Machine learning and plant lncRNAs

Using machine learning to predict long non-coding RNAs and exploring their evolutionary patterns and prevalence in plant transcriptomes

By Caitlin M.A. Simopoulos, BSc, MBinf

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University (2019) Doctor of Philosophy
Hamilton, Ontario (Department of Biology)

TITLE: Using machine learning to predict long non-coding RNAs and exploring their evolutionary patterns and prevalence in plant transcriptomes
AUTHOR: Caitlin M.A. SIMOPOULOS BSc, MBinf (McMaster University)
SUPERVISORS: Dr Elizabeth A. WERETILNYK and Dr G. Brian GOLDING
NUMBER OF PAGES: xv, 140

Abstract

Long non-protein coding RNAs (lncRNAs) represent a diverse and enigmatic classification of RNA. With roles associated with development and stress responses, these non-coding gene regulators are essential, and yet remain understudied in plants. Thus far, of just over 430 experimentally validated lncRNAs, only 13 are derived from plant systems and many of which do not meet the classic criteria of the RNA class. Without a solid definition of what makes a lncRNA, and few empirically validated transcripts, methods currently available for prediction fall short. To address this deficiency in lncRNA research, we constructed and applied a machine learning-based lncRNA prediction protocol that does not impose predefined rules, and utilises only experimentally confirmed lncRNAs in its training datasets.

Through model evaluation, we found that our novel lncRNA prediction tool had an estimated accuracy of over 96%. In a study that predicted lncRNAs from transcriptomes of evolutionary diverse plant species, we determined that molecular features of lncRNAs display different phylogenetic signal patterns compared to protein-coding genes. Additionally, our analyses suggested that stress-resistant species express fewer lncRNAs than more stress sensitive species. To expand on these results, we used the prediction tool in concert with a transcriptomic study of two natural accessions of the drought tolerant species *Eutrema salsugineum.* Previously reported to show little physiological differences in a first drought, but differ significantly in a second, we instead demonstrated that the two ecotypes displayed vastly different transcriptomic responses, including the expression of lncRNAs, to a first and second drought treatment. In conclusion, the prediction tool can be applied to studies to further our knowledge of lncRNA evolution and as an additional tool in classic transcriptomic studies. The suggested importance of lncRNAs in drought resistance, and evidence of expression in two natural *E. salsugineum* accessions, merits further studies on the molecular and evolutionary mechanisms of these putatively regulatory transcripts.

Acknowledgements

To my supervisors Dr Elizabeth Weretilnyk and Dr Brian Golding, thank you for the never-ending patience and support that you both have given me these past years. I am so fortunate to have had supervisors who welcomed all questions and who always encouraged me. Your mentoring fostered my love for, and confidence in, scientific research. Thank you to Dr Robin Cameron, my final supervisory committee member, for guiding my work and helping me grow professionally. Dr Peter Summers, thank you for your help and input throughout this part of my academic career. Your support did not go unnoticed. I am so lucky to have had four amazing and approachable mentors and to have been surrounded by strong leaders during my PhD experience.

Thank you to my external examiner, Dr Maheshi Dassanayake, for your invaluable feedback which contributed to the bettering of this thesis.

Thank you to all my peers at McMaster, but especially my colleagues in the Weretilnyk and Golding labs, for your comic relief and for accepting your fate as my “rubber ducks”. It has been rewarding working beside you all, and I already miss seeing you everyday. Daniella, our coffee walks were integral to surviving this chapter of my life. Thank you for always being there.

Mom and Dad, thank you for the love, hugs of support, millions of drives up to Guelph these past ten years, and especially for being there during the “three phone calls a day” weeks. I would not be who I am today without your encouragement. You are the best cheerleaders a daughter could ask for.

Luke, can you believe it? We did it! Imagine telling our twelve year old selves that we’d watch each other finish our PhDs? You are my favourite human and my very best friend. Thank you for your relentless support even when I did not deserve it. I truly don’t think I would have made it through this long and often lonely journey without you by my side.

Finally, Lucie, Earl and Sophia, I am eternally grateful for your unwavering love and companionship. There is no way to articulate how much I needed you three, but I hope my hugs and extra treats helped a bit.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Caitlin M.A. Simopoulos, declare that this thesis titled, “Using machine learning to predict long non-coding RNAs and exploring their evolutionary patterns and prevalence in plant transcriptomes” and the work presented in it are my own. I confirm that:

- Chapter 1 – I completed a literature search and wrote the manuscript that introduces the thesis. G. B. Golding and E. A. Weretilnyk contributed to the editing of the manuscript.

- Chapter 2 – I designed the experiment, compiled datasets, constructed and evaluated all models, and wrote the manuscript. G. B. Golding and E. A. Weretilnyk supervised the analysis and contributed to the editing of the manuscript. This chapter is published in BMC genomics.

- Chapter 3 – I designed the experiment, compiled datasets, trimmed and aligned RNA sequencing reads to appropriate genomes, predicted lncRNAs, completed all analyses and wrote the manuscript. G. B. Golding and E. A. Weretilnyk supervised the analysis and contributed to the editing of the manuscript. This chapter is submitted for publishing in G3.

- Chapter 4 – I completed all bioinformatic analyses described in the manuscript including read mapping, transcript assembly, novel transcript identification, lncRNA prediction, multivariate analyses and network analysis, and co-authored the manuscript with E. A. Weretilnyk and M. MacLeod. The original experiment was designed by M. MacLeod and E. A. Weretilnyk. M. MacLeod grew the *Eutrema salsugineum* plants used in the experiment and extracted RNA from 16 plants for RNA sequencing. S. Irani extracted RNA from 15 more plants for RNA sequencing and performed the RT-qPCR experiment. W. Sung and M. Champigny performed preliminary data analyses that I advanced with additional data and broadened computational tools. P. Summers contributed significantly to original experimental design and initial manuscript draft. G. B. Golding and E. A. Weretilnyk supervised the analyses and revision of the manuscript. This chapter is formatted for submission to BMC Genomics.

- Chapter 5 – I completed a literature search and wrote the manuscript that discusses the results of the thesis. G. B. Golding and E. A. Weretilnyk contributed to the editing of the manuscript.

