columbia 0 roots following a slightly modified protocol from (Staiger et al., 1991). We spun our cellular lysates at 2200 x g for 1 minute at 4°C before passing the supernatant through a series of progressively finer meshes (100 micron, 60 micron, 30 micron). Nuclei were washed and pelleted at 2200 x g for 10 minutes at 4°C. Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific) was used in place of the KCl in Buffer B. After dialysis, samples were prepared with the Qubit Protein Assay Kit (Invitrogen) per manufacturer instructions and total protein concentration was measured on a Qubit Fluorometer (Invitrogen). DNA probes were designed using selected PWMs and the flanking sequences from the associated TSS. Both the template strand and its reverse complement were labeled using the Pierce Biotin 3' End Labeling Kit (Thermo Fisher Scientific) and the complementary oligonucleotides were annealed. EMSAs were performed using the LightShift Chemiluminescent EMSA kit (Thermo Fisher

Scientific). Briefly, 15 μL of nuclear extract (containing 4 μg total protein) was incubated together at room temperature for 30 min with 80fmol of biotin-labeled probes in 25 μL reaction mixtures containing 1X Binding Buffer, 10 mM DTT, 40 ng/ μL poly(dI-dC), and 2% (v/v) glycerol and then separated on 6% native polyacrylamide gels in Tris-borate-EDTA buffer containing 45 mM Tris, 45 mM boric acid, and 1 mM EDTA, pH 8.3. Unlabeled probe was used as cold competitor in 300X excess. Labeled probes were detected using the Pierce Chemiluminescence Detection Kit (Thermo Fisher Scientific) according to manufacturer instructions and visualized on an Azure c600 imager (Azure Biosystems).

Evaluation of peak classification in same-tissue and other-tissue datasets

Two 3PEAT style models were constructed—one trained on 80% of the nanoCAGE-XL peaks from root and the other on 80% of the peaks from shoot. Models were tested on the other 20% of the peaks from both the same tissue and the opposite tissue. Then, we determined which peaks were misclassified by the models. For each test set, we divided the peaks which were classified incorrectly into four groups: “root-misclassified” (misclassified only by the root-trained model), “shoot-misclassified” (misclassified only by the shoot-trained model), “both-misclassified” (misclassified by both models), and “any-misclassified” (includes all peaks misclassified by at least one of the models). For each test-set-model pair, we compiled a list of genes that had at least one TSS peak which was misclassified by only that model. We observed that these lists greatly overlapped with the other list of their respective tissue-type, therefore we chose to exclude genes that were common to both lists. For each list of genes described above, we performed a GO enrichment analysis using GOATOOLS (Klopfenstein et al., 2018), a Python-based automated gene ontology enrichment analyzer. We used the genes represented in the full test set pertaining to the list as a population for comparison, limiting the scope of the analysis to “Biological Process” ontology terms, and limiting results to those with $p < 0.05$.

Additionally, a simple Plant-Ontology enrichment analysis (Jaiswal et al., 2005) was performed by creating a list of terms subordinate to each of the terms of interest, "root system" (PO:0025025), and "shoot system" (PO: 0009006). Genes with peaks misclassified exclusively by one of the models, which were annotated with terms subordinate to one of the terms of interest were considered to be matches to those terms. A Fisher's exact test was performed comparing the proportion of matches in genes misclassified by one model exclusively vs. genes misclassified by either model, with a cutoff of $p < 0.05$ for significant enrichment or depletion.

Construction of tissue of expression prediction (TEP) models

Feature Generation

Each TFBS feature represents an approximation of cumulative binding affinity that a particular TF has for a specific genomic region. Each open chromatin feature represents a percentage of nucleotides within the associated region that are open. In the model versions that used TFBS features, each ROE region as determined in previous section is divided into five overlapping sub-windows and two flanking windows as described in (Morton et al., 2014). The TFBS features are cumulative log-likelihood scores for each ROE sub-window on the same and opposite strands (as the gene in consideration). The chromatin features are computed as the percentage overlap between the open regions and each ROE sub-window. Only log-likelihood scores greater than zero are considered as potential binding sites and contribute to the sum, therefore the minimum value for a TFBS feature is 0 (this is a case where none of the nucleotides in the region represent a potential binding site with a greater-than-zero log-likelihood score). In the Tiled model, the entire region from 1 kb upstream to 500bp downstream of the TSS mode is divided into non-overlapping windows of 100bp in width. The TFBS features are computed as cumulative log-likelihood scores within each tile for both strands. The open chromatin features for Tiled model are computed as a percentage overlap between the open regions and each tile. In addition

to TFBS and chromatin features, we added sequence content features such as GCcontent (CG% within 100 bp upstream of TSS mode), CAcontent (CA% within 100 bp upstream of TSS mode), GAcontent (GA% within 100 bp upstream of TSS mode), ATcontent, and general promoter openness. ATcontent features were computed for each 20bp tiles within -200 to +40 bps from the TSS mode location.

Positional weight matrix set

Position Weight Matrices (PWMs) for TFs in *Arabidopsis thaliana* downloaded from TRANSFAC (Wingender, 2008), JASPAR (Bryne et al., 2008), AGRIS (Davuluri et al., 2003), and CIS-BP (Weirauch et al., 2014) databases. We developed a software program in Python in order to compute the element-wise distance between PWM pairs; pairs with a distance less than or equal to our empirical determined threshold of 0.9 were determined to be redundant.

Regions of enrichment and TF selection

TSS peaks identified as described above in root and shoot were collected (approximately 50,000 peaks) and 6 kb sequences were extracted (TSS - 3 kb, TSS + 3 kb, centered at each TSS mode) from the TAIR10 reference genome. As with the TSS-prediction models described earlier in the methods, the TFBS Scanner suite (Morton and Megraw, 2014; Morton et al., 2014) was used to scan each PWM over the extracted sequences, computing log-likelihood scores over these regions. Regions of enrichment (ROEs) were defined on both the forward and reverse strands by identifying the highest scoring region (the region with the largest sum of positive log-likelihood scores) for each PWM across all promoter examples (Morton et al., 2014). PWMs with cumulative log-likelihood score peaks up to 1 kb from the TSS mode were considered as regions of enrichment for that PWM. The ROEs for each PWM were computed using an updated version of ROEFinder software written in R. Our TEP-ROE model construction process is detailed in Supplementary Figure A19.

Promoter tiling

Using TSS peaks identified as described in previous sections, 6 kb sequences (TSS - 3 kb, TSS + 3 kb, centered at each TSS mode) were extracted from the TAIR10 reference genome. Sequences located from 1000 bp upstream of TSS mode up to 500 bp downstream of the TSS mode were divided into 100-nt-wide, non-overlapping tiles. The PWMs were then scanned over each tile to compute cumulative TFBS log-likelihood scores and percent overlap with open chromatin region in root and shoot tissues. Our TEP-Tiled model construction is detailed in Supplementary Figure A20.

Feature scaling

Open chromatin features, as described in the feature generation section, all share the same 0-1 range and are interpretable as an “openness proportion” without modification. TFBS features, however, are computed as a sum of positive log-likelihood scores over all nucleotides in the region, where each nucleotide is taken as the starting point of a potential TF binding site; the log-likelihood score is computed at this site using (1) the PWM associated with the TFs binding domain, and (2) a local background nucleotide distribution model. The maximum possible value for a TFBS feature is region length multiplied by the maximum possible log-likelihood value PWMscoreMax of any binding site (i.e. the score of the PWMs consensus sequence); this value represents a theoretical ‘maximal’ case in which every nucleotide in a region represents the consensus sequence of the PWM. To normalize TFBS features such that each feature conceptually approximates a proportion of the maximum binding affinity, each TFBS feature is divided by its region length. This puts the mean TFBS feature value on the same order of magnitude as the mean OC feature value and allows for comparison between feature regions of unequal length.

Model training and testing using nanoCAGE-XL TSSs

We constructed two classes of TSSs for use in training and testing of the ROE and Tiled models. The “root” class (class 0) consists of TSSs associated with transcripts that are strongly expressed in roots as compared to shoot, and the “shoot” class (class 1) consists of TSSs associated with transcripts that are strongly expressed in shoots as compared with roots. For our purposes, we defined “strongly expressed” as having an RNA-Seq data \log_2 fold-change value greater than 3. Additionally, TSSs present in both classes were screened to ensure that each TSS peak contained at least 300 reads in the tissue of its class label and each TSS-associated transcript had an RNA-Seq expression value in both roots and shoots of at least 30 TPM. This ensured that both values used for fold-change comparisons were reliably above background noise (Zavolan, 2015). The remaining TSS peaks were then randomly partitioned into 80% training and 20% independent held-out test sets. Each data set contains balanced number of labeled classes. The Python Scikit-learn library (Pedregosa et al., 2011) was used to implement L1-regularized logistic regression. L1-model weights were tuned on the training set with 5-fold cross-validation (Supplementary Figure A8), during which a range of parameter values was examined on the test partition. The average of parameter values resulting in the best performance across the folds was selected for final testing on the independent held-out test set (2.67 for TEP-ROE and 2.29 for TEP-Tiled). All auROC and auPRC values are reported on the independent held-out test set for each model.

Tissue of Expression Modeling Analyses

Model stability assessment

In order to evaluate the stability of our TEP models, we performed two types of assessments. First, we re-ran our TEP models 30 times on training and test sets with randomized 80/20 partitioning. For our second assessment, we removed the top N PWMs from our feature list ($n = 5, 25, 45$, etc). We used the new feature sets to re-

train and test the model. The results from both of these assessments can be found in Supplementary Figures A10, A11, A12 and A13.

Comparison to model using TAIR10 annotated TSSs

To investigate the importance of using precise, experimentally-obtained TSS locations in modeling, we generated a feature set as described above in the Feature Generation section of the Methods using only annotated TAIR10 TSS locations. We then trained the TEP-ROE and TEP-Tiled models on these annotated TSSs on nanoCAGE-XL data, as described in the Model Training and Testing Methods section above.

Hard-coded promoter analyses

For our final TEP models constructed using nanoCAGE-XL data (TEP-ROE and TEP-Tiled, including all feature types), promoter examples which were correctly classified with a high probability (≥ 0.9) were selected for our “hard-coded promoter” analysis. The sets of TFBS and chromatin feature values for each of these promoters were extracted, and for each set the

following formula was applied:

$$TFBS_SOP_{promoter_i} = \sum_{j=1}^{no_of_tfbs_features} TFBS_score_j \cdot W_j$$

$$OC_SOP_{promoter_i} = \sum_{k=1}^{no_of_oc_features} OC_openness_k \cdot W_k$$

W is the model weight vector for each feature after training, j and k are the number of TFBS features and OC features, respectively, and i is the number of promoters. $TFBS_SOP$ and OC_SOP are the sum of products for TFBS features and OC features, respectively. The distribution of the sum of products was computed for TFBS features

and for chromatin features, and the 5% tails of these distributions were considered for detecting putative hard-coded promoters (Supplementary Figure A21). Promoters for which TFBS_SOP fell above the 95th percentile and OC_SOP fell below the 5th percentile were labeled as putatively hard-coded. Additionally, we extracted the features with the largest products for each of the TSSs for further investigation. GO-enrichment analysis compared to the entire genome was performed for the genes with putatively hard-coded promoters using GOATOOLS (Klopfenstein et al., 2018).

In silico knockouts

The output of the classifier function for our trained L1-regularized logistic regression models is a probability between zero and one, which represents the predicted likelihood of differential expression in the two tissue types. Probabilities greater than 0.5 are labeled as “class 1” or “shoot”, and probabilities less than 0.5 are labeled as “class 0” or “root”. Logistic regression is a Generalized Linear Model, where the probability of belonging to class 1 is a function of sum of products of feature weights by feature values as follows:

$$P_i = f\left(\sum_{j=1}^{\# \text{ features}} W_j F_{ij}\right) = \frac{e^{\sum_{j=1}^{\# \text{ features}} W_j F_{ij}}}{1 + e^{\sum_{j=1}^{\# \text{ features}} W_j F_{ij}}}$$

where P_i is the probability that Promoter i belongs to class 1 (shoot), W_j is model weight, and F_{ij} is the j th feature value for promoter i . Negative values of the feature product sum yield $P_i(\text{shoot}) < 0.5$ (a root classification), and positive values of this feature product sum yield $P_i(\text{shoot}) > 0.5$ (a shoot classification). Our *in silico* knockout process “zeroes out” selected feature values and then computes the new model-predicted probability for a promoter. In the “single-knockout” experiment, only one feature was removed from the equation at a time, in order to determine effect on probability outcome for every promoter. Supplementary Tables A13 and A14 report cases with the largest probability ‘shifts’ across the 0.5 decision boundary,

indicating predicted high-probability ‘flips’ in tissue of greater expression upon *in silico* knockout of a single TF-feature region.

RESULTS

TSS-Seq, RNA-Seq, and OC-Seq dataset in *Arabidopsis* roots and shoots captures chromatin state together with promoter utilization in different plant organs

Our study uses “root” and “shoot” tissues harvested from 7-day old *Arabidopsis thaliana* seedlings, dissected immediately below the hypocotyl (Figure 2.1). Each tissue batch was separated into three portions, to which we applied Transcription Start Site Sequencing (TSS-Seq), Open Chromatin Sequencing (OC-Seq), and RNA-Seq expression profiling protocols. We used the nanoCAGE-XL protocol (Cumbie et al., 2015b) for performing TSS-Seq, the DNase-I-SIM protocol (Filichkin and Megraw) for performing OC-Seq (Cumbie et al., 2015a), and applied a standard RNA-Seq protocol (see “Dataset Generation” section in Materials and Methods). The nanoCAGE-XL and DNase-I-SIM protocols were developed by the lab to work efficiently with relatively low-volume plant tissues such as *Arabidopsis* seedling roots and shoots; these were vetted in publication using datasets that were generated by applying other current protocols to comparable tissue samples sequenced on the Hi-Seq 2000, which sequenced to sufficient depth and coverage to support our study (see Supplementary Figure A1). We also noted that outcomes for both nanoCAGE-XL and DNase-I-SIM were remarkably consistent with other TSS-Seq datasets (Morton et al., 2014) and DNase-I-Seq datasets (Zhang et al., 2012a) generated in similarly prepared *Arabidopsis* seedling tissue samples.

Figure 2.1 illustrates the data collection goal of our study. As described in the Introduction, precise TSSs are critical to any predictive modeling effort that seeks to relate TF binding sites to gene expression outcomes, particularly those in different tissues, organs, or cell types. Clearly a reasonable estimate of highly accessible or

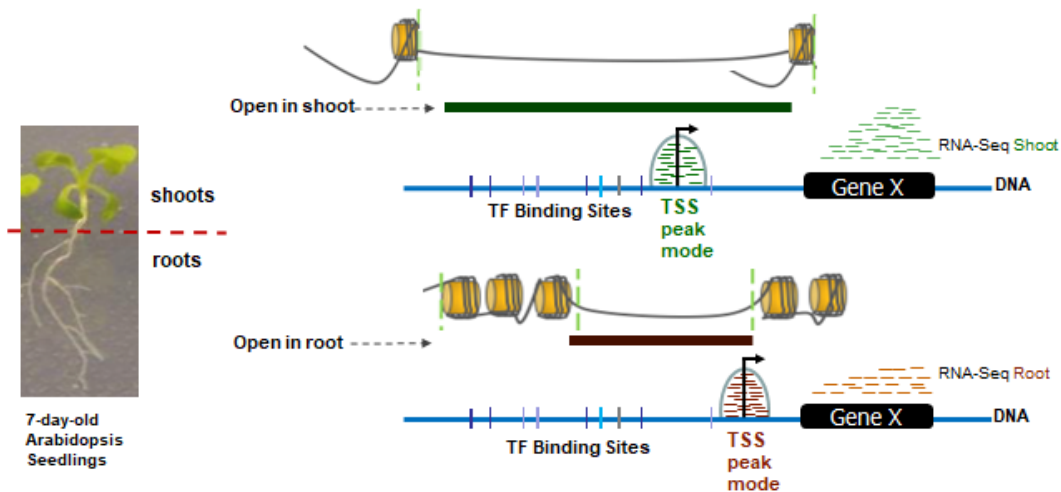


Figure 2.1 Datasets generated from 7-day-old wildtype *Arabidopsis thaliana* Columbia 0 roots and shoots. RNA-Seq reads align to annotated gene bodies to demonstrate gene expression, while OC-Seq reads highlight DNase-I Hypersensitive Sites. TSS-Seq reads align in peaks around the TSS; the mode of the peak is designated the location of the TSS.

“open” chromatin locations is also critical to our query, as it is plausible that TF binding sites within open chromatin regions are playing a large role in tissue-specific gene expression. Finally, we gathered RNA-Seq data for our samples as this form of expression profiling provides the most well-studied statistically robust estimate for gene expression levels, despite its implicit 3’ locational bias (Ross et al., 2013). We used these three data types to take an initial survey of apparent differences in gene expression program in our root and shoot samples (Supplementary Tables A1 – A24). Although we observed that TSS location and promoter accessibility are quite similar across tissues for many genes (Figure 2.2A), 2663 transcripts show strong expression in only one of the two tissue types (Figure 2.2C). Additionally, 525 differentially expressed transcripts are associated with a TSS peak in shoot only, while 707 transcripts have a TSS peak in root only. The 1431 differentially expressed transcripts with TSS peaks in both tissues results in 1632 peak pairs. Of these pairs, 222 (~14%) have very different TSS mode locations (TSS-mode-distance > 100 bp), and 471 cases have mode locations that differ by 10 to 100 nt. Out of the 939 cases which have a very similar TSS mode location (TSS-mode-distance < 10), we looked for

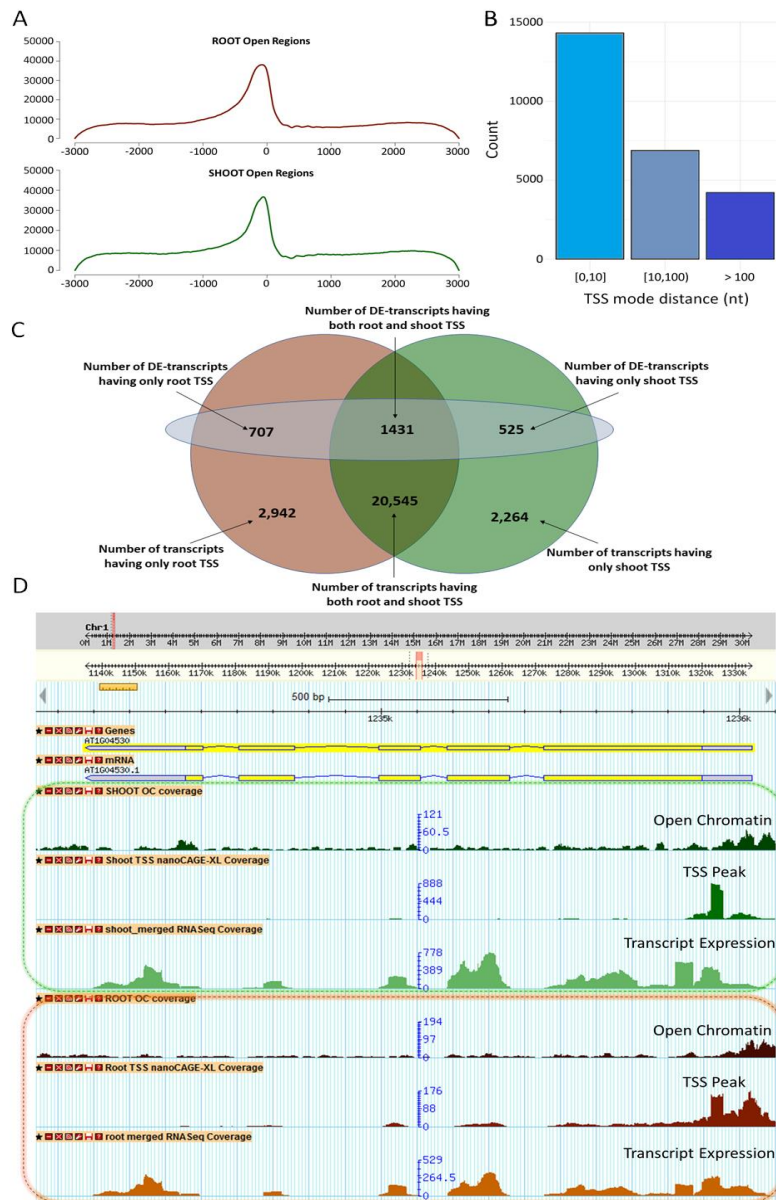


Figure 2.2 Summary of data outcomes. A) Charts comparing the general accessibility in roots (top) and shoots (bottom). B) C) Bar chart showing the difference in location between transcript-associated TSS modes in root and shoot. The majority of TSSs have similar locations in the two tissues, but ~16% have very different locations (>100 nt). C) Numbers in the top row of the Venn diagram represent differentially expressed (DE) transcripts, separated by tissue in which the transcripts have associated TSS peaks. The numbers at the bottom represent all transcripts associated with TSS peaks, separated by tissue in which they have TSS peaks. Total number of DE transcripts = 2,663; total number of transcripts = 24,928. D) A sample gene displayed in GBrowse, showing mapped read accumulations for each data type.

differences in patterns of open chromatin (% nucleotides disagreeing in chromatin state, Supplementary Figure A2). In the vast majority of cases, we were intrigued to see very little obvious difference in chromatin state—that is, we saw a large overlap in the percent of nucleotides agreeing in accessibility state. We observed that while major differences in TSS location or chromatin state might simply explain differential expression between the two tissues in perhaps 20% of the cases, in most cases the reason for differential expression could not be attributed to any obvious difference either in TF binding site usage or chromatin accessibility. We concluded that a quantitative modeling effort was necessary for further investigation.

Highly expressed TSSs can be accurately modeled in each tissue type using only DNA sequence

Past machine learning studies have shown that strongly expressing TSS locations in a tissue sample can be precisely predicted from DNA sequence alone, using surrounding TF binding sites as ‘features’—that is, numerical descriptors of the genomic location that one is inquiring about with the question “Is this location a highly expressing TSS or not” (Megraw et al., 2009; Morton et al., 2014). We hypothesized that if one could accurately model highly expressing TSS locations in the root sample and in the shoot sample individually, using TF binding site (TFBS) information as features, one could then inquire into model differences in binding site patterns that are “important to root expression” vs those that are “important to shoot expression”. For this task, we selected the 3PEAT model (Morton et al., 2014), as it remains the only high-performance plant TSS peak finder to date with features that can explicitly be interpreted as representing TF:promoter binding site interactions. Additionally, the 3PEAT model had previously been applied to an *Arabidopsis* root sample grown under nearly identical conditions to those in our current study, but where the sample was generated using a different TSS-Seq protocol known as “Paired-End Analysis of Transcription start sites” or “PEAT” (Ni et al., 2010); this enabled us to understand whether our nanoCAGE-XL TSS-Seq datasets would

support a similarly successful model to the PEAT root sample, which achieved an auROC in the high 90%'s. We applied the 3PEAT model to both root and shoot nanoCAGE-XL TSS-Seq samples from our study and found that we could predict strongly expressing root and shoot TSS locations in independent single-tissue-type models each with an auROC of 98%. We then examined the TFs associated with the top-weighted features (TF binding locations most important to model success) of the trained root and shoot 3PEAT models (Supplementary Table A5, Supplemental Data Set A1), to determine whether there were any obvious root-specific or shoot-specific differences. We observed that the two models shared the majority of their top-50 most important TFs, though a few differences in the top-10 indicated the possibility of a more important role for root-development-related TFs in the root model and shoot-development-related TFs in the shoot model (Supplementary Figure A3). We then looked for quantitative evidence of root-specific vs shoot-specific TF binding site pattern usage in the two models by applying the root-trained model to the shoot model's test set (i.e. locations that are either highly expressed TSSs or not highly expressed in shoot), and the shoot-trained model to the root model's test set (i.e. locations that are either highly expressed TSSs or not highly expressed in root) (see Materials and Methods for details). Surprisingly, both models performed essentially identically on test sets of TSSs in the 'other' tissue as on test sets of TSSs from the tissue in which the model was trained (Supplementary Table A6) – with the same 98% auROC and only a negligible drop below 80% in auPRC (area under the Precision Recall Curve, a complementary performance measure). However, we found a tendency for the models to produce approximately 20% more false positives (classifying non-TSS sites as TSSs) when applied to the test set derived from the 'other' tissue compared to the model that was trained on that tissue. This was compensated by an 8% decrease in false negatives by the shoot-trained model on the root test set compared to the root-trained model and a 15% decrease in false negatives by the root-trained model on the shoot test set compared to the shoot-trained model

(Supplementary Figure A4). Additionally, genes with TSSs that were misclassified by the shoot-trained model were statistically enriched for several root development GO-terms as compared to the full set of peaks that were tested (see Materials and Methods, Supplementary Table A7). Several additional GO-term enrichment and depletion observations, taken together with model feature-weight observations, strongly supported the idea that patterns of TF binding sites were likely to be well-explaining TSS expression in both tissues; yet the core of both models almost certainly described sequence information indicative of general transcription, as opposed to tissue-specific expression. We concluded that the 3PEAT TSS prediction concept provided an appropriate feature set that would potentially allow us to model the differences in tissue of expression based on TF binding site information but would need to be incorporated into a model that focused on differentially expressing genes.

TSSs enable meaningful TF binding-site-based feature set construction

The original 3PEAT model investigation (Morton et al., 2014) demonstrated that precise TSS locations were key to training a highly accurate TSS prediction model, with a substantial ~10% auROC performance drop if only annotated start sites were used. We wanted to investigate whether the 3PEAT model's TF binding site-based feature set construction was not only the key to predictive success in explaining strong TSS expression, but also carried plausible support for explaining pol-II transcription in reality. Specifically, we wanted to test whether the putative binding sites modeled as important to a gene's correct TSS location prediction were also likely to be TF-bound in the sample, therefore potentially functional. We also wanted to gain a basic indication of whether 'predictive sites' extracted from model features—that is, important TF binding site-enriched regions known in the PEAT model as “Regions of Enrichment” or ROEs (Figure 2.4B)—were likely to be highly sensitive to the specific dataset in terms of sample collection, TSS-Seq protocol, or informatic processing details such as selection of peak caller parameters. We selected

the root sample for testing, as it was then possible to compare 3PEAT TSS models built from two different datasets using very similar tissue samples, (i) the PEAT dataset (Morton et al., 2014) and (ii) the nanoCAGE-XL dataset generated for the present study.

We began by selecting putative cis-regulatory sites from the original 3PEAT model application to the PEAT dataset for *in vitro* TF protein:DNA binding interaction testing using the following procedure (see “Functional Binding Site Selection” in Materials and Methods for details). Cis-regulatory elements considered by the model included TF binding sites as well as core promoter elements such as TATA-box which facilitate direct interactions with the pol-II complex. First, likelihood scores for individual putative binding sites that contributed to each TSS prediction (i.e. each transcript detected in the *Arabidopsis* root) were calculated using their corresponding Positional Weight Matrix (PWM) binding domain representation and their position relative to the ROE for each element. The output of this pipeline was a genome-wide “master-list” of potentially functional cis-regulatory sites. These candidates were filtered by considering only strongly expressing “narrow-peak” TSSs, which have an enriched association with developmental genes responsible for tissue-specific patterning (Morton et al., 2014), and other restrictions to generate a stringent short-list of 500 sites (i.e. sites associated with the top 20% by importance-rank according to their 3PEAT model weight, then sites with highest likelihood scores located near the center of their ROE and within 120bp of their corresponding TSS). Finally, we selected five “high-scoring” candidate sites (INI-B, TATA box, Y-Patch, PIF3-binding element, and SQUAMOSA Promoter Binding element SQUA1) for evaluation using the Electrophoretic Mobility Shift Assay (EMSA or “gel shift” assay) and nuclear extracts prepared from *Arabidopsis* roots. We included several sites from the HSP90.2 promoter, one of the few genes with top-ranked sites in our list that had a known function, as well as one site each from the promoters of

ornithine carbamoyltransferase (OTC) and diacylglycerol kinase 2 (DGK2); in this first selection, we focused on sites that were located in regions of open chromatin in root, but did not account for expression level of the TF(s) corresponding to the PWM binding domain profile predicted to target the candidate sites. We observed gel-shifts for four out of the five candidate sites.

In troubleshooting the case that did not shift, we observed that PIF3, a circadian-controlled TF, had a very low level of expression as measured by the RNA-Seq outcome in the root sample of our current study. We also observed a general qualitative correlation between the intensity of the

shifted band and RNA-Seq expression level with the other candidates, suggesting that TF binding would be undetectable below a certain level in nuclear extracts. We then selected 6 additional top-scoring sites in the HSP90.2 promoter (Figure 2.3) but filtered out any sites associated with very lowly expressed TFs as measured by our

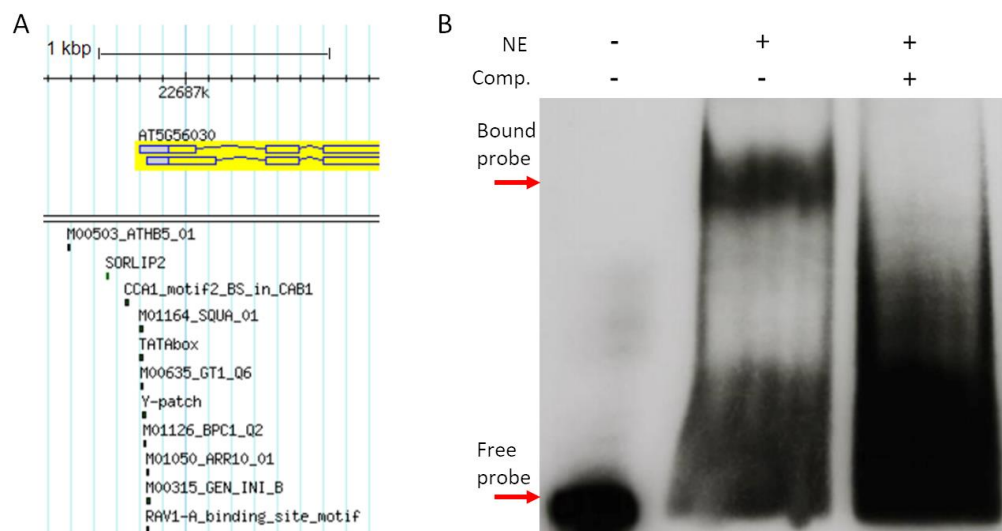


Figure 2.3 Selection and testing of putative functional binding sites. A) The HSP90.2 (AT5G56030) promoter region in GBrowse. Locations of the tested transcription factor binding sites in this promoter are displayed below the gene model. B) Y-patch electrophoretic mobility shift assay. The presence of a shifted band of probe indicates higher molecular weight than the free probe due to TF binding. The right-most lane contains >200X cold competitor to show that the shift is not an artifact.

RNA-Seq sample. Of these sites, all 6 resulted in a shifted band, indicating binding. We then considered whether applying the same process to the 3PEAT model, when trained using more recent peak-calling on the nanoCAGE-XL root sample, would include these same sites as relatively important to the expression of the target gene. We repeated the process (“3PEAT-style Model Construction” in Methods) using the 3PEAT model re-trained on the nanoCAGE-XL root sample that achieved an auROC of 98% described in the section above. We observed that all of our selected sites in the HSP90.2 promoter were included in the new top sites list, although some had a lower score; the tested site in the DGK2 promoter does not appear on our new list, though this gene had a very weak TSS peak in the sample, consistent with involvement in circadian function including lack of upregulation by PIF3. OTC had a moderate TSS peak in the sample, and its site was included (Supplementary Figure A5). In considering whether additional information could be obtained by performing traditional gel-shifts for these sites using purified TF protein, we concluded that if successful this would only demonstrate the ability to bind an oligo at unrealistically high concentrations of each TF, essentially confirming the TF’s PWM binding domain description as provided by a database. In total, the gel-shift of 10/11 predicted binding sites using nuclear extracts, taken together with qualitative correspondence between RNA-Seq level of the candidate TF and darkness of the shifted band, as well as the relatively stable predictive importance of these sites across TSS-Seq datasets, provided plausible support for binding of these sites at *in vivo* TF concentrations. We concluded that 3PEAT model’s ROE-based TF binding site features represent sites that are at least potentially functionally bound in a way that promotes their target gene’s transcription by pol-II.

TFBS locations and their chromatin state accurately predict tissue of expression for differentially expressed genes

Building on the successful TF Regions of Enrichment feature concept of the 3PEAT model for predicting TSS location, we constructed an analogous model that we called

the Tissue of Expression Prediction ROE model or TEP-ROE model. We reasoned that if patterns of TF binding site enrichments can predict the locations of strongly expressing TSSs on the genome, and high-affinity binding sites within important enrichment regions are plausibly functionally contributing to pol-II's frequent transcription initiation at these locations, then perhaps it is patterns of TF binding sites within these regions that can help well-distinguish a tissue in which a gene will express strongly from a tissue in which it will express to a much lesser extent. But it also seemed that the general accessibility of sites in these regions could prove important, as could general sequence enrichments such as AT-content and overall degree of openness in the vicinity of the TSS. Figure 2.4A shows the concept of the TEP-ROE model, with details provided in Methods. Like 3PEAT, TEP-ROE is an L1-regularized logistic regression classifier that takes as input (i) the DNA sequence surrounding a TSS (TSS - 1 kb, TSS + 500 nt) and (ii) chromatin accessibility state for both tissues in this region around the TSS, and returns the predicted tissue (root or shoot) in which that TSS will express most strongly. The most important concepts for understanding and interpreting the model (Figure 2.4A) are that (1) each TFBS feature represents a specific genomic region in relationship to a TSS where a particular TF binding domain has a high density of high affinity binding sites, (2) each OC feature represents the “openness” (degree of accessibility) of a corresponding TFBS feature region, as a percent of accessible nucleotides in this region, (3) the two OC_overall features OC_overall_root and OC_overall_shoot represent the percent openness of the ‘proximal’ region [TSS - 500 nt, TSS + 100 nt] around a TSS in root and in shoot, (4) sequence enrichment features (e.g. GC Content) represent the percent of certain nucleotides (e.g. G and C) in a 100 nt window around the TSS, and (5) the weight that a successfully trained model gives to each of these features represents an ‘importance value’—a large weight magnitude or “top-ranked feature” indicates a feature whose value contributes heavily toward the

decision about whether a TSS is predicted to express strongly in root or strongly in shoot.

We trained the TEP-ROE model on TSS locations associated with differentially expressed genes, using cross-validation for parameter selection and an independent held-out test set for reporting test performance (see “Model Training and Testing on nanoCAGE-XL TSSs” section in Methods). The model achieved an auROC of 92% and an auPRC (area under the Precision Recall Curve, an important co-indicator of performance) of 94% (Supplementary Figure A6).

In order to evaluate performance stability over a wide variety of dataset divisions (training vs testing) and algorithm seedings (different initial value settings of the optimization algorithm), we re-trained the model 30 times with different ‘seeds’ (Supplementary Figure A9). We observed that auROC and auPRC model performance outcomes were tightly distributed around means which were close to our TEP-ROE model’s performance values and concluded that our TEP-ROE model’s strong performance was representative. We then examined feature stability by looking at the feature weight ranking distribution for each of the 50 top-ranked features in the model, over the 30 models used in performance stability testing (Supplementary Figure A10); we found feature rankings to be acceptably stable in the sense that each of the 50 most important features stayed within the top-ranked 50 for all other test models, and most features’ rank remained within 5-10 ranking slots of

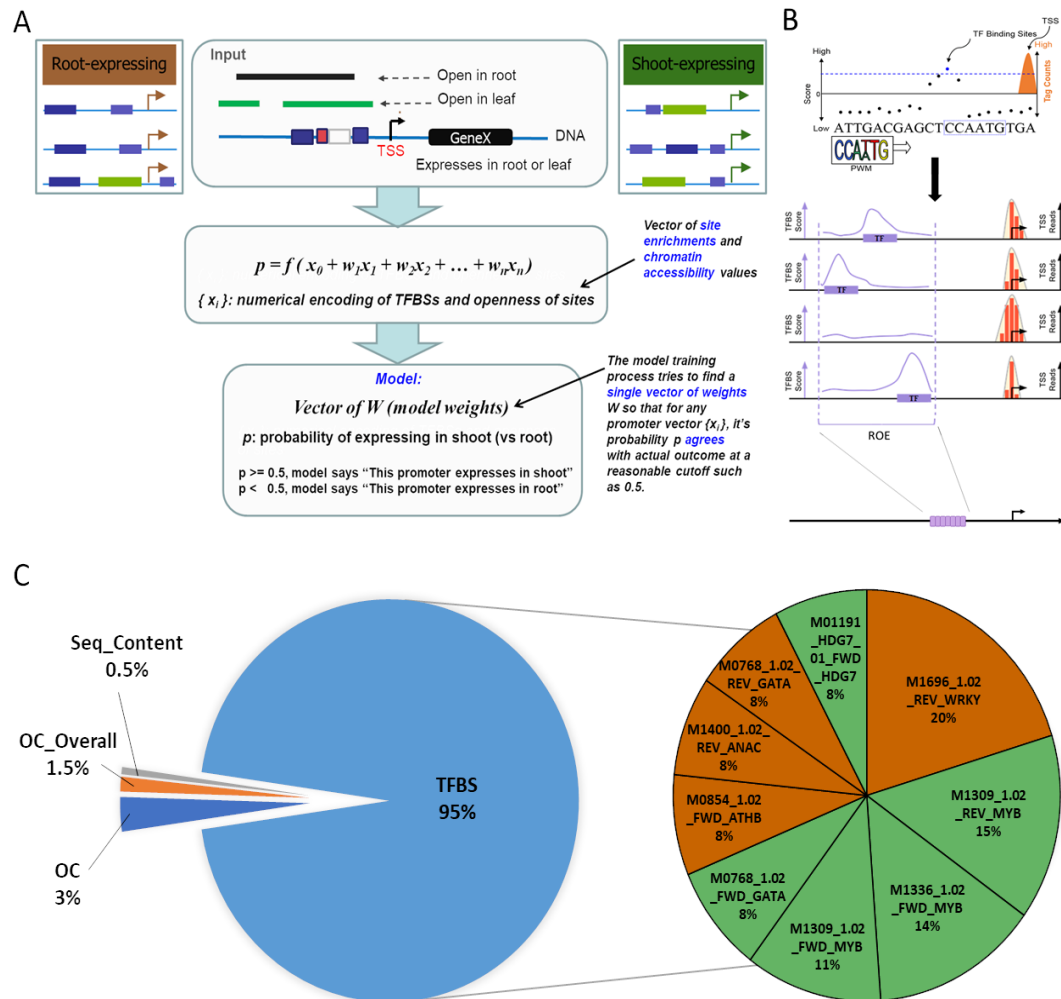


Figure 2.4. A) TEP model concept: Three data types (TSS-Seq, OC-Seq, RNA-Seq) are generated from roots and shoots; this allows us to numerically encode TFBS presence and chromatin accessibility. The numerical encoding is then used to train and test a machine learning model that outputs the probability of a transcript's expression in root or in shoot. B) TEP-ROE feature generation: Regions of enrichment (ROEs) are detected by scanning a PWM (TF binding profile) over each promoter region and calculating where the TFBS loglikelihood scores are significantly higher than background levels. These regions are then further divided into windows for feature scoring. C) In the chart to the left, TFBS features make up 95% of the total generated. The chart on the right highlights the features that the TEP-ROE model weighted most heavily. The names of these features contain quite a bit of relevant information (e.g. M1969_1.02_REV_WRKY: M1969_1.02 is the PWM designation from the database; REV indicates that the feature is located on the opposite strand from the gene; and WRKY is the associated TF family). Green pie wedges indicate that the model deemed this feature important for expression in shoots, while orange wedges indicate importance for expression in roots.

the mean in the vast majority of the test models.

Performance and stability indicated that it was meaningful to interpret the TEP-ROE model, as top-weighted features were likely to be important contributors to successful tissue prediction. The two OC_overall features were high-ranked contributors (Figure 2.4C, Supplemental Data Set A3.2), with a few general sequence content features falling into the top 300. The most striking aspect of the model outcome is that aside from the OC_overall features for root and shoot, and a small number of sequence content features, TFBS features comprised all of the top ~350 features, with the first OC feature appearing at rank 352. We performed a literature search on the top 100 TFBS feature binding domains and found that of the 20 which had functional annotation, four had literature support for activity in the same tissue whose weight sign (positive or negative) indicated that presence of this TFBS site density made expression in this tissue more likely. The locations of these important regions fell within 500 nt of the TSS, indicating the strong predictive role of TF site densities in this proximal region. Finally, when we re-trained a version of the TEP-ROE model using only the TAIR10 annotated start site for each differentially expressing gene rather than TSS-Seq peak locations, auROC dropped to 76%. This substantial ~15% auROC performance drop supports an important role for precise TSS locations in successful tissue of expression prediction model training.

Promoter ‘tiling’ model offers complementary view of important feature locations

The TEP-ROE model was constructed around the concept of “Regions of Enrichment”, which are special regions that one can think of as containing “high TF binding site densities” for a particular TF with respect to all TSSs in a sample type. Since this model style focuses only on a single Region of Enrichment for each TF, and not all TFs have these high binding site densities in our root and shoot samples, some TFs and their binding sites are omitted from consideration in the TEP-ROE model. We also wondered if the TEP-ROE concept was unnecessarily “confining”

important TF binding patterns that are considered by this model to locations very near to the TSS, just because this is where the highest binding site densities occur for most TFs. This led us to ask whether a model that simply “tiled” the same region surrounding the TSS with ‘tile regions’ (Figure 2.5A) would achieve similar or even greater performance—and if it did, would such a model select similar TF binding site density regions as important features. We constructed the TEP-Tiled model by following an identical procedure to the TEP-ROE model, except that ROEs were replaced by a series of non-overlapping 100 nt windows tiling the entire [TSS - 1 kb, TSS + 500 nt] region under consideration (see “Promoter Tiling” section in Methods). The TEP-Tiled model achieved nearly identical performance results (Supplementary Figure A7) to the TEP-ROE model, and its performance over a large number of seeded trials was similarly stable (Supplementary Figure A9). Building the model using only annotated TSSs from TAIR10 caused a similar ~15% auROC performance drop to that of the ROE model. The stability of top-weighted features decreased as compared to the TEP-ROE model (Supplementary Figure A11), but this is largely to be expected because the TEP-Tiled model has many thousands of additional features (many more tiles than ROE regions) and is therefore a very highly under-constrained model; that is, there are so many more features whose importance the model must consider than there are TSSs in the root and shoot classes that (1) there are many feature combinations that can potentially help the model to perform well, and (2) the model operates at the limit of the regularization process’ ability to identify meaningful feature combinations. It is this second issue that lead to declining performance when we examined models using tiles smaller than 100 nt wide. Nonetheless, the TEP-model’s strong and stable performance provided the ability to meaningfully examine the type and location of its most important features (Figure 2.5B). The OC_overall features play a similarly important role, and sequence content features have similar rankings in general. The first TFBS-associated OC region

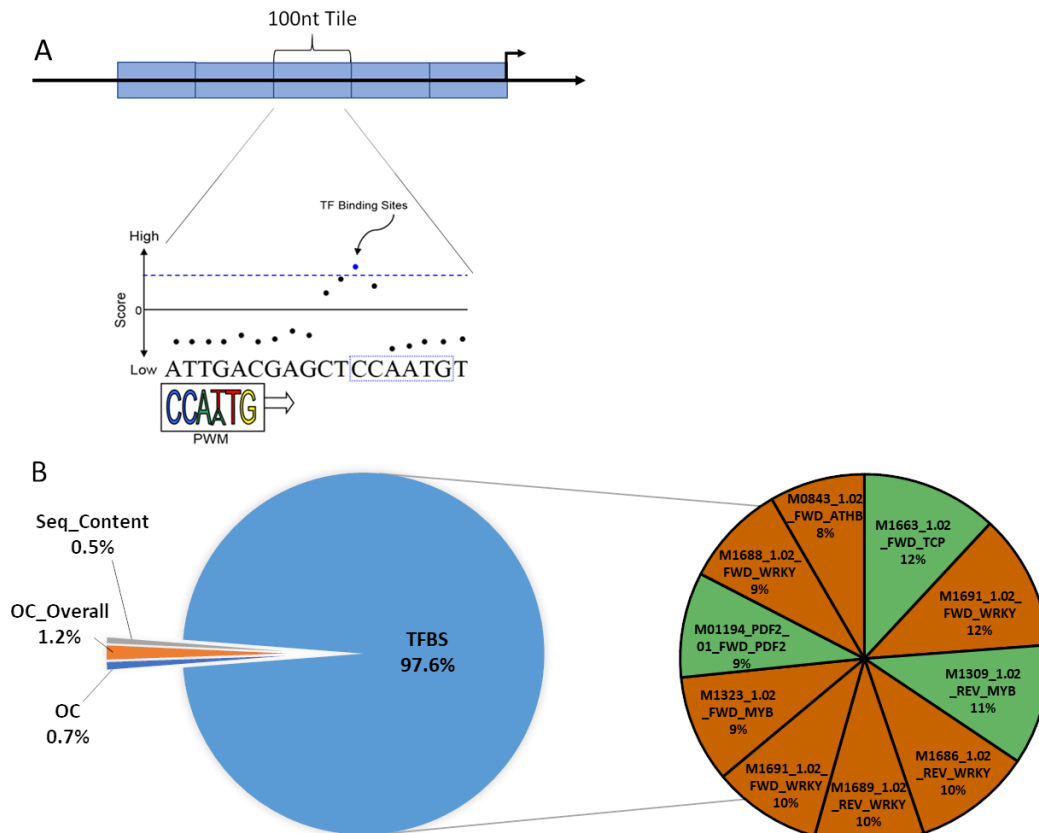


Figure 2.5. TEP-Tiled model. A) Cartoon schematic of TEP-Tiled model's feature generation. Instead of identifying ROEs and creating features within these smaller regions, this model generates 100 nt wide tiles over the entire promoter. B) As with the TEP-ROE model, TFBS features comprise the majority of the features generated by the TEP-Tiled model (97.5%). The pie chart to the right contains the top 10 most heavily weighted features. As with the TEP-ROE model, feature names have three parts (e.g. M1691_1.02_FWD_WRKY: M1961_1.02 is the database's identifier for the PWM; FWD means the feature is on the same strand as the gene; and WRKY is the associated TF family). Green pie wedges are features that the model deemed important for expression in shoots and the orange pie wedges are features the model deemed important for expression in roots.

appears at a lower importance rank (~740) than in the TEP-ROE model (~350), as TFBS features even more heavily dominated the top importance weight rankings.

To visually examine the relationships between top-ranked TFBS feature locations in the TEP-Tiled and TEP-ROE models, Figure 2.6 shows a heatmap overlay of the 100 top-weighted features in each model, displayed according to location with respect to the TSS. In general, for most of the TFBS features that the two models agree are in

the top 100 (those rows that contain both red and blue hues), there is some form of ‘telescoping effect’ or overlap in the regions that both models consider highly important. In these cases, typically the Tiled model agrees with at least one of the locations that the ROE model considers most important for a TF binding domain type, but also gives some lesser weight to at least one additional location. This seems to suggest that much of the time, when the models agree on an important TF binding domain, there is a tendency to agree on its most important location. But, clearly, there are many ‘only red’ or ‘only blue’ rows indicating that there is agreement on inclusion of a TF binding domain feature only about one third of the time. Supplementary Table A8 provides a quantitative look at whether the two models agree on what the most important TFBS or OC features are, location aside. In considering the top 10-weighted features in each model, about 30% are shared. However, all disagreements in this case appear to result from selection of different members of the same TF family by each model (among TF-associated features), as M1691_1.02_TFBS, M1686_1.02_TFBS, M1696_1.02_TFBS are all WRKY family transcription factor binding domains. In general, 20-30% features are identical between the two models in considering up to 200 top-weighted features, and dissimilarities appear to be due at least in part to the different models' inclusion of non-identical but relatively numerically similar binding domain profiles among TF families.

Finally, while Figure 2.6 shows that the most important locations in both models tend to fall within 500 nt of the TSS, the TEP-Tiled model indicates that occasionally an important TF binding domain location could be located nearly 1 kb upstream of the TSS. The most important ROEs from the TEP-ROE model all lie within 500 nt of the TSS, but we wondered if the TEP-Tiled model would select important tiles more than 1 kb upstream if given the opportunity. We re-trained the TEP-tiled model using [TSS - 2 kb, TSS + 500 nt] and observed a slight performance drop, with no top-10-ranked tiles in importance falling upstream of 1 kb (Supplementary Figure A14,

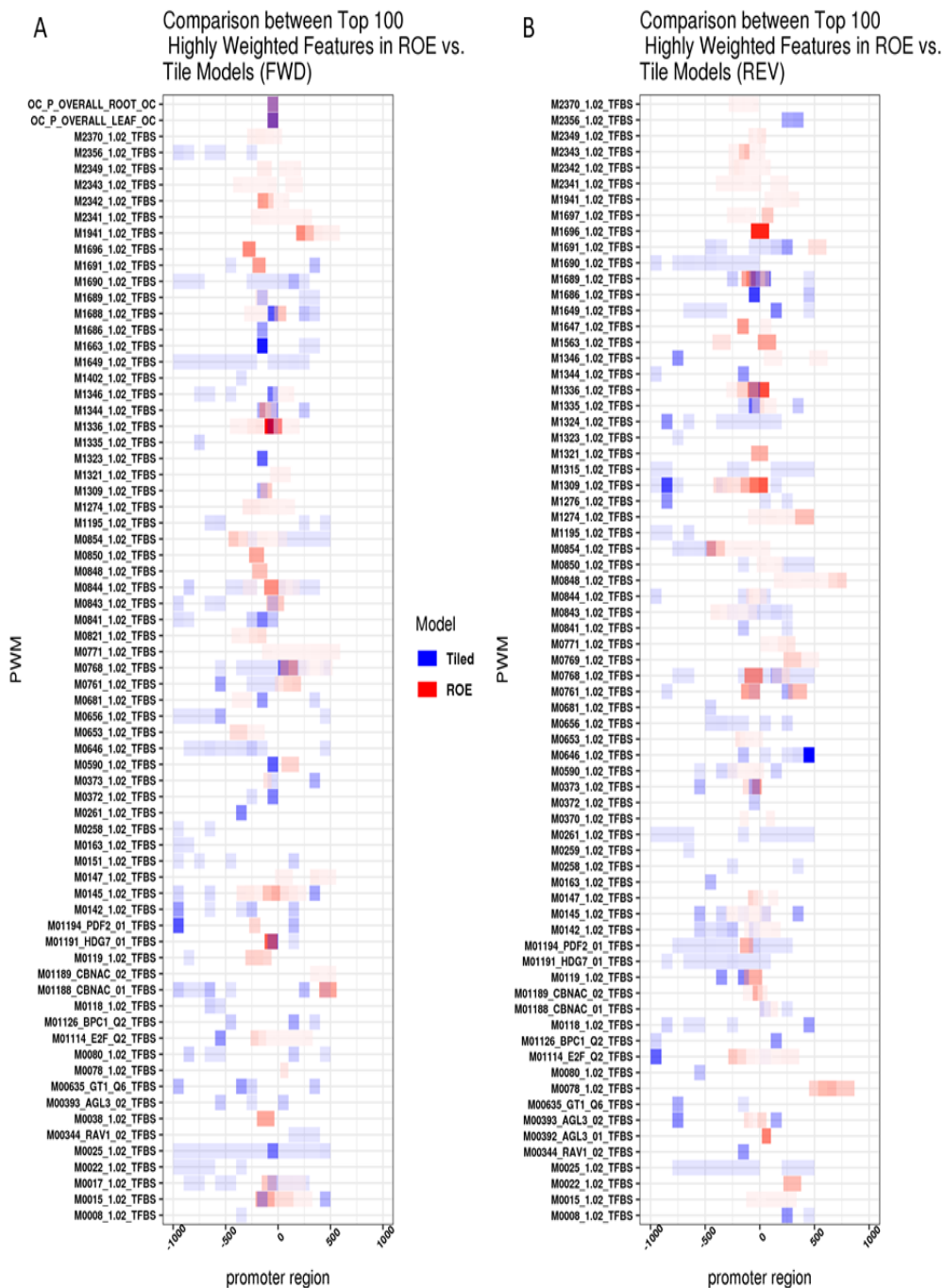


Figure 2.6. Comparisons of top PWMs between TEP models. A) Heatmap showing the differences between and shared PWMs on the same-as-gene (FWD) strand that are weighted highly by TEP-ROE (red) and TEP-Tiled (blue). B) Heatmap showing the differences between and shared PWMs on the opposite-from-gene (REV) strand that are weighted highly by TEP-ROE (red) and TEP-Tiled (blue).

Supplementary Table A9). Overall, comparisons between models support the idea that the TF binding locations which contribute most to model performance—that is, to the model’s ability to correctly predict the tissue of expression—lie within about 500 nt of the TSS.

TEP models suggest some promoters may express almost solely based on patterns of functionally bound sites

In examining the target genes of TFs that were identified by a model as very important to differential expression, we observed that the promoters of some TSSs seemed to contain high-affinity binding site densities in important regions for several different TFs. We were interested to investigate whether some promoters seemed to be “hard-coded”, in the sense that the promoter’s associated tissue of expression appeared to be entirely dictated by TF binding site patterns, with no influence from chromatin state according to a successfully trained Tissue of Expression Prediction (TEP) model. We ran an analysis to identify promoters where tissue of expression prediction was successful and nearly all of the promoter’s ‘important feature products’—meaning high-affinity TF binding site densities (TFBS features) or large chromatin accessibility values (OC features) that were associated with high model weights (feature importance values) — derived almost entirely from the presence of high-affinity TF binding site densities. We identified promoters for 18 genes that fell above the 95th percentile for ‘hard-codedness’ (see Methods) using the TEP-ROE model, and 43 genes using the TEP-Tiled model (Supplementary Tables A10 and A11). Both gene sets were enriched for GO terms associated with metabolite biosynthesis and transport (Supplementary Table A12), while the TEP-Tiled model ‘hard-coded’ genes were additionally associated with development.

The models suggest that genes whose promoters are ‘hard-coded’ by TF binding site content to express differentially in roots vs shoots (or vice versa) could be preferentially involved in very basic processes that need to be performed differently in one tissue vs another during development, based on TF presence alone. This

implies that chromatin state in these cases is perhaps directly modulated by one or more of the TFs involved in the important binding site patterns; we did observe that in the case of each model, at least one ‘top TF binding domain’ associated with the most important feature products was known to be involved in chromatin remodeling (Supplementary Table A12).

As a result of these inquiries, we wondered whether there existed cases of promoters such that ‘zeroing out’ a single TF binding site density would be predicted to ‘flip’ the tissue in which a gene was most highly expressed. The physical analog of this experiment would be a form of ‘in-silico knockout’, where high-affinity binding sites within an important region for a TF’s influence on tissue of expression are removed or occluded, so that this TF can no longer bind in this region with respect to the current TSS. We observed that the TEP-ROE model points to 8 genes whose promoters contain ‘knockout regions’ that would cause a 50% probability shift ‘across the decision boundary’ to change the predicted tissue of strongest expression (Supplementary Table A13). The TEP-Tiled model, which has many more regions than the TEP-ROE model, points to 28 such TF-tile ‘knockout’ locations that cause at least 50% probability shifts or greater to ‘flip’ predicted tissue of strongest expression (Supplementary Table A14). For the TEP models, several of these tissue-flip-causing TF binding site density ‘knockout’ regions were also among the top TF binding domains important to ‘hard-coded’ promoters. This outcome appears to corroborate the presence of highly influential binding region locations for specific TFs that may be serving in an important ‘master regulatory’ role for the tissue of expression of some promoters. Collectively, modeling experiments suggest that it is typically pairs or larger groups of these important TF binding density regions located in spatial patterns that most heavily influence tissue of expression within the ~500 nt upstream proximal promoter region of a transcript.

TF site presence and location are predominant explainers of tissue of expression

Although TFBS features were highly dominant in both the TEP-ROE and TEP-Tiled models, we unexpectedly observed that the chromatin regions surrounding the important TFBS feature binding site density locations were not considered important at all by the model. In fact, TFBS feature weight values assigned by each TEP model had virtually no correlation with model weight (importance) of the corresponding region of chromatin (Supplementary Figures A15 and A16). Additionally, we observed no correlation between the importance of regions containing large TFBS densities within individual TSS promoters and chromatin openness in these regions (Supplementary Figures A17 and A18). This was perplexing; because model structure means that important TFBS features represent locations where large TFBS densities contribute strongly to expression in one tissue vs another, we had anticipated that the state of chromatin accessibility of these regions at least in some cases would be correspondingly important. We noted that the OC_overall features, representing openness of the general proximal [TSS – 500 nt, TSS + 100 nt] region, received high weight in both tissues for both models. Both models also agreed on TFBS features as overwhelmingly more important than OC features as a collection. It seemed possible then that a higher degree of general openness of this TSS-proximal region was contributing the vast majority of chromatin state information in the models.

We decided to examine whether removal of OC features entirely would seriously hurt model performance. We therefore re-trained both the TEP-ROE and TEP-Tiled models on feature sets that were identical to the original models with the exclusion of any OC feature. We observed a ~5% drop in auROC in both cases (Figure 2.7), with the resulting “TFBS-Only” models performing surprisingly strongly at 87% auROC, with 88% and 91% auPRC respectively. We then wondered if a TFBS-Only model with root and shoot OC_overall features included, but no other OC features, would perform as well as the original model. We tested this idea, and both model types







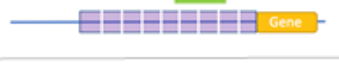

Model	auROC	auPRC
TFBSs + Chromatin state – ROE 	92%	94%
TFBSs only – ROE 	87%	88%
Chromatin state only – ROE 	84%	86%
TFBSs + OC General – ROE 	91%	92%
TFBSs + Chromatin state – Tiled 	93%	94%
TFBSs only – Tiled 	87%	91%
Chromatin state only – Tiled 	85%	88%
TFBSs + OC General – Tiled 	91%	93%

Figure 2.7 Performance summary. The removal of chromatin information from our feature set results in a performance decrease of 5-6% in both of our TEP model types. Removal of sequence-based information decreases performance by 8-9% in the TEP models. Addition of a general openness feature for the promoter region to the TFBS only TEP models (blue-highlighted lines) restores performance essentially to the same level as the original models containing all chromatin features.

achieved essentially the same performance as the original TEP-ROE and TEP-Tiled models achieved with all feature types present (Figure 2.7). Thus, by returning only a single general measure of openness of the [TSS - 500 nt, TSS + 100 nt] proximal promoter region in each tissue to a DNA-sequence-based-features-only version of each model, performance was restored to essentially the same level as when all OC features were included.

Finally, we expected that the ‘inverse’ experiment-- removal of all TFBS features-- would not seriously hurt model performance, because there is a strong tendency of expressing promoters to be more open (Figure 2.2A) within ~1 kb upstream or so of the TSS; thus, open chromatin in general should be a strong predictor of differential

gene expression. We went forward with this experiment and observed that both models performed with an auROC of ~85% (Figure 2.7), a ~7% drop in auROC from the original TEP-ROE and TEP-Tiled models. This result was surprising only in that it suggests that chromatin-state patterns, at least on a regional scale, can predict tissue of expression in strongly differentially expressing cases only about as well as TFBS and other DNA sequence content information alone. By far, the highest weighted coefficients in this model were the OC_overall root and shoot coefficients, corroborating the predictive power of general chromatin openness within the [TSS – 500 nt, TSS + 100 nt] region.

While it is possible that examination of OC patterns on a binding-site-scale might contribute some additional information if it were to become technically feasible to train such a model, all outcomes in our set of experiments clearly suggest that TF binding site information within promoter DNA is the predominant explainer of the tissue of expression. In particular, outcomes support the concept that patterns of TFBS densities within the ‘more-open’ ~500 nt proximal promoter region provide the largest influence on the tissue(s) in which a gene will preferentially express. Outcomes support a secondarily influential role for the degree to which a promoter is generally open in this region within a tissue, although of course the modeling experiments cannot inform whether this higher degree of proximal-promoter openness results from a chromatin remodeling process that is TF-dependent. Surprisingly, modeling outcomes did not support any direct association between important TF site density locations and the importance of chromatin state in these locations.

DISCUSSION

Model success in predicting tissue of expression fundamentally derives from precise TSS information

The Tissue-of-Expression-Prediction or “TEP” models constructed in our study are conceptually straightforward classical machine learning models— they use L1-regularized logistic regression to find specific high-affinity TF binding site regions positioned in relationship to a TSS, along with chromatin accessibility values in these regions, that collectively identify the correct tissue of greater expression for the vast majority of differentially expressed genes in our sample set (auROC ~90%). Previous attempts at this specific task in similarly genomically complex organisms have made good progress but achieved middling results at best (auROC ~75%). Why then did the TEP models perform so well as compared to past models, including a recent deep learning model in cell lines (Agarwal and Shendure, 2020) that did in fact have precise genome-wide TSS-Seq information available? The construction of a first-of-its-kind dataset with generation of both Transcription Start Site sequencing data and Open Chromatin sequencing data in two different tissues of the same healthy individuals likely contributed to success. However, all of our experiments clearly indicate that the most important contributor to predictive success was modeling with accurate TSSs in each plant organ. Our outcome is consistent with similar classical machine learning studies including (Vandenbon and Nakai, 2010) in demonstrating that TF binding site information alone is capable of achieving relatively strong predictive performance, and (Natarajan et al., 2012) in confirming that chromatin accessibility information boosts inference of genes that are expressing differently in different tissues/cell-types. In relationship to the (Vandenbon and Nakai, 2010) study, which performed an identical task with auROC of 75%, the TEP models’ ~15% auROC performance drop when only annotated start sites were used is consistent with the idea that accurate TSS information within each tissue is largely responsible for dramatic performance boost.

However, precise TSS information alone is unlikely to be solely responsible for high sensitivity and specificity given that a sophisticated DNA-sequenced-based deep learning model had the benefit of TSS-seq data but achieved only ~65% auROC on

this task; (Agarwal and Shendure, 2020) concludes in fact that the model's performance is not boosted by the use of TSS-seq data instead of annotated start sites. Given the observations in our study, it is very likely that explicit use of important biological information such as TF binding profiles in feature set construction confers a large benefit in predicting tissue of expression. TEP model feature sets are carefully constructed to use high-affinity TF binding site densities as opposed to thresholding, and to encompass positional relationships of these densities to the TSS. Additionally, despite its relative simplicity as a classical machine learning model, regularized logistic regression is a time-tested method that still routinely outperforms deep-learning approaches in genomic classification of phenotype from transcriptomics data (Smith et al., 2020). While deep learning models hold exciting promise, it seems likely that current architectures are as-yet unable to learn TF binding site features with enough precision to take advantage of patterns in their positional relationships to each other and to the TSS.

We would hypothesize in this context that much of the “missing mass” in performance to bring auROC up near 100% with a TEP-style model is contained in the incomplete TF binding site domain profile collection presently available even in a well-characterized model species such as *Arabidopsis*. It is certainly possible that an unconsidered influence such as DNA methylation status plays a role, although this seems largely correlative with chromatin status and not necessarily definitively causal. Finally, it is possible that important micro-scale chromatin accessibility patterns are not able to be well-captured at present by our model, given that the potentially relevant set of binding site locations and their associated chromatin state is enormous as compared to the set of highly expressed TSS locations in the genome; regularization algorithms do have limits on their ability to select the most predictive features from a vast sea of uninformative values using a relatively small number of examples. Yet we see little indication of this ‘micro-scale chromatin accessibility

pattern' transcriptional control concept within the fairly broad regions of accessible chromatin in the *Arabidopsis* proximal promoter.

Accurate TSSs implicate proximal cis-regulatory regions as primary determinants of tissue-specific gene expression

Our modeling outcomes strongly suggest that DNA sequence, within about 500 nt directly upstream of the TSS, is by far the most influential feature in successfully predicting tissue expression level differences, as opposed to distal chromatin status. Specifically, our study suggests a paradigm shift in the way we generally assume plant promoters to operate: for the vast majority of differentially expressed genes in developing *Arabidopsis* organs, it is the pattern of cis-regulatory sites in the TSS-proximal DNA of these regions, regardless of chromatin state, that is most explanatory of the tissue of expression. The presence of TF binding site Regions of Enrichment, and the ability to predict both TSS location and tissue of expression primarily from binding sites within these regions, underscore the important tissue specificity role of TF binding site patterns within the [TSS – 500 nt, TSS + 100 nt] 'proximal' promoter region in developing *Arabidopsis* seedlings.

It is surprising that TF binding site patterns in relatively accessible proximal promoter regions could largely dictate a gene's tissue of expression, though there is a growing body of genome-scale evidence that this may well be the case in higher eukaryotes (Vandenbon and Nakai, 2010; Huminiecki and Horbańczuk, 2017; Chereji et al., 2019). Studies such as (Maher et al., 2018) emphasize specific groups of TFs that appear to act as 'control modules' within different plant tissues and cell types; it may be the case that a relatively small and distinct group of TF master regulators tends to work in-concert within each tissue to help orchestrate chromatin remodeling, ensuring that promoter regions are largely accessible in the important proximal locations.

Additionally, several studies suggest the intriguing possibility that distal enhancer regions may in fact be playing a significant role in tissue specific gene expression, but

that Pol-II interactions with these enhancers are dictated to a high degree by proximal promoter sequences. The (Taher et al., 2013) study entitled “Sequence signatures extracted from proximal promoters can be used to predict distal enhancers” provides substantive computational evidence for this concept. (Ong and Corces, 2011) provides a literature synthesis of studies on enhancer function in tissue specific gene regulation, noting cumulative evidence that chromatin looping between enhancer and promoter regions is likely to be dictated at least in part by specific groups of TFs. Our study is consistent with the possibility that distal enhancers are indeed playing a substantial role, but are interacting with specific patterns of TFs which bind the proximal promoter to mediate chromatin looping.

Implications for synthetic biology: systematic design of tissue-specific promoters

A recent study (Cai et al., 2020) strongly supports the concept that the specific locational arrangements of endogenous binding sites within a plant promoter can have a dramatic effect on overall expression level. The construction and outcomes from our Tissue-of-Expression-Prediction from Regions of Enrichment or “TEP-ROE” model carry two practical implications along these lines for additionally directing strong expression in one tissue vs another. Firstly, the TEP-ROE model identifies specific TSS-proximal TF binding site regions as important to differential gene expression in each tissue sample— in our study, developing *Arabidopsis* roots and shoots. Our gel-shift analysis provides plausible support for the idea that when these model-identified patterns of high-affinity TF sites are located upstream of a specific promoter, then these sites may be bound and functional, serving in the context of surrounding sequence to preferentially upregulate gene expression in a particular tissue. Secondly, when certain TF binding densities are given a zero-coefficient or ‘removed’ from a promoter, this can produce a large shift across the decision boundary, indicating a model prediction that removal of high-affinity sites in this region would change the tissue in which a gene expresses most strongly.

In other words, in-silico "knockouts" identify TF binding regions that alter the predicted tissue in which the gene is differentially expressed. There are hundreds of cases in which a single TF-region 'knockout' is predicted to cause such a shift, and thousands of cases in which a 'double-knockout' is predicted to cause such a shift. Taken together, these results suggest strong potential for tissue-specific promoter design. For this application, rather than focusing on differentially expressed genes, one would re-train a TEP-ROE model to classify the tissue of expression for genes that expressed very highly in one tissue and very little in the other, in terms of absolute transcript counts. This would allow identification of specific high-affinity TF binding sites that, when removed from the context of a certain promoter, change the tissue of expression for a gene entirely. In summary, our model presents the exciting possibility that tissue-specific synthetic promoters can be systematically constructed using endogenous cis-regulatory sites whose presence/absence in specific locations leads to a predicted shift in tissue of expression.

DATA AND MODEL AVAILABILITY

The full dataset of mapped, annotated nanoCAGE-XL TSS peaks, DNase I SIM peaks, and RNA-Seq expression levels for root and shoot samples in our study is available on GBrowse at

<http://megraw.cgrb.oregonstate.edu/suppmats/TissueOfExpressionPredictionDatasets>.

All raw datasets, processed datasets, and model coefficient files are also made available for download. Model training and evaluation pipelines are available upon request; these are designed to run on a Sun Grid Engine computing cluster and require user familiarity with the Unix/Linux operating system, Make, Java, R, and Python; the pipelines cannot be supported by the authors on other hardware systems.

ACCESSION NUMBERS

All raw reads have been deposited in the National Center for Biotechnology Information Sequence Read Archive repository under the following accession numbers: OC-Seq (DNase I SIM) – PRJNA285928; TSS-Seq (nanoCAGE-XL) – PRJNA658605; RNA-Seq – PRJNA658596.

ACKNOWLEDGEMENTS

We thank Dr. Uwe Ohler for ideas and discussions with M.M that inspired the conception of this study. We thank Dr. Ashok Prasad for his idea to investigate the existence of “hard-coded” promoters. We thank Dr. Jason Cumbie for his help in preliminary evaluation of dataset quality, Natalie Brewer for her help in the PWM literature search, and Jordan Holdaway and Teresa Tran for their help in tissue preparation for the study. Data generation for the study was supported by an NIH K99-R00 Pathway to Independence Award GM097188 to M.M. Algorithm design and computational analysis for the study was supported by an NSF CAREER Award 1750698 to M.M.

Chapter Three

Metabolomics analysis reveals both plant variety and choice of hormone treatment modulate vinca alkaloid production in *Catharanthus roseus*

Valerie N. Fraser, Benjamin Philmus, Molly Megraw

Plant Direct

September 28, 2020

<https://doi.org/10.1002/pld3.267>

ABSTRACT

The medicinal plant *Catharanthus roseus* produces numerous secondary metabolites of interest for the treatment of many diseases—most notably for the terpene indole alkaloid (TIA) vinblastine, which is used in the treatment of leukemia and Hodgkin’s lymphoma. Historically, methyl jasmonate (MeJA) has been used to induce TIA production, but in the past, this has only been investigated in either whole seedlings, cell culture, or hairy root culture. This study examines the effects of the phytohormones MeJA and ethylene on the induction of TIA biosynthesis and accumulation in the shoots and roots of 8-day old seedlings of two varieties of *C. roseus*. Using LCMS and RT-qPCR, we demonstrate the importance of variety selection, as we observe markedly different induction patterns of important TIA precursor compounds. Additionally, both phytohormone choice and concentration have significant effects on TIA biosynthesis. Finally, our study suggests that several early-induction pathway steps as well as pathway-specific genes are likely to be transcriptionally regulated. Our findings highlight the need for a complete set of ’omics resources in commonly used *C. roseus* varieties and the need for caution when extrapolating results from one cultivar to another.

INTRODUCTION

Many plant-derived secondary metabolites have chemical properties that give them therapeutic value for the treatment of cancers, hypertension, and other illnesses (Balunas and Kinghorn, 2005). In the medicinal plant *Catharanthus roseus* (L.) G. Don, the terpene indole alkaloid (TIA) family of natural products include many valuable medicinal compounds such as the clinically used antineoplastic agents vinblastine and vincristine, as well as the antihypertensive agent ajmalicine (Figure 3.1). Vinblastine and vincristine, used in the treatment of lymphoblastic leukemia (Noble et al., 1958; Johnson et al., 1963), are naturally produced at low levels in the leaves of the plant, which makes the chemical extraction of the two alkaloids difficult and time consuming (Tyler, 1988). Pharmaceutical scientists generally extract the more abundant precursor compounds from the leaf and perform an *in vitro* coupling to increase the yield of vinblastine and vincristine, which is then isolated (Potier, 1980; Ishikawa et al., 2008); this process, however, can be cost prohibitive. While MeJA is too expensive for practical use in a large-scale agricultural production, ethephon (a commercially available ethylene derivative) is a viable and cost-effective option for increasing alkaloid yields prior to chemical extraction.