Acronyms

ABA abscisic acid

ASCO-lncRNA alternative Splicing Competitor-lncRNA

ASL Antisense Long

AUC area under the curve

BAUM By Adaptive Unique Mapping

BC1 Brain Cytoplasmic RNA 1

bZIP basic leucine zipper

CAGE cap-analysis gene expression

CCAT-1 L colorectal cancer-associated transcript 1-long isoform

CER1 Cerberus 1

circRNA circular RNA

CNCI Coding-Non-Coding Index

COLDAIR cold assisted intronic noncoding RNA

COLDWRAP cold of winter-induced noncoding RNA from the promoter

CREMA Classifying RNA by Ensemble Machine learning Algorithm

DEG differentially expressed gene

DEWAX DECREASE WAX BIOSYNTHESIS

DHS DROUGHT HYPERSENSITIVE

DRIR drought induced lncRNA

EMSA Electrophoretic Mobility Shift Assay

ENCODE ENCyclopedia Of DNA Elements

ENOD40 early nodulin-40

eRNA enhancer RNA

FANTOM Functional ANnoTation Of the Mammalian genome

FDR false discovery rate

FISH fluorescent *in situ* hybridization

FLC FLOWERING LOCUS C

FPKM fragment per kilobase per million mapped reads

FTSW fraction of transpirable soil water

GO gene ontology

HAR Human-Accelerated Region

HID1 hidden treasure 1

HNRNPU heterogeneous nuclear ribonucleoprotein U

HOTAIR HOX transcript anitisense RNA

HULC highly upregulated in liver cancer

IGF2 insulin-like growth factor

IPS1 Induced by Phosphate Starvation 1

IPS2 Induced by Phosphate Starvation 2

KCR1 Very-long-chain 3-oxoacyl-CoA reductase

KCS2 3-ketoacyl-CoA synthase

LAIR LRK Antisense Intergenic RNA

LBR lamin B receptor

lncRNA long non-protein coding RNA

Lnx3 Ligand of numb-protein X 3

MCC Matthews correlation coefficient

miPEP miRNA-encoded peptide

miRNA microRNA

NAT natural antisense transcript

NAT lncRNA natural antisense transcript lncRNA

ncRNA non-protein coding RNA

NORAD noncoding RNA activated by DNA damage

ORF open reading frame

PCA principal component analysis

PCAT1 prostate cancer-associated transcript 1

PIF3 phytochrome-interacting factor 3

PNRD Plant Non-coding RNA Database

PP2C Protein Phosphatase 2C

PRC2 Polycomb Repressive Complex 2

QTL quantitative trait locus

RBM15 RNA binding motif protein 15

RIDL Repeat Insertion Domains of LncRNAs

RIKEN Rikagaku Kenkyusho

SINE short interspersed nuclear element

SNP single nucleotide polymorphism

SnRK2 SNF1-related protein kinase 2

SPEN spen family transcriptional repressor

SRA Sequence Read Archive

SVM support vector machine

tapRNA topological anchor point RNA

TE transposable element

Uchl1 ubiquitin carboxy-terminal hydrolase L1

UCR Ultra Conserved Region

WGCNA weighted gene co-expression network analysis

WTAP WT1 associated protein

Xic X-inactivation center

Xist X-inactive specific transcript

Chapter 1

Introduction

Caitlin M. A. Simopoulos

1.1 What is a long non-protein coding RNA?

Transcription of *H19*, the first identified eukaryotic long non-protein coding RNA (lncRNA), was observed in 1990 (Brannan et al. 1990). Without formal functional characterization, *H19* was identified to co-express with its neighbouring protein coding gene, insulin-like growth factor (*IGF2*), during embryogenesis suggesting a connection to development. While reminiscent of a protein-coding gene, *H19* was a unique locus that did not associate with ribosomes, but instead contained multiple open reading frames (ORFs) that remained untranslated (Brannan et al. 1990). Like a functional protein-coding mRNA, *H19* was transcribed by RNA polymerase II, post-transcriptionally modified by splicing and polyadenylation, and found to be conserved within human, mice and chicken genomes. In the end, the maternally imprinted locus *H19* (Bartolomei et al. 1991), became the first representative of a novel class of mRNA-like non-protein coding transcripts, which would later be referred to as lncRNAs.

Functioning as gene regulators, lncRNAs are now more generally defined as transcripts longer than 200nt with a lack of protein coding ability (Derrien et al. 2012). Like *H19*, lncRNAs are most often transcribed by RNA polymerase II, spliced and post-transcriptionally modified with a $5'$ cap and $3'$ poly-A tail (Derrien et al. 2012). However this lncRNA definition is rather fluid as there exists putative lncRNAs that do not meet the required 200nt threshold (Wang et al. 2005), encode small translated ORFs (Lauressergues et al. 2015), or are not always polyadenylated having been identified using a combination of poly(A)- RNASeq and total RNASeq (Di et al. 2014). LncRNAs act as tight gene expression regulators and are typically expressed in a low, condition- and tissue-dependent manner (Derrien et al. 2012). The characteristic low abundance and specific expression patterns of lncRNAs has led to difficulties in lncRNA identification. Furthermore, when lncRNA expression is detected, post-transcriptional modifications can make the non-coding transcripts appear indistinguishable from protein coding mRNAs.

Extensive transcription of lncRNAs in mammalian genomes was first described by the Functional ANnoTation Of the Mammalian genome (FANTOM) consortium and Rikagaku Kenkyusho (RIKEN) Institute (Okazaki et al. 2002; Carninci et al. 2005) who identified tens of thousands of transcribed non-protein coding loci. Using cap-analysis gene expression (CAGE) technology, the ENCyclopedia Of DNA Elements (ENCODE) Project aimed to further the research started by FANTOM and RIKEN by continued identification of DNA elements in the human genome with goals to eventually characterise all transcribed loci (The ENCODE Project Consortium 2012). During this expansion, The ENCODE Project Consortium (2012) identified that while 80% of the human genome is actively transcribed, only 3% of the genome accounts for protein coding genes. This discovery suggested that non-protein coding genes were 24 times more abundant in the human genome compared to protein coding genes, inciting a movement towards understanding non-protein coding RNA (ncRNA) functionality.

1.2 The fast evolution and fuzzy conservation of lncRNAs

Functionality of non-translated loci has been disputed, and ncRNAs being more abundantly transcribed compared to protein-coding loci does not necessarily imply that all ncRNAs have function. Some have argued that ncRNA, particularly lncRNAs, represent "transcriptional noise", or "non-functional transcription" (Hüttenhofer et al. 2005). For example, nucleotide and amino acid sequence similarity is often used to infer homology and to identify putative function of protein-coding genes (Pearson 2013). LncRNAs, however, do not display the same patterns of sequence or structural similarity observed in protein-coding genes (Nelson et al. 2016; Rivas et al. 2017). Instead, lncRNA conservation, if existing, is typically characterized by synteny or molecular function (Diederichs 2014). Expressed at low levels and demonstrating little sequence conservation even between closely related species (Derrien et al. 2012), one could argue that lncRNAs are the product of leaky transcription by a non specific RNA polymerase (Kung et al. 2013). Ponjavic et al. (2007), however, have identified trends in mouse lncRNAs that imply purifying selection and indicate functionality rather than transcriptional noise. For example, while mouse lncRNAs display a higher rate of insertions, deletions and substitutions than in protein-coding genes, the rate is lower than what is observed in purely intergenic regions. Similarly, Hangauer et al. (2013) describe five fold more trait-associated single nucleotide polymorphisms (SNPs) in humans located in lncRNAs than non-transcribed regions and suggest that the enrichment of SNPs in lncRNAs implies an influence on phenotypes and diseases in humans. More recently, Nelson et al. (2016) observed a similar trend in lncRNAs predicted in multiple plant species, where lncRNAs showed more sequence identity than intergenic loci.

Sequence similarity of lncRNAs between species may be inconsistent, but lncRNAs can often instead display syntenic conservation. For example, Mohammadin et al. (2015) identified that a subset of Brassicaceae lncRNAs that appear to be species-specific at nucleotide sequence level were positionally conserved by proximity to nearest protein-coding genes. Similarly, Hezroni et al. (2015) identified lncRNAs that display synteny with human lncRNAs in other animal species. Recently, Amaral et al. (2018) confirmed that syntenic lncRNAs have similar functions in many mammalian species and found evidence of positionally conserved lncRNAs that associated with conserved promoter regions. These syntenic lncRNAs also display conserved tissue-specific expression profiles between humans and mice, and a subset of the transcripts contain conserved secondary structures that associate with promoters. In total, Amaral et al. (2018) identified 764 of these conserved lncRNAs and have named this subclass of lncRNAs as topological anchor point RNAs (tapRNAs). This suggests that while lncRNAs may lack conservation in nucleotide sequence, other molecular traits, such as expression patterns or promoter sequences, may display evolutionarily conserved patterns. Additionally, the lack of detectable conservation in lncRNAs may merely be a result of lncRNA research infancy. Specifically, researchers may not have identified enough functional lncRNAs to detect conservation, or there may be subclasses of lncRNAs yet to be identified that display conserved molecular traits.

When considering lncRNAs as a whole, genome position is not the only molecular trait known to exhibit conservation in lncRNAs. Haerty and Ponting (2015) provide evidence of selection on exonic splicing enhancers and GC content of tested animal lncRNAs. Haerty and Ponting (2015) discuss that selection on nucleotide composition related traits may indicate conserved functionality, and in particular, selection on GC content may play a role in secondary structure conservation of RNA molecules. However, lncRNA structural conservation remains controversial. Rivas et al. (2017) found little evidence for structural conservation even in sequence-conserved lncRNAs such as X-inactive specific transcript (*Xist*) and HOX transcript anitisense RNA (*HOTAIR*), while a recent preprint manuscript by Tavares et al. (2018) shows contradicting conclusions found using different parameters in an RNA structure prediction program, R-scape. As such, conservation of lncRNAs is complex, and the evolutionary relationships of plant lncRNA molecular traits has yet to be fully explored.

Overall, lncRNAs display evidence of fast evolution of which the mechanisms are still unclear and under investigation (Pang et al. 2006). Nevertheless, select lncRNAs have been shown to evolve *de novo* from transposable elements (TEs), their molecular progenitors (Kapusta et al. 2013). LncRNA evolution from TEs led to the development of the Repeat Insertion Domains of LncRNAs (RIDL) hypothesis to explain and predict lncRNA function. The RIDL hypothesis suggests that it is the ancient functions of the TE-functional domain sequences in lncRNA that allow "newly evolved" lncRNAs to be functional. For example, TEs can provide their lncRNA descendants with DNA binding domains (Johnson and Guigo 2014).

TE-directed lncRNA evolution is not the only hypothesis surrounding how lncRNAs develop. Protein coding genes often evolve through gene duplication which is evident through the identification of gene families (Zhang 2003). This process can also give rise to functional ncRNA (Romito and Rougeulle 2011). An interesting example of pseudogenization through gene duplication involves the eutherian X chromosome inactivation and the lncRNA *Xist* (Romito and Rougeulle 2011). The X-inactivation center (Xic) is an area on the X chromosome in animals with X and Y sex chromosomes and is essential in X chromosome dosage compensation. The dosage compensation process occurs through X chromosome inactivation in XX females to prevent excessive expression of both inherited X chromosomes. Genes located in the Xic, including *Xist*, are not homologous in all animals, and inconsistent conservation of the Xic may indicate that many genes in this location are pseudogenes. Interestingly, *Xist* shares the most sequence homology with a protein-coding gene in chicken Ligand of numb-protein X 3 (*Lnx3*). Mutations and transposon sequence contributions consequently led to the loss of protein coding ability, and the evolution of the now essential lncRNA, *Xist*.