Over the last 50 years, laboratory studies of vinca alkaloid production has been induced *in planta* with MeJA via root-uptake from growth medium or through exposure to vapor in an enclosed system (Aerts et al., 1994; Rijhwani and Shanks, 1998; El-Sayed and Verpoorte, 2004). Ethylene and its derivative ethephon (EPTN) have more recently been identified as an induction agent for the TIA pathways (Pan et al., 2010; Wang et al., 2016). Foliar application of ETPN, a compound that is quickly converted to ethylene upon uptake into the cell, does not require any special equipment and is a method that can be straightforwardly transferred from a laboratory setting into a greenhouse setting for agricultural-scale production of these desirable compounds. If large scale biopharmaceutical production is the ultimate goal, foliar

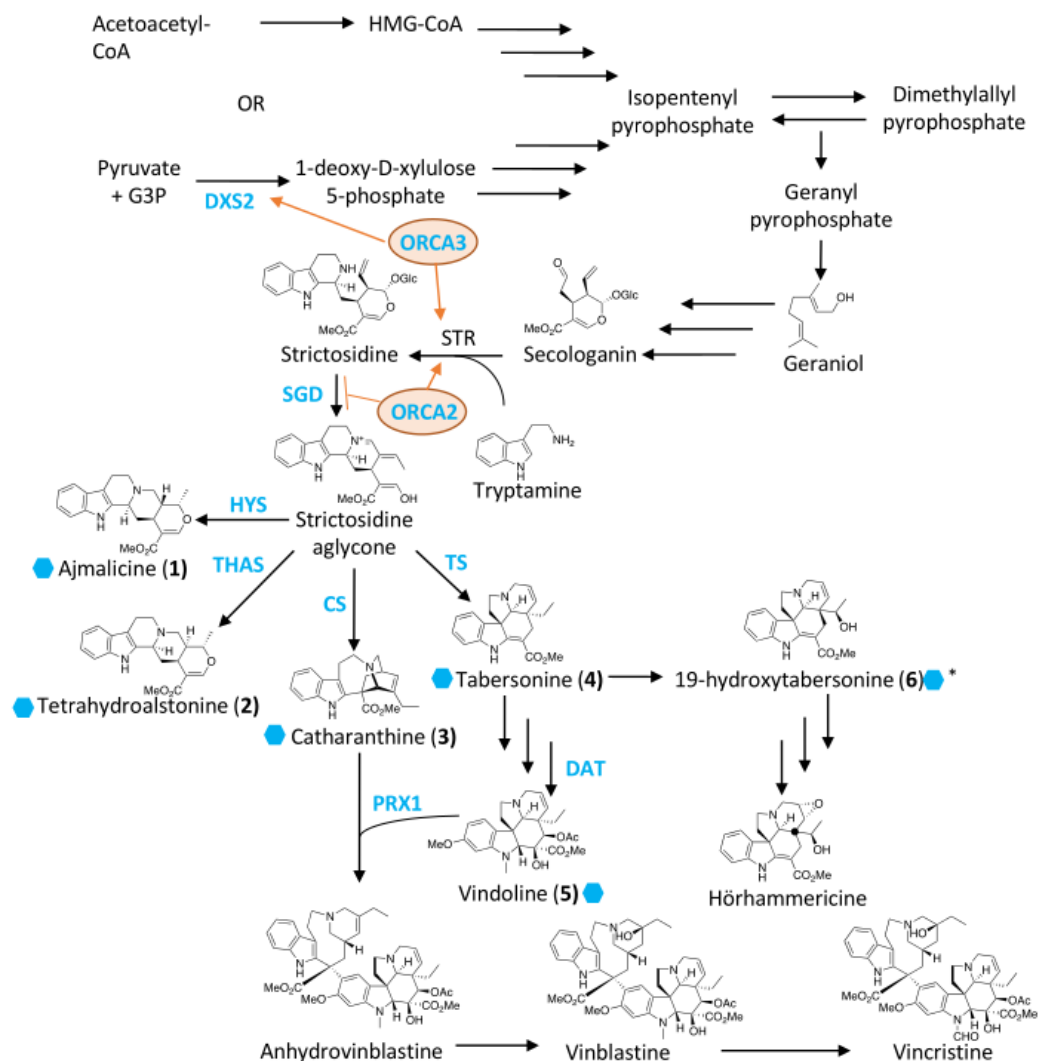


Figure 3.1 Pathway diagram from MVA and MEP to TIA. Three arrows symbolize multiple enzymatic steps and intermediates. Abbreviations in blue are the genes selected for RT-qPCR. Orange ovals represent the TFs selected for RT-qPCR. Compounds marked with a hexagon were quantified on LCMS. An asterisk over the hexagon denotes our hypothesis for the identity of the uncharacterized compound.

ETPN treatment is ideal since it is inexpensive and does not need to be reapplied to obtain the desired result.

Hairy root culture is a commonly studied system with strong potential for *C. roseus* for alkaloid production and extraction; however, it is a technically challenging system

(Williams and Doran, 2000). In particular, this and similar culture systems require special equipment and impeccable sterile technique to prevent contamination. Additionally, not all precursor alkaloids of interest in the TIA pathway can be found in the roots at levels that would make extraction viable (e.g. vindoline) in the absence of further genetic engineering developments in this system (O'Keefe et al., 1997; St-Pierre et al., 1999; Laflamme et al., 2001; Besseau et al., 2013), and those that are present are regulated differently than in seedlings (Pan et al., 2016). Alternatively, *C. roseus* seeds are easy to germinate and are relatively fast-growing in soil. Gently uprooting seedlings from the soil and thoroughly washing in deionized water allows collection of all parts of the plant in a relatively short amount of time and with minimal concern regarding contamination. These considerations make plants a good system not only for biological studies but also provides potential for greenhouse-level scale-up of alkaloid precursor production.

The biosynthetic pathway for terpene indole alkaloid (TIA) production in *C. roseus* has been the focus of investigation for many years starting with precursor labeling experiments (Verpoorte et al., 1997) to the more recent identification of biosynthetic and regulatory genes. The TIA biosynthetic pathway begins with the coupling of the isoprene building blocks dimethylallyl pyrophosphate (DMAP) and isopentenyl pyrophosphate (IPP) to form geranyl diphosphate (GPP). Reduction of GPP to geraniol followed by a multi-enzyme conversion results in secologanin, a common precursor for many plant natural products (Leete, 1967; Verpoorte et al., 1997). A Pictet-Spengler reaction coupling secologanin and tryptamine (derived from decarboxylation of tryptophan by tryptophandecarboxylase (TDC)) by strictosidine reductase (STR) yields strictosidine. Deglycosylation by strictosidine (SGD) followed by a spontaneous rearrangement yields strictosidine aglycone, a branch point for the biosynthesis of many TIAs including ajmalicine, tetrahydroalstonine, catharanthine, and tabersonine. Catharanthine and tabersonine are formed from strictosidine aglycone via a series of shared reactions facilitated by enzymes named

precondylocarpine acetate synthase (PAS) and dihydroprecondylocarpine synthase (DPAS), followed by separate conversions by either catharathine synthase (CS) or tabersonine synthase (TS) respectively (Caputi et al., 2018). These four enzymes have also been described as geissoschizine synthase, O-acetylstemmadenine oxidase, hydrolase 1, and hydrolase 2, respectively (Qu et al., 2018). Tabersonine is converted in multiple steps to vindoline, which is then coupled to catharanthine by (PRX1) to yield anhydrovinblastine, which is then subsequently converted to vinblastine and vincristine (Money et al., 1968; Verpoorte et al., 1997). Early steps in the TIA biosynthetic pathway are transcriptionally regulated by ORCA2 and ORCA3 (Liu et al., 2011; Pan et al., 2012; Li et al., 2013). Upregulation of ORCA2 inhibits SGD expression while upregulating STR expression (Liu et al., 2011; Li et al., 2013). ORCA3 upregulation results in the upregulation of both DXS2 (non-mevalonate isoprenoid biosynthesis) and STR (Pan et al., 2012). Regulation of the later biosynthetic steps in the TIA pathway is currently unknown.

Many different cultivars of *C. roseus* have been developed for ornamental uses and, of these, some have also been evaluated for their utility in alkaloid production. Among these genetically diverse varieties, however, only a few have been selected for genomic and transcriptomic resource development (Góngora-Castillo et al., 2012; Verma et al., 2014; Kellner et al., 2015; Pan et al., 2018). “Little Bright Eye” (LBE) is a variety that has been commonly used for plant pathology research and was used in the initial efforts to identify the TIA biosynthetic genes. More recently, other varieties have been investigated in transcriptional and metabolomic studies (Góngora-Castillo et al., 2012; Verma et al., 2014; Kellner et al., 2015; Pan et al., 2018). “SunStorm Apricot” (SSA) was developed for horticultural use and recently was selected for genome sequencing (Kellner et al., 2015); it remains the only sequenced *C. roseus* variety to date. Given that no single *C. roseus* variety currently has a complete set of genomic, transcriptomic, and metabolomic data available (Supplementary Table B1), we wanted to investigate how alkaloid production and

response to stimuli differ between the two varieties associated with the most widely used 'omics resources (LBE and SSA). With this in mind, we designed a study of the alkaloid induction patterns of ethylene and MeJA in these two varieties of *C. roseus* (LBE vs. SSA). Some precursor alkaloids are restricted to certain tissues (O'Keefe et al., 1997; St-Pierre et al., 1999); thus, we chose to perform all assays in both roots and shoots rather than in whole seedlings, which has not been addressed in previous *C. roseus* work. Additionally, testing hormonal induction in LBE has allowed us to compare observations with previous studies, while including SSA provides an opportunity for future genomic investigation into the regulation of important induction pathways.

In this work, the *in planta* effects of foliar MeJA or ETPN treatments on the metabolomic profiles in roots and shoots on alkaloid levels was investigated in both varieties. The natural differences in alkaloid levels between these varieties in roots and shoots were also investigated. Finally, this work examines the transcriptional effect of these phytohormones on the expression of genes involved in the terpene indole alkaloid biosynthetic pathway and examines the relationship between transcriptional and metabolic profiles. We show that not only do varietal differences play a major role in alkaloid response to hormonal stimuli, but also that genetic variation between SSA and LBE is substantial enough to affect wildtype levels of alkaloids in both roots and shoots.

MATERIALS AND METHODS

Plant material and growth

Two *Catharanthus roseus* varieties were selected for these experiments: “SunStorm Apricot” (obtained from www.expressseed.com) and “Little Bright Eyes” (obtained from www.neseeds.com). 10-12 seeds of a single variety were planted in 4-inch plastic pots filled to 1 cm below the top with MetroMix potting mix (35%-45%

Sphagnum moss, bark, pumice, dolomite limestone). Pots were arranged on labeled trays, which were covered with plastic domes to increase humidity until seedlings emerged through the soil. The plants were grown in an environmentally controlled growth room under a 12-hour light/12-hour dark photo-cycle with a 22°C ambient temperature.

Extraction protocol validation

Prior to beginning the bulk of this study, we validated our extraction techniques to ensure that technical and biotic influences were minimized using SSA. To test our technical reproducibility, we pulverized up 10-20 shoots in liquid nitrogen and then, after thorough mixing of the resulting powder, the powder was divided into three approximately equal portions of plant material. The replicate plant material was extracted and analyzed via LCMS (in technical duplicate) as described in the section “**LCMS quantitation of *C. roseus* alkaloid**”. The vindoline concentration was determined to be 12.2 ± 2.7 µg/mg wet weight. To test for biological variability, 20 plants were grown under identical conditions and randomly allocated to three samples. The samples were pulverized in liquid nitrogen, extracted with methanol and analyzed by LCMS (in technical duplicate) as described in the section “**LCMS quantitation of *C. roseus* alkaloid**”. In these samples the vindoline concentration was determined to be 10.9 ± 3.0 µg/mg wet weight. This demonstrated that our extraction protocol was reproducible and accurate.

Phytohormone treatments and sample collection

The “SunStorm Apricot” variety of *C. roseus* seedlings were germinated in soil and grown to 8 days post-germination (Aerts et al., 1994; El-Sayed and Verpoorte, 2004), at which time they were sprayed with 5 mL of DI water or 100 µM or 1 mM ethephon (dissolved in DI water). After treatment, plants were sealed inside 2-gallon zip-top bags and returned to the growth chamber for 24 hours. On the next day, the plants

were carefully uprooted, washed with DI water, separated at the hypocotyl into roots and shoots with a surgical blade, and flash-frozen in liquid nitrogen. Samples were stored at -80 °C until they could be processed, a minimum of 24 hours.

These concentrations of ethephon were chosen as they are the manufacturer (Monterey Lawn & Garden) recommended concentration for agricultural applications (1 mM) or identical to concentration of MeJa applied (100 μ M). Ethephon was mixed in DI water alone for the treatments, while the control treatment consisted of DI water. The plants were then handled as described above. Each sample from both of these experiments consisted of all the plants from a single pot; there were 12 pots for each variety. At the 1 mM concentration, the plants began showing signs of senescence, becoming yellow and wilted.

Using the data obtained from the optimization trials, we designed our larger experiment as shown in Figure 3.2. For this experiment, both SSA and LBE seedlings were grown to 8 days after germination. 6 pots of each variety were selected at random from the trays, sprayed with a combined volume of 5 mL of DI water (ethephon control), DI water + 0.02% DMSO (methyl jasmonate control), 100 μ M ethephon, 1 mM ethephon, or 100 μ M methyl jasmonate + 0.02% DMSO. After treatment, the plants were handled as described above. We processed 6 replicates for each treatment. Each sample contained all the plants from a single pot (~8 plants).

Alkaloid Extraction

Shoots were ground in liquid nitrogen with a mortar and pestle; roots were macerated by hand with a metal spatula directly in the methanol solution to prevent sample loss during the grind and transfer process due to the small amount of tissue. Shoot extractions were performed using 1 mL methanol containing 10 μ M ajmaline (internal standard) per 100 mg tissue. Root extractions were performed using 1 mL of methanol containing 1 μ M ajmaline (internal standard) per 10 mg tissue. The extracts

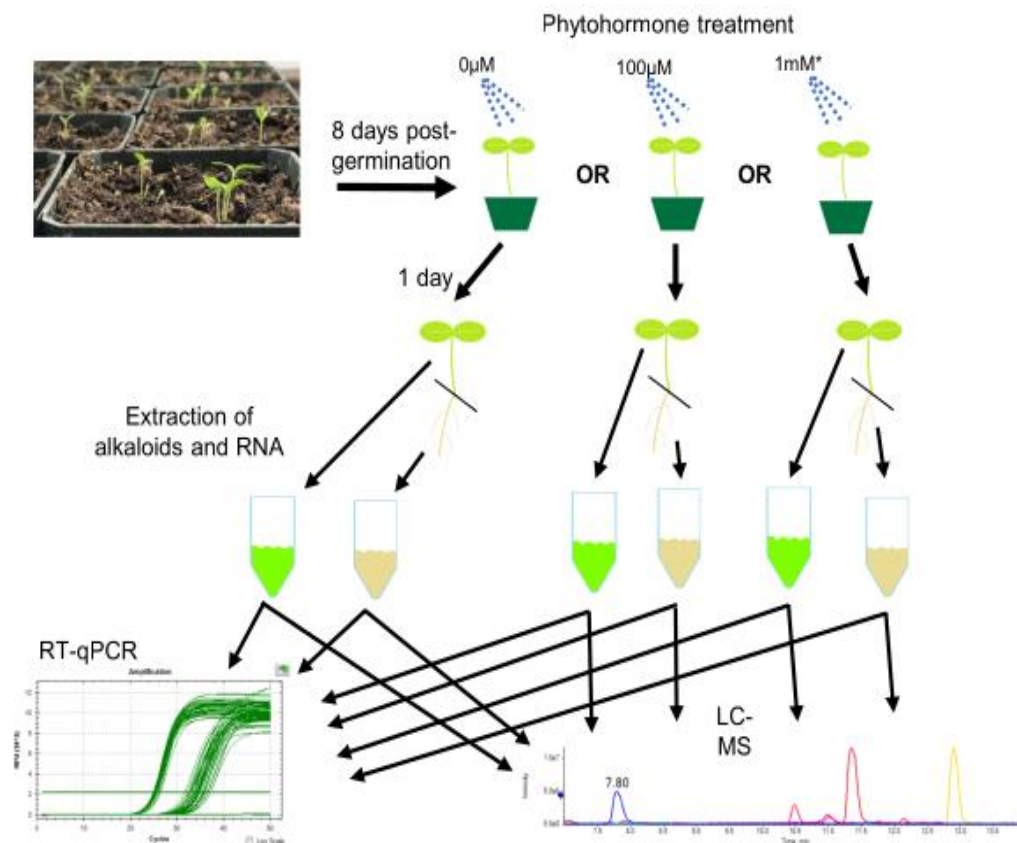


Figure 3.2 Experimental design of this study. Seeds of two *Catharanthus roseus* varieties (LBE and SSA) were sown and grown to 8 days post-germination. At that time, the plants were treated with various concentrations of either ethephon or MeJA. The asterisk on the 1mM denotes that the concentration was only used for ethephon. Seedlings were harvested 24 hours after treatment, divided into roots and shoots, and flash frozen. Total alkaloid content and RNA were extracted from the frozen tissues, which were then used for LCMS and RT-qPCR analyses.

were then allowed to stand at room temperature ($\sim 22\text{ }^{\circ}\text{C}$) for 20 minutes and then the cellular debris was pelleted by centrifugation ($3,220 \times g$, $22\text{ }^{\circ}\text{C}$, 20 min). The cleared extracts were then filtered through a $0.22\text{ }\mu\text{m}$ nylon syringe filter to remove remaining particulate. The shoot alkaloid extracts were diluted 1:10 in methanol and $20\text{ }\mu\text{l}$ was transferred to HPLC vials containing glass sample inserts. The root extracts were used undiluted. The filled

HPLC vials were stored at -80°C until they could be analyzed via LC-MS (described below). Unfortunately, some samples were lost during processing. In the end, the final replications for each treatment were as follows: LBE E0-S = 6; LBE E0-R = 5; LBE E1 = 6 and 6; LBE E4 = 6 and 6; LBE M0 = 6 and 6; LBE M1 = 6 and 6; SSA E0-S = 6; SSA E0-R = 3; SSA E1-S = 6; SSA E1-R = 5; SSA E4 = 6 and 6; SSA M0-S = 6; SSA M0-R = 5; SSA M1 = 6 and 6.

LCMS quantitation of *C. roseus* alkaloids

LCMS analysis was achieved using a Shimadzu Prominence HPLC (consisting of a degasser, two LC-10AD HPLC pumps, an autosampler, and system controller) upstream of a 3200 QTrap mass spectrometer (AbSciex). Separation was achieved using Luna C18 (2) column (150 x 2.00 mm, 3 µm) at a flow rate of 0.2 ml/ min and the following gradient, where line A was water with 0.1% (v/v) formic acid and line B was acetonitrile with 0.1% (v/v) formic acid. The column was pre-equilibrated with 85% A/15% B. Upon injection (2 µL of prepared HPLC sample) the mobile phase composition was maintained for 1 minute followed by changing the mobile phase to 60% A/40% B over 14 minutes using a linear gradient. The mobile phase was then changed to 0% A/100% B over the next 1 minute and held at this ratio for 8 minutes. The mobile phase was changed to 85% A/15% B over 1 minute and the column was equilibrated at 85% A/15% B for 7 minutes prior to the next injection. The mass spectrometer settings were as follows: MS (EMS positive mode, 50-1500 m/z), Curtain gas, 40.0; Collision gas, Medium; IonSpray voltage, 4500.0; Temperature, 400.0; Ion Source Gas 1, 35.0; Ion Source Gas 2, 35.0; Interface heater, ON; Declustering potential, 45.0; Entrance potential, 4.0; Collision energy, 5.0, number of scans to sum, 2; scan rate, 4000 Da/sec. MS/MS (MRM mode) For catharanthine (Q1, 337.3; Q3, 144.2; time 40 msec, CE (volts) 20.0); for tabersonine (Q1, 337.3; Q3, 305.3; time 40 msec, CE (volts) 20.0); for vinblastine (Q1, 406.2; Q3, 271.9; time 40 msec, CE (volts) 30.0); for vincristine (Q1, 413.2; Q3, 353.4; time 40 msec, CE

(volts) 30.0). Curtain gas, 40.0; Collision gas, Medium; IonSpray voltage, 4500.0; Temperature, 400.0; Ion Source Gas 1, 35.0; Ion Source Gas 2, 35.0; Interface heater, ON; Declustering potential, 45.0; Entrance potential, 10.0; Collision cell exit potential, 3.0. Data was acquired using the Analyst software package (AbSciex). LCMS grade H₂O, acetonitrile, methanol and were purchased from MilliporeSigma. LCMS grade formic acid was purchased from Fisher Chemicals. All other chemicals were purchased from Sigma-Aldrich and used without further purification unless otherwise specified.

Standard curves were generated by analyzing commercial standards at known concentrations using the identical LCMS settings. Vindoline, vinblastine sulfate, vincristine sulfate, and catharanthine were obtained from Cayman Chemicals. Ajmaline and tetrahydroalstonine were obtained from Extrasynthese, while ajmalicine was obtained from Millipore-Sigma. Lochnericine and 16-hydroxytabersonine (aka 11-hydroxytabersonine) were obtained from MuseChem.

RNA extraction and qRT-PCR

Stored tissues were ground with mortars and pestles that had been treated with RNase Zap to prevent sample degradation. The ground tissues were divided into two 2 mL microfuge tubes, which were used immediately to extract total RNA using the RNeasy Mini Kit (Qiagen) in conjunction with their RNase-Free DNase Set (Qiagen) as directed. The total RNA for each sample was quantified on a Nanodrop (Thermo Scientific) and integrity was confirmed on a Bioanalyzer 2100 (Agilent) in the Center for Genome Research and Biocomputing Core Facilities at Oregon State University. Only samples with RINs ≥ 8.0 were used for two-step qRT-PCR. Each biological replicate was used for two technical replicates, bringing the total to four replicates for each sample. 300 ng of input RNA from each sample was reverse transcribed using the SuperScript RT kit (Invitrogen). qPCR and melt curve analyses were performed using the SYBR PCR kit (Qiagen) on a BioRad C1000 Touch thermocycler with a

BioRad CFX96 detection system (BioRad). Transcript data was extracted using CFX Manager software (BioRad). Primers not sourced from literature were designed using PrimerQuest tool (Integrated DNA Technologies); all primers were ordered from Sigma-Aldrich with standard desalting.

Data Analyses

Relative intensities for each were determined from LCMS data by calculating the area under the peak (AUC) using Peakview version 2.2 (AbSciex) and then dividing that value by the AUC of our internal standard, ajmaline. Absolute concentrations were calculated from the AUC and a standard curve for each alkaloid; each quantity was then normalized using the original wet weight of the sample. We performed Welch's t-tests to determine the significance of differences in alkaloid concentrations between varieties and two-way ANOVA followed by Tukey pairwise comparison post-hoc analyses to determine the significance of treatments. For qPCR data analysis, LinRegPCR (Ruijter et al., 2009) was used to determine primer efficiencies. Absolute copy numbers of transcripts were determined ($\bar{X}_{0_s} = \Delta T * \hat{E}_s^{\left[\bar{b}_a * \log_{\hat{E}_s}(\bar{E}_a) - \bar{c}_{qs}\right]}$) and then normalized to the absolute copy number of 40S ribosomal protein S9 (RPS9), our control gene, from the same sample. The resulting data was analyzed using ANOVA and Welch's t-tests. All statistical analyses were performed in R (version 3.4.3). Values are considered significant below $p = 0.1$.

RESULTS

The terpene indole alkaloid biosynthetic pathway of *Catharanthus roseus* (TIA, Figure 3.1) is central to the production of its medically relevant natural products. We have designed a large study to examine the transcriptional regulation of vinca alkaloid production in the roots and shoots of seedlings from two varieties that are of interest to medicinal chemistry and genomics researchers (Figure 3.2). Here, we present our findings.

Alkaloid levels substantially differ between varieties

A comparison of the spectrometric results of the untreated control plants highlights notable differences between the plant varieties themselves. As these control plants were only sprayed with deionized water and the pots were arranged randomly to avoid positional effects, changes observed are attributable to variety. Vinblastine and vincristine were below the limit of detection of our LCMS system and thus are not discussed here. Additionally, we were unable to separate ajmalicine and tetrahydroalstonine despite multiple attempts, as they have identical masses, fragmentation patterns, and retention times. Therefore, we report these two compounds here as a single value relative to the internal standard.

In shoots, untreated SSA plants have markedly higher levels of tabersonine (Welch's t-test, $p \leq 0.01$), while LBE has a higher concentration of vindoline (Figure 3.3A). The mean vindoline concentration is greater in LBE than in SSA, but the difference is not significant due to LBE having much more intra-varietal variation (Welch's t-test, $p = 0.1033$). In roots, untreated LBE has higher concentrations of catharanthine and tabersonine, but not statistically significant (Welch's t-test, $p = 0.13$ and 0.2 , respectively). (Figure 3.3B). An alkaloid with a mass to charge ratio of 353 is present at high levels in the roots of untreated LBE and at lower levels in the roots of untreated SSA (Figure 3.3C). Overall, we observe that important differences arise in alkaloid concentration between varieties and between the tissues of these varieties, even without the application of an induction agent.

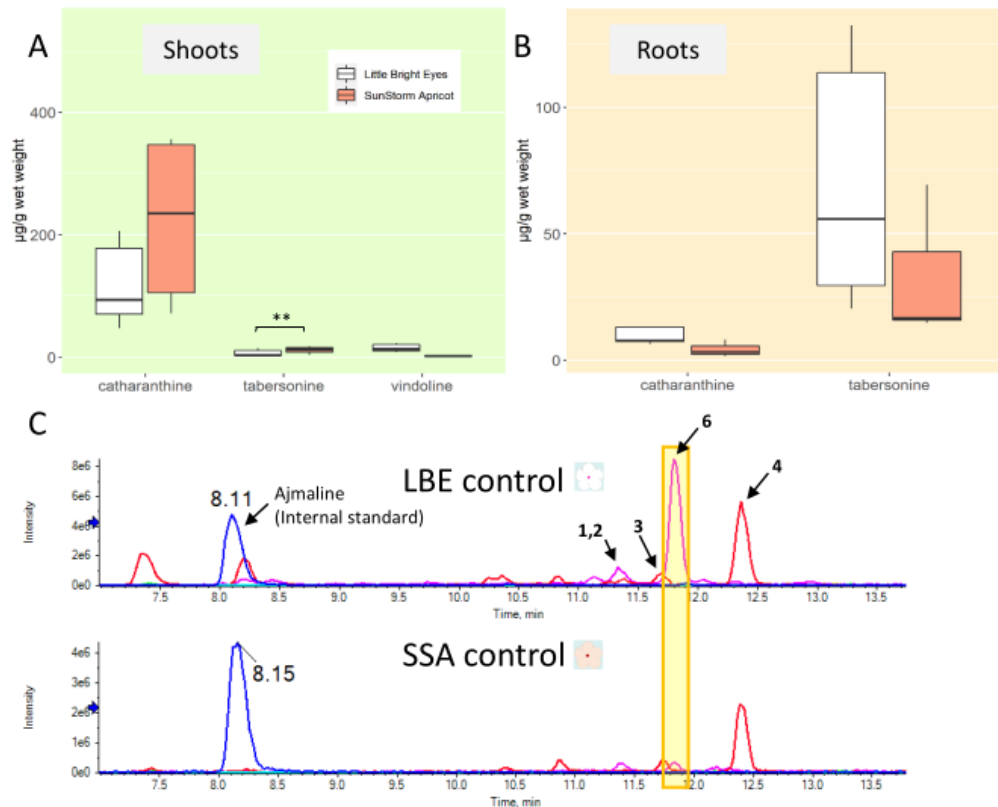


Figure 3.3 Alkaloid concentrations differ greatly between untreated plants of the two varieties. Alkaloids are labeled as follows: **1** = ajmalicine; **2** = tetrahydroalstonine; **3** = catharanthine; **4** = tabersonine; **5** = vindoline; **6** = 19-hydroxytabersonine (putative identification). ** denotes a p-value ≤ 0.01 (A) In shoots, SunStorm Apricot has a higher concentration of catharanthine and tabersonine, while Little Bright Eye has a much greater concentration of vindoline. (B) In roots, Little Bright Eye has higher concentrations of catharanthine, tabersonine, and ajmalicine/tetrahydroalstonine. (C) Representative LCMS traces from the $0 \mu\text{M}$ ethephon treatment group; even in the control treatment, there are obvious differences between the varieties. The white flower represents LBE, and the peach flower represents SSA.

Induction of alkaloid levels differs markedly based on which phytohormone is used

In both shoots and roots, both methyl jasmonate (MeJA) and ethephon (ETPN) either caused an increase in alkaloid level or had no effect; there was no evidence of a significant decrease in any of the alkaloids examined.

For catharanthine and the combined ajmalicine/tetrahydroalstonine peak, treatment with MeJA increased the concentration in the shoots of both varieties but not significantly (Figure 3.4, Supplementary Figure B1). Application of MeJA significantly increased the concentration of tabersonine in SSA, but not in LBE (Figure 3.4; Welch's t-test, $p \leq 0.01$). For vindoline, there was a small increase in LBE, which was not significant, and no increase in SSA.

Treatment of LBE with ETPN increased the concentration of catharanthine, tabersonine, and vindoline at both concentrations (Figure 3.4A, C, E). For the ajmalicine/tetrahydroalstonine peak, a small increase in the mean was observed that was not statistically significant. In SSA shoots, ETPN only significantly increases the levels of tabersonine at both treatment concentrations (Figure 3.4C). None of the other alkaloids examined showed increases in concentration.

In the roots of LBE, MeJA treatment only increased the concentration of tabersonine (Figure 3.4, and Supplementary Figure B4). For SSA, MeJA treatment significantly increased the concentration of tabersonine (Figure 3.4); the mean amount of ajmalicine/tetrahydroalstonine increased, though the increase was not significant (Supplementary Figure B1). The mean concentration of catharanthine was not significantly changed by MeJA in the roots of either variety.

Treatment of LBE with ethephon did not significantly alter the concentrations of any of the alkaloids examined in this study in the roots. In the case of SSA, treatment with 1 mM ethephon significantly increased concentrations of catharanthine, ajmalicine/tetrahydroalstonine, and an unidentified alkaloid at $m/z = 353$ in the roots (vide infra; Figure 3.4, Supplementary Figure B1, and Supplementary Table B5).

Treatment of SSA with ethephon increases tabersonine four-fold in roots but cannot be considered statistically significant (Figure 3.4D).

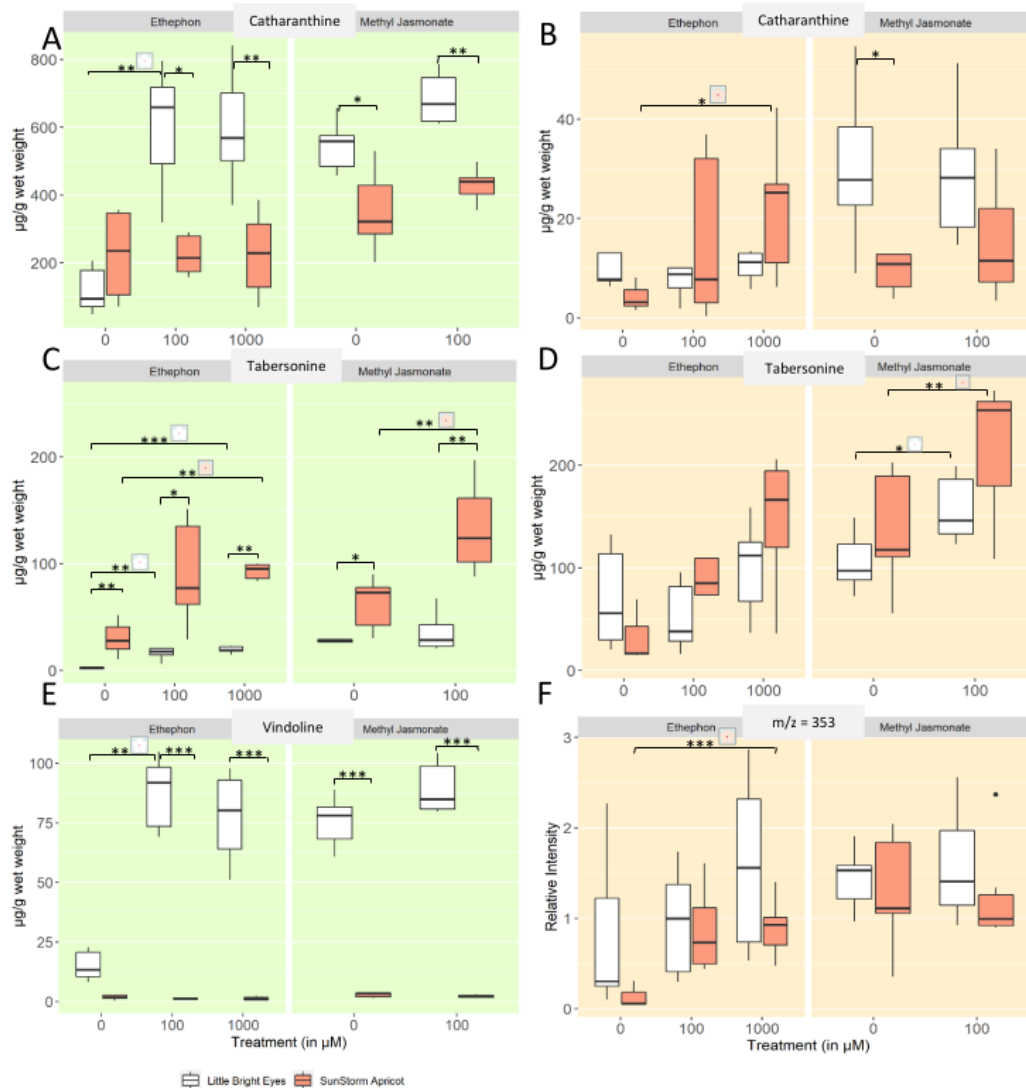


Figure 3.4 Alkaloid concentrations differ between varieties and treatment. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; all represented statistics are from Welch's t-test post-hoc analyses. Significance markers with a white flower represent treatment differences in LBE, while those with a peach flower represent treatment differences in SSA. (A) Catharanthine concentrations increased in LBE shoots after treatment with both hormones, but only with methyl jasmonate in SSA shoots. (B) In roots, catharanthine increased markedly in SSA after treatment with ethephon. (C) In shoots, tabersonine levels increase greatly in SSA upon treatment with either phytohormone, but only after treatment with ethephon in LBE. (D) Tabersonine levels increase significantly in the roots of both varieties after treatment with either phytohormone. (E) Vindoline concentration increases significantly in LBE shoots after treatment with ethephon. (F) The amount of the unidentified alkaloid present relative to the internal standard increased significantly in the roots of SSA after treatment with ethephon.

Overall, ethephon significantly increased the levels of a catharanthine, tabersonine, and vindoline in LBE shoots while MeJA did not significantly increase the amount of any alkaloid examined. In the SSA shoot samples only tabersonine was significantly increased in both the ETPN and MeJA treatments. All of the alkaloids mentioned in this study are precursors in the TIA pathway. Additionally, the interaction between treatment and variety was significant for some of the alkaloids. In shoots, catharanthine and vindoline were increased significantly by the interaction of ETPN and variety while the interaction of MeJA and variety significantly affected tabersonine. In the roots, however, none of the alkaloids had significant interaction effects. This latter result may be due to the foliar application of the phytohormones.