1.3 LncRNAs can regulate all levels of gene expression

In a review on lncRNA classifications, St. Laurent et al. (2015) curated over 50 overlapping classes of lncRNAs from the literature, confirming the complexity and fluidity of lncRNA definitions, conservation and evolution. The classifications range from descriptors of genomic location, to length of transcript, to molecular function. The review alerts researchers to the problems that stem from a lack of clear, consensus-driven lncRNA subclasses, and the need to update the definition of the RNA class. Due to the lack of recognized functional classifications, lncRNAs are typically described by their location to the nearest protein coding gene as: intergenic, intronic, sense, bidirectional antisense, promoter associated, upstream and enhancer-contained RNA (Wang and Cheksnova 2017). However, these classifications offer little insight on the molecular mechanisms of the transcript which can vary greatly and affect gene expression at levels of transcription or translation. LncRNAs do not have a single molecular function and their cellular localization depends on the specific function of the transcript. In this section, I describe the most broadly recognized molecular roles for functionally characterized lncRNAs.

1.3.1 Natural antisense transcripts

A natural antisense transcript (NAT) is a locus that is transcribed from the opposite strand of a sense gene. Sense-antisense NAT pairs are two genes that overlap at least partly and are located on opposite DNA strands. In this particular arrangement, the antisense partner specifically regulates its sense-overlapping gene. Typically, NATs regulate their sense-partner gene in *cis* although expression levels of the NAT are on average three times lower than sense expression levels (Zinad et al. 2017). NATs can also regulate distant genes in *trans*, for example when the NAT is a small ncRNA precursor. Katayama et al. (2005) estimated that just over 70% of all human transcriptional units show evidence of bidirectional transcription, suggesting that NATs are not a rare event and may have essential functionality.

NATs have been associated with abiotic stress responses in plants (Borsani et al. 2005; Xu et al. 2017a). Because NATs are so pervasive in both animals and plants, Yin et al. (2007) created the first NAT database, antiCODE, to help further research on this gene class. Yin et al. (2007) curated sense-antisense pairs from both animal and plant sources classifying gene partners as protein-coding or ncRNAs. antiCODE contributed to natural antisense transcript lncRNAs (NAT lncRNAs) identification across kingdoms and helped focus future NAT lncRNAs research.

NAT lncRNAs are known to regulate the expression of their target genes in two ways: concordant (induction of sense transcript) and discordant (downregulation of sense transcript) regulation (Faghihi and Wahlestedt 2009). Mirroring the *de novo* evolution of many lncRNAs, transcription of NATs is often activated by the insertion of TEs and TE-associated promoters at the $3'$ region of the sense-partner (Conley et al. 2008). Interestingly, NATs also seem to regulate the insertion of TEs and other mobile genetic elements, suggesting roles in genome stability (Zinad et al. 2017).

As alluded to above, NAT lncRNAs do not have a single mode of action and are classified by how gene pairs overlap, and not molecular function. Typically, NATs are referred to as "head-to-head" where gene pairs have overlapping $5'$ regions, "tail-to-tail" where gene pairs have overlapping $3'$ regions, and "full overlap" where one locus completely overlaps its pair (Latge et al. 2018). NAT lncRNAs have been implicated in transcriptional interference by promoter competition, RNA polymerase collision, promoter occlusion or RNA polymerase dislodgement. However, NAT lncRNAs typically regulate gene expression by chromatin modification, as observed in two well-studied mammalian NAT lncRNAs, *Xist* and *HOTAIR* (Monfort et al. 2015; Wu et al. 2013).

1.3.2 lncRNA-directed gene regulation via chromatin modification

Inside the nucleus, lncRNAs are able to regulate transcription not only by transcriptional interference, but also by DNA methylation, chromatin remodelling and histone modifications (Akhade et al. 2017) The two most well-studied examples of lncRNAs regulating methylation of their target genes are *Xist* and *HOTAIR*. *Xist*, an extremely long lncRNA, is a 17kb transcript encoded by a gene on the X chromosome and, as mentioned previously, is involved in X chromosome inactivation. *Xist* is localised to the nucleus and is constitutively expressed in XX females in order to coat the inactive chromosome and prevent transcription (Clemson et al. 1996).

H3K27me3 methylation, activated by Polycomb Repressive Complex 2 (PRC2), has been proposed as a key component to X inactivation (Monfort et al. 2015). The potential importance of H3K27me3 methylation led researchers to suggest that *Xist* interacts, directly or indirectly, with PRC2 to induce methylation and consequently repress the expression of a copy of the X chromosome. However, although first identified in 1992 (Brown et al. 1992) and described as "the best characterised lncRNA to date" (Pintacuda et al. 2017), the exact mechanisms of X inactivation by methylation and *Xist* remain unclear. A review by Rocha and Heard (2017) discusses spen family transcriptional repressor (SPEN), RNA binding motif protein 15 (RBM15) and WT1 associated protein (WTAP), and heterogeneous nuclear ribonucleoprotein U (HNRNPU) and lamin B receptor (LBR) as putative interaction partners contributing to methylation of the inactive X chromosome, but the exact mechanism of action remains unknown.

HOTAIR, also involved in methylation of its target gene *HoxD*, directly recruits the PRC2 complex (Wu et al. 2013). Although the *HOTAIR* transcript is 2.2kb long, it is merely an 89-mer domain of *HOTAIR* that interacts with a heterodimer of two PRC2 subunits initiating methylation (Wu et al. 2013). As reviewed by Bhan and Mandal (2015), over-expression and dysregulation of *HOTAIR* is known to have negative health effects on humans, particularly involvement in a variety of cancers. Similarly, lncRNAs from plant systems are also involved in chromatin modification for gene regulation. For example, multiple lncRNAs located in the FLOWERING LOCUS C (*FLC*) locus control flowering time are mediated by cold temperature and function by indirect interaction with PRC2 (discussed in detail in Section 1.6).

1.3.3 LncRNAs as endogenous miRNA "sponges"

LncRNAs do not always localise to the nucleus. Those that are functional in the cytoplasm can alter gene expression after their target gene has been transcribed via post-transcriptional regulation. For example, endogenous microRNA (miRNA) sponges are a type of lncRNA that can regulate their target gene indirectly. An *Arabidopsis thaliana* lncRNA, Induced by Phosphate Starvation 1 (*IPS1*), is a natural miRNA target decoy (Franco-Zorrilla et al. 2007). *IPS1* contains a conserved 23nt region that is complementary to its target miRNA, miR-399. *IPS1* functions by binding to miR-399 and preventing the miRNA from acting on its target gene, *PHO2*, which in turn regulates the expression of phosphate transporters. Although miRNAs in plant systems typically silence gene expression by cleaving its mRNA target gene, *IPS1* is uncleavable by virtue of an internal mismatch between *IPS1* and its target. Thus, *IPS1* continuously binds miR-399 preventing the normal gene silencing activity by the miRNA transcript (Franco-Zorrilla et al. 2007).

1.3.4 Translation of small ORFs in lncRNAs

While extensive translation of lncRNAs is controversial (Banfai et al. 2012), miRNA primary transcripts have been observed to encode small *cis*-regulatory peptides (Lauressergues et al. 2015) invalidating their "non-coding" namesake. miRNA-encoded peptides (miPEPs) translated *in vivo* were found to up-regulate the transcription of their respective miRNAs. Synthetically made peptides had similar regulatory and physiological effects as endogenous miPEP products where miRNA accumulation was found to increase with miPEP production or application, and developmental changes in root formation were observed (Lauressergues et al. 2015). Furthermore, translation of "non-coding" ORFs is not unique to the primary transcripts of miRNAs. Using a combination of RNASeq and RiboSeq, Bazin et al. (2017) identified over half of the predicted lncRNAs in *A. thaliana* roots experiencing phosphate limitation to be occupied by ribosomes. Additionally, the lncRNA *Aw112010* was recently identified to encode a translated small ORF that is involved in mucosal immunity in mammals (Jackson et al. 2018).