Master regulators are upregulated by hormonal induction

In the shoots of both varieties, the higher concentration of ETPN induces an increase in the number of ORCA2 transcripts (Figure 3.5A; ANOVA, $p \leq 0.001$). ORCA2 transcripts in roots, however, respond oppositely in the two varieties: increasing with ETPN treatment in SSA while decreasing in LBE and decreasing with MeJA treatment in SSA increasing in LBE (Figure 3.5B; ANOVA, $p \leq 0.001$). In untreated plants, ORCA3 transcripts are present at significantly higher levels in the shoots of SSA than in the same tissue in LBE (Figure 3.5C; Welch's t-test, $p \leq 0.001$). These levels in SSA, however, do not respond to treatment with ETPN—unlike in LBE, where they are significantly decreased (ANOVA, $p \leq 0.01$). In roots, treatment with MeJA increases transcripts in both varieties; however, we see opposing effects depending on variety in shoots (Figure 3.5D). The changes in ORCA3 in SSA do mirror the changes seen in catharanthine and ajmalicine in roots, though this is not true of ORCA3 in LBE. Treatment with phytohormones does have an effect on the expression of these master regulators, though both the basal level of expression and

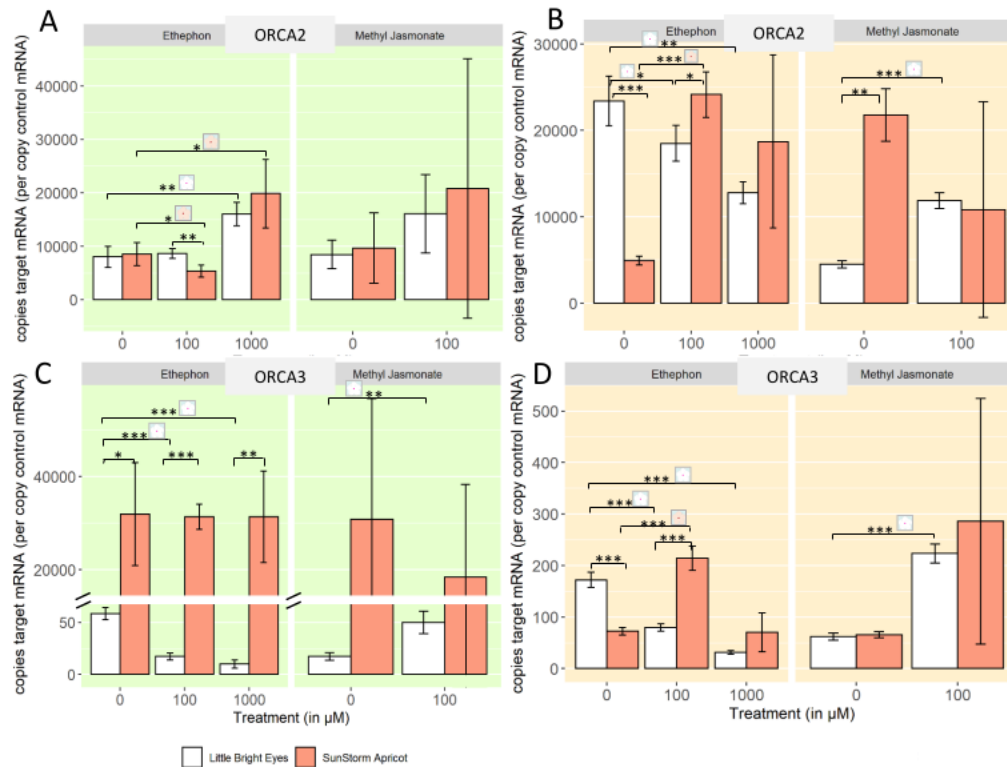


Figure 3.5 Expression of key regulatory genes are transcriptionally regulated upon phytohormone treatment. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; all represented statistics are from Welch's t-test post-hoc analyses. Significance markers with a white flower represent treatment differences in LBE, while those with a peach flower represent treatment differences in SSA. (A) ORCA2 transcripts in shoots (B) ORCA2 transcripts in roots (C) ORCA3 transcripts in shoots (+ zoomed in panel) (D) ORCA3 transcripts in roots.

the induction effect appears to vary between the two varieties.

Evidence supports transcriptional regulation of key pathway steps

We found two key primary metabolic enzymes in upstream pathways change upon phytohormone treatment. In the mevalonate-independent pathway (MEP) pathway, the DXS2 transcript level increased in the roots of our young plants treated with MeJA (Supplementary Figure B2C); in the mevalonate (MVA) meanwhile, HMGS transcript abundance decreased in the shoots of the MeJA-treated plants (Supplementary Figure B2A).

In shoots, the catharanthine synthase (CS) transcript levels increase in all groups except for ETPN treated SSA, which is consistent with the trends we observed in catharanthine production (Figure 3.6A; ANOVA, $p \leq 0.001$). While catharanthine appears to be transcriptionally regulated in LBE roots by both ETPN and MeJA, CS levels in SSA roots are only increased by ETPN (Figure 3.6B; ANOVA, $p \leq 0.001$). In the shoots of both varieties, tabersonine concentrations do not appear to be transcriptionally regulated, as tabersonine synthase (TS) transcript levels increase after treatment with ETPN and decrease after treatment with MeJA, which is not at all consistent with the trends in alkaloid concentration (Figure 3.6C). In roots, on the other hand, tabersonine does appear to be transcriptionally regulated in both varieties, as a significant induction of TS transcripts after treatment with ETPN is observed which are correlated with observed changes in the tabersonine concentrations (Figure 3.6D).

Interestingly, although vindoline concentrations increase significantly in LBE shoots upon treatment with ETPN, there is no associated increase in deacetylvindoline *O*-acetyltransferase (DAT) mRNA levels; in fact, we observe a significant decrease in both varieties (Figure 3.6E; ANOVA, $p \leq 0.01$). Shoots treated with MeJA, however, do have similar increases in DAT mRNA and vindoline (Welch's *t*-test, $p \leq 0.1$). In roots and shoots of both of the varieties, ETPN treatment caused a statistically significant increase in the transcription of PRX1 (Figure 3.6F, Supplementary Table B6; ANOVA, $p \leq 0.001$). Meanwhile, MeJA decreased transcription levels in all tissues of SSA and, significantly, in the roots of LBE (Figure 3.6F, Supplementary Table B7; Welch's *t*-test, $p \leq 0.01$). The change in ajmalicine/tetrahydroalstonine is consistent with the patterns that we observed in SSA roots and the increase that we saw in SGD transcripts is seen in all SSA tissues (Supplementary Table B6; ANOVA, $p \leq 0.001$).

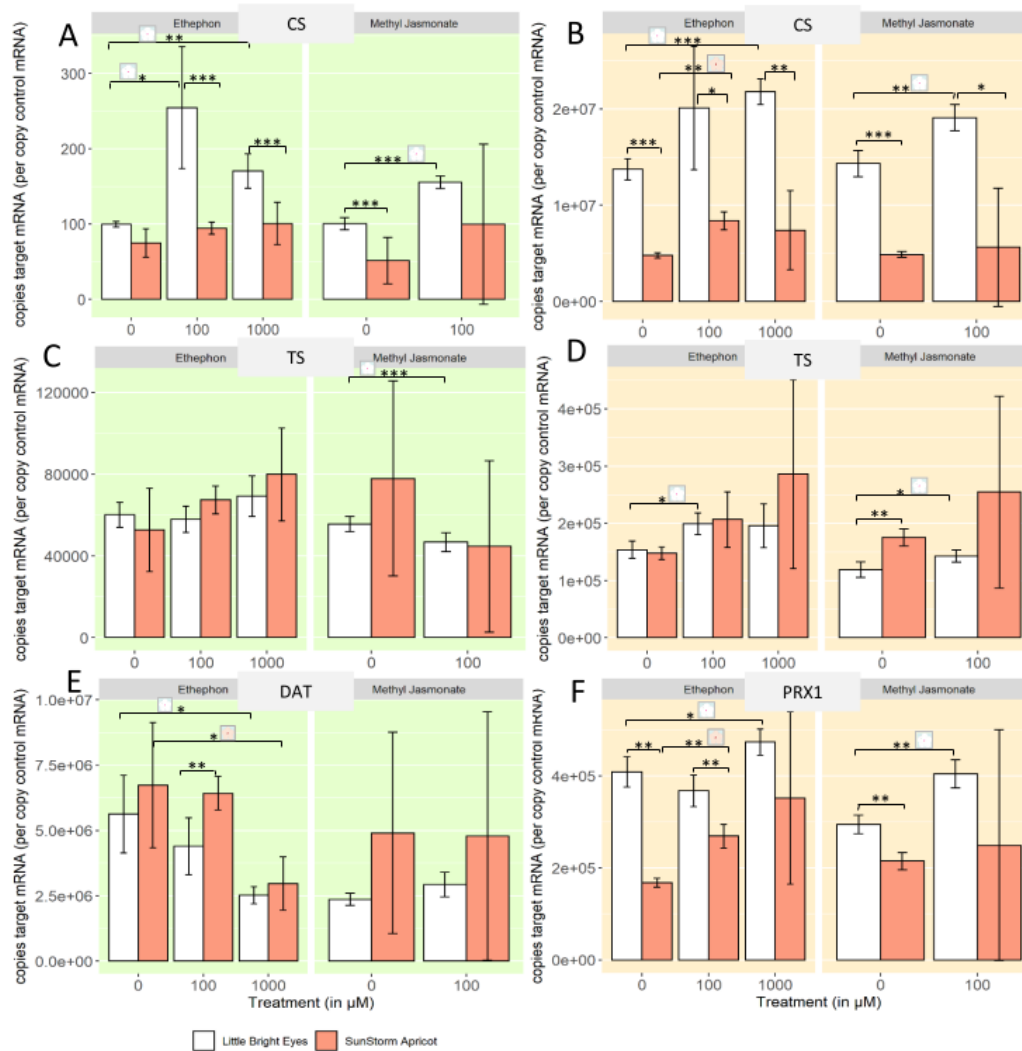


Figure 3.6 Expression of some key enzymes in the TIA pathway are transcriptionally regulated upon phytohormone treatment. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; all represented statistics are from Welch's t-test post-hoc analyses. Significance markers with a white flower represent treatment differences in LBE, while those with a peach flower represent treatment differences in SSA. (A) CS shoots (B) CS roots. (C) TS shoots (D) TS roots (E) DAT (F) PRX1 transcript levels in roots increase after treatment with ethephon in both varieties and decrease after treatment with MeJA, but only in LBE.

DISCUSSION

Previous studies have investigated the genomics or metabolomics of different *C. roseus* varieties (Magnotta et al., 2006; Kim et al., 2007; Chung et al., 2011). While

many of the examined varieties in these works are used mainly for ornamental purposes, several are used medicinally. None of the studies, however, include both LBE and SSA—which would allow for the utilization of available genomic and transcriptomic resources to inform future bioengineering efforts (Supplementary Table B1). We selected these two varieties for this reason.

Questions generated by our study

One of the compounds that exhibited a clear concentration difference between LBE and SSA was the uncharacterized alkaloid (compound **6**). Due to this alkaloid's mass-to-charge ratio ($m/z=353$) and its retention time with respect to the other identified peaks, we suspect that this compound is a hydroxylated tabersonine. We endeavored to confirm our hypothesis about our uncharacterized alkaloid's identity using commercially available standards for alkaloids with the appropriate molecular weight (including 11-hydroxytabersonine, yohimbine, and lochnericine); however, none of the commercially available standards exhibited the same retention time as this peak. We therefore posit that this compound is 19-hydroxytabersonine, which is only present in *C. roseus* roots (Shanks et al., 1998) and for which we were unable to identify a commercial supplier or create our own standard. If this is, in fact, 19-hydroxytabersonine, a significant increase is interesting, though not necessarily a desirable, result as it channels tabersonine to hörhammericine, echitovenine, or minovincine—none of which are clinically used (Shanks et al., 1998).

Similarly to observations in previous studies (Aerts et al., 1994; El-Sayed and Verpoorte, 2004; Jaleel et al., 2009; Pan et al., 2010; Wang et al., 2016; Zhang et al., 2018), treatment of both LBE and SSA seedlings with either methyl jasmonate or ethephon induced the production of various precursor alkaloids. These hormones also affected the expression levels of several key biosynthetic enzymes and transcription factors. Previous studies concluded that ethylene induces the MVA pathway while jasmonate induces the MEP in older seedlings (Pan et al., 2018; Zhang et al., 2018). It

is interesting to note that our results for key enzymes from these pathways (DXS2 in MEP and HMGS in MVA) in roots and shoots treated through foliar application of the phytohormones show different induction patterns than were observed in (Pan et al., 2018; Zhang et al., 2018), where treatments were applied to entire seedlings via hydroponic supplementation. A different physiological outcome in roots and shoots may be expected given that the immediate uptake happens through a different tissue; in some aspects, however, our results are not directly comparable, as outcome specific to the individual plant parts was not investigated in these past studies. Regardless, both DXS2 and HMGS are key enzymes in the formation of the indole component of terpene indole alkaloids, so a change in transcript abundance due to phytohormone treatment could have downstream effects on concentration of each alkaloid.

Examining varietal differences in the roles of master regulators

The ORCA family of transcription factors have been documented as central regulators of early stage TIA intermediate production in *C. roseus* (Liu et al., 2011; Pan et al., 2012; Li et al., 2013). Previous studies in hairy root culture have shown that the overexpression of ORCA2 significantly increases concentrations of catharanthine and vindoline levels while decreasing tabersonine levels (Liu et al., 2011; Li et al., 2013), but our results do not appear to have the same correlations. These results underscore the need for broader investigation in different varieties and the care that must be taken when extrapolating results from one variety of *C. roseus* to inform results or pathway engineering of another variety. As ORCA3 positively regulates two key genes in the TIA pathway (Pan et al., 2012), the significant difference between its expression in SSA and LBE makes its promoter an interesting target for further investigation and future bioengineering efforts. Our study also supports ORCA2 as a candidate for engineering, as its transcript levels increased upon treatment. As with any master regulator, however, there is the possibility of the

generation of off-target effects such as activation of potential repressors, so a careful investigation into genes controlled by these two TFs would be necessary.

Linking transcriptional changes to metabolite production

Additionally, we selected seven biosynthetic genes that encode key enzymes directly related to the biosynthesis of terpene indole alkaloids. Of these genes, five perform important reactions in the path toward vinblastine; the remaining two genes are involved in reactions that branch off from the vinblastine biosynthesis pathway but catalyze the formation of other medicinally relevant alkaloids (e.g. reserpine, etc.). Given the evidence from past studies that there are important pathway differences between roots and shoots, we felt that it was necessary to investigate the expression of these genes in these separate plant organs, as this information will be useful for engineering alkaloid production in biopharmaceutical settings. The observed results are intriguing.

Catharanthine synthase (CS) and tabersonine synthase (TS) produce catharanthine and tabersonine, respectively, and were recently determined to be two of the four missing enzymes in the TIA pathway (Caputi et al., 2018; Qu et al., 2018). We were particularly interested in discovering how the various phytohormone treatments affected their transcription, since relatively little research has been published on these genes since their discovery in the last two years. In shoots, the changes in the concentration of catharanthine are consistent with the observed changes in CS transcript number, suggesting that this particular step in the TIA pathway is transcriptionally regulated; this does not, however, appear to be the case in roots. Tabersonine production, meanwhile, does not appear to be transcriptionally regulated, as observed changes in TS transcript levels do not correspond with the changes in the alkaloid concentrations. One explanation for this behavior could be that an enzyme directly upstream acts as a bottleneck in the pathway, while the amount of TS present remains consistent because it is expressed at a level that is sufficient to handle an

increased amount of substrate. Alternatively, post-transcriptional or translational changes caused by the hormone treatments could be responsible for the observed increases in tabersonine concentration.

The changes in vindoline and the associated biosynthetic enzyme DAT are puzzling. SSA shoots treated with MeJA have similar increases in DAT mRNA and vindoline (Welch's t-test, $p \leq 0.1$), which is consistent with the changes in vindoline concentration observed in previous studies of *C. roseus* plants over-expressing DAT (Wang et al., 2012). In SSA, shoots treated with ethephon, however, the concentration of the alkaloid increased dramatically upon induction even though the number of DAT transcripts decrease. Perhaps the ethephon caused a post-translational modification that increased the efficiency of the enzyme (Chen and Bleecker, 1995). Further investigation into this response is needed. Although α -3',4'-anhydrovinblastine and vinblastine levels were below the detection limit of our mass spectrometer, they are still key alkaloids, which is why we chose to examine PRX1. Previous work in cell culture demonstrated a correlation between the over-expression of PRX1, an increase in the number of SGD transcripts, and an increase in ajmalicine accumulation (Jaggi et al., 2011). The results observed for these genes and for the ajmalicine/tetrahydroalstonine peak in SSA are consistent with these patterns. Re-examination of the alkaloid extracts with a higher-sensitivity mass spectrometer would allow us to examine the changes in α -3',4'-anhydrovinblastine, vinblastine, and vincristine concentrations caused by the hormone treatments and how they relate to the increases observed in PRX1.

Overall, production of many key TIAs appear to be transcriptionally regulated in at least one tissue. ETPN and MeJA induce approximately equal numbers of the biosynthetic genes that code for key enzymes in the TIA pathway. They also induce genes upstream of the TIA pathway, which may be useful information for future bioengineering attempts. When taken in conjunction with the changes observed in

alkaloid concentrations and with consideration of the cost of large-scale application, ETPN appears to be a viable option with considerable potential for alkaloid production in a biopharmaceutical setting.

Conclusions

In summary, our work demonstrates that choice of *C. roseus* variety, phytohormone type, and treatment concentration all have an impact on the levels of key alkaloids in each plant organ. Either a genomic or transcriptomic resource is available for the two varieties investigated here, but neither variety has both. The differing baseline metabolic profile as well as the differing responses to phytohormone treatment emphasize the importance of choosing an appropriate variety for one's desired outcomes. Additionally, optimization of treatments is crucial; timing of phytohormone application and harvest, as well as the concentration applied, can have significant effects on both the health of the plants and the induced changes in alkaloid concentrations. Finally, this study suggests that ethephon is a viable and agriculturally relevant induction agent for key alkaloids in a large-scale biopharmaceutical production setting.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Foundation of Oregon Award #2018-1954 to MM and Research Startup Funds from the College of Pharmacy to BP. We also wish to thank the Oregon State University Mass Spectrometry Facility (OSUMSC) for assistance.

Chapter Four

Conclusions

Valerie N. Fraser

In this dissertation, I have presented two distinct investigations into gene expression. First, we developed a series of machine learning models and their associated analyses for tissue specific gene expression prediction in *Arabidopsis thaliana*. Both of the models that were built for this study performed remarkably well with auROC values above 90%; that said, however, our tissue of expression (TEP) model that utilized computationally determined regions of enrichments (ROEs) for each transcription factor binding site (TFBS) performed marginally better. When compared to other relevant studies (Natarajan et al., 2012; Huminiecki and Horbańczuk, 2017; Agarwal and Shendure, 2020; N'Diaye et al., 2020), our models achieve the strongest performance outcomes for tissue specific gene expression. Additionally, we found that the majority of the top features being used by the TEP models to accurately predict tissue of expression is sequence information. The addition of a general openness feature alone to the TFBS sequence information increased accuracy of prediction nearly as much as the inclusion of specific chromatin accessibility features. This suggests that, though chromatin features do inform the models' tissue of expression prediction capabilities, we are mostly capturing the open state of promoters of actively transcribed genes.

This work also demonstrated that there are promoters that can be classified as “hard-coded”, meaning that the associated tissue of expression for those promoters appears to be almost solely dictated by TFBS patterns rather than chromatin accessibility features. The transcripts associated with these promoters are enriched for GO terms associated with metabolite biosynthesis and transport. Taking this idea that the patterns of TFBSs present in a promoter are primarily responsible for tissue of expression and that the ROEs generated for our models represent likely locations of functional TFBSs, we developed a system for “knocking down” features to evaluate the potential of a gene switching its tissue of expression. The removal of a single feature flips the tissue of expression prediction across the decision boundary for hundreds of transcripts; removal of two features (in the vein of a double-knockout),

causes thousands of transcripts to switch. This system has significant implications for synthetic biology; a recent study (Cai et al., 2020) supports the hypothesis that the specific positional arrangements of binding sites within a promoter can have a dramatic effect on overall expression level. With the use of ROEs and our *in silico* knockout system, tissue-specific synthetic promoters can be systematically constructed using rational combinations of endogenous cis-regulatory sites for targeted, tissue-specific gene expression.

In the second part of my dissertation, I described a study of the metabolomic and transcript level analyses for the production of vinblastine and its intermediates in *Catharanthus roseus*. We found that hormonal treatment, choice of variety, and tissue type all can have significant impact on the production of monoterpene indole alkaloids and on the expression of both master regulators and key biosynthetic enzymes. Our work truly underscores the need for careful consideration when selecting a plant variety to work in, as the natural variation we observed between “Sunstorm Apricot” (SSA) and “Little Bright Eye” (LBE) was non-trivial and, in some cases, statistically significant.

I selected ORCA family TFs as our master regulators of interest, as they are well studied (Liu et al., 2011; Pan et al., 2012; Li et al., 2013). We observed a significant difference in ORCA3 transcript levels between SSA and LBE, both in the control group and after treatment. Additionally, there is a significant difference in expression level between roots and shoots in LBE. Very little research has been done on the determinants of tissue specific gene expression in *C. roseus*; in fact, only one recent paper has been released on this particular topic (Duge de Bernonville et al., 2020). This study found that tissue-specific DNA methylation generally correlates with tissue specific gene expression for some of the genes in the monoterpene indole alkaloid biosynthetic pathway, as well as for some of the transcription factors known to regulate alkaloid production. Interestingly, the genes that code for ORCA2 and

ORCA3 were more highly expressed in roots than in leaves in this study; my results, however, were the exact opposite. Regardless of these conflicting outcomes, ORCA3 positively regulates two key genes in the TIA pathway (Pan et al., 2012) and appears to exhibit tissue specific expression, which makes its promoter an interesting target for further investigation and future bioengineering efforts.

My original concept for this dissertation was to start with machine learning models in a model eudicot, move into the medicinal plant world, and then apply what I learned from the modeling project to the transcriptional control of natural product biosynthesis pathways. Investigating transcriptional regulation of specialized gene expression is the thread that runs through all of my work. The production of vinblastine in *C. roseus* is a perfect example of how modeling of tissue specific gene expression and natural products research can align; the gene for at least one master regulator and multiple biosynthetic enzymes in the TIA biosynthesis pathway are transcribed at significantly higher level in the leaves of the plant.

Future directions for investigating transcriptional regulation of specialized metabolism

Future directions for this research involve applying the tissue of expression prediction machine learning model to natural product pathway elucidation in *C. roseus*. Datasets similar to those featured in the TEP modeling project should be generated for *C. roseus*, ideally for both of the varieties we used in our natural products research. TSS-seq of SSA leaves is currently underway, with plans to expand into SSA roots and both tissues of LBE. We have switched from nanoCAGE-XL to the STRIPE-Seq protocol (Policastro et al., 2020) for the library prep and will likely continue with it moving forward. Eventually, we would like to generate TSS-seq data and OC-seq data for the root and shoot tissues of both varieties that have been treated with ethephon and methyl jasmonate, as in Chapter 3. For the chromatin accessibility data, the lab is looking to switch to ATAC-seq (Bajic et al., 2018), due to the fact that sites

designated as “closed” by DNase I SIM can potentially be obscured by proteins other than nucleosomes, such as transcription factors or even polymerases.

The genome of LBE also needs to be sequenced; our lab has proposed to generate PacBio long read sequencing data (Eid et al., 2009) for both SSA and LBE to address this gap. This data will be used in conjunction with the TSS-seq data to precisely locate promoter regions. The sequences for the promoters upstream of TIA pathway genes and known TFs can be extracted and TFBS composition and promoter architecture comparisons can be made between the two varieties. Of particular interest will be the differences between TFBS patterns in the promoters of the genes that displayed tissue specific gene expression, as well as whether there are significant differences for genes associated with our observed changes in alkaloid concentrations. Additionally, an investigation could be launched into the possible enrichment of ethylene and/or jasmonate responsive elements in the promoters for genes upstream of precursor alkaloids that experienced a significant change in concentration due to hormonal treatment.

At the bench, the lab could attempt to remove some of the bottlenecks in the vinca alkaloid biosynthesis pathway. Using the *C. roseus* TSS-seq data and the feature weights from an appropriately trained TEP model, it would be easy to apply the “*in silico* knockout” framework from Chapter 2 to the promoter sequences upstream of the genes responsible for the conversion of tabersonine to vindoline. Other similar knockouts can be done on the promoters of master regulators. From there, targeted synthetic promoters can be rationally designed using combinations of the TFBSs that caused major shifts across the tissue of expression prediction decision boundary. These synthetic promoters can be cloned into a plasmid upstream of the gene for a reporter molecule such as a fluorescent protein or GUS, and the entire vector can be transformed into seedlings using the EASI protocol for transient expression (Mortensen et al., 2019).

In addition to the projects mentioned above, a model could be constructed to predict whether a gene is involved in a natural product pathway using TFBS patterns in promoter regions. In brief, the model would be trained to discriminate between groups of genes that are members of the pathway and those that are not. There is substantial evidence in the literature that TFs tend to be heavily involved in the regulatory cascades of stress pathways (Singh et al., 2002; von Koskull-Doring et al., 2007; Ohama et al., 2016), therefore it is a reasonable expectation that pathway-specific patterns of TFs are very likely to be present. Our lab did a preliminary trial of this kind of modeling to predict terpene synthases in *Arabidopsis*. The model used curated subsets of the root TSS-seq and RNA-seq data from Chapter 2 and performed well, having an auROC around 0.6, though the list of genes for training and testing was smaller than we would have liked. As far as medicinal plants go, *C. roseus* would be an ideal organism to use as a proof of concept for this particular research project; the TIA biosynthetic pathway is well studied and the final genes have recently been identified (Caputi et al., 2018; Qu et al., 2018). Once the patterns of cis-regulatory elements that distinguish biosynthetic genes have been determined, this process could be applied to any of the numerous other medicinal plants with incomplete pathways—provided the appropriate data can be generated.

LITERATURE CITED

- Aerts, R.J., Gisi, D., Carolis, E., De Luca, V., and Baumann, T.W.** (1994). Methyl jasmonate vapor increases the developmentally controlled synthesis of alkaloids in *Catharanthus* and *Cinchona* seedlings. *Plant J* **5**, 635-643.
- Agarwal, V., and Shendure, J.** (2020). Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Reports* **31**, 107663.
- Andersson, R.** (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37**, 314-323.
- Backstrom, S., Elfving, N., Nilsson, R., Wingsle, G., and Bjorklund, S.** (2007). Purification of a plant mediator from *Arabidopsis thaliana* identifies PFT1 as the Med25 subunit. *Mol Cell* **26**, 717-729.
- Bajic, M., Maher, K.A., and Deal, R.B.** (2018). Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq. *Methods Mol Biol* **1675**, 183-201.
- Balunas, M.J., and Kinghorn, A.D.** (2005). Drug discovery from medicinal plants. *Life Sci* **78**, 431-441.
- Besseau, S., Kellner, F., Lanoue, A., Thamm, A.M., Salim, V., Schneider, B., Geu-Flores, F., Hofer, R., Guirimand, G., Guihur, A., Oudin, A., Glevarec, G., Foureau, E., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., Werck-Reichhart, D., Burlat, V., De Luca, V., O'Connor, S.E., and Courdavault, V.** (2013). A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. *Plant Physiol* **163**, 1792-1803.
- Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A., Schiltz, R.L., Miranda, T.B., Sung, M.H., Trump, S., Lightman, S.L., Vinson, C., Stamatoyannopoulos, J.A., and Hager, G.L.** (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* **43**, 145-155.
- Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S.** (2008). F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537-2538.
- Breiling, A., Turner, B.M., Bianchi, M.E., and Orlando, V.** (2001). General transcription factors bind promoters repressed by Polycomb group proteins. *Nature* **412**, 651-655.
- Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A.** (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36**, D102-106.
- Cai, Y.M., Kallam, K., Tidd, H., Gendarini, G., Salzman, A., and Patron, N.J.** (2020). Rational design of minimal synthetic promoters for plants. *Nucleic Acids Res.*
- Caputi, L., Franke, J., Farrow, S.C., Chung, K., Payne, R.M.E., Nguyen, T.D., Dang, T.T., Soares Teto Carqueijeiro, I., Koudounas, K., Duge de Bernonville, T., Ameyaw, B., Jones, D.M., Vieira, I.J.C., Courdavault, V., and O'Connor, S.E.** (2018). Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* **360**, 1235-1239.

- Charron, J.B., He, H., Elling, A.A., and Deng, X.W.** (2009). Dynamic landscapes of four histone modifications during deetiolation in *Arabidopsis*. *Plant Cell* **21**, 3732-3748.
- Chen, Q.H.G., and Bleecker, A.B.** (1995). Analysis of Ethylene Signal-Transduction Kinetics Associated with Seedling-Growth Response and Chitinase Induction in Wild-Type and Mutant *Arabidopsis*. *Plant Physiol.* **108**, 597-607.
- Cheng, C., Yan, K.K., Yip, K.Y., Rozowsky, J., Alexander, R., Shou, C., and Gerstein, M.** (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15.
- Chereji, R.V., Eriksson, P.R., Ocampo, J., Prajapati, H.K., and Clark, D.J.** (2019). Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. *Genome research* **29**, 1985-1995.
- Chung, I.M., Kim, E.H., Li, M., Peebles, C.A., Jung, W.S., Song, H.K., Ahn, J.K., and San, K.Y.** (2011). Screening 64 cultivars *Catharanthus roseus* for the production of vindoline, catharanthine, and serpentine. *Biotechnol Prog* **27**, 937-943.
- Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S.** (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* **9**, 279-289.
- Colinas, M., and Goossens, A.** (2018). Combinatorial Transcriptional Control of Plant Specialized Metabolism. *Trends Plant Sci* **23**, 324-336.
- Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A., and Jaenisch, R.** (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936.
- Cumbie, J.S., Filichkin, S.A., and Megraw, M.** (2015a). Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in *Arabidopsis thaliana*. *Plant Methods* **11**, 42.
- Cumbie, J.S., Ivanchenko, M.G., and Megraw, M.** (2015b). NanoCAGE-XL and CapFilter: an approach to genome wide identification of high confidence transcription start sites. *BMC Genomics* **16**, 597.
- Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E.** (2003). AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinformatics* **4**, 25.
- Dellino, G.I., Schwartz, Y.B., Farkas, G., McCabe, D., Elgin, S.C., and Pirrotta, V.** (2004). Polycomb silencing blocks transcription initiation. *Mol Cell* **13**, 887-893.
- Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigó, R., Birney, E., and Weng, Z.** (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology* **13**, R53.
- Duge de Bernonville, T., Maury, S., Delaunay, A., Daviaud, C., Chaparro, C., Tost, J., O'Connor, S.E., and Courdavault, V.** (2020). Developmental Methylome of the Medicinal Plant *Catharanthus roseus* Unravels the Tissue-Specific Control of the Monoterpene Indole Alkaloid Pathway by DNA Methylation. *Int J Mol Sci* **21**.

- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.
- El-Sayed, M., and Verpoorte, R. (2004). Growth, metabolic profiling and enzymes activities of *Catharanthus roseus* seedlings treated with plant growth regulators. *Plant Growth Regul* **44**, 53-58.
- Ernst, J., Kheradpour, P., Mikkelson, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B.E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49.
- Feuerborn, A., and Cook, P.R. (2015). Why the activity of a gene depends on its neighbors. *Trends in Genetics* **31**, 483-490.
- Filichkin, S.A., and Megraw, M. DNase I SIM: A Simplified In-Nucleus Method for DNase I Hypersensitive Site Sequencing.
- Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., Andersson, R., Mungall, C.J., Meehan, T.F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y.A., Plessy, C., Vitezic, M., Severin, J., Semple, C.A., Ishizu, Y., Young, R.S., Francescato, M., Alam, I., Albanese, D., Altschuler, G.M., Arakawa, T., Archer, J.A.C., Arner, P., Babina, M., Rennie, S., Balwiercz, P.J., Beckhouse, A.G., Pradhan-Bhatt, S., Blake, J.A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Maxwell Burroughs, A., Califano, A., Cannistraci, C.V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H.C., Dalla, E., Davis, C.A., Detmar, M., Diehl, A.D., Dohi, T., Drabløs, F., Edge, A.S.B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M.C., Faulkner, G.J., Favorov, A.V., Fisher, M.E., Frith, M.C., Fujita, R., Fukuda, S., Furlanello, C., Furuno, M., Furusawa, J.-i., Geijtenbeek, T.B., Gibson, A.P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K.J., Ho Sui, S.J., Hofmann, O.M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B.R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A.S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y.I., Kawashima, T., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Peter Klinken, S., Knox, A.J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A.T., Laros, J.F.J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L.,

- Mackay-sim, A., Manabe, R.-i., Mar, J.C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D.A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C.L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Nozaki, T., Ogishima, S., Ohkura, N., Ohmiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J.G.D., Rackham, O.J.L., Ramilowski, J.A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M.B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E.A., Schulze-Tanzil, G.G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J.W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R.K., 't Hoen, P.A.C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyoda, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L.M., Verardo, R., Vijayan, D., Vorontsov, I.E., Wasserman, W.W., Watanabe, S., Wells, C.A., Winteringham, L.N., Wolvetang, E., Wood, E.J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S.E., Zhang, P.G., Zhao, X., Zucchelli, S., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D.A., Carninci, P., Hayashizaki, Y., The, F.C., the, R.P., and Clst. (2014). A promoter-level mammalian expression atlas. *Nature* **507**, 462-470.
- Góngora-Castillo, E., Childs, K.L., Fedewa, G., Hamilton, J.P., Liscombe, D.K., Magallanes-Lundback, M., Mandadi, K.K., Nims, E., Runguphan, W., Vaillancourt, B., Varbanova-Herde, M., DellaPenna, D., McKnight, T.D., O'Connor, S., and Buell, C.R. (2012). Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLOS ONE* **7**, e52506.
- Guertin, M.J., and Lis, J.T. (2010). Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet* **6**, e1001114.
- Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl, S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J.A., Ecker, J.R., and Ren, B. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-491.
- Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**, 311-318.
- Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-728.

- Hoopes, G.M., Hamilton, J.P., Wood, J.C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N.J., and Buell, C.R.** (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *Plant J* **97**, 1154-1167.
- Huang, J., Zheng, J., Yuan, H., and McGinnis, K.** (2018). Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. *BMC Plant Biol* **18**, 111.
- Huminiecki, Ł., and Horbańczuk, J.** (2017). Can We Predict Gene Expression by Understanding Proximal Promoter Architecture? *Trends in Biotechnology* **35**, 530-546.
- Ibrahim, M.M., Lacadie, S.A., and Ohler, U.** (2015). JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics* **31**, 48-55.
- International Rice Genome Sequencing, P.** (2005). The map-based sequence of the rice genome. *Nature* **436**, 793-800.
- Isah, T., Umar, S., Mujib, A., Sharma, M.P., Rajasekharan, P.E., Zafar, N., and Fruk, A.** (2018). Secondary metabolism of pharmaceuticals in the plant in vitro cultures: strategies, approaches, and limitations to achieving higher yield. *Plant Cell Tiss Org* **132**, 239-265.
- Ishikawa, H., Colby, D.A., and Boger, D.L.** (2008). Direct coupling of catharanthine and vindoline to provide vinblastine: total synthesis of (+)- and ent-(-)-vinblastine. *J Am Chem Soc* **130**, 420-421.
- Jaggi, M., Kumar, S., and Sinha, A.K.** (2011). Overexpression of an apoplastic peroxidase gene CrPrx in transgenic hairy root lines of *Catharanthus roseus*. *Appl Microbiol Biotechnol* **90**, 1005-1016.
- Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., and Zapata, F.** (2005). Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comp Funct Genomics* **6**, 388-397.
- Jaleel, C.A., Gopi, R., Gomathinayagam, M., and Panneerselvam, R.** (2009). Traditional and non-traditional plant growth regulators alters phytochemical constituents in *Catharanthus roseus*. *Process Biochem* **44**, 205-209.
- John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A.** (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**, 264-268.
- Johnson, I.S., Armstrong, J.G., Gorman, M., and Burnett, J.P., Jr.** (1963). The Vinca Alkaloids: A New Class of Oncolytic Agents. *Cancer Res* **23**, 1390-1427.
- Juven-Gershon, T., and Kadonaga, J.T.** (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **339**, 225-229.
- Kadonaga, J.T.** (2004). Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell* **116**, 247-257.
- Kadonaga, J.T.** (2012). Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**, 40-51.

- Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M.** (2010). Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-2931.
- Kellner, F., Kim, J., Clavijo, B.J., Hamilton, J.P., Childs, K.L., Vaillancourt, B., Cepela, J., Habermann, M., Steuernagel, B., Clissold, L., McLay, K., Buell, C.R., and O'Connor, S.E.** (2015). Genome-guided investigation of plant natural product biosynthesis. *Plant J* **82**, 680-692.
- Kim, S.W., Ban, S.H., Jeong, S.-C., Chung, H.-J., Ko, S.M., Yoo, O.J., and Liu, J.R.** (2007). Genetic discrimination between *Catharanthus roseus* cultivars by metabolic fingerprinting using ¹H NMR spectra of aromatic compounds. *Biotechnol Bioprocess Eng* **12**, 646.
- Kim, T.-K., and Shiekhhattar, R.** (2015). Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948-959.
- Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramirez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., and Tang, H.** (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Sci Rep* **8**, 10872.
- Ko, J.Y., Oh, S., and Yoo, K.H.** (2017). Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. *Molecules and cells* **40**, 169-177.
- Kumari, S., and Ware, D.** (2013). Genome-Wide Computational Prediction and Analysis of Core Promoter Elements across Plant Monocots and Dicots. *PLoS ONE* **8**, e79011.
- Laflamme, P., St-Pierre, B., and De Luca, V.** (2001). Molecular and biochemical analysis of a Madagascar periwinkle root-specific minovincinine-19-hydroxy-*O*-acetyltransferase. *Plant Physiol.* **125**, 189-198.
- Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A., and Huala, E.** (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**, D1202-1210.
- Langmead, B., and Salzberg, S.L.** (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359.
- Lee, L.R., Wengier, D.L., and Bergmann, D.C.** (2019). Cell-type-specific transcriptome and histone modification dynamics during cellular reprogramming in the Arabidopsis stomatal lineage. *Proc Natl Acad Sci U S A* **116**, 21914-21924.
- Leete, E.** (1967). Alkaloid Biosynthesis. *Annual Review of Plant Physiology* **18**, 179-196.
- Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., and Kendziorski, C.** (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035-1043.
- Leung, M.K., Xiong, H.Y., Lee, L.J., and Frey, B.J.** (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics* **30**, i121-129.
- Li, B., and Dewey, C.N.** (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* **12**.

- Li, C., Leopold, A.L., Sander, G.W., Shanks, J.V., Zhao, L., and Gibson, S.I.** (2013). The ORCA2 transcription factor plays a key role in regulation of the terpenoid indole alkaloid pathway. *BMC Plant Biol* **13**, 155.
- Li, Y., Liu, X., Chen, R., Tian, J., Fan, Y., and Zhou, X.** (2019). Genome-scale mining of root-preferential genes from maize and characterization of their promoter activity. *BMC Plant Biol* **19**, 584.
- Liu, D.H., Ren, W.W., Cui, L.J., Zhang, L.D., Sun, X.F., and Tang, K.X.** (2011). Enhanced accumulation of catharanthine and vindoline in *Catharanthus roseus* hairy roots by overexpression of transcriptional factor ORCA2. *Afr J Biotechnol* **10**, 3260-3268.
- Liu, Y., Patra, B., Pattanaik, S., Wang, Y., and Yuan, L.** (2019). GATA and Phytochrome Interacting Factor Transcription Factors Regulate Light-Induced Vindoline Biosynthesis in *Catharanthus roseus*. *Plant Physiol* **180**, 1336-1350.
- Loh, Y.H., Zhang, W., Chen, X., George, J., and Ng, H.H.** (2007). Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes Dev* **21**, 2545-2557.
- Lu, Z., Marand, A.P., Ricci, W.A., Ethridge, C.L., Zhang, X., and Schmitz, R.J.** (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants* **5**, 1250-1259.
- Magnotta, M., Murata, J., Chen, J., and De Luca, V.** (2006). Identification of a low vindoline accumulating cultivar of *Catharanthus roseus* (L.) G. Don by alkaloid and enzymatic profiling. *Phytochemistry* **67**, 1758-1764.
- Maher, K.A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D.A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M.W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S.M., and Deal, R.B.** (2018). Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. *The Plant Cell* **30**, 15.
- Malik, S., and Roeder, R.G.** (2000). Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. *Trends Biochem Sci* **25**, 277-283.
- Margueron, R., and Reinberg, D.** (2011). The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343-349.
- Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., and Hatzigeorgiou, A.G.** (2009). A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res* **19**, 644-656.
- Mejía-Guerra, M.K., Li, W., Galeano, N.F., Vidal, M., Gray, J., Doseff, A.I., and Grotewold, E.** (2015). Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites. *The Plant Cell* **27**, 3309.
- Money, T., Wright, I.G., McCapra, F., Hall, E.S., and Scott, A.I.** (1968). Biosynthesis of indole alkaloids. Vindoline. *Journal of the American Chemical Society* **90**, 4144-4150.
- Mortensen, S., Bernal-Franco, D., Cole, L.F., Sathitloetsakun, S., Cram, E.J., and Lee-Parsons, C.W.T.** (2019). EASI Transformation: An Efficient Transient Expression Method for Analyzing Gene Function in *Catharanthus roseus* Seedlings. *Front Plant Sci* **10**, 755.
- Morton, T., and Megraw, M.** (2014). 3PEAT TFBS-Scanner Toolset.

- Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A., Benfey, P.N., Ohler, U., and Megraw, M.** (2014). Paired-End Analysis of Transcription Start Sites in Arabidopsis Reveals Plant-Specific Promoter Signatures. *The Plant Cell* **26**, 2746-2760.
- Murata, J., and De Luca, V.** (2005). Localization of tabersonine 16-hydroxylase and 16-OH tabersonine-16-O-methyltransferase to leaf epidermal cells defines them as a major site of precursor biosynthesis in the vindoline pathway in *Catharanthus roseus*. *Plant J* **44**, 581-594.
- Myers, L.C., and Kornberg, R.D.** (2000). Mediator of transcriptional regulation. *Annu Rev Biochem* **69**, 729-749.
- N'Diaye, A., Byrns, B., Cory, A.T., Nilsen, K.T., Walkowiak, S., Sharpe, A., Robinson, S.J., and Pozniak, C.J.** (2020). Machine learning analyses of methylation profiles uncovers tissue-specific gene expression patterns in wheat. *Plant Genome* **13**, e20027.
- Nakayama, J., Rice, J.C., Strahl, B.D., Allis, C.D., and Grewal, S.I.** (2001). Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113.
- Nascimento, N.C.d., and Fett-Neto, A.G.** (2010). Plant Secondary Metabolism and Challenges in Modifying Its Operation: An Overview. In *Plant Secondary Metabolism Engineering: Methods and Applications*, A.G. Fett-Neto, ed (Totowa, NJ: Humana Press), pp. 1-13.
- Natarajan, A., Yardimci, G.G., Sheffield, N.C., Crawford, G.E., and Ohler, U.** (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome research* **22**, 1711-1722.
- Ni, T., Corcoran, D., Rach, E., Song, S., Spana, E., Gao, Y., Ohler, U., and Zhu, J.** (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature methods* **7**, 521-527.
- Noble, R.L., Beer, C.T., and Cutts, J.H.** (1958). Role of chance observations in chemotherapy: *Vinca rosea*. *Ann N Y Acad Sci* **76**, 882-894.
- O'Keefe, B.R., Mahady, G.B., Gills, J.J., Beecher, C.W.W., and Schilling, A.B.** (1997). Stable vindoline production in transformed cell cultures of *Catharanthus roseus*. *J Nat Prod* **60**, 261-264.
- Ohama, N., Kusakabe, K., Mizoi, J., Zhao, H., Kidokoro, S., Koizumi, S., Takahashi, F., Ishida, T., Yanagisawa, S., Shinozaki, K., and Yamaguchi-Shinozaki, K.** (2016). The Transcriptional Cascade in the Heat Stress Response of Arabidopsis Is Strictly Regulated at the Level of Transcription Factor Expression. *Plant Cell* **28**, 181-201.
- Ong, C.-T., and Corces, V.G.** (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12**, 283-293.
- Orphanides, G., Lagrange, T., and Reinberg, D.** (1996). The general transcription factors of RNA polymerase II. *Genes Dev* **10**, 2657-2683.
- Pajoro, A., Madrigal, P., Muiño, J.M., Matus, J.T., Jin, J., Mecchia, M.A., Debernardi, J.M., Palatnik, J.F., Balazadeh, S., Arif, M., Ó'Maoiléidigh, D.S., Wellmer, F., Krajewski, P., Riechmann, J.-L., Angenent, G.C., and Kaufmann, K.** (2014). Dynamics of

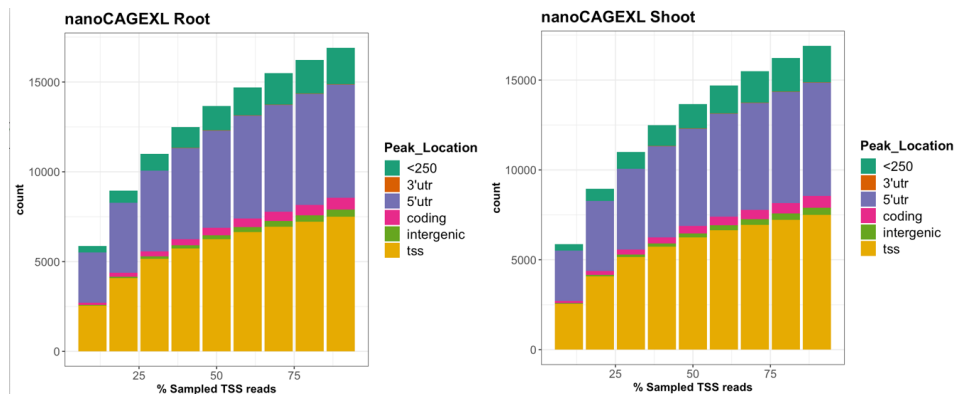
- chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biology* **15**, R41.
- Pan, Q., Chen, Y., Wang, Q., Yuan, F., Xing, S., Tian, Y., Zhao, J., Sun, X., and Tang, K.** (2010). Effect of plant growth regulators on the biosynthesis of vinblastine, vindoline and catharanthine in *Catharanthus roseus*. *Plant Growth Regul* **60**, 133-141.
- Pan, Q., Wang, Q., Yuan, F., Xing, S., Zhao, J., Choi, Y.H., Verpoorte, R., Tian, Y., Wang, G., and Tang, K.** (2012). Overexpression of ORCA3 and G10H in *Catharanthus roseus* plants regulated alkaloid biosynthesis and metabolism revealed by NMR-metabolomics. *PLOS ONE* **7**, e43038.
- Pan, Q.F., Mustafa, N.R., Tang, K.X., Choi, Y.H., and Verpoorte, R.** (2016). Monoterpenoid indole alkaloids biosynthesis and its regulation in *Catharanthus roseus*: a literature review from genes to metabolites. *Phytochem Rev* **15**, 221-250.
- Pan, Y.-j., Lin, Y.-c., Yu, B.-f., Zu, Y.-g., Yu, F., and Tang, Z.-H.** (2018). Transcriptomics comparison reveals the diversity of ethylene and methyl-jasmonate in roles of TIA metabolism in *Catharanthus roseus*. *BMC Genomics* **19**, 508.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.** (2011). Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830.
- Policastro, R.A., Raborn, R.T., Brendel, V.P., and Zentner, G.E.** (2020). Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq. *Genome Res* **30**, 910-923.
- Potier, P.** (1980). Synthesis of the antitumor dimeric indole alkaloids from *Catharanthus* species (Vinblastine Group). *J Nat Prod* **43**, 72-86.
- Qu, Y., Safonova, O., and De Luca, V.** (2018). Completion of the canonical pathway for assembly of anticancer drugs vincristine/vinblastine in *Catharanthus roseus*. *Plant J* **97**, 257-266.
- Quinlan, A.R., and Hall, I.M.** (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.
- Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., and Jenuwein, T.** (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593-599.
- Ricci, W.A., Lu, Z., Ji, L., Marand, A.P., Ethridge, C.L., Murphy, N.G., Noshay, J.M., Galli, M., Mejía-Guerra, M.K., Colomé-Tatché, M., Johannes, F., Rowley, M.J., Corces, V.G., Zhai, J., Scanlon, M.J., Buckler, E.S., Gallavotti, A., Springer, N.M., Schmitz, R.J., and Zhang, X.** (2019). Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants* **5**, 1237-1249.
- Rijhwani, S.K., and Shanks, J.V.** (1998). Effect of elicitor dosage and exposure time on biosynthesis of indole alkaloids by *Catharanthus roseus* hairy root cultures. *Biotechnol Prog* **14**, 442-449.

- Robertson, A.G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., Cezard, T., Fejes, A.P., Wederell, E.D., Cullum, R., Euskirchen, G., Krzywinski, M., Birol, I., Snyder, M., Hoodless, P.A., Hirst, M., Marra, M.A., and Jones, S.J.** (2008). Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* **18**, 1906-1917.
- Rodgers-Melnick, E., Vera, D.L., Bass, H.W., and Buckler, E.S.** (2016). Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3177-E3184.
- Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B.** (2013). Characterizing and measuring bias in sequence data. *Genome Biology* **14**, R51.
- Ruijter, J.M., Ramakers, C., Hoogaars, W.M., Karlen, Y., Bakker, O., van den Hoff, M.J., and Moorman, A.F.** (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res* **37**, e45.
- Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A.** (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**, 424-436.
- Shanks, J.V., Bhadra, R., Morgan, J., Rihwani, S., and Vani, S.** (1998). Quantification of metabolites in the indole alkaloid pathways of *Catharanthus roseus*: Implications for metabolic engineering. *Biotechnol Bioeng* **58**, 333-338.
- Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J.A., Lenhard, B., Crawford, G.E., and Furey, T.S.** (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research* **23**, 777-788.
- Singh, K., Foley, R.C., and Onate-Sanchez, L.** (2002). Transcription factors in plant defense and stress responses. *Curr Opin Plant Biol* **5**, 430-436.
- Singh, R., Lanchantin, J., Robins, G., and Qi, Y.** (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32**, i639-i648.
- Smale, S.T., and Kadonaga, J.T.** (2003). The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449-479.
- Smith, A.M., Walsh, J.R., Long, J., Davis, C.B., Henstock, P., Hodge, M.R., Maciejewski, M., Mu, X.J., Ra, S., Zhao, S., Ziemek, D., and Fisher, C.K.** (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics* **21**, 119.
- Snyder, Matthew W., Kircher, M., Hill, Andrew J., Daza, Riza M., and Shendure, J.** (2016). Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57-68.
- Sonawane, A.R., Platig, J., Fagny, M., Chen, C.Y., Paulson, J.N., Lopes-Ramos, C.M., DeMeo, D.L., Quackenbush, J., Glass, K., and Kuijjer, M.L.** (2017). Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**, 1077-1088.
- Soutourina, J., Wydau, S., Ambroise, Y., Boschiero, C., and Werner, M.** (2011). Direct interaction of RNA polymerase II and mediator required for transcription in vivo. *Science* **331**, 1451-1454.

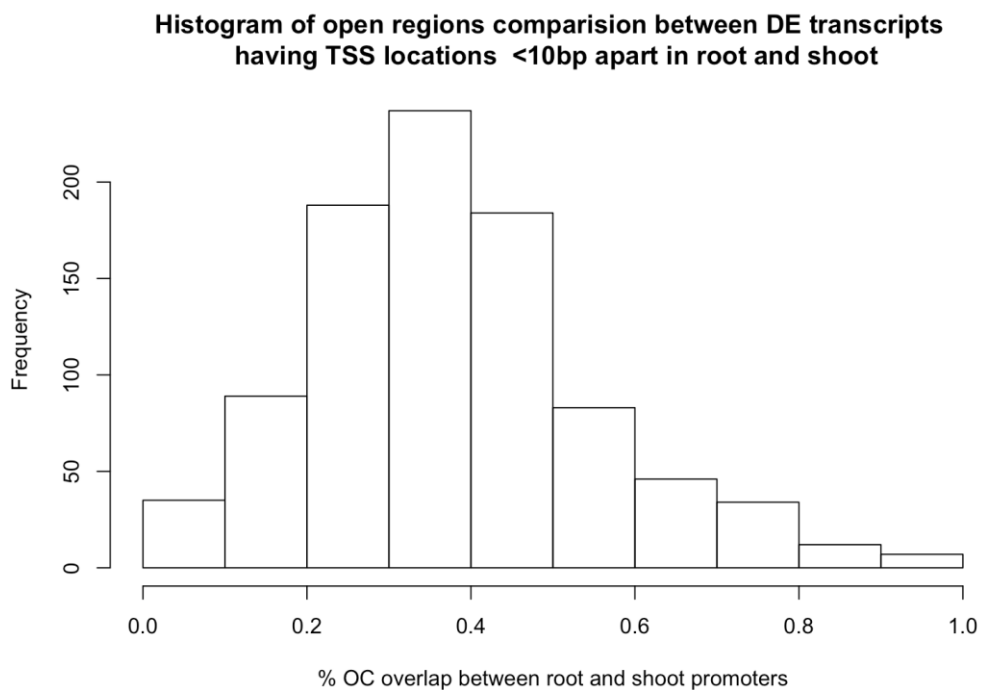
- Spitz, F., and Furlong, E.E.M.** (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**, 613-626.
- St-Pierre, B., Vazquez-Flota, F.A., and De Luca, V.** (1999). Multicellular compartmentation of catharanthus roseus alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. *Plant Cell* **11**, 887-900.
- Staiger, D., Becker, F., Schell, J., Koncz, C., and Palme, K.** (1991). Purification of tobacco nuclear proteins binding to a CACGTG motif of the chalcone synthase promoter by DNA affinity chromatography. *Eur J Biochem* **199**, 519-527.
- Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P., Stergachis, A.B., Vernet, B., Johnson, A.K., Haugen, E., Sullivan, S.T., Thompson, A., Neri, F.V., 3rd, Weaver, M., Diegel, M., Mnaimneh, S., Yang, A., Hughes, T.R., Nemhauser, J.L., Queitsch, C., and Stamatoyannopoulos, J.A.** (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep* **8**, 2015-2030.
- Taher, L., Smith, R.P., Kim, M.J., Ahituv, N., and Ovcharenko, I.** (2013). Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. *Genome biology* **14**, R117-R117.
- Thomas, M.C., and Chiang, C.-M.** (2006). The General Transcription Machinery and General Cofactors. *Critical Reviews in Biochemistry and Molecular Biology* **41**, 105-178.
- Tyler, V.E.** (1988). Medicinal plant research: 1953-1987. *Planta Med* **54**, 95-100.
- Van der Heijden, R., Verpoorte, R., and Tenhoopen, H.J.G.** (1989). Cell and Tissue-Cultures of *Catharanthus-Roseus* (L) Don,G. - a Literature Survey. *Plant Cell Tiss Org* **18**, 231-280.
- Vandenbon, A., and Nakai, K.** (2010). Modeling tissue-specific structural patterns in human and mouse promoters. *Nucleic acids research* **38**, 17-25.
- Vera, D.L., Madzima, T.F., Labonne, J.D., Alam, M.P., Hoffman, G.G., Girimurugan, S.B., Zhang, J., McGinnis, K.M., Dennis, J.H., and Bass, H.W.** (2014a). Differential Nuclease Sensitivity Profiling of Chromatin Reveals Biochemical Footprints Coupled to Gene Expression and Functional DNA Elements in Maize. *The Plant Cell* **26**, 3883.
- Vera, D.L., Madzima, T.F., Labonne, J.D., Alam, M.P., Hoffman, G.G., Girimurugan, S.B., Zhang, J., McGinnis, K.M., Dennis, J.H., and Bass, H.W.** (2014b). Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. *Plant Cell* **26**, 3883-3893.
- Verma, M., Ghangal, R., Sharma, R., Sinha, A.K., and Jain, M.** (2014). Transcriptome analysis of *Catharanthus roseus* for gene discovery and expression profiling. *PLOS ONE* **9**.
- Verpoorte, R., van der Heijden, R., and Moreno, P.R.H.** (1997). Chapter 3 Biosynthesis of Terpenoid Indole Alkaloids in *Catharanthus roseus* Cells, pp. 221-299.
- von Koskull-Doring, P., Scharf, K.D., and Nover, L.** (2007). The diversity of plant heat stress transcription factors. *Trends Plant Sci* **12**, 452-457.
- Wang, Q., Li, W., Liu, X.S., Carroll, J.S., Janne, O.A., Keeton, E.K., Chinnaiyan, A.M., Pienta, K.J., and Brown, M.** (2007). A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Mol Cell* **27**, 380-392.

- Wang, Q., Xing, S., Pan, Q., Yuan, F., Zhao, J., Tian, Y., Chen, Y., Wang, G., and Tang, K.** (2012). Development of efficient *Catharanthus roseus* regeneration and transformation system using *Agrobacterium tumefaciens* and hypocotyls as explants. *BMC Biotechnol* **12**, 34.
- Wang, X., Pan, Y.-J., Chang, B.-W., Hu, Y.-B., Guo, X.-R., and Tang, Z.-H.** (2016). Ethylene-induced vinblastine accumulation is related to activated expression of downstream TIA pathway genes in *Catharanthus roseus*. *Biomed Res Int* **2016**.
- Weirauch, Matthew T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, Hamed S., Lambert, Samuel A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M.G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, John S., Govindarajan, S., Shaulsky, G., Walhout, Albertha J.M., Bouget, F.-Y., Ratsch, G., Larrondo, Luis F., Ecker, Joseph R., and Hughes, Timothy R.** (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431-1443.
- Wilken, M.S., Brzezinski, J.A., La Torre, A., Siebenthal, K., Thurman, R., Sabo, P., Sandstrom, R.S., Vierstra, J., Canfield, T.K., Hansen, R.S., Bender, M.A., Stamatoyannopoulos, J., and Reh, T.A.** (2015). DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenetics & Chromatin* **8**, 8.
- Williams, G.R., and Doran, P.M.** (2000). Hairy root culture in a liquid-dispersed bioreactor: characterization of spatial heterogeneity. *Biotechnol Prog* **16**, 391-401.
- Wingender, E.** (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics* **9**, 326-332.
- Xi, H., Shulha, H.P., Lin, J.M., Vales, T.R., Fu, Y., Bodine, D.M., McKay, R.D.G., Chenoweth, J.G., Tesar, P.J., Furey, T.S., Ren, B., Weng, Z., and Crawford, G.E.** (2007). Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS genetics* **3**, e136-e136.
- Yamamoto, Y.Y., Yoshioka, Y., Hyakumachi, M., and Obokata, J.** (2011). Characteristics of Core Promoter Types with respect to Gene Structure and Expression in *Arabidopsis thaliana*. *DNA Research* **18**, 333-342.
- Zavolan, M.** (2015). Inferring gene expression regulatory networks from high-throughput measurements. *Methods* **85**, 1-2.
- Zhang, W., Zhang, T., Wu, Y., and Jiang, J.** (2012a). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *Plant Cell* **24**, 2719-2731.
- Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J.** (2012b). High-resolution mapping of open chromatin in the rice genome. *Genome Res* **22**, 151-162.
- Zhang, X.-N., Liu, J., Liu, Y., Wang, Y., Abozeid, A., Yu, Z.-G., and Tang, Z.-H.** (2018). Metabolomics analysis reveals that ethylene and methyl jasmonate regulate different branch pathways to promote the accumulation of terpenoid indole alkaloids in *Catharanthus roseus*. *J Nat Prod* **81**, 335-342.

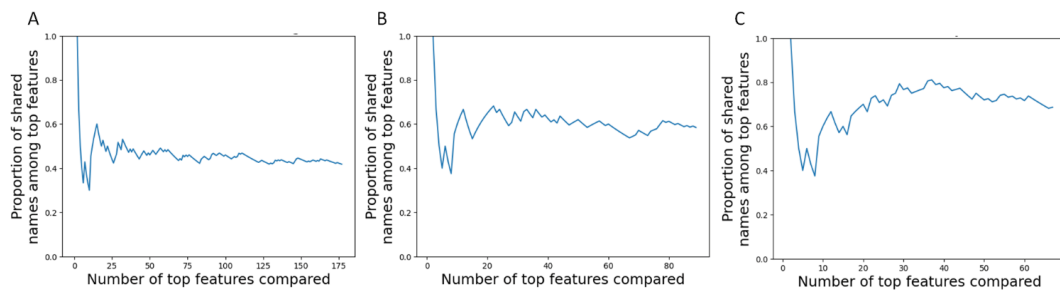
Appendix A: Supplementary Materials for Chapter 2



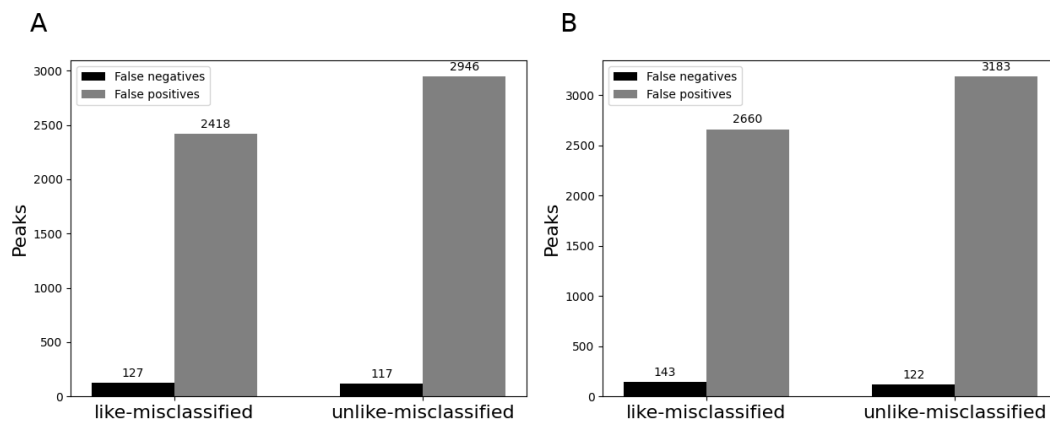
Supplementary Figure A1: Sequencing depth analysis of nanoCAGE-XL root and shoot data to determine whether sequencing depth accurately represented gene expression in our pooled root and shoot samples. Starting with stringently-mapping reads, reads were randomly sampled and peaks called according to our TSS peak calling procedure. As subsample size increases, fewer new genes are observed in each dataset. In each dataset, the change in number of genes has leveled off to nearly zero as the percentage of reads sampled increases from 90% to 100%. This indicates that the sampling depth we achieved in terms of gene coverage was within a few percent of the maximum sampling depth that could be achieved.



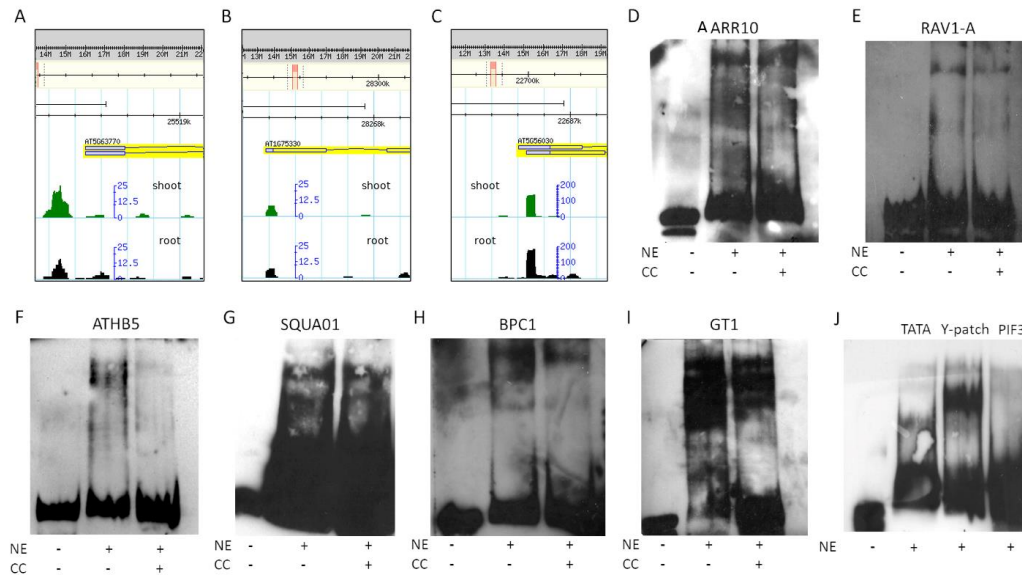
Supplementary Figure A2: Histogram showing the percent overlap in chromatin openness between the differentially expressed root and shoot transcripts with mapped TSS modes in close proximity (less than 10 nt apart). Percent overlap is computed as the percentage of nucleotides in the region surrounding each TSS [TSS - 3 kb, TSS + 3 kb] that agree in chromatin accessibility state (open vs closed).



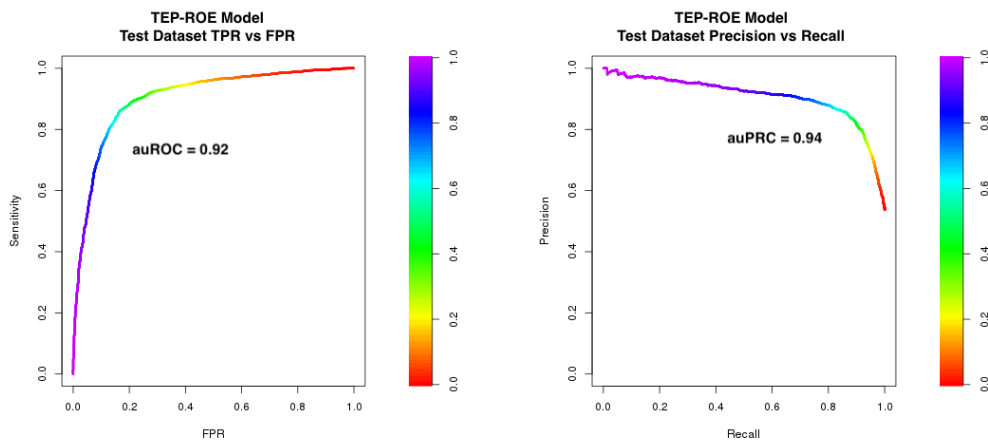
Supplementary Figure A3: Proportion of shared names in the top N features of both root and shoot 3PEAT models, calculated as $P = \frac{|TopN_{root} \cup TopN_{shoot}|}{N}$, i.e., the number of shared names among the top N names in both lists divided by N. A) Feature names including PWM (Positional Weight Matrix), strand and TSS. B) Feature names including PWM and strand. C) Feature names including PWM only.



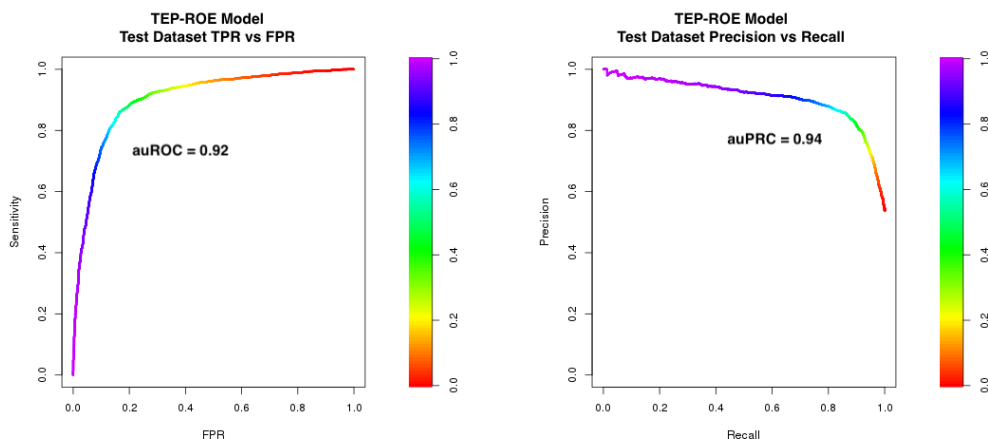
Supplementary Figure 4: The “like-misclassified” peaks are peaks that were misclassified by the model trained on the same tissue that the peaks were expressed in (root or shoot). The “unlike-misclassified” peaks were misclassified by the model trained on the other tissue. A) Peaks misclassified from the root test-set. B) Peaks misclassified from the shoot test-set model.

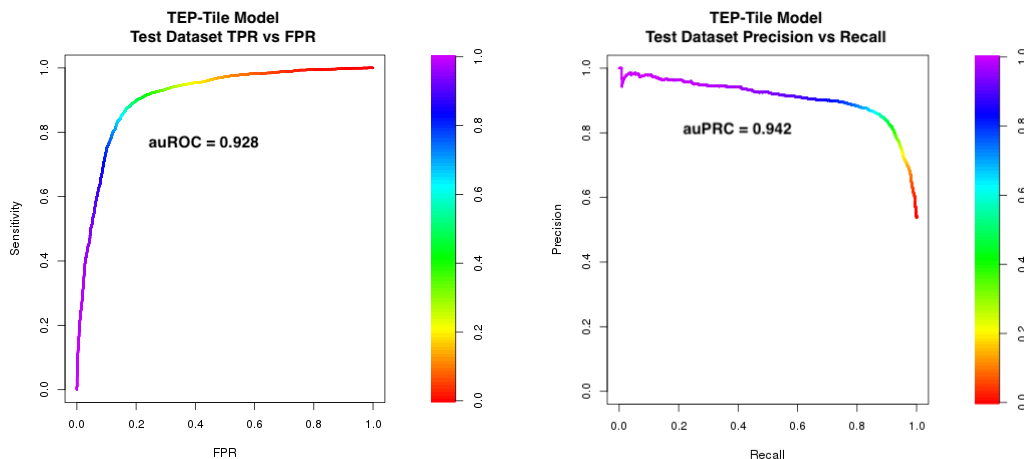


Supplementary Figure A5: A) GBrowse screenshot highlighting the small TSS-Seq peaks for DGK2. B) GBrowse screenshot highlighting the small TSS-Seq peaks for OTC. C) GBrowse screenshot of the large TSS-Seq peaks for HSP90.2, illustrating why we selected our second set of sites for testing from the promoter of this gene. D-I) EMSAs testing selected sites from the HSP90.2 promoter. NE = Nuclear Extract from 7-day old *Arabidopsis thaliana* Col0 roots; CC = cold competitor, >200X. J) EMSA showing binding of three sites. TATA from the HSP90.2 promoter and Y-Patch from the OTC promoter show a shift resulting from binding. PIF3 from the DGK2 promoter does not. The RNA-Seq abundance for PIF3 in the root sample was 0.1 TPM; TBP, however, had 45 TPM. No cold competitor was used for this particular gel.

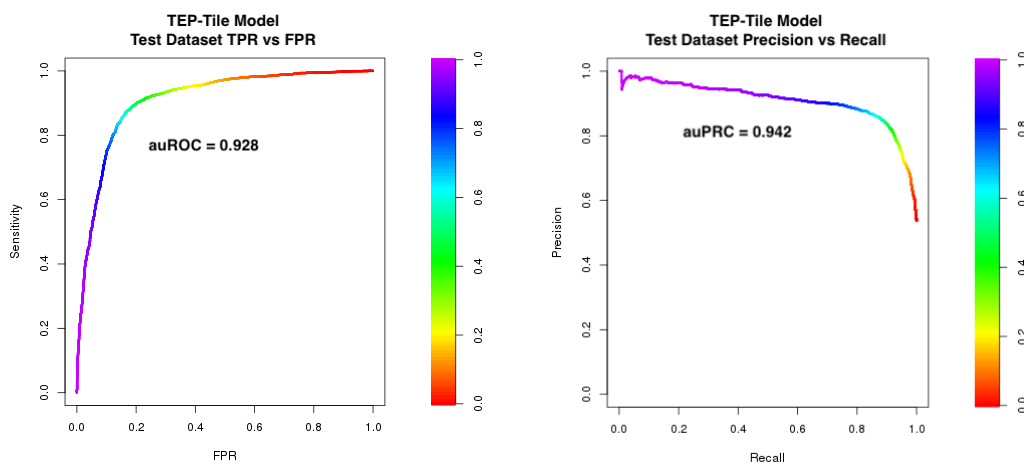


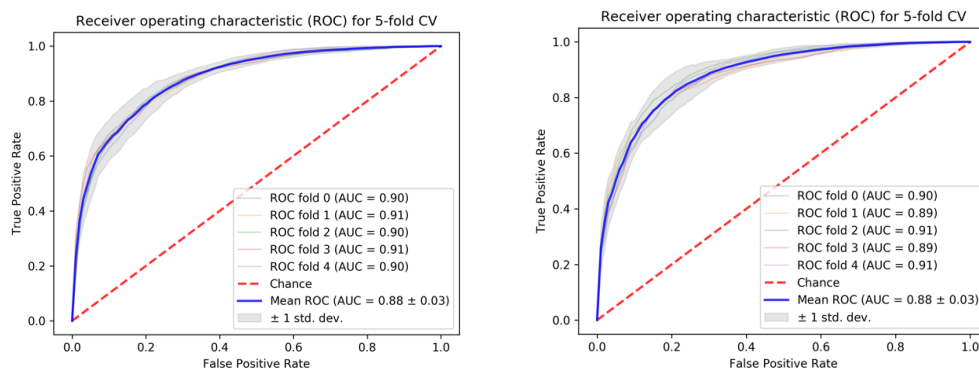
Supplementary Figure A6: Plots displaying true positive rate vs false positive rate (ROC) and precision vs recall (PRC) show the performance of the TEP-ROE model on an independent, held-out test set. The color gradient shows the probability threshold at each point on an auROC or auPRC curve (i.e. the probability threshold that produces the given FPR, TPR/Sensitivity values, or the given Recall, Precision values, at that point on the curve). The area under Sensitivity-Specificity curve indicates auROC performance value, and the area under Precision-Recall curve indicates the auPRC performance value. A ‘perfect model’ is associated with an auROC and auPRC equal to 1.0 (100% of the area is under the curve). A model that places examples into a class at-random will have an auROC and auPRC equal to 0.5 (50% of the area is under the curve).