1.3.5 Regulation of translation by lncRNAs

Other cytoplasmic lncRNAs are capable of regulating gene expression post-transcriptionally by affecting translation or disrupting RNA stability. Brain Cytoplasmic RNA 1 (*BC1*), associated with epileptic seizures in humans, localises to dendrites and represses translation by preventing initiation of the small ribosomal subunit (Wang et al. 2005). Conversely, antisense-ubiquitin carboxy-terminal hydrolase L1 (*Uchl1*) is a lncRNA that increases translation of its sense protein UCHL1 (Carrieri et al. 2012). Antisense-*Uchl1* interacts with a short interspersed nuclear element (SINE), specifically a SINEB2 element, located inside sense-*Uchl1* and by a mechanism that remains unknown, results in an increase of *uchl1* translation by polysomes.

1.4 Functional validation of lncRNAs

Empirical characterization of lncRNAs is not trivial. Although a variety of lncRNA functions are presented above, we still do not know the full extent of lncRNA functionality. Because lncRNAs are known to interact with DNA, RNA and proteins, there is no single assay or experiment that can be used to identify binding partners or target sequences of a lncRNA. Typically, functional validation of a lncRNA starts with computational prediction after sequencing using prediction software discussed in depth in Section 1.7.2. To confirm that the putative lncRNA is non-coding, RiboSeq, or ribosome footprinting is often used to identify ribosome-protected fragments of RNA (Tichon et al. 2016). In combination with computational methods, wet lab experiments can be used to determine localization and to identify binding partners and putative functions of novel lncRNAs as discussed in this Section.

1.4.1 Co-expression studies

Co-expression networks are used to identify clusters of genes that co-express with lncRNAs of interest. Typically, genes comprising a single cluster are enriched in a particular function or pathway. Using a “guilt-by-association” approach, one can make functional predictions for un-annotated genes using functional enrichment of annotated genes in a certain cluster (Langfelder and Horvath 2008). Liao et al. (2011) used this co-expression network and “guilt-by-association” approach with microarray data to make function predictions for 340 mouse lncRNAs. Using three different network-based functional enrichment methods, the putative mouse lncRNAs were annotated with organ and tissue development, cellular transportation and metabolic process functionality.

Pellegrina et al. (2017) used a similar network-based technique to identify putative functions of lncRNAs involved in the human immune response to sepsis. The authors identified a group of co-expressed putative lncRNAs predicted to be involved in cellular respiration. Additionally, lncRNAs that were differentially expressed during sepsis compared to control individuals displayed conserved regulatory motifs that indicate methylation and transcription factor binding. Similarly, Fang et al. (2017) used a co-expression approach to identify function and pathways of lncRNAs associated with Alzheimer’s in mice models. This computational approach to lncRNA characterization is used to predict function and identify putative pathway involvement, and produces information that may be invaluable to experimental methods used in lncRNA validation studies.

1.4.2 Subcellular localization of lncRNA expression

Subcellular localization of lncRNA expression can be used to narrow the search for putative candidate functions. Before empirical studies, cellular location of lncRNAs can be predicted from sequence motifs in lncRNA transcripts using a deep neural network algorithm (Gudenas and Wang 2018). After computational prediction, researchers often use fluorescent *in situ* hybridization (FISH) for empirical studies on transcript subcellular location identification. For example, Qin et al. (2017) used FISH to demonstrate that a drought associated lncRNA, drought induced lncRNA (*DRIR*) localized to root cell nuclei. A nuclear lncRNA ruled out many possible functions, such as miRNA target mimicry, because of its functional location in the nucleus of a cell. Interestingly, a human lncRNA, noncoding RNA activated by DNA damage (*NORAD*), was found to localise to both the cytoplasm and nucleus in normal conditions (Munschauer et al. 2018). However, when human cancer cells were stressed with DNA damage, using FISH the genome stabilizing lncRNA was identified to preferentially localize to the nucleus. The transition to a primarily nuclear-located lncRNA after DNA damage induction suggests that *NORAD* is more involved in DNA stabilization when cells are under stress.

1.4.3 RNA-protein interactions

As previously mentioned, the lncRNA *HOTAIR* methylates its target gene by directly interacting with PRC2. Electrophoretic Mobility Shift Assay (EMSA) was used to identify the exact subunit of PRC2, the ERH2-EED heterodimer, that directly interacts with *HOTAIR* (Wu et al. 2013). After identifying the interacting protein subunit, gradual deletions in *HOTAIR* were also used, in conjunction with EMSA, to identify an 89-mer domain of *HOTAIR* that was essential for the interaction with the ERH2-EED heterodimer (Wu et al. 2013). Using a similar technique, proteins interacting with *NORAD* were identified using RNA antisense purification and RNA-protein specific cross-linking (Munschauer et al. 2018). Interacting proteins were enriched in functions related to DNA unwinding and damage repair, eventually leading to identification of an interacting topoisomerase-complex assembly. The authors confirmed the predicted genome stability functionality of *NORAD* using a combination of FISH and a RNA-protein interaction assay.

1.4.4 CRISPRi non-coding libraries

Liu et al. (2017b) used a unique CRISPR-based approach to systematically repress transcription of a large catalogue of predicted human lncRNAs in three cell lines. This genome-wide approach was used to identify phenotypes associated with lncRNA knockdown, which in turn was used to predict function. In total, although exact molecular mechanisms were not identified, Liu et al. (2017b) identified 499 lncRNAs essential for growth in human cell lines. This confirmed that a modified CRISPR protocol could be used in future lncRNA studies to identify additional functions of lncRNAs in other species.

1.5 Implications of lncRNAs in human diseases

To date, the majority of functionally characterized lncRNAs are of human origin partly due to their involvement in human diseases. The connection of lncRNAs to diseases, disorders, and illnesses is generally due to the essential roles of lncRNAs in mammalian development (Parna et al. 2017). Parna et al. (2017) cite over 70 lncRNAs essential to the development of mammals, and 43 diseases, disorders or syndromes associated with dysregulation of lncRNAs. For example, rheumatoid arthritis, Alzheimer's disease, Autism spectrum disorders and various cardiac disorders have been associated with disordered expression of at least one associated lncRNA. Because lncRNAs are gene expression regulators, it is not surprising that these transcripts are also involved in cancer. LncRNAs associated to disease can also be used a biomarkers or potential gene therapy targets and may be useful in a clinical setting. However Gomes et al. (2017) suggest further studies are required before application as many lncRNAs demonstrate currently uncharacterised pleiotropic effects.

As discussed previously, lncRNA functionality is not straight forward as lncRNAs are able to regulate genes on all possible levels and the extent of lncRNA function remains unknown. *H19*, an example of a lncRNA with multiple functions, is associated with cancer and its expression is induced by the *MYC* oncogenic transcription factor with consequent down-regulation of the tumor-suppressive p53 (Barsyte-Lovejoy et al. 2006). *H19* is not the only lncRNA known to interact with *MYC*. Transcribed upstream of *MYC*, the lncRNA colorectal cancer-associated transcript 1-long isoform (*CCAT-1 L*) is located within an enhancer. *CCAT-1 L* promotes chromatin looping and directly regulates the *MYC* transcription which is constitutively expressed in multiple cancers (Xiang et al. 2014) *HOTAIR*, the lncRNAs involved in methylation of *HoxD*, can form transcription factor triplets and has been observed to directly affect gene regulation in glioblastoma (Li et al. 2016b).

LncRNAs also regulate cancer pathways that are not *MYC*-associated. For example, some cancer-associated lncRNAs can act as natural miRNA target mimics, reminiscent of *IPS1* in *A. thaliana*. The lncRNA highly upregulated in liver cancer (*HULC*) acts as an endogenous miRNA sponge for multiple miRNAs, including miR-372, a cancer-associated small ncRNA (Wang et al. 2010). LncRNAs have also been shown influence genome stability, such as prostate cancer-associated transcript 1 (PCAT1), which represses *BRCA2* and impairs double stranded DNA break repair (Prensner et al. 2014). Since 2010, the connection of cancer and lncRNA expression had been an extremely active area of both cancer and lncRNA research (Renganathan and Felley-Bosco 2017). While understanding the molecular mechanisms of cancer is of interest, many studies also focus on using cancer associated-lncRNAs as therapeutic targets.

1.6 The roles of lncRNAs in plant development and stress response

LncRNAs remain a relatively novel classification of gene regulators and this area of research is currently dominated by lncRNAs found in animal systems with a large focus on humans. Of just over 440 empirically functionally characterized lncRNAs in all species, merely 13 have been identified in plants to date (Wang and Cheksnova 2017; Nejat and Mantri 2018; Zhao et al. 2018b). Because so few validated plant lncRNAs exist, little is known of the functionality and biogenesis differences between plant and animal lncRNA transcripts. However, differences in small ncRNAs between plant and animal systems suggests there may be fundamental differences between how lncRNAs function in each system. For example, miRNAs in plants associate with different proteins during miRNA-mediated gene silencing, and silence genes by mRNA cleavage rather than translational interference in animals (Millar and Waterhouse 2005). We do know, however, that while most lncRNAs in all species are transcribed by RNA polymerase II, the polymerase also responsible for transcription of protein coding mRNAs, some lncRNAs can also be transcribed by plant specific RNA polymerases Pol IV and Pol V (Wierzbicki et al. 2008). As we continue to validate plant lncRNAs we will be better equipped to identify additional differentiating features between plant- and animal-derived lncRNA transcripts.

Similar to animal systems, lncRNAs play essential roles in the development of plants. Two functionally characterized lncRNAs from plant systems, alternative Splicing Competitor-lncRNA (*ASCO-lncRNA*) and early nodulin-40 (*ENOD40*), are both involved in plant root development and interact directly with speckle-binding proteins (Bardou et al. 2014). Nuclear speckle-binding proteins localize to the nuclear-speckle, an organelle that contains most splicing machinery. Nuclear speckle-binding proteins are defined as a specific type of RNA binding protein that function as alternative splicing regulators and act on specific mRNA targets. Both *ASCO-lncRNA* and *ENOD40* are involved in alternative splicing of their appropriate nuclear speckle-binding protein target mRNAs and are involved in root architecture changes.

Bardou et al. (2014) proposed that *ASCO-lncRNA* “hijacks” the nuclear speckle-binding protein to modify the splicing events of the target mRNAs. This “hijacking” event causes shifts in lateral root formation. *ENOD40*, a conserved and small ORF containing lncRNA, differs from *ASCO-lncRNA* and is functional in the cytoplasm rather than the nucleus (Campalans et al. 2004). After interaction with a nuclear speckle-binding protein, *ENOD40* localizes to the cytoplasm in root nodules during development and is involved in nodule organogenesis (Campalans et al. 2004). Hidden treasure 1 (*HID1*) is another conserved plant lncRNA involved in plant development (Wang et al. 2014b). Constitutively expressed in all tissues, *HID1* promotes seedling photomorphogenesis in *A. thaliana* by repressing the transcription of phytochrome-interacting factor 3 (*PIF3*) via chromatin binding.

Akin to *Xist*, *COOLAIR* transcripts may be the most well-studied and well-characterized lncRNAs in plant systems. *COOLAIR* transcripts are two of many lncRNAs that negatively regulate *FLC* during vernalization in *A. thaliana*. *COOLAIR* transcripts are joined by a suite of three other lncRNAs, cold assisted intronic noncoding RNA (*COLDAIR*), cold of winter-induced noncoding RNA from the promoter (*COLDWRAP*) and Antisense Long (*ASL*) transcript, that all play unique roles in vernalization and/or *FLC* regulation (Heo and Sung 2011; Kim and Sung 2017; Shin and Chekanova 2014).

The lncRNAs controlling vernalization modulate the expression of *FLC*, a MADS-box transcription factor that induces multiple genes required for flowering (Bastow et al. 2004). *COOLAIR*, *COLDAIR* and *COLDWRAP* are all transcribed from the *FLC* locus and interestingly, *COOLAIR* represents two alternatively spliced NAT lncRNAs. The exact functions of *COOLAIR* transcripts and *COLDAIR* are difficult to discern, however researchers have shown that while *COOLAIR* transcripts may not be required for vernalization, these transcripts are responsible for an increased rate in the vernalization response. *COOLAIR* transcripts were also found to bind to the genomic sequence of *FLC* and directed a change in chromatin to H3K27me3 methylation (Csorba et al. 2014). Although H3K27me3 methylation is typically associated to PRC2, there is little evidence to suggest direct interaction of PRC2 with *COOLAIR*. Alternatively, *COLDWRAP*, expressed from FLC's promoter, works directly with *COLDAIR* to form a chromatin loop that does interact with PRC2 to mediate epigenetic silencing of *FLC* (Kim and Sung 2017). *COLDAIR* differs from its other vernalization-associated lncRNAs as it transcribed from *FLC*'s first intron (Heo and Sung 2011). *ASL* is another lncRNA that regulates *FLC*. Also acting in *cis*, *ASL* is transcribed from a region overlapping with *COOLAIR*, however, *ASL* is involved in regulation of *FLC* in *A. thaliana* ecotypes that do not require vernalization (Shin and Chekanova 2014).

While many lncRNAs are involved in organ development, plant lncRNAs are also commonly associated with a stress response. As RNASeq becomes ubiquitous with molecular biology research, researchers can use sequencing technology to identify novel lncRNA sequences in addition to how expression changes during stress. Such novel studies have identified thousands of predicted lncRNAs in many plant species associated with multiple stressful conditions (*e.g.* Zhang et al. 2014, Zhao et al. 2018b, Liu et al. 2018b, Zhang et al. 2018) and developmental stages (*e.g.*

Huang et al. 2018, Liu et al. 2018a, Kiegle et al. 2018), however most lncRNAs remain without empirical functional characterization.

1.7 Current lncRNA prediction methods and databases

Currently, lncRNA research remains a bioinformatic challenge due to the lack of empirically validated lncRNAs. Typically, researchers rely either on databases of predicted lncRNAs or are required to predict lncRNAs from sequencing data. Tools for lncRNA-focused research exist, but few are applicable to plant systems. In particular, there is an under-representation of empirically validated plant lncRNAs sequences in comprehensive lncRNAs repositories, and a lack of plant lncRNAs included in training datasets of machine learning algorithms for lncRNA prediction.

1.7.1 LncRNA databases

To date, there are two plant lncRNA databases currently available with associated peer reviewed prediction methods: GREENC (Paytuvi-Gallart et al. 2016) and CANTATAdb (Szczesniak et al. 2016). Both GREENC and CANTATAdb use transcript filtering methods for lncRNA prediction, where transcript features must meet thresholds to be classified as putative lncRNAs. GREENC uses transcriptome annotation provided by Phytozome v10.3 (Goodstein et al. 2012) leaving predictions only possible for known transcriptional units in each plant species according to reference genome annotation. GREENC's filtering protocol does not consider a transcript to be a putative lncRNA if they: 1. are $<$ 200nt in length, 2. have an ORF $>$120 amino acids 3. align to a protein in the SwissProt database (Bairoch and Apweiler 2000), 4. are predicted to be coding by the Coding Potential Calculator (Kong et al. 2007) or, 5. are known/are predicted to be other classifications of non-coding RNA. GREENC's methodology results in lncRNA predictions for 43 plant species that are restricted to reference annotations and a "classic" definition of a lncRNA. Additionally, the protocol discounts lncRNAs that align to protein coding sequences and can filter out lncRNAs that overlap with protein-coding genes, such as *COOLAIR*.

CANTATAdb (Szczesniak et al. 2016) also uses a filtering method, but is not restricted to annotated sequences in plant transcriptomes and instead uses transcripts assembled from RNASeq data. CANTATAdb's prediction protocol first removes transcripts previously annotated as protein-coding, rRNA, tRNA, or small ncRNAs. Transcripts are then classified as protein coding or non-coding by the Coding-Non-Coding Index (CNCI) (Sun et al. 2013). Transcripts with BLAST hits in the Rfam database, identified as plastid transcripts or are $>$ 200 nt are removed from lncRNA predictions. Similarly to GREENC, predictions made available by CANTATAdb are restricted to arbitrary definitions of lncRNAs previously set to distinguish from small ncRNAs, and which may not be truly representative of lncRNAs transcripts.

Not all plant-focused ncRNA databases rely on their own methods for transcript classification. The Plant Non-coding RNA Database (PNRD) (Yi et al. 2015) uses a data combing approach to aggregate predicted and validated ncRNA from a combination of other databases, literature, and their own data. PNRD contains multiple classifications of ncRNA for over 150 plant species, of which only 21 species include lncRNA predictions. However, the vast majority of lncRNAs available from PNRD remain computationally predicted.

NONCODEv5 (Fang et al. 2018), a ncRNA database focused on lncRNAs, contains lncRNA predictions from 17 species, of which only a single species is from the plant kingdom (*A. thaliana*). NONCODEv5 contains lncRNAs predicted by all other versions of NONCODE, as well as novel lncRNAs from recently published articles and lncRNA and genomic databases. NONCODEv5 is arguably the most extensive lncRNA repository containing ncRNA sequences identified by: Ensembl (Zerbino et al. 2018), RefSeq (O'Leary et al. 2016), lncRNAdb (Quek et al. 2015), LNCipedia (Volders et al. 2013), TAIR (Lamesch et al. 2011), FlyBase (Gramates et al. 2017), Lnc2Cancer (Ning et al. 2016), MNDR (Wang et al. 2013b), and LncRNAWiki (Ma et al. 2015). However, while NONCODEv5 offers a large amount of information on each predicted lncRNA, such as expression in multiple tissue types and predicted 3D structure, it is difficult to discern the status of empirical validation of each transcript, and similar to PNRD, the vast majority of lncRNA sequences available in the database remain computationally predicted.