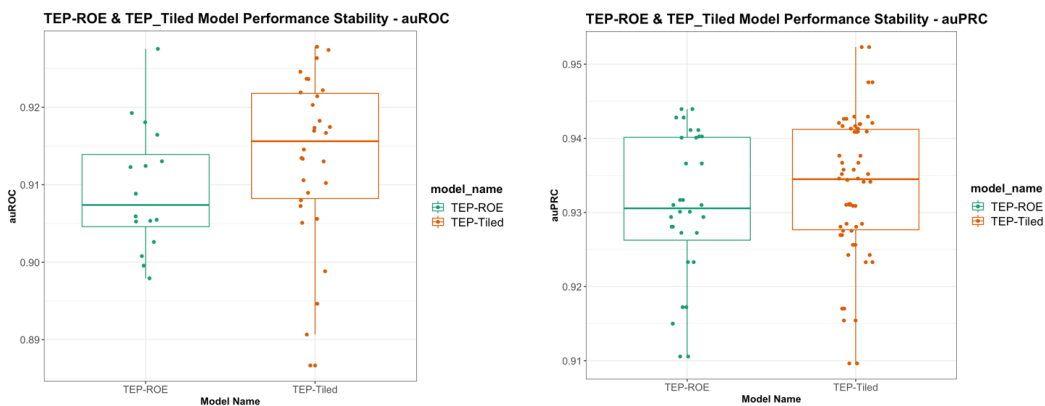


Supplementary Figure A7: Plots displaying true positive rate vs false positive rates (ROC) and precision vs recall (PRC) show the performance of the TEP-Tiled model on an independent, held-out test set. The color gradient shows the probability threshold at each point on an auROC or auPRC curve (i.e. the probability threshold that produces the given FPR, TPR/Sensitivity values, or the given Recall, Precision values, at that point on the curve). The area under Sensitivity-Specificity curve indicates auROC performance value, and the area under Precision-Recall curve indicates the auPRC performance value. A ‘perfect model’ is associated with an auROC and auPRC equal to 1.0 (100% of the area is under the curve). A model that places examples into a class at-random will have an auROC and auPRC equal to 0.5 (50% of the area is under the curve).

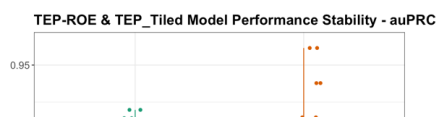
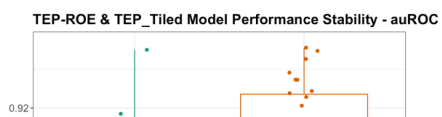
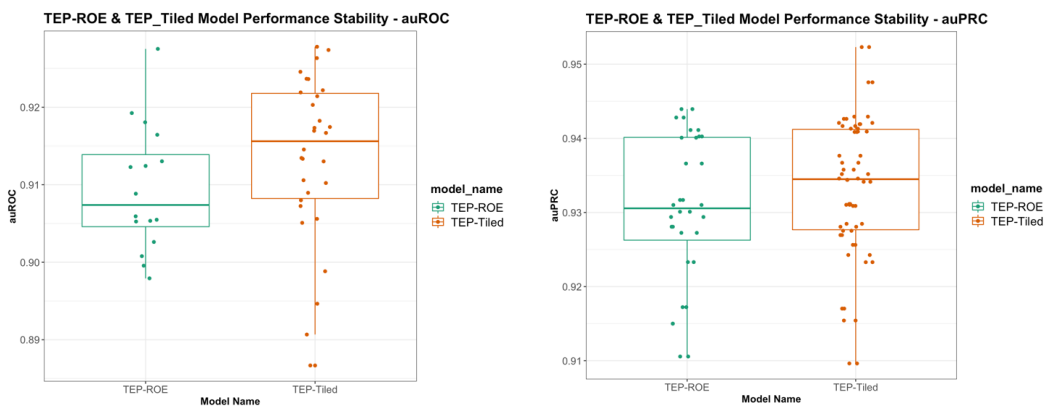


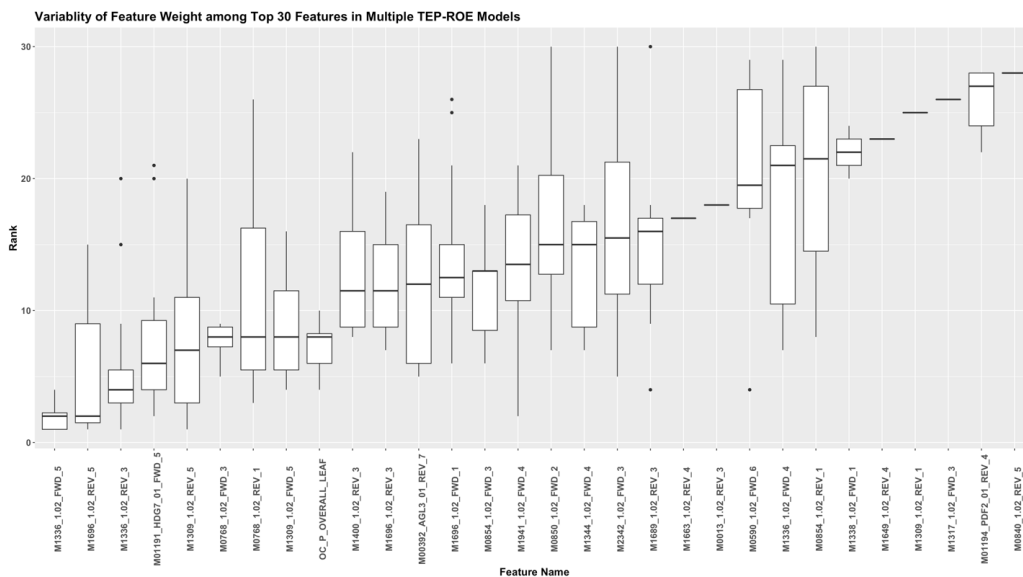


Supplementary Figure A8: 5-fold cross-validation was performed to determine the optimal regularization parameter for the TEP L1-regularized logistic regression models (Left: TEP-ROE, Right: TEP-Tiled). The training samples were divided into 5 partitions of equal size; a model was trained on 4 partitions, and auROC was computed on the 5th test partition over an L1 parameter range; the optimally performing L1 parameter was selection for this partitioning. This process was performed on all 5 possible partitionings. Plots show the auROC curve on each fold in the ROE model (Left) and the Tiled model (right). (The average of optimally performing L1 parameters over the 5 cross-validation partitions for each model was then used for the final model, trained on the entire cross-validation set and tested on the independent held-out test set for reporting of model performance results in the main manuscript).

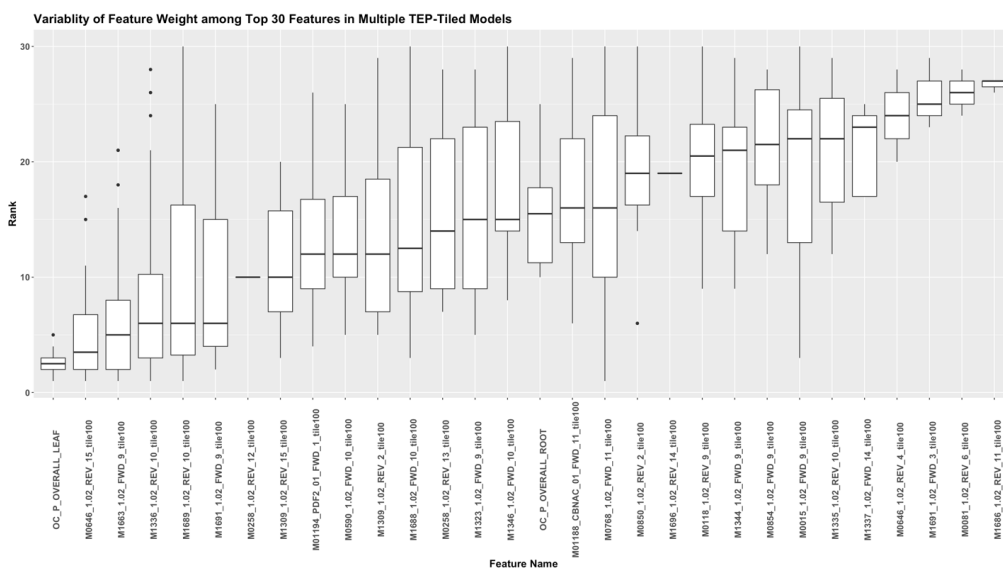


Supplementary Figure A9: Plots displaying auROC and auPRC for 30 model-runs each of the TEP-ROE and TEP-Tiled models. Training datasets were randomly partitioned 30 times and 5-fold cross-validation was performed, followed by testing on independent held-out samples. The auROC and auPRC were computed for each of 30 runs, for each model type. The TEP-Tiled model shows a slightly higher median auROC than the TEP-ROE model, however the variability in the Tiled model’s performance is also higher— both in auROC and auPRC.

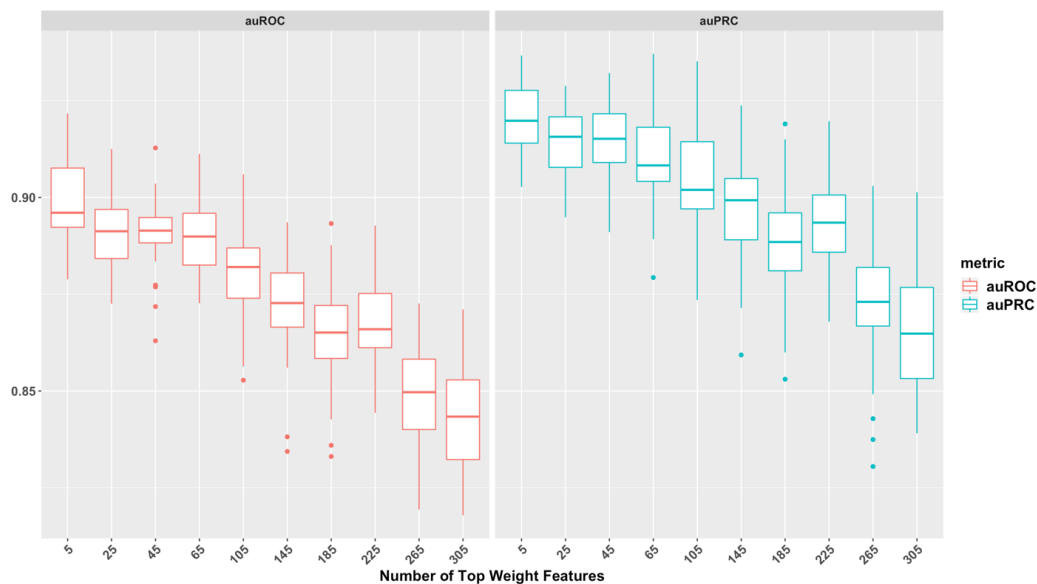




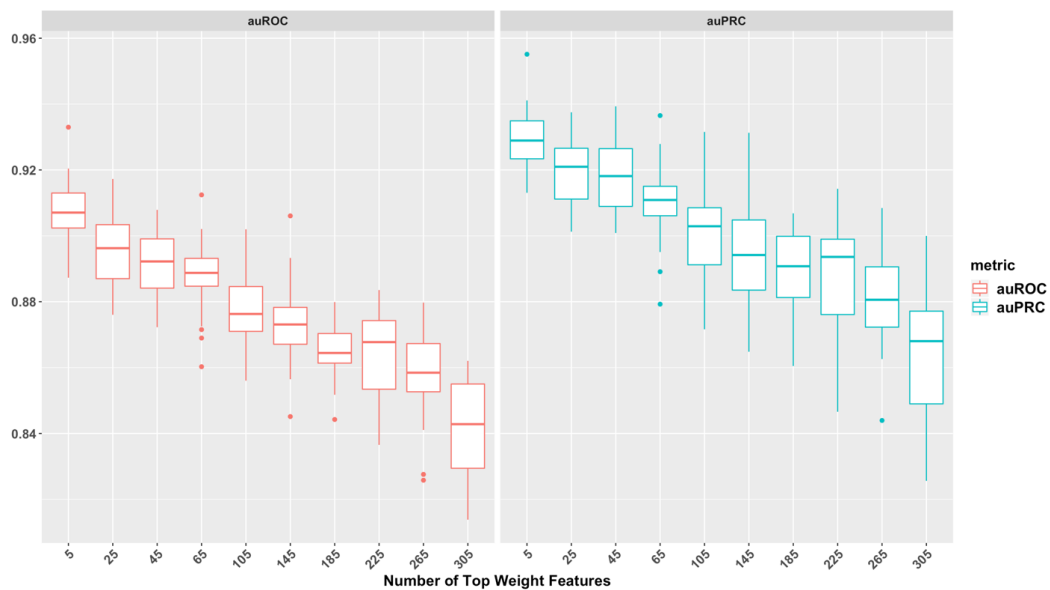
Supplementary Figure A10: Plot showing the variability in feature weight for the top 30 features over 30 runs of the TEP-ROE model. The top 30 most heavily weighted features from ROE model were considered, along with the rank of features in all 30 model-runs. This plot shows that some features display a substantially higher rank variability.



Supplementary Figure A11: Plot showing the variability in feature weight for top 30 features over 30 runs of the TEP-ROE model. The top 30 most heavily weighted features from tiled model were considered, along with the rank of features in all 30 model-runs. This plot shows that some features display a substantially higher rank variability.



Supplementary Figure A12: Plots showing the drop in auROC and auPRC of the TEP-ROE model's performance after removing top weighted features from the training data. Sets of highly weighted PWMs (TF binding domain profiles) and their associated features were removed in multiple stages (top 5, 25, 45, ..., 345) from training data. The TEP-ROE model's performance after removing each set of PWMs was recorded. The decrease in both auROC and auPRC indicates that the PWM sets are contributing unique information to the model, as other features do not appear to 'bubble up' to compensate for their removal in a way that maintains model performance.

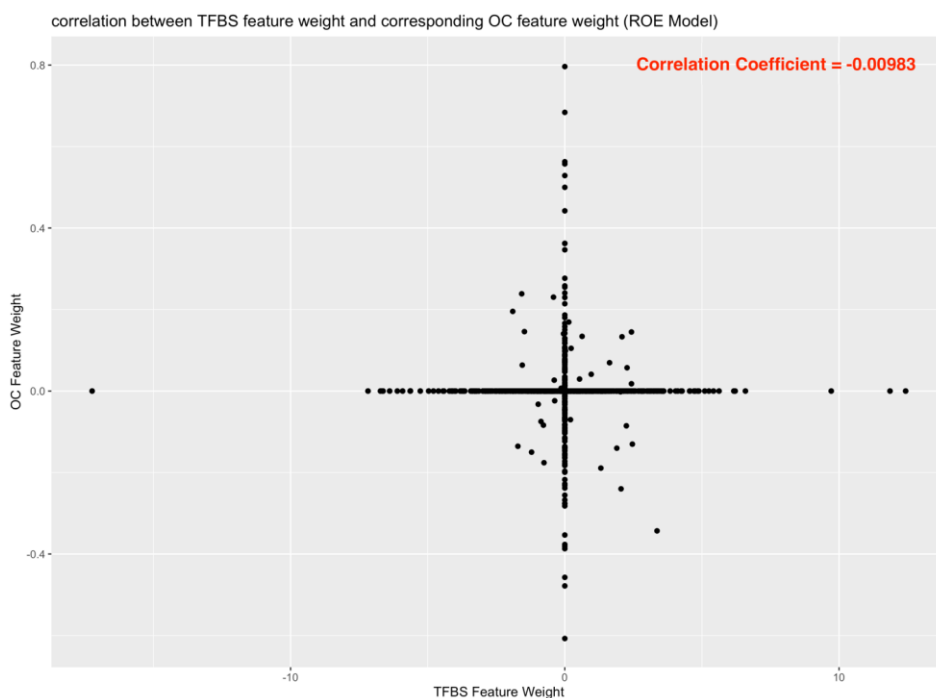


Supplementary Figure A13: Plots showing the drop in auROC and auPRC of the TEP-Tiled model's performance after removing top weighted features from the training data. Sets of highly weighted PWMs (TF binding domain profiles) and their associated features were removed in multiple stages (top 5, 25, 45, ..., 345) from training data. The TEP-Tiled model's performance after removing each set of PWMs was recorded. The decrease in both auROC and auPRC indicates that the PWM sets are contributing unique information to the model, as other features do not appear to 'bubble up' to compensate for their removal in a way that maintains model performance.

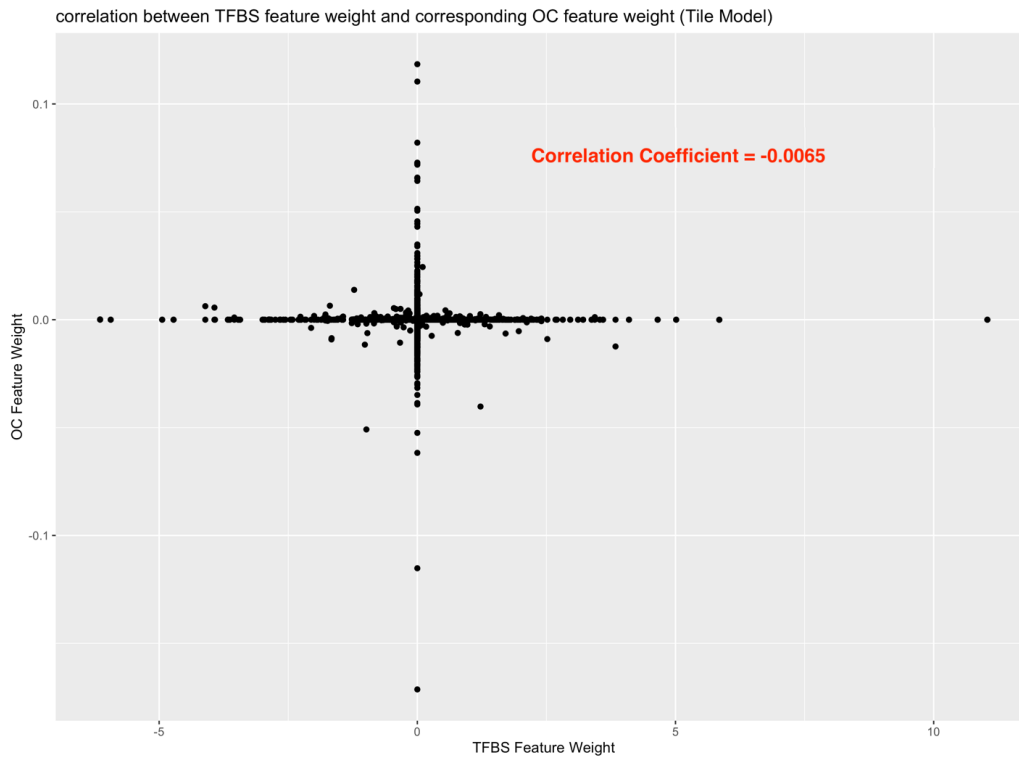


Supplementary Figure A14: Map of top-weighted feature comparison between the [TSS - 1 kb, TSS + 500 nt] Tiled model (TEP-Tiled, the “original model”) and a [TSS - 2 kb, TSS + 500 nt] version of the Tiled model (“enhancer model”). Features in the TEP-Tiled original model are shown in blue, and in the enhancer model in red. This map shows fairly high overall agreement between the two models despite the large additional array of features associated with upstream tiles that the enhancer model could identify as important. Agreement between each model’s set of top-30 most important features is close to 90%, and the vast majority of these features lie within [TSS - 1 kb, TSS + 500 nt] for both

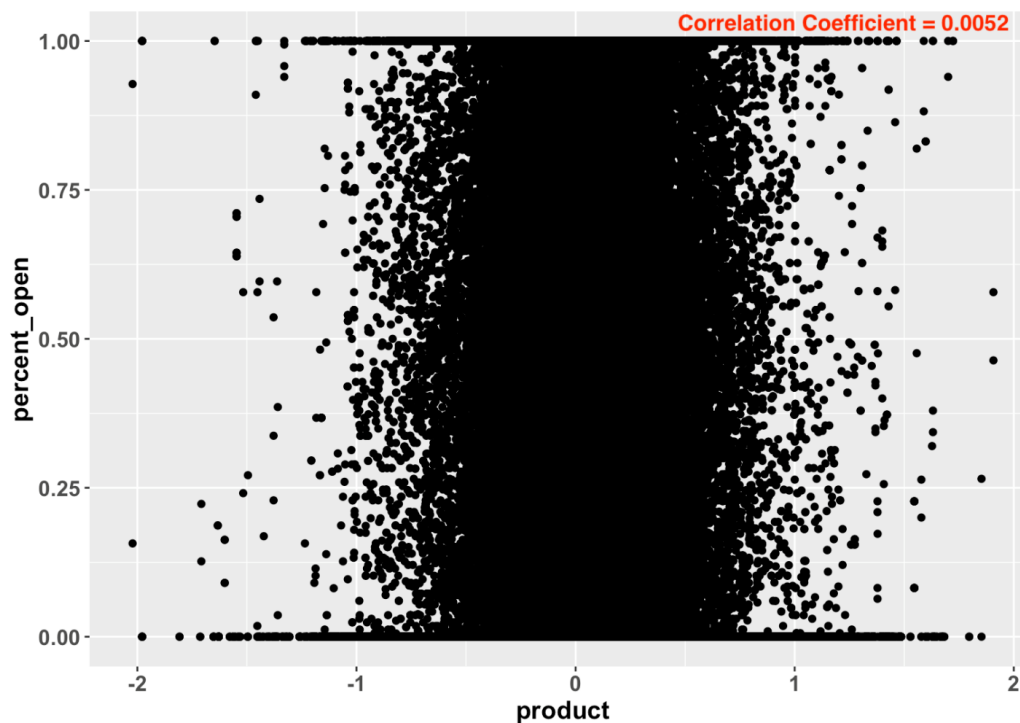
models. The few features in the enhancer model that are located further than 1kb upstream of the TSS are not among top10 most heavily weighted features (see **Supplementary Table 1**).



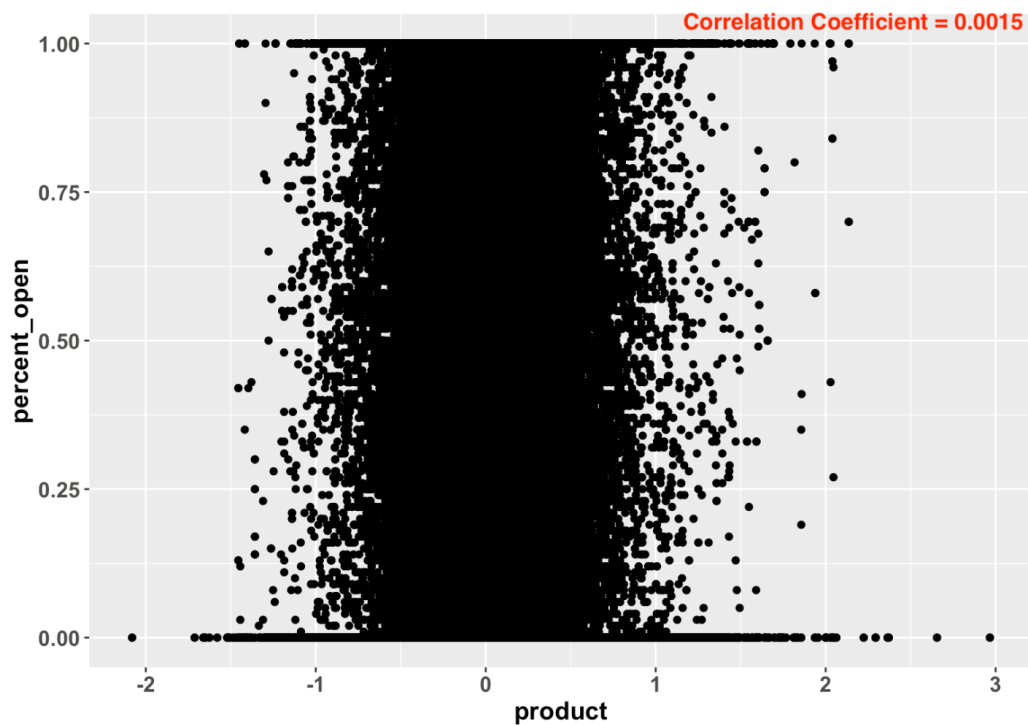
Supplementary Figure A15: Plot showing the correlation between TFBS and OC feature weights in the TEP-ROE model. This figure shows that TFBS features have no correlation with OC features in terms of model importance. The apparent visual positive correlation along the x and y axis derives from features with near-0 coefficients (these features were deemed ‘unimportant’ and essentially set to 0 by the regularization process in model training).



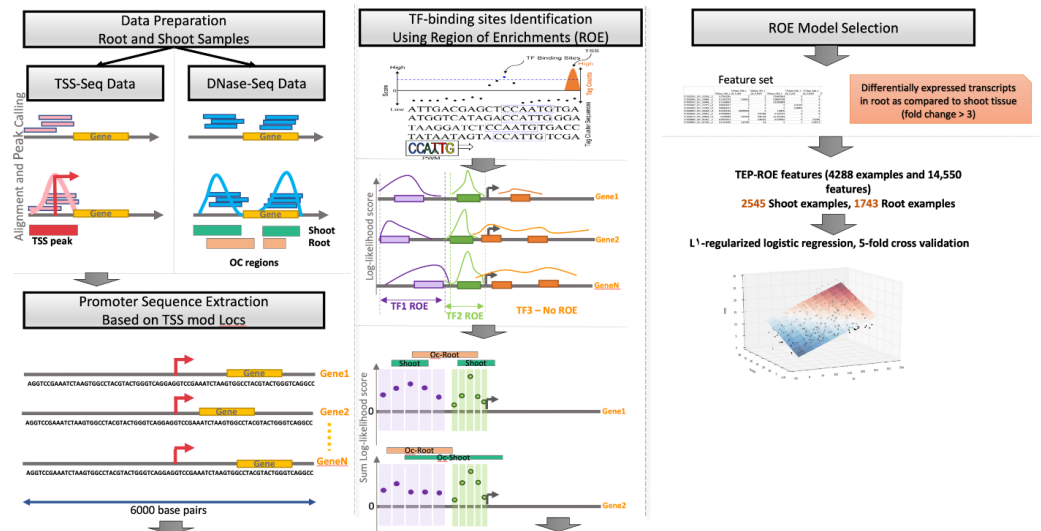
Supplementary Figure A16: Plot showing the correlation between TFBS and OC feature weights in the TEP-Tiled model. This figure shows that TFBS features have no correlation with OC features in terms of model importance. The apparent visual positive correlation along the x and y axis derives from features with near-0 coefficients (these features were deemed ‘unimportant’ and essentially set to 0 by the regularization process in model training).



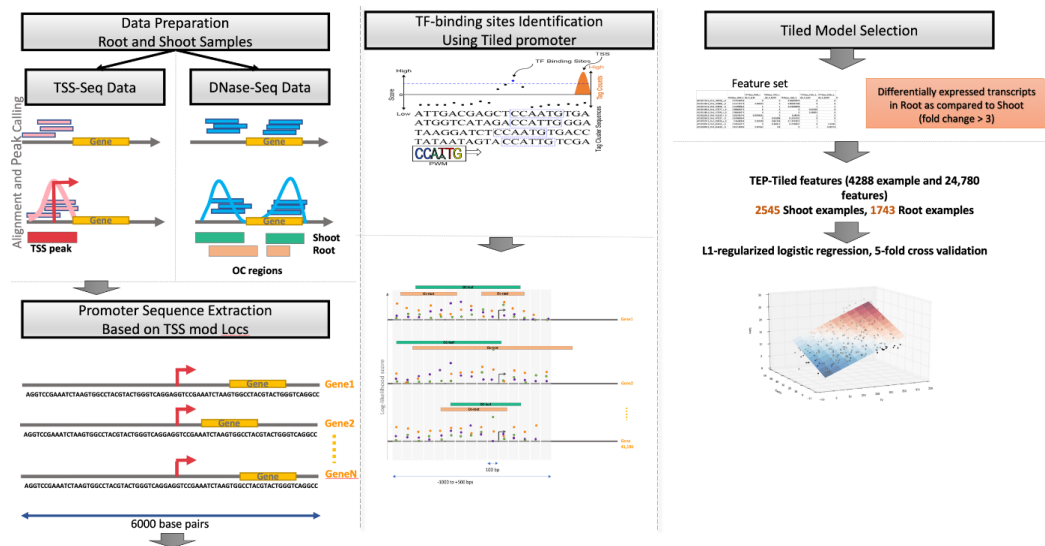
Supplementary Figure A17: Plot showing the correlation between products of TFBS feature value and TFBS model weight on the x-axis, and chromatin % openness on the y-axis, for the top 20 TFBS features in the TEP-ROE model. Each point represents an instance of the product vs. % openness in an individual promoter region. % openness was computed as the percentage of nucleotides within the TFBS feature region that are open (accessible). The correlation value is close to zero, indicating little/no relationship between TFBS feature importance and chromatin accessibility in the TFBS regions of individual promoters.



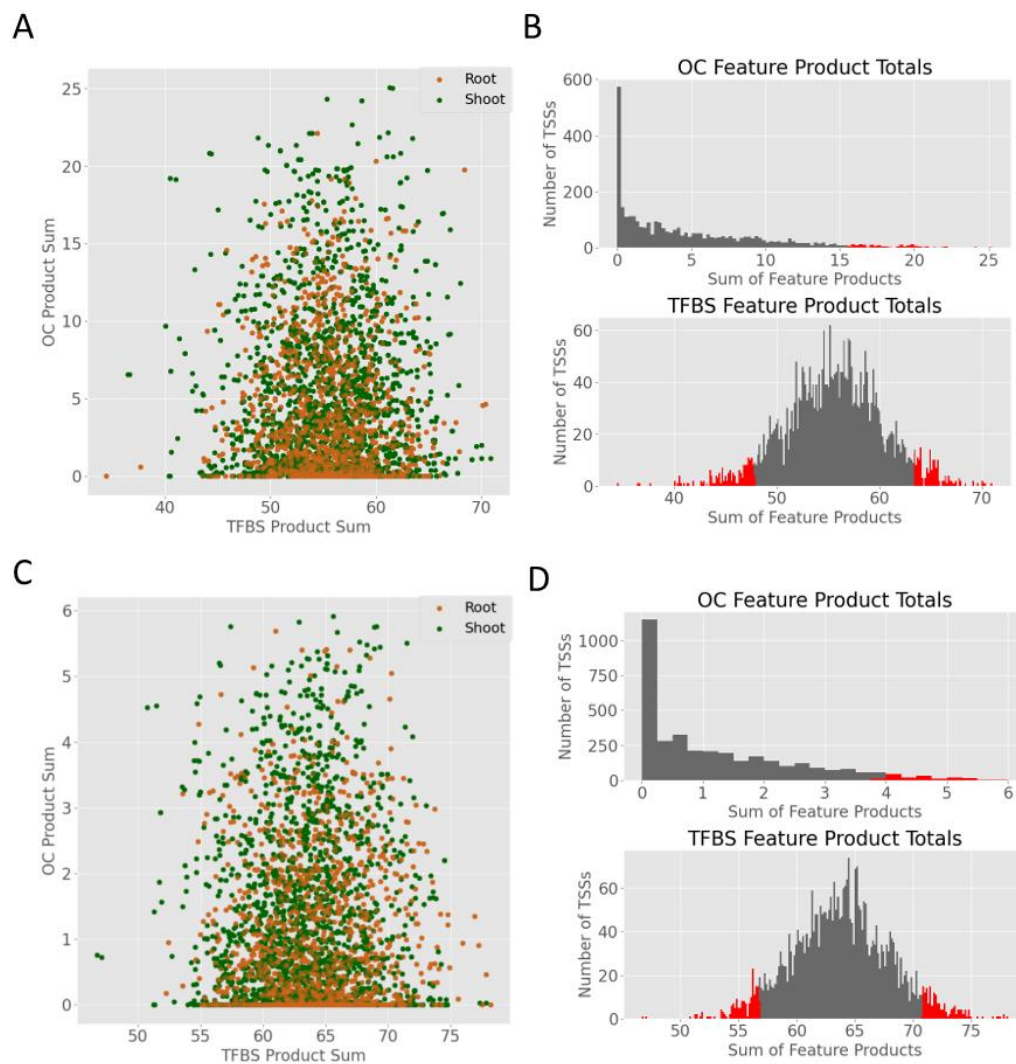
Supplementary Figure A18: Plot showing the correlation between products of TFBS feature value and TFBS model weight on the x-axis, and chromatin % openness on the y-axis, for the top 20 TFBS features in the TEP-Tiled model. Each point represents an instance of the product vs. % openness in an individual promoter region. % openness was computed as the percentage of nucleotides within the TFBS feature region that are open (accessible). The correlation value is close to zero, indicating little/no relationship between TFBS feature importance and chromatin accessibility in the TFBS regions of individual promoters.



Supplementary Figure A19: The modeling process begins with raw dataset processing for TSS-seq, DNase-seq and RNA-seq datasets. This includes mapping reads to genome, calling peaks for both OC-openness and TSS-peak identification, and detecting differentially expressed transcripts to define class labels (root vs shoot). 6kb sequences were extracted (TSS - 3 kb, TSS + 3 kb, centered at each TSS mode), and TFBS log-likelihood summed scores (log-likelihood values above zero) were computed within this region for all PWMs (TF binding profiles). A Region of Enrichment (ROE) was identified for each TF (if present), and model features were generated within ROE regions. The TEP-ROE model training dataset contains 4288 samples, including 2545 shoot-expressed TSS promoters and 1743 root expressed TSS promoters. Finally, L1-regularized logistic regression was used to train and test the model. Using 5-fold cross-validation, the optimal regularization parameter was computed. Final auROC and auPRC metrics were reported on an independent held-out test set comprising 20% of the dataset.



Supplementary Figure A20: The modeling process begins with raw dataset processing for TSS-seq, DNase-seq and RNA-seq datasets. This includes mapping reads to the genome, calling peaks for both OC-openness and TSS-peak identification, and detecting differentially expressed transcripts to define class labels (root vs shoot). 6kb sequences were extracted (TSS - 3 kb, TSS + 3 kb centered at each TSS mode), and TFBS log-likelihood summed scores (log-likelihood values above zero) were computed within this region for all PWMs (TF binding profiles). The region from 1000 nt upstream to 500 nt downstream of the TSS mode was divided into 100nt-wide tiles, and TFBS feature scores (log-likelihood sum scores) were computed within each tile. Model features were generated using TFBS scores and % OC openness in each tile (feature generation process within each Tile is identical to that performed within each ROE region in the ROE model). The training data in TEP-Tiled model contains 4288 samples, including 2545 shoot-expressed promoters and 1743 root expressed promoters. Finally, L1-regularized logistic regression was used to train and test the model. Using 5-fold cross-validation, the optimal regularization parameter was computed. Final auROC and auPRC metrics were reported on an independent held-out test set comprising 20% of the dataset.



Supplementary Figure A21: A) Dot-plot displaying the sums of the TFBS feature products (feature value * feature weight assigned by model) and sums of the OC feature products (feature value * feature weight assigned by model) for each correctly classified TSS from the TEP-ROE model. Green dots indicate shoot; brown dots indicate root. B) Top histogram is a distribution of OC feature product sums from the TEP-ROE model. Bottom histogram is a distribution of TFBS feature product sums from the TEP-ROE model. The red bars represent TSSs that fell into the 5th and 95th percentiles. C) Dot-plot demonstrating the TFBS feature product sums and the OC feature product sums for each correctly classified TSS from the TEP-Tiled model. Green dots indicate shoot; brown dots indicate root. B) Top histogram is a distribution of OC feature product sums from the TEP-Tiled model. Bottom histogram is a distribution of TFBS feature product sums from the TEP-Tiled model. The red bars represent TSSs that fell into the 5th and 95th percentiles.

	No. TSS peaks	Mapped Location		
		<250	5' utr	tss
Root	26,040	2223	6047	17,770
Shoot	24,595	1636	5152	17,807

Supplementary Table A1: Table showing the number of total valid TSS peaks identified in each TSS-seq dataset, along with their mapped locations in the TAIR10 genome. The JAMM peak finder was used to identify TSS peaks from nanoCage-XL root and shoot samples. Only peaks having more than 50 reads and located in the region immediately upstream of gene body (within 500 nt) are considered “valid TSS peaks”.