Unlike the previously mentioned databases, lncRNAdb v2.0 is a lncRNA database manually curated from literature and contains only empirically validated lncRNAs from multiple species. However, while the database does contain sequences from plant systems, it is not plant focused and the large majority of sequences are animal-derived. Similarly, LncRNAdisease (Chen et al. 2013), a database focused on the association of lncRNAs and human diseases, contains only experimentally confirmed lncRNA sequences and does not contain plant sequences.

1.7.2 lncRNA prediction software

There are fundamental issues with lncRNA prediction pipelines that use transcript filtering, or alignment- and homology-based methods. For example, transcript filtering often restricts prediction of lncRNAs to transcripts longer than 200nt and those that do not code for proteins or peptides. However, there are functionally characterized lncRNAs that violate both of these arbitrary rules. For example, *BC1* is a 152nt lncRNA, shorter than the classic 200nt cutoff for lncRNA prediction (DeChiara and Brosius 1987). Additionally, *ENOD40*, a lncRNA identified in soybean, encodes two functional small peptides (Rohrig et al. 2002). There is little evidence for extensive sequence conservation of lncRNA sequences, even in species of the same family (Nelson et al. 2016), making prediction via homology difficult. Machine learning, therefore, is an alternative method that researchers have used to help predict this heterogeneous class of RNA where predictions are not restricted to hard thresholds and instead use patterns in data to identify sequences of interest.

A brief introduction to machine learning in computational biology

While machine learning is not only used in a biological setting, the extensive amount of data that a single experiment can generate makes biology an ideal subject area for the application and extension of machine learning algorithms. The objective of machine learning is to identify patterns and relationships in data without having to specify which traits are the most important and informative features (Angermueller et al. 2016). This is particularly useful in biological systems because predictions made by machine learning algorithms are not reliant on a set of rules or thresholds and can come from incomplete data (Angermueller et al. 2016). In machine learning, algorithms typically fall under two main categories: supervised and unsupervised learning. Supervised learning approaches "learn" from paired input variables containing features and classification outcomes, where the algorithm is trained on known positive and negative data. Using these training datasets, a machine "learns" a function that classifies the training information that can be applied to new and unknown data points. Unsupervised learning approaches do not include the training step and instead identify patterns in input features that are used to cluster data sets into groupings based on similarity.

The earliest mention of machine learning in biology was described by Rosenblatt (1958) and referred to the theory of "the perceptron" algorithm. The perceptron, a detailed mathematical model that attempted to explain how animals think and process information, became the backbone to the more recent field of artificial neural networks and was eventually applied in 1982 to identify translational start-sites in *E. coli* (Stormo et al. 1982). More recently, machine learning has been applied to "big data" for pattern identification and clustering to ultimately make experimental procedures more time and cost efficient. For example, secondary protein structure can be predicted through deep convolutional neural fields (Wang et al. 2016), enabling more accurate functional predictions that are used for drug design. The `Basset` software uses convolution neural networks to identify DNA motifs, even in non-coding regions, that predict genome accessibility and protein-interaction potential (Kelley et al. 2016). The tool can be used to identify mutations in a single DNA sequencing experiment that may affect the genome's ability to bind with proteins, and Kelley et al. (2016) hope the tool can contribute to research on medicine personalized to an individual's genome.

As modern machine learning approaches are developed, biologists are able to make use of large datasets to solve even more complex biological problems. Often researchers choose ensemble methods for robust predictions. This class of predictors uses multiple machine learning algorithms to make a single prediction, and relies on many learners to make accurate final predictions (Polikar 2012). While these traditional machine learning algorithms are able to identify patterns in given features to make predictions in data, deep learning, a new group of learning algorithms, is unique in that they do not always require the input of scientists to curate distinguishing features for prediction. For example, when considering deep learning in genomics, one can use genomic or transcriptomic sequences as input, and allow a deep neural network to identify essential distinguishing features in nucleotide sequences without researcher intervention (Angermueller et al. 2016).

Machine learning for lncRNA prediction

There are many machine learning frameworks, and a multitude of potential combinations of traits to use to distinguish lncRNAs from all other transcripts in a species. Firstly, while not necessarily developed to identify lncRNAs, protein coding predictors are used to discriminate between ncRNA and protein coding genes. CPC2 (Kang et al. 2017), CPAT (Wang et al. 2013a) and CNCI (Sun et al. 2013) are all tools that quantify the likelihood of a transcript to be translated into a functional protein. The results and features of protein coding predictors are often used as features in other lncRNA prediction software.

There are however, specific tools created to identify lncRNAs from nucleotide sequences. PLEK, or predictor of long non-coding RNAs and messenger RNAs is based on a k-mer scheme and uses a support vector machine (SVM) algorithm trained on k-mer usage in lncRNAs (Li et al. 2014). LncRScan-SVM (Sun et al. 2015), another SVM based software, uses features based on transcript length, codon usage, nucleotide composition, and protein coding ability. LncADeep (Yang et al. 2018) takes a different approach and uses deep neural networks to not only predict lncRNAs, but also to identify potential interacting proteins using pathway enrichment and KEGG annotations. The information given by these software tools can be used to distinguish potential non-peptide producing lncRNAs from protein coding mRNA, although they are not built to differentiate between functional lncRNAs and transcriptional noise.

PLEK, LmcRScan-SVM and LncADeep have high accuracies on test data, however each algorithm was only trained on animal-derived sequences. PLEK was trained on data from nine animal species and tested on human cell line data (Li et al. 2014). LncRScan-SVM, trained on both human and mice lncRNAs, was only tested on human datasets (Sun et al. 2015). Lastly, LncADeep was trained solely on human data and tested on both mouse and human transcripts (Yang et al. 2018). While these software are useful for animal lncRNA prediction, animal-sequence biases may prevent accurate predictions when they are used on plant-derived sequences. As there exists differences between animals and plants in other ncRNAs, such as miRNAs (Axtell et al. 2011), the inclusion of plant transcript sequences in training and testing datasets of machine learning methods is imperative for accurate plant lncRNA prediction.

In recognition of the lack of plant-focused lncRNA classifiers, Singh et al. (2017) created PLncPRO. PLncPRO, a random forest model, is trained on plant sequences and was tested on both animal and plant derived transcripts. However, like PLEK, LncRScan-SVM and lncADeep, the vast majority of lncRNA sequences on which the PLncPro algorithm was trained are yet to be empirically validated and remain merely computationally predicted. In other words, the software may have been trained on sequences that are transcribed and meet the classic criteria for distinguishing lncRNAs but remain without a validated regulatory role. The report of Simopoulos et al. (2018) represents the first machine learning based lncRNA classifier that is trained and tested on plant transcripts and uses a training dataset with only empirically validated lncRNAs.

1.8 Drought is a major cause of crop loss

As mentioned in Section 1.6, lncRNAs have been implicated in plants' responses to stress (Zhang et al. 2014; Zhao et al. 2018b; Liu et al. 2018b; Zhang et al. 2018). Drought contributes the most of all abiotic stressors to a reduction in crop yield (Boyer 1982), thus of interest to researchers focused on crop improvement. Predictions regarding temperature and precipitation changes due to climate change are difficult to make and they vary depending on global location (Schlaepfer et al. 2017). However, models tend to agree that reduced soil moisture, or aridity, is predicted to expand during the 21st century (Schlaepfer et al. 2017). Schlaepfer et al. (2017) also predict that the timespan of ecological droughts, a long-term reduction in water resources that creates overall stress in an ecosystem, will increase in deep soil layers. Lesk et al. (2016) identified that recent droughts (from 1985-2007) cause, on average, a greater reduction in cereal yield (13.7% yield reduction) than droughts that occurred between 1964 and 1984 (6.7% yield reduction), which the authors suggest may be due to many reasons including an increased drought sensitivity in recent cultivars, or longer and more severe droughts. Lobell et al. (2014) made similar conclusions regarding crops and discuss problems with modern maize cultivars grown in the U.S Midwest that have increased in yield but have not become more drought tolerant. Currently, even countries using technological advances to farming are at risk of crop yield reduction directly related to climate change and drought (Schlaepfer et al. 2017). Understanding the molecular mechanisms of how both stress sensitive and resistant plant species respond to drought stress is imperative to maintain high yielding species that are less adversely impacted by drought stress.

1.8.1 Molecular responses of plants to drought stress

As dehydration stress is commonly encountered in agriculture, and the stress has a major impact on yield, the classic molecular responses of plants to drought are well studied. Like other environmental stresses, drought stress first activates the biosynthetic pathways for the production of abscisic acid (ABA), a stress hormone (Xiong and Zhu 2003). ABA then induces an extensive downstream signalling pathway that ultimately results in phosphorylation of target proteins and activation of stress-related transcription factors. ABA is first perceived by binding to PYRABACTIN RESISTANCE1/PYR1-LIKE/REGULATORY COMPONENTS OF ABA RECEPTORS (PYR/PYL/RCAR), a group of receptors (Danquah et al. 2014). Once bound to ABA, PYR/PYL/RCAR receptors experience a conformational change allowing binding of Protein Phosphatase 2Cs (PP2Cs). The complex formation between receptors and PP2C, in turn, inhibits PP2C and ultimately indirectly activates SNF1-related protein kinase 2 (SnRK2) a major regulator of the plant stress response (Umezawa et al. 2013). Activated SnRK2 targets abiotic stress associated genes, such as transcription factors that affect the expression of other genes, ion channels, and membrane proteins (Umezawa et al. 2013).

Altered gene expression accompanies the physiological changes of plants experiencing dehydration, and many of these changes are directly affected by SnRK2 targets. For example, the ABA-responsive element binding factor, a basic leucine zipper (bZIP) transcription factor, is known to be activated in the ABA/SnRK2 regulation cascade and increased expression has been shown to improve abiotic stress tolerance in multiple plant species (Huang et al. 2010; Lee et al. 2010; Muniz-Garcia et al. 2012) There is also evidence that SnRK2 targets are associated with physiological changes to drought stress such as stomatal closure, an important strategy used by plants for water conservation (Grondin et al. 2015)

While lncRNAs are known to play roles in stress responses of plants, only recently was a drought-response associated lncRNA identified. Expression of *DRIR* was first identified in an RNASeq experiment of *A. thaliana* seedlings experiencing salt stress, but increased expression was also confirmed in plants experiencing a drought stress (Qin et al. 2017). *DRIR* is an *A. thaliana*-specific 755nt long lncRNA that does not share homology with transcripts in any other species. Like other lncRNAs identified in *A. thaliana*, *DRIR* contains a short ORF. *DRIR* displays tissue-specific expression with high expression identified in leaves and roots and little expression in stems, inflorescences or siliques. While the lncRNA is known to remain in the nucleus, the exact molecular mechanisms regulated by *DRIR* are difficult to identify. However, *DRIR* may be involved in the ABA response to drought stress as expression of this lncRNA was induced by ABA, and well-characterized genes in the ABA signalling cascade, such as *AB15*, are induced during *DRIR*-overexpression. The work by Qin et al. (2017) underlines that even well-known signalling pathways, like the drought-induced ABA signalling cascade, may have previously unidentified lncRNA players, such as *DRIR*.

1.9 *E. salsugineum* as a model system for studying abiotic stress

Eutrema salsugineum is a halophytic plant that is naturally tolerant to multiple abiotic stresses (Amtmann 2009). *E. salsugineum*'s innate tolerance to salt (Gong et al. 2005), drought (MacLeod et al. 2014), cold (Griffith et al. 2007), and nutritional deficiencies (Velasco et al. 2016) suggests that *E. salsugineum* demonstrates characteristics of an extremophile. As such, *E. salsugineum* has been used as a model organism for stress tolerance studies. There exists two commonly-studied natural accessions of *E. salsugineum*, Shandong and Yukon (reviewed by Kazachkova et al. 2018), named for the location of natural habitat. Shandong plants, naturally found in temperate and monsoonal Shandong, China, have different molecular and physiological responses to drought stress in comparison to Yukon plants, naturally found in the sub-arctic and semi-arid Yukon, Canada (Champigny et al. 2013; MacLeod et al. 2014). Wang et al. (2018c) recently estimated the divergence time of all known *E. salsugineum* ecotypes at around 34 kya. Interestingly, the species is naturally found in many diverse environments, for example, in China, Russia, and in North America as far south as Mexico (German and Koch 2017). The analysis by Wang et al. (2018c) also revealed little genetic diversity between ecotypes, around a quarter less than has been observed in *A. thaliana*. This leads to questions on how each ecotype can have unique gene expression patterns when experiencing stress with so little detectable genetic variation. However, as Wang et al. (2015) discuss, extreme stress has been associated with high purifying selection, potentially explaining the genetic uniformity of the species.

To further molecular studies on this model organism, the *E. salsugineum* genome was sequenced by Yang et al. (2013). Since then, researchers have continued progress into understanding the genetic mechanisms behind this plant's stress tolerance abilities. In fact, studies by Yin et al. (2018) and Champigny et al. (2013) have identified an additional 65 and 665 transcripts from RNASeq data of *E. salsugineum*, some associated to a stress response. Thus *E. salsugineum* represents a potential untapped resource of stress-associated genes, likely including seemingly species-specific lncRNAs that merit identification and functional analysis.

1.9.1 *E. salsigineum* in comparison to *A. thaliana*

A. thaliana is the most commonly used model plant for molecular studies (Provart et al. 2016). Its small size and short life cycle enable researchers to grow plants quickly with few resources. *E. salsugineum* is similar to *A. thaliana* and shares many morphological features, such as a small size and short life cycle, and ability to self-pollinate. Although having an estimated divergence time of 43.2 million years, both species have maintained a significant amount of genetic similarity and share 80% sequence homology and 70% of the *E. salsugineum* genome is in synteny with that of *A. thaliana* (Yang et al. 2013). The two species belong to the *Brassicaceae* family but are found in two separate phylogenetic clades (Yang et al. 2013).

Although the two species are very genetically similar, *A. thaliana* lacks the extensive stress tolerance abilities demonstrated by *E. salsugineum*. Adapted to a high salt environment, and naturally found in environments with saline soil, *E. salsugineum* is a halophyte (Champigny et al. 2013). However, similarities in genome sequence and structure of the two species means that the extensive amount of genetic methods developed for *A. thaliana* can often be applied to *E. salsugineum*. Additionally, combinations of comparative "-omic" and physiological studies between the two species can help reveal the specific adaptations held by *E. salsugineum* that confer superior stress tolerance. As there is large genetic overlap between *A. thaliana* and *E. salsugineum*, one can infer that large scale changes do not underlie the higher stress tolerance in *E. salsugineum*, and instead that the *A. thaliana* genome, and many other plant species, potentially already contain genes essential to stress tolerance. For example, Zhu et al. (2014) isolated and characterized *EsWAX1* from *E. salsugineum* for the creation of transgenic *A. thaliana* plants. *EsWAX1*, associated with cuticular wax formation, increased drought tolerance when expressed in *A. thaliana*, suggesting that the difference in the cuticular waxes of *E. salsugineum* plays important roles in the increased drought tolerance of the species. Additionally, overexpression of *EsWAX1*, a MYB transcription factor, was associated differential expression of other *A. thaliana* genes such as Cerberus 1 (CER1), 3-ketoacyl-CoA synthase (KCS2), Very-long-chain 3-oxoacyl-CoA reductase (KCR1) demonstrating that research on one species can be applied to the other.

1.9.2 Molecular mechanisms required for *E. salsugineum* stress tolerance

E. salsugineum, although naturally resistant to abiotic stress, often expresses “classic” stress-associated genes when experiencing abiotic stress. Gong et al. (2005) found that during salt stress, around 40% of genes differentially expressed in *E. salsugineum* were also differentially expressed in *A. thaliana*. For example, ABA responsive genes and other general stress-related genes such as *RAB18*, *RD21A* and *RD19A* were upregulated in both *A. thaliana* and *E. salsugineum* plants grown in saline conditions (Gong et al. 2005). However, the differences in molecular responses between the two species may explain *E. salsugineum*’s stress response strategies. Furthermore, differential gene expression studies may not show the entire picture of *E. salsugineum*’s stress response. Instead of regulating stress responsive genes, *E. salsugineum* constitutively expresses stress associated gene genes, even without being exposed to salinity, a process referred to as “priming” (Taji et al. 2004; Wong et al. 2006).

Constitutive expression of abiotic stress genes is not unique to *E. salsugineum*’s salinity response. Consistent high expression of stress related genes has also been observed for *E. salsugineum* when exposed to phosphate deprivation (Velasco et al. 2016). Of interest is also how the commonly studied Yukon and Shandong *E. salsugineum* ecotypes uniquely respond to stress. During drought, some plants can acclimate physiologically to survive in water deprived environments. Cuticular waxes, and expression of associated genes, can change to help the plant reduce water loss (LeProvost et al. 2013). Xu et al. (2014) identified differences in cuticular wax content and differential gene expression of genes associated with cuticular waxes in Yukon and Shandong ecotypes. More cuticle waxes correlated to a reduced water loss in Yukon plants, demonstrating that the Yukon ecotype is superior at tolerating drought stress (Xu et al. 2014).