Tissue	#TSS peaks mapped to individual Transcripts				
	1 Transcript	2 Transcripts	3 Transcripts	4 Transcripts	5 Transcripts
Root	21,260	1950	236	35	4
Shoot	21,229	1416	128	30	6

Supplementary Table A2: Table showing the number of TSS peaks associated with one or more transcripts. Most TSS peaks are assigned to unique transcripts (first column left, “1 Transcript”). However, there are few TSS peaks in the vicinity of more than one transcripts in TAIR10 genome (within 500 nt).

	ROOT	SHOOT	TOTAL
No. Of TSS peaks (No. of Promoters)	26,040	24,595	50,635
No. of Transcripts with mapped TSS peak upstream	23,487 (67% coverage)	22,809 (65% coverage)	25,751 (73%)
No. of Genes with mapped TSS peak upstream	16,841(62%)	16,086 (59%)	18,580 (68%)
No. of closed Promoters	656 (1.3%)	639 (1.2%)	123 closed in both (0.24%)
No. of DE Transcripts	575	1037	1612
No. of DE Genes	562	963	1525
No. of Tissue-specific Promoters	1743	2545	4288

Supplementary Table A3: Total number of TAIR10 transcripts is 35,176 transcripts, which are associated with 27,206 protein coding genes. For simplicity in the row descriptions, a “Promoter” is equated to the [TSS - 3 kb, TSS + 3 kb] region surrounding a TSS. The first row shows the total number of TSS peaks used to extract [TSS - 3 kb, TSS + 3 kb] regions centered at TSS peak mode. The second row in the table shows the number of transcripts with TSS peaks upstream of the TAIR10 gene body (within 500 nt). The third row shows the number of TSS peaks covering TAIR10-annotated protein-coding genes. 68% of total TAIR annotated genes are assigned with at least one TSS peak in at least one of the tissues. The fourth row shows the number of TSS peaks with closed chromatin over the [TSS - 3 kb, TSS + 3 kb] region. Only 1% of the [TSS - 3 kb, TSS + 3 kb] regions, out of ~50,635 TSSs, are closed in one of the two tissues, and less than 0.3% are closed in both tissues. The last three rows show RNA-seq related information. A transcript is considered a tissue-specific or differentially expressed (DE) transcript if it expressed in both tissues (the expression value in one tissue is greater than 300 and the minimum expression value in the other tissue is greater than 30) and has log-fold-change above 3 (computed using RSEM). Read abundance was calculated on transcript level; we report differentially expressed genes as the number of genes that have at least one DE transcript. Finally, 4288 tissue specific TSSs and their [TSS - 3 kb, TSS + 3 kb] surrounding regions were used to train/test the ML model. Since the surrounding sequences are extracted from the set of TSS peaks, the TSSs assigned to DE transcripts are considered as tissue-specific (and their surrounding regions are considered tissue-specific promoters).

	No. Transcripts (% Coverage)	No. Genes (% Coverage)
Low/No-expression in both tissues	2,355 (7%)	1,117 (4%)
Expressed in both tissues	13,789 (40%)	12,447 (46%)
Not-Differentially expressed	12,177 (35%)	10,922 (40%)

Supplementary Table A4: Low or No-expression transcripts are those for which the mean normalized expression value is less than 30 in both tissue types. Genes that have low or no expression are those whose transcripts are all low/non-expressing. Transcripts expressed in both tissues are those for which the mean normalized expression values were greater than 300 in one tissue and greater than 30 in the other tissue. Percent coverage is out of 35,176 TAIR10 transcripts (for No. of Transcripts), or 27,206 protein coding genes (for No. of Genes).

Shoot PWMs	Shoot coeffs	Root coeffs	Root PWMs
GAcontent	0.90291 3	0.55416 7	GAcontent
GCcontent	0.50632 6	0.45186 1	GCcontent
CAcontent	0.35932 2	0.45063 3	GA
M00502_TEIL_01	0.28499 8	0.37589 3	M00702_SPF1_Q2
M01126_BPC1_Q2	0.23044 1	0.25785 6	TATAbox
RAV1-A_binding_site_motif	0.22113 6	0.22231 8	CAcontent
M00503_ATHB5_01	0.20687 2	0.20711 5	Inr
M00355_PBF_01	0.20172 8	0.19535 3	TEF-box_promoter_motif
Inr	0.19520 6	0.18954 3	RAV1-A_binding_site_motif
TATAbox	0.19012 3	0.17911 7	Bellringer_replumless_pennywise_BS1_I N_AG
M01136_DOF_Q2	0.18255	0.16839 3	M00355_PBF_01
Bellringer_replumless_pennywise_BS1_I N_AG	0.17712 2	0.16662 6	MYB4_binding_site_motif
M00439_C1_Q2	0.17623	0.15994	M00314_GEN_INI3_B
Y_Patch	0.17142 5	0.15883 7	M01006_AGP1_01
M00506_LIM1_01	0.16536 1	0.15491 9	M01126_BPC1_Q2
M01135_GAMYB_Q2	0.14844 4	0.15297 9	M01050_ARR10_01
GA	0.14656 3	0.14662	M00439_C1_Q2
TEF-box_promoter_motif	0.13626 7	0.13738 1	M00653_OCSBF1_01
MYB4_binding_site_motif	0.13102 2	0.13094 6	M01054_BHLH66_01
SORLIP2	0.12592	0.12888 2	Y_Patch
M00344_RAV1_02	0.12241 1	0.12503 3	M00952_PCF5_01
M00702_SPF1_Q2	0.11960 7	0.12295 6	M00502_TEIL_01
M01057_ERF2_01	0.11398 7	0.12095 3	SORLIP2
M01164_SQUA_01	0.11333 2	0.12084 7	BoxII_promoter_motif
M00653_OCSBF1_01	0.10963 8	0.11939 8	AtMYC2_BS_in_RD22
M00315_GEN_INI_B	0.10081 6	0.10727 6	GCbox
M01050_ARR10_01	0.09994 2	0.10244 1	M00506_LIM1_01

Hexamer_promoter_motif	0.09898 4	0.09830 4	M01057_ERF2_01
M01006_AGP1_01	0.09303 2	0.09755 8	M00344_RAV1_02
ATHB2_binding_site_motif	0.08882 4	0.09406	MYB3_binding_site_motif
M00440_CG1_Q6	0.08875 8	0.09264 2	ATHB2_binding_site_motif
SORLIP5	0.08736 8	0.09127 7	AG_BS_in_SPL_NOZ
T-box_promoter_motif	0.08591 6	0.08891 7	M01135_GAMYB_Q2
BoxII_promoter_motif	0.08443 2	0.08490 6	M00635_GT1_Q6
M01194_PDF2_01	0.08308 6	0.08465 7	M01164_SQUA_01
GCbox	0.07911 6	0.08159 4	M00503_ATHB5_01
RAV1-B_binding_site_motif	0.07635 5	0.08141 5	M01194_PDF2_01
DRE-like_promoter_motif	0.07515 1	0.08009 4	JASE2_motif_in_OPR1
TELO-box_promoter_motif	0.07469 7	0.07948 6	DRE-like_promoter_motif
M00700_ROM_Q2	0.07298 4	0.07928 8	M00353_DOF2_01
M00438_ARF_Q2	0.07115	0.07647 7	M01136_DOF_Q2
M01188_CBNAC_01	0.07080 3	0.07634 4	M00654_OSBZ8_Q6
M00952_PCF5_01	0.06969 8	0.07610 7	CCA1_binding_site_motif
M00370_CPRF3_Q2	0.06749 9	0.07584 3	Hexamer_promoter_motif
E2F_DP_BS_in_AtCDC6	0.06647 1	0.07483 7	M01133_AG_Q2
M01130_PBF_Q2	0.06604 8	0.07284 3	M00313_GEN_INI2_B
SORLREP3	0.06239 9	0.07272 1	EveningElement_promoter_motif
AG_BS_in_SPL_NOZ	0.06208 1	0.07157 6	T-box_promoter_motif
M00376_TGA1A_Q2	0.06079 4	0.06797 1	M00404_MADSB_Q2
M01021_ID1_01	0.05948 8	0.06609 9	AtMYB2_BS_in_RD22

Supplementary Table A5. Top 50 most heavily weighted features for the 3PEAT nanoCAGE-XL root-trained model and shoot-trained model, with their respective model coefficient weights.

Tissue of Training Dataset	Tissue of Test Dataset	
	Shoot	Root
Root	auROC: 0.98 auPRC:0.79	auROC:0.98 auPRC:0.80
Shoot	auROC: 0.98 auPRC: 0.80	auROC:0.98 auPRC: 0.78

Supplementary Table A6. 3PEAT model performance outcomes for TSS location prediction in a given tissue sample. The training dataset is the collection of genomic sites (locations that are highly expressed TSSs in a tissue, or are not TSSs in a tissue) on which a classifier model was trained; the test dataset is the collection of genomic sites on which a model was tested to yield a performance measure.

GO-term	Description	p-value
GO: 0010449	root meristem growth	0.024373259
GO: 0010102	lateral root morphogenesis	0.045790547
GO: 0010101	post-embryonic root	0.045790547
GO: 0010449	root meristem growth	0.03363946
GO: 2000280	regulation of root development	0.029422773
GO: 0010101	post-embryonic root	0.043180084
GO: 0010102	lateral root morphogenesis	0.043180084
GO: 0010449	root meristem growth	0.022479093
GO: 0090057	root radial pattern formation	0.047136064
GO: 0090057	root radial pattern formation	0.044521176
GO: 0010101	post-embryonic root	0.040432343
GO: 0010102	lateral root morphogenesis	0.040432343
GO: 0010449	root meristem growth	0.021048648

Supplementary Table A7: GO-term enrichment analysis of the TSSs misclassified by the 3PEAT-style single tissue models. All of the significant terms for the misclassified TSSs are related to root development. Only terms with a p-value < 0.05 are reported here.

No. Top Features	No. of PWMs (TEP-ROE)	No. of PWMs (TEP-Tiled)	No. of Common PWMs	No. of Distinct PWMs	% common
2	1	1	0	2	0.00000
5	3	4	1	6	14.28571
7	5	6	2	8	20.00000
10	7	9	4	9	30.76923
12	9	11	4	13	23.52941
15	12	13	5	16	23.80952
20	17	16	6	21	22.22222
50	36	40	15	50	23.07692
100	75	72	28	92	23.33333
150	99	92	43	112	27.74194
200	122	111	57	127	30.97826
300	151	144	82	147	35.80786
400	171	161	112	155	41.94757
500	182	173	135	144	48.38710
1000	189	189	157	101	60.85271

Supplementary Table A8: Differences among top weighted features between TEP-ROE and TEP-Tiled models. For a given number of top features from both the TEP-ROE and TEP-Tiled models, the number of PWMs representing only the features that agree between the two models was calculated; this number can be found in the “%common” column. As the number of features examined increases, top-feature agreement between the two models also increases.

feature	pwm	strand	win	coef	type	left	right
OC_P_OVERALL_LEAF	OC_P_OVERALL_LEAF	FWD	20	7.216611	enhancer	-100	0
M1663_1.02_FWD_19	M1663_1.02	FWD	19	5.885361	enhancer	-200	-100
M1691_1.02_FWD_19	M1691_1.02	FWD	19	-5.864451	enhancer	-200	-100
M1309_1.02_REV_12	M1309_1.02	REV	12	5.223288	enhancer	-900	-800
M1686_1.02_REV_20	M1686_1.02	REV	20	-5.126922	enhancer	-100	0
M1689_1.02_REV_20	M1689_1.02	REV	20	-4.728434	enhancer	-100	0
M1691_1.02_FWD_21	M1691_1.02	FWD	21	-4.722934	enhancer	0	100
M1323_1.02_FWD_19	M1323_1.02	FWD	19	-4.649934	enhancer	-200	-100
M01194_PDF2_01_FWD_11	M01194_PDF2_01	FWD	11	4.556978	enhancer	-1000	-900
M1688_1.02_FWD_20	M1688_1.02	FWD	20	-4.449704	enhancer	-100	0
OC_P_OVERALL_ROOT	OC_P_OVERALL_ROOT	FWD	20	-4.440923	enhancer	-100	0
M0843_1.02_FWD_16	M0843_1.02	FWD	16	-4.135187	enhancer	-500	-400
M1410_1.02_FWD_8	M1410_1.02	FWD	8	-4.005008	enhancer	-1300	-1200
M1620_1.02_REV_4	M1620_1.02	REV	4	3.864812	enhancer	-1700	-1600
M0267_1.02_FWD_17	M0267_1.02	FWD	17	-3.863250	enhancer	-400	-300
M1686_1.02_FWD_19	M1686_1.02	FWD	19	-3.746127	enhancer	-200	-100
M0854_1.02_FWD_19	M0854_1.02	FWD	19	-3.693703	enhancer	-200	-100
M1335_1.02_REV_20	M1335_1.02	REV	20	-3.687715	enhancer	-100	0
M1309_1.02_REV_20	M1309_1.02	REV	20	3.676828	enhancer	-100	0
M0119_1.02_REV_19	M0119_1.02	REV	19	-3.535757	enhancer	-200	-100
M0163_1.02_REV_25	M0163_1.02	REV	25	-3.526353	enhancer	400	500
M01194_PDF2_01_FWD_19	M01194_PDF2_01	FWD	19	3.507581	enhancer	-200	-100
M1195_1.02_FWD_13	M1195_1.02	FWD	13	-3.506908	enhancer	-800	-700
M0680_1.02_REV_4	M0680_1.02	REV	4	-3.460753	enhancer	-1700	-1600
M1702_1.02_FWD_22	M1702_1.02	FWD	22	-3.440453	enhancer	100	200
M1403_1.02_FWD_11	M1403_1.02	FWD	11	-3.437669	enhancer	-1000	-900
M01191_HDG7_01_FWD_20	M01191_HDG7_01	FWD	20	3.436033	enhancer	-100	0
M0646_1.02_REV_25	M0646_1.02	REV	25	-3.334678	enhancer	400	500
M0259_1.02_REV_2	M0259_1.02	REV	2	-3.279065	enhancer	-1900	-1800
M1688_1.02_FWD_18	M1688_1.02	FWD	18	-3.198103	enhancer	-300	-200

Supplementary Table A9: Top 30 features from the Tiled enhancer model, which computes features over tiles in the region [TSS - 2 kb, TSS + 500 nt]. The first column contains the full feature name. The “coef” column contains the feature’s model weight. A positive weight indicates shoot class association, and a negative weight indicates root class association. The right-most two columns show the genome coordinates (in nt) of the feature’s tile relative to TSS mode.

TSS ID	Tissue	Top TFBS	Associated TFs
AT1G31050.1_Chr1_11079176_-_0	Root	M1940_1.02_FWD_5	IDD1, IDD4, IDD6, IDD12, MGP, JKD
AT1G54940.1_Chr1_20481654_+_0	Root	M0263_1.02_FWD_4	bZIP17, bZIP49
AT1G78660.2_Chr1_29585876_+_0	Root	M0844_1.02_FWD_4	ATHB-22
AT2G16980.1_Chr2_7376364_+_0	Root	M0142_1.02_REV_7	“AT-hook containing protein”
AT2G16980.1_Chr2_7376366_+_0	Root	M0010_1.02_FWD_5	ERF9, ERF10, ERF110, ATERF14, ATABI4, RAP2.6
AT2G16980.2_Chr2_7376364_+_0	Root	M0142_1.02_REV_7	“AT-hook containing protein”
AT2G16980.2_Chr2_7376366_+_0	Root	M0010_1.02_FWD_5	ERF9, ERF10, ERF110, ATERF14, ATABI4, RAP2.6
AT4G08555.1_Chr4_5448141_+_0	Root	M0371_1.02_REV_6	ZFP8
AT4G35380.1_Chr4_16819824_+_0	Root	M0582_1.02_REV_4	CAMTA2
AT5G36970.1_Chr5_14605243_-_0	Root	M1576_1.02_REV_1	YAB3, YAB5
AT3G09162.1_Chr3_2808252_-_0	Shoot	M1344_1.02_FWD_4	KUA1
AT3G14330.1_Chr3_4782437_-_0	Shoot	M2343_1.02_REV_3	BZR1
AT3G14330.1_Chr3_4782444_-_0	Shoot	M2343_1.02_REV_3	BZR1
AT4G26555.1_Chr4_13406181_-_0	Shoot	M1309_1.02_REV_5	MYBD, MYBH
AT4G26555.1_Chr4_13406188_-_0	Shoot	M1309_1.02_REV_5	MYBD, MYBH
AT5G13730.1_Chr5_4430798_-_0	Shoot	M0370_1.02_FWD_6	ZAT9, STZ, AZF3
AT5G60040.2_Chr5_24173327_+_0	Shoot	M0119_1.02_REV_4	AHL13, HAP3

Supplementary Table A10: Table containing TSS IDs that have putatively hard-coded promoters, computed from weights and feature values from the TEP-ROE model. Also listed is the top weighted TFBS for each TSS and the TFs that are associated with it.

TSS ID	Tissue	Top TFBS	Associated TF
AT1G12160.1_Chr1_4126070_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT1G33280.1_Chr1_12072606_+0	Root	M0142_1.02_FWD_1_tile100	“AT-hook containing protein”
AT1G45015.1_Chr1_17021589_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT1G45015.2_Chr1_17021589_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT1G53680.1_Chr1_20038359_+0	Root	M0142_1.02_REV_14_tile100	“AT-hook containing protein”
AT1G62990.1_Chr1_23337372_+0	Root	M0015_1.02_FWD_9_tile100	DREB2, DREB2C, DREB2D
AT1G68150.1_Chr1_25543988_+0	Root	M0646_1.02_REV_15_tile100	DOF1.5
AT2G16970.1_Chr2_7369541_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT2G16980.1_Chr2_7376364_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT2G16980.1_Chr2_7376366_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT2G16980.2_Chr2_7376364_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT2G16980.2_Chr2_7376366_+0	Root	M0118_1.02_REV_9_tile100	HMGA
AT3G05770.1_Chr3_1712207_-0	Root	M01126_BPC1_Q2_REV_12_tile100	BPC1
AT3G09270.1_Chr3_2849273_-0	Root	M0118_1.02_REV_9_tile100	HMGA
AT4G18550.1_Chr4_10226979_-0	Root	M0118_1.02_REV_9_tile100	HMGA
AT5G12420.1_Chr5_4026782_-0	Root	M0118_1.02_REV_9_tile100	HMGA
AT5G40730.1_Chr5_16301102_+0	Root	M1663_1.02_FWD_9_tile100	TCP1, TCP10, TCP13
AT5G45920.1_Chr5_18622757_+0	Root	M0015_1.02_REV_15_tile100	DREB2, DREB2C, DREB2D
AT5G45920.1_Chr5_18622778_+0	Root	M1274_1.02_REV_15_tile100	ASL5, LBD3, LBD4
AT5G65160.1_Chr5_26034000_-0	Root	M0080_1.02_REV_2_tile100	EDF3
AT1G51805.1_Chr1_19225648_-0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT1G51805.2_Chr1_19225648_-0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT1G52000.1_Chr1_19336089_-0	Shoot	M0372_1.02_REV_11_tile100	ZAT1, ZAT4, ZAT9, AZF2
AT1G64860.1_Chr1_24098017_+0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT1G64860.2_Chr1_24098017_+0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT1G76960.1_Chr1_28920925_-0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT1G79040.1_Chr1_29736063_+0	Shoot	M01188_CBNAC_01_FWD_11_tile100	CBNAC
AT3G01060.1_Chr3_18779_-0	Shoot	M0583_1.02_REV_10_tile100	ATGRP2B, CSDP2
AT3G01060.2_Chr3_18779_-0	Shoot	M0583_1.02_REV_10_tile100	ATGRP2B, CSDP2
AT3G03341.1_Chr3_790618_-0	Shoot	M0142_1.02_FWD_1_tile100	“AT-hook containing protein”
AT3G51820.1_Chr3_19219005_-0	Shoot	M1274_1.02_REV_15_tile100	ASL5, LBD3, LBD4
AT3G51820.1_Chr3_19219011_-0	Shoot	M1274_1.02_REV_15_tile100	ASL5, LBD3, LBD4
AT3G52150.1_Chr3_19342053_+0	Shoot	M0372_1.02_REV_11_tile100	ZAT1, ZAT4, ZAT9, AZF2
AT3G52150.1_Chr3_19342057_+0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT3G52150.2_Chr3_19342053_+0	Shoot	M0372_1.02_REV_11_tile100	ZAT1, ZAT4, ZAT9, AZF2
AT3G52150.2_Chr3_19342057_+0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT4G17560.1_Chr4_9780335_+0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT4G17560.1_Chr4_9780336_+0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT4G26950.2_Chr4_13534252_-0	Shoot	M1274_1.02_REV_15_tile100	ASL5, LBD3, LBD4
AT5G13730.1_Chr5_4430798_-0	Shoot	M1309_1.02_REV_15_tile100	MYBD, MYBH
AT5G24150.1_Chr5_8175404_-0	Shoot	M0118_1.02_REV_9_tile100	HMGA
AT5G24150.2_Chr5_8175404_-0	Shoot	M0118_1.02_REV_9_tile100	HMGA

Supplementary Table A11: Table containing TSS IDs that have putatively hard-coded promoters, computed from weights and feature values from the TEP-Tiled model. Also listed is the top weighted TFBS for each TSS and the TFs that are associated with it.

GO-Term	Name	p-value	Model Type
GO:0015904	tetracycline transmembrane transport	0.000608	TEP-ROE
GO:0046900	tetrahydrofolylpolyglutamate metabolic process	0.001823	TEP-ROE
GO:0006855	drug transmembrane transport	0.001823	TEP-ROE
GO:0015893	drug transport	0.002127	TEP-ROE
GO:0032012	regulation of ARF protein signal transduction	0.00243	TEP-ROE
GO:0032774	RNA biosynthetic process	0.002451	TEP-ROE
GO:0046578	regulation of Ras protein signal transduction	0.002733	TEP-ROE
GO:0051056	regulation of small GTPase mediated signal transduction	0.002733	TEP-ROE
GO:0046483	heterocycle metabolic process	0.005007	TEP-ROE
GO:1900865	chloroplast RNA modification	0.005763	TEP-ROE
GO:0006725	cellular aromatic compound metabolic process	0.005993	TEP-ROE
GO:0016554	cytidine to uridine editing	0.00667	TEP-ROE
GO:1901360	organic cyclic compound metabolic process	0.006758	TEP-ROE
GO:0016070	RNA metabolic process	0.006912	TEP-ROE
GO:0015850	organic hydroxy compound transport	0.007576	TEP-ROE
GO:0016553	base conversion or substitution editing	0.008482	TEP-ROE
GO:0034654	nucleobase-containing compound biosynthetic process	0.010517	TEP-ROE
GO:0045492	xylan biosynthetic process	0.011194	TEP-ROE
GO:0006760	folic acid-containing compound metabolic process	0.011194	TEP-ROE
GO:0042558	pteridine-containing compound metabolic process	0.011796	TEP-ROE
GO:0009059	macromolecule biosynthetic process	0.012875	TEP-ROE
GO:0009863	salicylic acid mediated signaling pathway	0.013599	TEP-ROE
GO:0090304	nucleic acid metabolic process	0.014502	TEP-ROE
GO:0006352	DNA-templated transcription, initiation	0.016598	TEP-ROE
GO:0045491	xylan metabolic process	0.017496	TEP-ROE
GO:1902531	regulation of intracellular signal transduction	0.017795	TEP-ROE
GO:0034641	cellular nitrogen compound metabolic process	0.018046	TEP-ROE
GO:0018130	heterocycle biosynthetic process	0.019841	TEP-ROE
GO:0019438	aromatic compound biosynthetic process	0.024284	TEP-ROE
GO:0006139	nucleobase-containing compound metabolic process	0.024824	TEP-ROE
GO:0070592	cell wall polysaccharide biosynthetic process	0.02495	TEP-ROE
GO:0070589	cellular component macromolecule biosynthetic process	0.026138	TEP-ROE
GO:0044038	cell wall macromolecule biosynthetic process	0.026138	TEP-ROE
GO:0071482	cellular response to light stimulus	0.028214	TEP-ROE
GO:0006575	cellular modified amino acid metabolic process	0.028806	TEP-ROE
GO:1901362	organic cyclic compound biosynthetic process	0.029586	TEP-ROE
GO:0071478	cellular response to radiation	0.02999	TEP-ROE
GO:0098656	anion transmembrane transport	0.032058	TEP-ROE
GO:0010410	hemicellulose metabolic process	0.033533	TEP-ROE
GO:0010383	cell wall polysaccharide metabolic process	0.043214	TEP-ROE
GO:0050794	regulation of cellular process	0.047963	TEP-ROE
GO:0015711	organic anion transport	0.048458	TEP-ROE
GO:0015850	organic hydroxy compound transport	0.000144	TEP-Tiled

GO:0006638	neutral lipid metabolic process	0.000285	TEP-Tiled
GO:0006639	acylglycerol metabolic process	0.000285	TEP-Tiled
GO:0006749	glutathione metabolic process	0.000451	TEP-Tiled
GO:0009407	toxin catabolic process	0.000493	TEP-Tiled
GO:0006352	DNA-templated transcription, initiation	0.000704	TEP-Tiled
GO:0046462	monoacylglycerol metabolic process	0.000709	TEP-Tiled
GO:0052651	monoacylglycerol catabolic process	0.000709	TEP-Tiled
GO:0046340	diacylglycerol catabolic process	0.000709	TEP-Tiled
GO:0009404	toxin metabolic process	0.00081	TEP-Tiled
GO:0019748	secondary metabolic process	0.001285	TEP-Tiled
GO:0071461	cellular response to redox state	0.001418	TEP-Tiled
GO:0015904	tetracycline transmembrane transport	0.001418	TEP-Tiled
GO:0009058	biosynthetic process	0.001494	TEP-Tiled
GO:0010029	regulation of seed germination	0.00183	TEP-Tiled
GO:1900140	regulation of seedling development	0.001953	TEP-Tiled
GO:0006790	sulfur compound metabolic process	0.001969	TEP-Tiled
GO:0071482	cellular response to light stimulus	0.002038	TEP-Tiled
GO:0006575	cellular modified amino acid metabolic process	0.002124	TEP-Tiled
GO:0010270	photosystem II oxygen evolving complex assembly	0.002127	TEP-Tiled
GO:0080005	photosystem stoichiometry adjustment	0.002127	TEP-Tiled
GO:0071478	cellular response to radiation	0.002302	TEP-Tiled
GO:0098754	detoxification	0.002393	TEP-Tiled
GO:0046461	neutral lipid catabolic process	0.002835	TEP-Tiled
GO:0046464	acylglycerol catabolic process	0.002835	TEP-Tiled
GO:0080148	negative regulation of response to water deprivation	0.003543	TEP-Tiled
GO:1901362	organic cyclic compound biosynthetic process	0.003586	TEP-Tiled
GO:0051775	response to redox state	0.00425	TEP-Tiled
GO:0046339	diacylglycerol metabolic process	0.00425	TEP-Tiled
GO:0006855	drug transmembrane transport	0.00425	TEP-Tiled
GO:0044249	cellular biosynthetic process	0.004261	TEP-Tiled
GO:0046503	glycerolipid catabolic process	0.004956	TEP-Tiled
GO:0015893	drug transport	0.004956	TEP-Tiled
GO:0008150	biological_process	0.005045	TEP-Tiled
GO:1901576	organic substance biosynthetic process	0.005138	TEP-Tiled
GO:0046486	glycerolipid metabolic process	0.006851	TEP-Tiled
GO:0104004	cellular response to environmental stimulus	0.007546	TEP-Tiled
GO:0071214	cellular response to abiotic stimulus	0.007546	TEP-Tiled
GO:0015918	sterol transport	0.008482	TEP-Tiled
GO:0009987	cellular process	0.008497	TEP-Tiled
GO:0044237	cellular metabolic process	0.009157	TEP-Tiled
GO:1901001	negative regulation of response to salt stress	0.009185	TEP-Tiled
GO:0034641	cellular nitrogen compound metabolic process	0.010425	TEP-Tiled
GO:0048829	root cap development	0.010591	TEP-Tiled
GO:0071704	organic substance metabolic process	0.011325	TEP-Tiled
GO:0008152	metabolic process	0.01244	TEP-Tiled

GO:0032774	RNA biosynthetic process	0.013147	TEP-Tiled
GO:0018130	heterocycle biosynthetic process	0.014533	TEP-Tiled
GO:1901259	chloroplast rRNA processing	0.014797	TEP-Tiled
GO:2001141	regulation of RNA biosynthetic process	0.015672	TEP-Tiled
GO:0010187	negative regulation of seed germination	0.017592	TEP-Tiled
GO:0019432	triglyceride biosynthetic process	0.017592	TEP-Tiled
GO:0051252	regulation of RNA metabolic process	0.018883	TEP-Tiled
GO:0010207	photosystem II assembly	0.018986	TEP-Tiled
GO:0010192	mucilage biosynthetic process	0.018986	TEP-Tiled
GO:0019438	aromatic compound biosynthetic process	0.019332	TEP-Tiled
GO:0019915	lipid storage	0.019683	TEP-Tiled
GO:0046460	neutral lipid biosynthetic process	0.019683	TEP-Tiled
GO:0046463	acylglycerol biosynthetic process	0.019683	TEP-Tiled
GO:0010025	wax biosynthetic process	0.020379	TEP-Tiled
GO:0006629	lipid metabolic process	0.02054	TEP-Tiled
GO:2000652	regulation of secondary cell wall biogenesis	0.021074	TEP-Tiled
GO:0010166	wax metabolic process	0.021074	TEP-Tiled
GO:0019219	regulation of nucleobase-containing compound metabolic process	0.021157	TEP-Tiled
GO:0006641	triglyceride metabolic process	0.021769	TEP-Tiled
GO:1901570	fatty acid derivative biosynthetic process	0.021769	TEP-Tiled
GO:0010556	regulation of macromolecule biosynthetic process	0.022065	TEP-Tiled
GO:0016126	sterol biosynthetic process	0.022464	TEP-Tiled
GO:0010191	mucilage metabolic process	0.022464	TEP-Tiled
GO:0048868	pollen tube development	0.023158	TEP-Tiled
GO:1901000	regulation of response to salt stress	0.023158	TEP-Tiled
GO:0010089	xylem development	0.024545	TEP-Tiled
GO:2000070	regulation of response to water deprivation	0.02593	TEP-Tiled
GO:0031326	regulation of cellular biosynthetic process	0.025991	TEP-Tiled
GO:0009889	regulation of biosynthetic process	0.027707	TEP-Tiled
GO:0019761	glucosinolate biosynthetic process	0.028003	TEP-Tiled
GO:0016144	S-glycoside biosynthetic process	0.028003	TEP-Tiled
GO:0019758	glycosinolate biosynthetic process	0.028003	TEP-Tiled
GO:0047484	regulation of response to osmotic stress	0.029384	TEP-Tiled
GO:0010109	regulation of photosynthesis	0.029384	TEP-Tiled
GO:0015995	chlorophyll biosynthetic process	0.030073	TEP-Tiled
GO:0048856	anatomical structure development	0.030401	TEP-Tiled
GO:1901568	fatty acid derivative metabolic process	0.030762	TEP-Tiled
GO:1903338	regulation of cell wall organization or biogenesis	0.032826	TEP-Tiled
GO:0048580	regulation of post-embryonic development	0.033762	TEP-Tiled
GO:0006779	porphyrin-containing compound biosynthetic process	0.034885	TEP-Tiled
GO:0044271	cellular nitrogen compound biosynthetic process	0.036955	TEP-Tiled
GO:0033014	tetrapyrrole biosynthetic process	0.037625	TEP-Tiled
GO:1901360	organic cyclic compound metabolic process	0.038226	TEP-Tiled
GO:0051171	regulation of nitrogen compound metabolic process	0.039126	TEP-Tiled
GO:0006694	steroid biosynthetic process	0.041039	TEP-Tiled

GO:0015994	chlorophyll metabolic process	0.042401	TEP-Tiled
GO:0009834	plant-type secondary cell wall biogenesis	0.042401	TEP-Tiled
GO:1901659	glycosyl compound biosynthetic process	0.043081	TEP-Tiled
GO:0080090	regulation of primary metabolic process	0.044226	TEP-Tiled
GO:0006518	peptide metabolic process	0.045872	TEP-Tiled
GO:0044248	cellular catabolic process	0.045971	TEP-Tiled
GO:0048519	negative regulation of biological process	0.046071	TEP-Tiled
GO:0006807	nitrogen compound metabolic process	0.047825	TEP-Tiled
GO:0016125	sterol metabolic process	0.04986	TEP-Tiled

Supplementary Table A12: GO-term enrichment analysis for the genes associated with the putatively “hard-coded” promoters. Included in this table are the GO-term, the description of the term, the p-value (all $p < 0.05$), and an indication of which model the set “hard-coded” promoters came from. Both GO-term enrichment experiments are contained in this single table.

shift_amount	feature_id	Post-knockout prob1	Pre-knockout prob1	tss_name
-0.525761141	M0119_1.02_REV_4	0.803841825	0.278080684	AT5G52040.4_Chr5_21130346_+_0
-0.525761141	M0119_1.02_REV_4	0.803841825	0.278080684	AT5G52040.3_Chr5_21130346_+_0
-0.525761141	M0119_1.02_REV_4	0.803841825	0.278080684	AT5G52040.2_Chr5_21130346_+_0
-0.525761141	M0119_1.02_REV_4	0.803841825	0.278080684	AT5G52040.1_Chr5_21130346_+_0
-0.515625035	M0119_1.02_REV_4	0.81283711	0.297212075	AT5G03700.1_Chr5_967421_-_0
-0.504363411	M0119_1.02_REV_4	0.781646375	0.277282964	AT5G59590.1_Chr5_24010616_-_0
-0.501480944	M0119_1.02_REV_4	0.758376904	0.256895961	AT5G11810.1_Chr5_3808743_+_0
-0.500727139	M0119_1.02_REV_4	0.731933861	0.231206722	AT5G11810.1_Chr5_3808742_+_0
-0.475457796	M0119_1.02_REV_4	0.727074621	0.251616825	AT1G59640.2_Chr1_21911147_-_0
-0.475457796	M0119_1.02_REV_4	0.727074621	0.251616825	AT1G59640.1_Chr1_21911147_-_0
0.460226045	M0011_1.02_REV_7	0.231872302	0.692098347	AT5G62720.2_Chr5_25191868_+_0
0.460226045	M0011_1.02_REV_7	0.231872302	0.692098347	AT5G62720.1_Chr5_25191868_+_0
-0.4578229	M1696_1.02_REV_5	0.728059863	0.270236963	AT3G16565.2_Chr3_5642734_-_0
-0.4578229	M1696_1.02_REV_5	0.728059863	0.270236963	AT3G16565.1_Chr3_5642734_-_0
0.449305754	M0011_1.02_REV_7	0.259173323	0.708479077	AT5G62720.2_Chr5_25191858_+_0
0.449305754	M0011_1.02_REV_7	0.259173323	0.708479077	AT5G62720.1_Chr5_25191858_+_0
-0.444530928	M0119_1.02_REV_4	0.893084762	0.448553834	AT5G03700.1_Chr5_967404_-_0
-0.442292868	M0119_1.02_REV_4	0.716685211	0.274392343	AT5G28640.1_Chr5_10649583_-_0
-0.431771759	M0119_1.02_REV_4	0.905423334	0.473651575	AT5G52040.4_Chr5_21130352_+_0
-0.431771759	M0119_1.02_REV_4	0.905423334	0.473651575	AT5G52040.3_Chr5_21130352_+_0
-0.431771759	M0119_1.02_REV_4	0.905423334	0.473651575	AT5G52040.2_Chr5_21130352_+_0
-0.431771759	M0119_1.02_REV_4	0.905423334	0.473651575	AT5G52040.1_Chr5_21130352_+_0
-0.430064813	M0119_1.02_REV_4	0.646933728	0.216868915	AT2G23030.1_Chr2_9806673_-_0
0.428419104	M1309_1.02_REV_5	0.245943985	0.674363089	AT2G45170.2_Chr2_18624291_+_0
0.428419104	M1309_1.02_REV_5	0.245943985	0.674363089	AT2G45170.1_Chr2_18624291_+_0
-0.426744547	M0119_1.02_REV_4	0.570128701	0.143384154	AT1G59640.2_Chr1_21911140_-_0
-0.426744547	M0119_1.02_REV_4	0.570128701	0.143384154	AT1G59640.1_Chr1_21911140_-_0
0.426717636	M1309_1.02_REV_5	0.335557429	0.762275065	AT2G45170.2_Chr2_18624284_+_0
0.426717636	M1309_1.02_REV_5	0.335557429	0.762275065	AT2G45170.1_Chr2_18624284_+_0
-0.426571534	M0119_1.02_REV_4	0.4749933	0.048421766	AT5G15190.2_Chr5_4933562_-_0
-0.426571534	M0119_1.02_REV_4	0.4749933	0.048421766	AT5G15190.1_Chr5_4933562_-_0
-0.414274796	M2347_1.02_REV_2	0.888924088	0.474649292	AT3G14100.1_Chr3_4672952_+_0
-0.404004807	M0119_1.02_REV_4	0.827577491	0.423572685	AT3G56380.1_Chr3_20905295_+_0

Supplementary Table A13: The results from our in silico knockout experiments for the TEP-ROE model, sorted by the absolute value of the change in probability of expression in shoot ($|\text{shift_amount}|$). The “feature_id” column contains the name of the feature that was zeroed out for the experiment, while the “tss_name” column contains the name of the TSS that crossed the model’s decision boundary. “Pre-knockout prob1” and “Post-knockout prob1” are the pre- and post- knockout probabilities of expression in shoot for the given TSS ($1 - \text{prob_value}$ gives probability of expression in root for the given TSS).

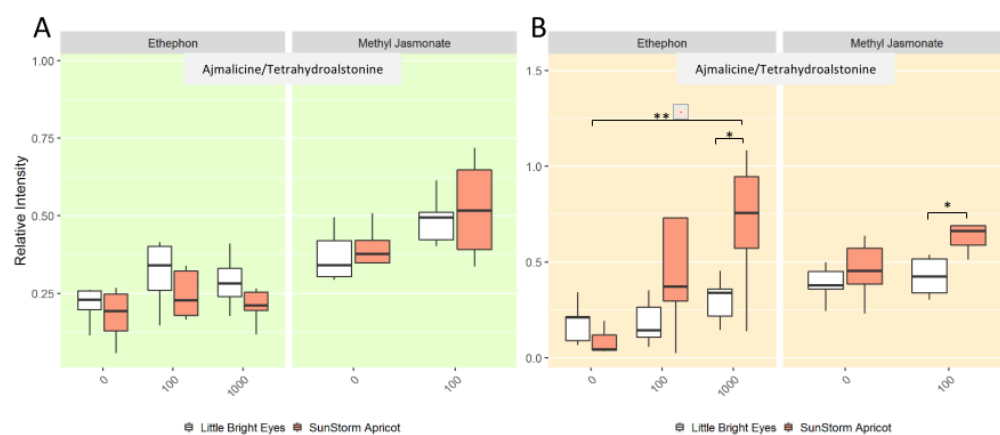
shift_amount	feature_id	Post-knockout prob1	Pre-knockout prob1	tss_name
-0.525761141	M0119_1.02_REV_4	0.80384183	0.27808068	AT5G52040.4_Chr5_21130346_+_0
-0.525761141	M0119_1.02_REV_4	0.80384183	0.27808068	AT5G52040.3_Chr5_21130346_+_0
-0.525761141	M0119_1.02_REV_4	0.80384183	0.27808068	AT5G52040.2_Chr5_21130346_+_0
-0.525761141	M0119_1.02_REV_4	0.80384183	0.27808068	AT5G52040.1_Chr5_21130346_+_0
-0.515625035	M0119_1.02_REV_4	0.81283711	0.29721208	AT5G03700.1_Chr5_967421_-_0
-0.504363411	M0119_1.02_REV_4	0.78164638	0.27728296	AT5G59590.1_Chr5_24010616_-_0
-0.501480944	M0119_1.02_REV_4	0.7583769	0.25689596	AT5G11810.1_Chr5_3808743_+_0
-0.500727139	M0119_1.02_REV_4	0.73193386	0.23120672	AT5G11810.1_Chr5_3808742_+_0
-0.475457796	M0119_1.02_REV_4	0.72707462	0.25161682	AT1G59640.2_Chr1_21911147_-_0
-0.475457796	M0119_1.02_REV_4	0.72707462	0.25161682	AT1G59640.1_Chr1_21911147_-_0
0.460226045	M0011_1.02_REV_7	0.2318723	0.69209835	AT5G62720.2_Chr5_25191868_+_0
0.460226045	M0011_1.02_REV_7	0.2318723	0.69209835	AT5G62720.1_Chr5_25191868_+_0
-0.4578229	M1696_1.02_REV_5	0.72805986	0.27023696	AT3G16565.2_Chr3_5642734_-_0
-0.4578229	M1696_1.02_REV_5	0.72805986	0.27023696	AT3G16565.1_Chr3_5642734_-_0
0.449305754	M0011_1.02_REV_7	0.25917332	0.70847908	AT5G62720.2_Chr5_25191858_+_0
0.449305754	M0011_1.02_REV_7	0.25917332	0.70847908	AT5G62720.1_Chr5_25191858_+_0
-0.444530928	M0119_1.02_REV_4	0.89308476	0.44855383	AT5G03700.1_Chr5_967404_-_0
-0.442292868	M0119_1.02_REV_4	0.71668521	0.27439234	AT5G28640.1_Chr5_10649583_-_0
-0.431771759	M0119_1.02_REV_4	0.90542333	0.47365158	AT5G52040.4_Chr5_21130352_+_0
-0.431771759	M0119_1.02_REV_4	0.90542333	0.47365158	AT5G52040.3_Chr5_21130352_+_0
-0.431771759	M0119_1.02_REV_4	0.90542333	0.47365158	AT5G52040.2_Chr5_21130352_+_0
-0.431771759	M0119_1.02_REV_4	0.90542333	0.47365158	AT5G52040.1_Chr5_21130352_+_0
-0.430064813	M0119_1.02_REV_4	0.64693373	0.21686892	AT2G23030.1_Chr2_9806673_-_0
0.428419104	M1309_1.02_REV_5	0.24594399	0.67436309	AT2G45170.2_Chr2_18624291_+_0
0.428419104	M1309_1.02_REV_5	0.24594399	0.67436309	AT2G45170.1_Chr2_18624291_+_0
-0.426744547	M0119_1.02_REV_4	0.5701287	0.14338415	AT1G59640.2_Chr1_21911140_-_0
-0.426744547	M0119_1.02_REV_4	0.5701287	0.14338415	AT1G59640.1_Chr1_21911140_-_0
0.426717636	M1309_1.02_REV_5	0.33555743	0.76227506	AT2G45170.2_Chr2_18624284_+_0
0.426717636	M1309_1.02_REV_5	0.33555743	0.76227506	AT2G45170.1_Chr2_18624284_+_0
-0.426571534	M0119_1.02_REV_4	0.4749933	0.04842177	AT5G15190.2_Chr5_4933562_-_0
-0.426571534	M0119_1.02_REV_4	0.4749933	0.04842177	AT5G15190.1_Chr5_4933562_-_0
-0.414274796	M2347_1.02_REV_2	0.88892409	0.47464929	AT3G14100.1_Chr3_4672952_+_0
-0.404004807	M0119_1.02_REV_4	0.82757749	0.42357268	AT3G56380.1_Chr3_20905295_+_0

Supplementary Table A14: The results from our in silico knockout experiments for the TEP-Tiled model, sorted by the absolute value of the change in probability of expression in shoot ($|\text{shift_amount}|$). The “feature_id” column contains the name of the feature that was zeroed out for the experiment, while the “tss_name” column contains the name of the TSS that crossed the model’s decision boundary. “Pre-knockout prob1” and “Post-knockout prob1” are the pre- and post- knockout probabilities of expression in shoot for the given TSS ($1 - \text{prob_value}$ gives probability of expression in root for the given TSS).

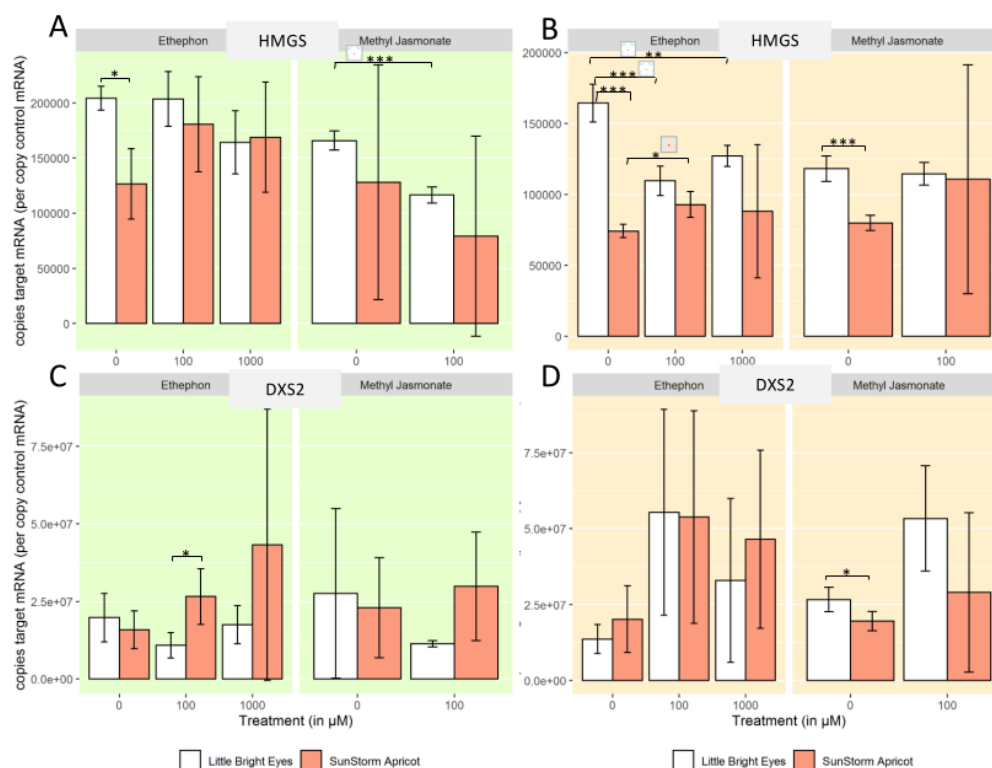
Supplementary Datasets A1 and A2 will be made available online at:

<https://oregonstate.box.com/s/15t3yhatp2ok2hjln0nopn9ahb6qqnw>

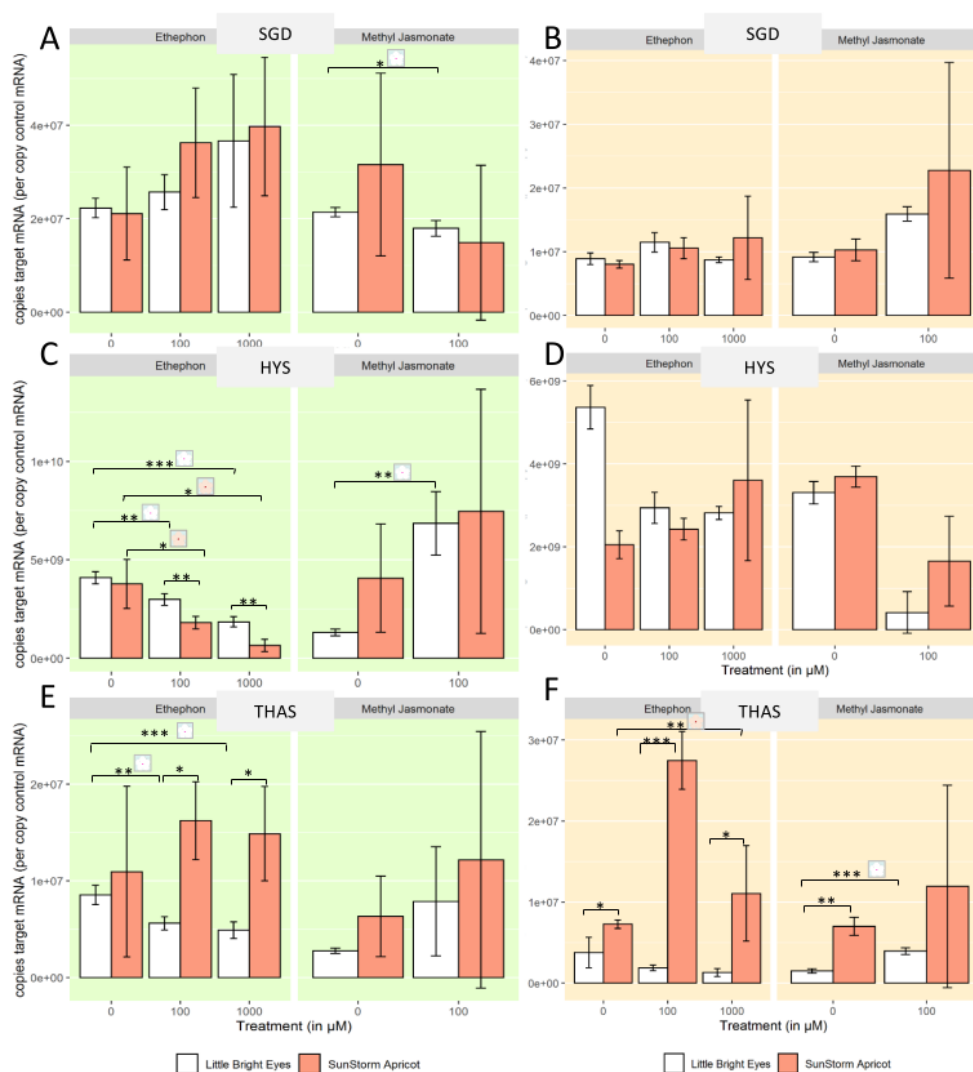
Appendix B: Supplementary Materials for Chapter 3



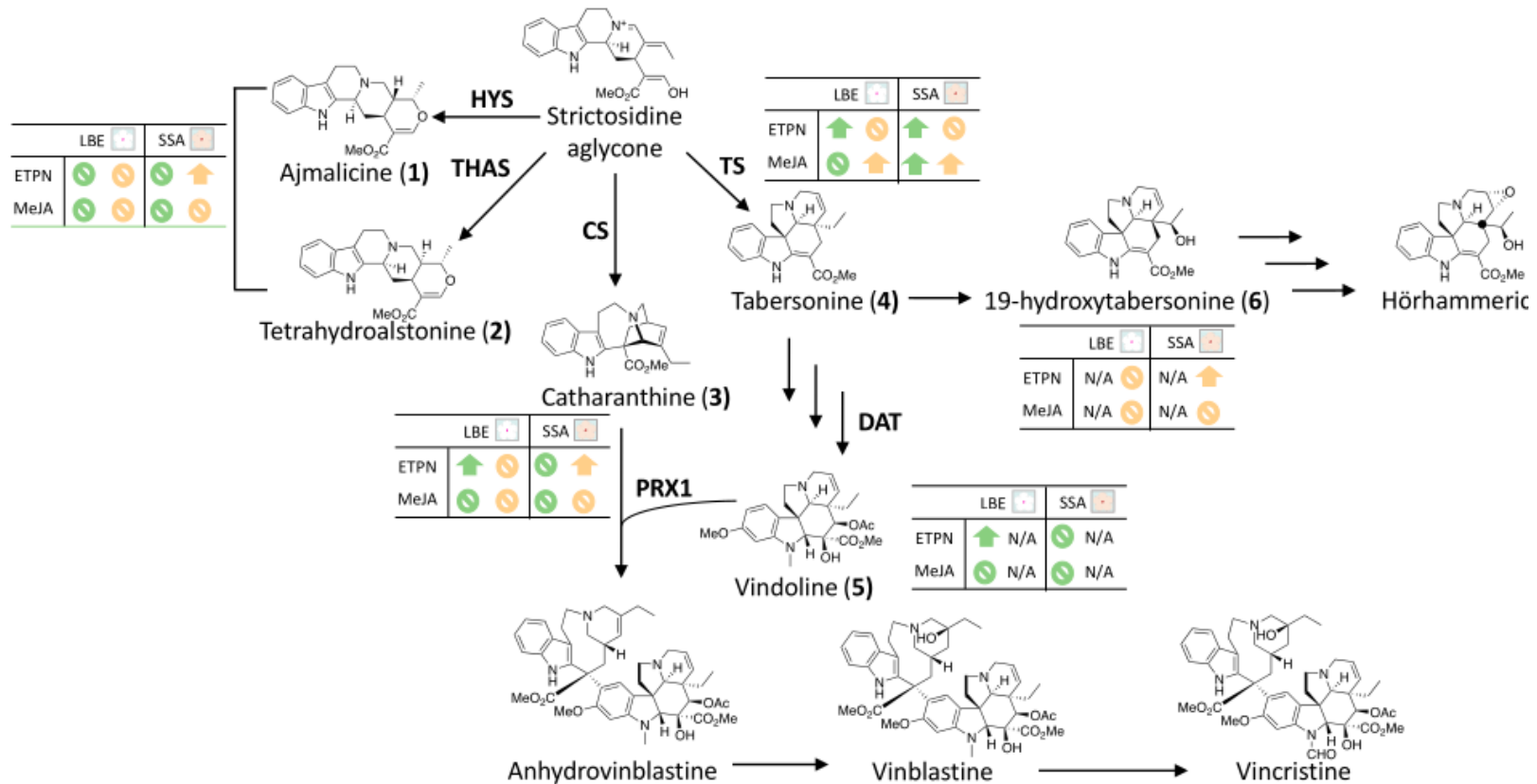
Supplementary Figure B1: The concentration of ajmalicine/tetrahydroalstonine relative to our internal standard (ajmaline) appears to increase upon hormone treatment. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; all represented statistics are from Welch's t-test post-hoc analyses. Significance markers with a white flower represent treatment differences in LBE, while those with a peach flower represent treatment differences in SSA. (A) Peak intensity relative to internal standard in shoots; (B) peak intensity relative to internal standard in roots.



Supplementary Figure B2: Expression of genes from pathways upstream of the TIA pathway appear to be transcriptionally regulated by hormone treatment. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; all represented statistics are from Welch's t-test post-hoc analyses. Significance markers with a white flower represent treatment differences in LBE, while those with a peach flower represent treatment differences in SSA. (A) HMGS/MVA pathway in shoots (B) HMGS/MVA pathway in roots (C) DXS2/MEP pathway in shoots (D) DXS2/MEP pathway in roots.



Supplementary Figure B3: Expression of some key enzymes in the TIA pathway are transcriptionally regulated upon phytohormone treatment. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; all represented statistics are from Welch's t-test post-hoc analyses. Significance markers with a white flower represent treatment differences in LBE, while those with a peach flower represent treatment differences in SSA. (A) SGD expression in shoots (B) SGD expression in roots. (C) HYS expression in shoots (D) HYS expression in roots (E) THAS expression in shoots (F) THAS expression in roots.



Supplementary Figure B4 Summary of changes in metabolite levels in response to phytohormone treatment. Colored arrows pointing up reflect an increase in either concentration or peak intensity relative to ajmaline (internal standard). A symbol reflects no change. The color of the arrow or symbol reflects the tissue: green = shoot, tan = root.

Variety	Genome	Transcriptome	Metabolome	Source
SunStorm Apricot	Yes	No	Yes	Kellner 2015; Chung 2011; Magnotta 2006
Little Bright Eye	No	Yes	No	Góngora-Castillo 2012
Prabal	No	Yes	No	Verma 2014
63 other varieties	No	No	Yes	Chung 2011
49 other varieties	No	No	Yes	Magnotta 2006

Supplementary Table B1 The 'omics data available for selected varieties of *Catharanthus roseus*.

(µg/g wet wt)	Little Bright Eye					SunStorm Apricot				
	0 µM ETPN	100 µM ETPN	1 mM ETPN	0 µM MeJA	100 µM MeJA	0 µM ETPN	100 µM ETPN	1 mM ETPN	0 µM MeJA	100 µM MeJA
Tabersonine										
Shoots	3.64 ± 3.77	18.22 ± 8.55	21.35 ± 6.77	29.73 ± 7.91	35.65 ± 18.29	30.03 ± 15.77	108.66 ± 83.40	94.38 ± 31.16	63.02 ± 25.02	133.40 ± 43.37
Roots	70.40 ± 53.35	69.93 ± 72.42	100.07 ± 46.39	105.26 ± 28.84	156.99 ± 33.52	33.59 ± 26.85	108.57 ± 98.05	147.04 ± 64.93	135.07 ± 77.26	217.91 ± 67.9
Catharanthine										
Shoots	174.45 ± 189.09	601.37 ± 181.58	595.06 ± 171.18	545.68 ± 75.48	637.64 ± 165.61	224.03 ± 134.52	287.11 ± 196.48	224.77 ± 125.45	352.43 ± 121.26	429.60 ± 50.29
Roots	13.28 ± 10.93	11.57 ± 11.48	17.26 ± 18.26	30.38 ± 15.88	28.91 ± 13.69	4.30 ± 3.19	16.02 ± 16.68	22.13 ± 13.78	11.85 ± 8.96	15.33 ± 11.84
Vindoline										
Shoots	25.03 ± 28.64	81.39 ± 27.24	77.51 ± 19.08	75.66 ± 10.55	81.60 ± 26.18	1.76 ± 0.99	1.68 ± 1.15	1.26 ± 0.76	2.82 ± 1.08	2.21 ± 0.53

Supplementary Table B2 Mean alkaloid concentrations detected (shown ± standard deviation) in each tissue under each treatment condition.

All treatments:

Shoots:

	Treatment	Variety	Interaction
Tetrahydroalstonine	5.35e-05 ***	0.483	0.730
Ajmalicine	0.00155 **	0.39483	0.79135
Catharanthine (MRM)	2.67e-05 ***	1.83e-06 ***	0.0128 *
Tabersonine (MRM)	0.000234 ***	1.31e-09 ***	0.032702 *
Vindoline	0.00026 ***	< 2e-16 ***	0.00029 ***

Signif. codes: 0 '***' | 0.01 '**' | 0.05 '*' | 0.1 '.'

Roots:

	Treatment	Variety	Interaction
Tetrahydroalstonine	0.009913 **	0.000816 ***	0.096387 .
Ajmalicine	0.11075	0.00126 **	0.16752
Catharanthine (MRM)	0.203	0.108	0.162
Tabersonine (MRM)	0.000198 ***	0.048320 *	0.495856

Signif. codes: 0 '***' | 0.01 '**' | 0.05 '*' | 0.1 '.'

Supplementary Table B3 p-values for absolute concentrations of alkaloids from ANOVA.

A

Variety comparisons

Shoots:

	Eth 0 μ M (control)	Eth 100 μ M	Eth 1mM	MeJA 0 μ M (control)	MeJA 100 μ M
Catharanthine	0.6133	0.01656 *	0.001983 **	0.009982 **	0.02625 *
Tabersonine	0.008407 **	0.0449 *	0.001858 **	0.02097 *	0.001604 **
Vindoline	0.1033	0.000813 ***	0.0001867 ***	1.143e-05 ***	0.0006946 ***

Roots:

	Eth 0 μ M (control)	Eth 100 μ M	Eth 1mM	MeJA 0 μ M (control)	MeJA 100 μ M
Catharanthine	0.1388	0.6362	0.6143	0.03932 *	0.0954 .
Tabersonine	0.2476	0.4871	0.1831	0.3548	0.08791 .

B

Treatment effects

Shoots:

		Catharanthine	Tabersonine	Vindoline
LBE	Eth (0:100)	0.002572 **	0.006743 **	0.00582 **
	Eth (0:1000)	0.002411 **	0.0005549 ***	0.004941 **
	MeJA (0:100)	0.2558	0.4912	0.623
SSA	Eth (0:100)	0.5329	0.06897 .	0.893
	Eth (0:1000)	0.9923	0.002391 **	0.3502
	MeJA (0:100)	0.1951	0.008793 **	0.2516

Roots

		Catharanthine	Tabersonine
LBE	Eth (0:100)	0.8036	0.9903
	Eth (0:1000)	0.6636	0.341
	MeJA (0:100)	0.8765	0.01715 *
SSA	Eth (0:100)	0.2052	0.1716
	Eth (0:1000)	0.0239 *	0.009398 **
	MeJA (0:100)	0.5842	0.06147 .

Supplementary Table B4 p-values for absolute concentrations of alkaloids analyses from Welch's t-test pairwise comparisons post-hoc. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; (A) p-values for pairwise comparisons between varieties; (B) p-values for pairwise comparisons of treatments for each variety.

A Variety comparisons

Shoots:

	Eth 0 μ M (control)	Eth 100 μ M	Eth 1mM	MeJA 0 μ M (control)	MeJA 100 μ M
Ajmalicine/ Tetrahydroalstonine	0.2879	0.8004	0.318	0.8737	0.4245
Catharanthine	0.0491 *	0.0111 *	0.003289 **	0.0105 *	0.0341 *
Tabersonine	0.0252 *	0.03674 *	0.002858 **	0.02719 *	0.001498 **
Vindoline	0.00638 **	4.117e-5 ***	1.997e-6 ***	8.102e-8 ***	2.591e-5 ***

Roots:

	Eth 0 μ M (control)	Eth 100 μ M	Eth 1mM	MeJA 0 μ M (control)	MeJA 100 μ M
Ajmalicine/ Tetrahydroalstonine	0.2417	0.1926	0.03341 *	0.4341	0.01938 *
Unknown 353	0.169	0.8337	0.1536	0.6517	0.3356
Catharanthine	0.2881	0.6361	0.616	0.04065 *	0.1079
Tabersonine	0.1592	0.4952	0.1926	0.385	0.07137 .

B Treatment effects

Shoots:

		Ajmalicine/ Tetrahydroalstonine	Catharanthine	Tabersonine	Vindoline
LBE	Eth (0:100)	0.2536	0.1021	0.08171 .	0.4615
	Eth (0:1000)	0.4328	0.1353	0.01179 *	0.3351
	MeJA (0:100)	0.3265	0.353	0.5384	0.9569
SSA	Eth (0:100)	0.2455	0.4596	0.05323 .	0.97
	Eth (0:1000)	0.3365	0.7701	0.003449 **	0.5153
	MeJA (0:100)	0.08547 .	0.194	0.008402 **	0.2849

Roots

		Ajmalicine/ Tetrahydroalstonine	Catharanthine	Tabersonine	Uncharacterized m/z = 353
LBE	Eth (0:100)	0.9717	0.8488	0.9638	0.7975
	Eth (0:1000)	0.1226	0.617	0.2827	0.2184
	MeJA (0:100)	0.5281	0.953	0.01491 *	0.6319
SSA	Eth (0:100)	0.1265	0.2159	0.1933	0.06465 .
	Eth (0:1000)	0.005827 **	0.02762 *	0.01057 *	0.001765 **
	MeJA (0:100)	0.05296 .	0.5081	0.04527 *	0.9346

Supplementary Table B5 p-values for peak intensity of alkaloids relative to internal standard (ajmaline) from Welch's t-test pairwise comparisons post-hoc. * denotes a p-value ≤ 0.05 ; ** denotes a p-value ≤ 0.01 ; *** denotes a p-value ≤ 0.001 ; red boxes are around the uncharacterized m/z = 353 peak. (A) p-values for pairwise comparisons between varieties; (B) p-values for pairwise comparisons of treatments for each variety.

qPCR stats (normalized)

Shoots:

	Treatment	Variety	Interaction
SGD	1.09e-06 ***	0.5798	0.0305 *
CS	1.32e-07 ***	3.52e-12 ***	0.00152 **
TS	1.71e-13 ***	0.00928 **	1.73e-06 ***
HYS	5.18e-09 ***	0.0017 **	3.84e-08 ***
THAS	0.028701 *	0.000266 ***	0.006317 **
DAT	0.00329 **	0.18643	0.01432 *
ORCA2	2.73e-08 ***	0.0763 .	0.7814
ORCA3	0.0130 *	4.44e-11 ***	0.0378 *
PRX1	0.000301 ***	0.465309	0.014337 *
DXS2	0.122	0.623	0.518
HGMS	5.26e-05 ***	2.80e-06 ***	0.232

Signif. codes: 0 '***' | 0.01 '**' | 0.05 '*' | 0.1 '.'

Roots:

	Treatment	Variety	Interaction
SGD	2.57e-06 ***	0.00299 **	0.00122 **
CS	5.49e-08 ***	1.05e-12 ***	0.00225 **
TS	1.98e-10 ***	3.02e-07 ***	1.21e-05 ***
HYS	< 2e-16 ***	0.0662 .	4.62e-16 ***
THAS	4.93e-10 ***	3.64e-15 ***	3.39e-10 ***
ORCA2	7.42e-12 ***	1.30e-05 ***	9.96e-13 ***
ORCA3	5.08e-08 ***	0.005728 **	0.000266 ***
PRX1	1.54e-11 ***	2.86e-08 ***	0.00062 ***
DXS2	0.00164 **	0.88019	0.26284
HGMS	1.52e-09 ***	3.79e-10 ***	0.000998 ***

Signif. codes: 0 '***' | 0.01 '**' | 0.05 '*' | 0.1 '.'

Supplementary Table B6 p-values for normalized RT-qPCR from ANOVA.

Shoots (at 95% confidence):

	SGD	CS	TS	HYS	THAS
LBE					
E0 vs E1	0.1828	0.0312 *	0.6502	0.00217 **	0.0041 **
E0 vs E4	0.1355	0.0076 **	0.1804	3.777e-05 ***	0.0016 **
M0 vs M1	0.0166 *	7.517e-05 ***	0.0260 *	0.00601 **	0.1682
SSA					
E0 vs E1	0.0976 .	0.1282	0.2501	0.0473 *	0.3360
E0 vs E4	0.0885 .	0.1895	0.1257	0.0125 *	0.4746
M0 vs M1	0.2403	0.4377	0.3751	0.3716	0.4516

	DAT	ORCA2	ORCA3	PRX1	DXS2	HMGS
LBE						
E0 vs E1	0.2345	0.5960	7.656e-05 ***	0.2316	0.1039	0.9682
E0 vs E4	0.0221 *	0.0017 **	2.460e-05 ***	0.7947	0.6560	0.0625 .
M0 vs M1	0.0863 .	0.1268	0.0059 **	0.0116 *	0.3203	1.380e-04 ***
SSA						
E0 vs E1	0.8151	0.0484 *	0.9285	0.4976	0.1209	0.0944 .
E0 vs E4	0.0433 *	0.0330 *	0.9417	0.0125 *	0.3004	0.2115
M0 vs M1	0.9716	0.4330	0.4768	0.7450	0.5855	0.5112

Varieties

	Eth 0µM (control)	Eth 100µM	Eth 1mM	MeJA 0µM (control)	MeJA 100µM
SGD	0.8267	0.1690	0.7763	0.3730	0.7357
CS	0.0772 .	0.0282 *	0.0093 **	0.0454 *	0.3712
TS	0.5309	0.0917 .	0.4365	0.4203	0.9399
HYS	0.6547	0.0017 **	0.0013 **	0.1385	0.8606
THAS	0.6229	0.0120 *	0.0246 *	0.1875	0.5828
DAT	0.4679	0.0252 *	0.4589	0.2783	0.4920
ORCA2	0.7190	0.0045 **	0.3258	0.7555	0.7308
ORCA3	0.0102 *	0.00018 ***	0.00784 **	0.0975 .	0.1610
PRX1	0.0493 *	0.149645	0.0727 .	0.9298	0.9335
DXS2	0.4954	0.0309 *	0.3236	0.7836	0.1249
HMGS	0.01204 *	0.4017	0.8788	0.5297	0.4735

Supplementary Table B7 p-values for normalized RT-qPCR in shoots from Welch's t-test pairwise comparisons post-hoc.

Roots (at 95% confidence):

Treatment

	SGD	CS	TS	THAS	HYS
LBE					
E0 vs E1	0.0352 *	0.1404	0.0102 *	0.1318	0.00044 ***
E0 vs E4	0.7443	1.0464e-04 ***	0.1158	0.0736 .	0.00128 **
M0 vs M1	0.00014 ***	0.00255 **	0.03599 *	0.00013 ***	0.00025 ***
SSA					
E0 vs E1	0.0482 *	0.0026 **	0.0908 .	0.0012 **	0.1311
E0 vs E4	0.2969	0.2932	0.1919	0.2899	0.2070
M0 vs M1	0.2366	0.8262	0.4141	0.4890	0.0294 *

	ORCA2	ORCA3	PRX1	DXS2	HMGS
LBE					
E0 vs E1	0.0361 *	0.00021 ***	0.1349	0.0893 .	0.00078 ***
E0 vs E4	0.0022 **	0.00016 ***	0.0248 *	0.2474	0.00528 **
M0 vs M1	8.820e-05 ***	0.00012 ***	0.0016 **	0.0509 .	0.5620
SSA					
E0 vs E1	0.00051 ***	0.00062 ***	0.00219 **	0.1488	0.0180 *
E0 vs E4	0.0706 .	0.9302	0.1439	0.1717	0.5944
M0 vs M1	0.1769	0.1621	0.8028	0.5235	0.5005

Variety

	Eth 0 μ M (control)	Eth 100 μ M	Eth 1mM	MeJA 0 μ M (control)	MeJA 100 μ M
SGD	0.1741	0.4535	0.3710	0.3053	0.4782
CS	0.000278 ***	0.03403 *	0.0037 **	0.00052 ***	0.0195 *
TS	0.5280	0.7979	0.3571	0.0014 **	0.2743
HYS	0.00011 ***	0.0698 .	0.4782	0.0835 .	0.1028
THAS	0.02918 *	0.00067 ***	0.0447 *	0.00156 **	0.2910
ORCA2	0.00079 ***	0.0166 *	0.3216	0.0013 **	0.8774
ORCA3	0.00016 ***	0.00073 ***	0.1297	0.4201	0.6366
PRX1	0.00033 ***	0.00460 **	0.2854	0.0012 **	0.3041
DXS2	0.334845	0.95096	0.5236	0.0344 *	0.1805
HMGS	0.000305 ***	0.0507 .	0.19485	0.000798 ***	0.9302

Supplementary Table B8 p-values for normalized RT-qPCR in roots from Welch's t-test pairwise comparisons post-hoc.

Gene	Forward primer	Reverse primer	Genbank Accession	Product length
strictosidine β -D-glucosidase (SGD)	GGAGGATCTGCTTATCAGTGTG	TGGCTGGATATCGGTTTGT	AF112888.1	91nt
catharanthine synthase (CS)	CTCCTGGCGGGATGAATAAC	GGAAACCAGGGTAACCAACA	MF770512	139nt
tabersonine synthase (TS)	AGATGCTCCTGGTGAAATG	CAACCATGGAAATCAGCAACC	MF770513	104nt
heteroyohimbine synthase (HYS)	AGCAATCAGATTTGCCAAGG	GGTTACTGTTGAGCAAGAAAG	KU865325.1	120nt
class III peroxidase 1 (PRX1)	TTCCATTGGGAAGAAGAGATGG	TTAGGAGGGCACTTGTGTTG	AM236087.1	96nt
tetrahydroalstonine synthase (THAS)	TTTAGGTGCACCAGAAATGC	TTCCTTCACTACTCCAGCAG	KM524258.1	97nt
6-17-O-acetylvindoline O-acetyltransferase (DAT)	AGAGACCTAGTCCTTCCCAAAC	AAAGCAACCGCCAAACCT	AF053307.1	101nt
ORCA2	CGTTTCAACTCCGTGGTTCTA	ATCGGCGTCTAGGACTTACT	AJ238740.1	92nt
ORCA3	CCAGCTCGGAATTGACTTCTAC	GGCTACCGGTTTCTGTATTT	EU072424.1	96nt
1-deoxy-D-xylulose-5-phosphate synthase (DSX2)	CGAATGGGGTTTTAATGAGG	GAGTGGAGAAATGGGAGGAA	DQ848672.1	61nt
hydroxymethylglutaryl-CoA synthase (HMGS)	CTCAATGAGTATGACGGCAGTT	AGACGACCAATTTGCTTTGG	JF739871.1	72nt
RPS9 (reference gene)	TCAGTTTCTACCGGAACATG	GCTTCAACTCTGCATCCAATC	AJ749993	84nt

Supplementary Table B9 qPCR primers; CS and TS primer sequences were obtained from Caputiet *al*,2018. DSX2 and HMGS primer sequences were obtained from Zhang *et al*,2012.