MacLeod et al. (2014) also found differences in the physiological, metabolic and molecular responses of Yukon and Shandong plants to a progressive drought treatment where plants were subjected to two water deprivation conditions separated by a re-watering recovery period. Instead of using time intervals during during both the dehydration and hydration components of the study, the authors determined the fraction of transpirable soil water (FTSW). This method allowed the authors to account for, and quantify, the progression of water deficit for each ecotype. Plants were observed at a control, or well-watered condition, at two points during a first drought (40% and 10% FTSW), a re-watering condition, and again at two points during a second drought (40% and 10% FTSW). Although both ecotypes grew during the experimental period, Yukon plants were able to reach and maintain a lower solute potential than Shandong plants. Select dehydrins showed different expression patterns in each ecotype with Yukon plant dehydrins typically being expressed at higher baseline levels in control plants and larger changes in relative gene expression when exposed to drought stress. However, this study did not include analysis of global molecular changes of *E. salsugineum* ecotypes experiencing a progressive drought.

1.9.3 Applying drought response strategies of extremophyte species to crops

Drought responses of plants are typically associated with two main strategies: drought avoidance and drought tolerance (Fang and Xiong 2015). Drought avoidance strategies describe methods plants use to maintain physiological processes under mild stress, such as adjusting growth rate or adjusting certain morphological traits like increased cuticular wax or a changed root architecture (Xu et al. 2014; Ogburn and Edwards 2010). Drought tolerance, on the other hand, describes methods taken by a plant that can sustain high levels of stress by up-regulating genes and metabolic pathways that can help mitigate potential damage by stress (Mitra 2001). MacLeod et al. (2014) found that Yukon and Shandong *E. salsugineum* plants each display one of these drought response strategies with Shandong plants demonstrating drought avoidance and Yukon plants drought tolerance. *E. salsugineum* genotypes employing different strategies for drought tolerance offer a unique perspective to drought research. Specifically, one can use comparative studies to delineate tolerance and avoidance traits and their associated genetic determinants which is knowledge that could be applied towards crop improvement. Furthermore, *E. salsugineum* Yukon plant's capacity for "priming" in response to salt and phosphate stress (Gong et al. 2005; Velasco et al. 2016), where classic stress-responsive genes are constitutively expressed under non-stressed conditions, has been identified as a drought response strategy in resurrection plants (Costa et al. 2016). While "priming" has not yet been identified in *E. salsugineum* plants undergoing dehydration stress, other species using this approach which suggests that the "priming" strategy is conserved throughout many species.

In a recent review, Bechtold (2018) discusses how, while many studies have identified drought-associated genes, few studies have successfully used these genes to make improvements in crop plants, and suggests that researchers consider a more "systems biology" based approach for crop improvement. Fang and Xiong (2015) instead suggest that our current poor state of knowledge of the molecular cross-talk between various drought and other abiotic stress response strategies is holding back efforts to improve drought tolerance in crops. As such, continued work into understanding the molecular mechanisms behind drought response strategies is needed to enable researchers to identify usable genes for crop improvement. For example, using genes that are involved in more conserved molecular pathways may have more success than pathways unique to certain species.

1.10 Thesis objectives

The state of our knowledge on lncRNAs is overwhelmed with studies on transcripts derived from animal systems. As such, information on plant lncRNAs is severely lacking. Although there is evidence to support the important roles that lncRNAs play in plants, such as in development and stress responses (Bardou et al. 2014; Heo and Sung 2011), the limited number of experimentally validated plant lncRNAs restricts research on the molecular mechanisms and potential applications of this RNA class. Current computational lncRNA prediction methods are not adequate for lncRNA identification in plant transcriptomes. For example, plant-geared databases, such as GREENC and CANTATAdb, use filtering methods for lncRNA prediction that cannot identify "non-canonical" lncRNAs shorter than 200nt or those that code for small, functional peptides. Similarly, the machine learning-based software tools that are currently available are often trained mostly on non-validated animal sequences. Furthermore, our current knowledge of lncRNA evolution and conservation suggests that limited sequence and secondary structure conservation exists (Nelson et al. 2016; Rivas et al. 2017). Other than positional conservation and GC content patterns, few other molecular features of lncRNAs have been tested for conservation explained by phylogenetic relationships. Finally, although lncRNAs have been implicated in abiotic stress, there remains a single validated drought-associated lncRNA, *DRIR*, that has only been identified in the stress-sensitive *A. thaliana*. The gap in knowledge on plant lncRNAs and their connections to plant stress responses led to the following hypotheses:

1. An accurate and appropriate lncRNA prediction tool is required to reduce the gap in knowledge on plant lncRNAs compared to non-coding transcripts identified in animal transcriptomes. I hypothesize that an ensemble machine learning framework will allow the most accurate plant lncRNA predictions when using small training datasets comprised only of empirically validated lncRNA sequences.

2. Our understanding of lncRNA evolution is limited. The contributions of molecular traits, other than nucleotide sequences, to conservation and functionality of long non-coding transcripts is currently unknown. If lncRNAs display limited nucleotide sequence conservation yet evolve from the same ancestor, then lncRNAs should display detectable evolutionary patterns in other molecular features through phylogenetic signal estimation. In addition, I hypothesize that lncRNA sequences should display different phylogenetic patterns than protein coding genes.

3. Yukon and Shandong *E. salsugineum* ecotypes have different physiological responses to a progressive drought treatment that indicate tolerance and avoidance drought response strategies, respectively (MacLeod et al. 2014). The different responses to drought may be a result of local adaptations to the different natural environments of each ecotype. I hypothesize that the global molecular changes of both *E. salsugineum* ecotypes undergoing a progressive drought will reflect their respective drought response strategies and hence will be different, analogous to the physiological differences reported by MacLeod et al. (2014). Additionally, I predict that each ecotype will express unique lncRNAs in response to drought stress that act as gene expression regulators dictating the molecular changes required for acclimation to drought.

1.10.1 Brief experimental objectives

This thesis aims to address the gaps in plant lncRNA research by firstly constructing an appropriate lncRNA prediction tool. Secondly, this thesis will describe the applications of said tool to further evolutionary studies on lncRNAs, an important RNA classification. Finally, this thesis will highlight how researchers can use a lncRNA prediction tool alongside transcriptome studies with specific applications to understanding the drought responses of the halophytic plant, *E. salsugineum*. The thesis will seek to achieve the following three main objectives:

1. Create a flexible machine learning-based lncRNA prediction tool trained only on all empirically validated lncRNAs from all species, and test software on plant RNA sequencing data. The tool should rank predictions to help researchers prioritize high scoring lncRNAs for future empirical validation, particularly for experimental studies on lncRNAs identified in plants.

2. (a) Apply the lncRNA prediction tool to transcriptomes of plants with diverse evolutionary histories and evaluate the extent to which putative lncRNAs are included among reference annotations of selected species.

   (b) Estimate phylogenetic signal, or evolutionary patterns in molecular traits of lncRNAs predicted from transcriptomes of diverse plant species.

3. (a) Describe global transcriptional changes that occur during a progressive drought treatment in two *E. salsugineum* genotypes that display two different drought tolerance strategies. In addition, identify the putative lncRNA contributions to *E. salsugineum*'s molecular responses to drought stress.

   (b) Discuss the potential connection of local adaptation to environmental stress, specifically drought.

Chapter 2

Prediction of plant lncRNA by ensemble machine learning classifiers

Caitlin M. A. Simopoulos, Elizabeth A. Weretilnyk, G. Brian Golding

As published in BMC Genomics, 2018, 19:316
https://doi.org/10.1186/s12864-018-4665-2

2.1 Preface

Chapter 2 describes the construction and testing of a long non-protein coding RNA (lncRNA) prediction tool. As described in Chapter 1, the framework of the available software for lncRNA prediction were not adequate, particularly in the application to plant systems, due to: 1. A lack of training on plant-derived transcript sequences, 2. The use of invalid transcript filtering methods, and 3. Training datasets consisting of mainly computationally predicted sequences. In this work, we constructed and evaluated 24 different ensemble machine learning algorithms. In addition, we compared the predictions of the best performing model to a known plant lncRNA prediction database, GreeNC. This chapter is published in BMC Genomics as: C. Simopoulos et al. (2018). Prediction of plant lncRNA by ensemble machine learning classifiers. *BMC Genomics* 19, 316. i I made significant contributions to this study. I conceived the experiment jointly with E.A. Weretilnyk and G.B. Golding. I curated training datasets and constructed and tested a total of 24 machine learning algorithms. I evaluated all 24 models for selection of a final, most accurate model for release as a software tool. I developed the code for the software tool which is made available for public use at: `https://github.com/gbgolding/crema`. I wrote the first version of this manuscript, which was edited and approved by E.A. Weretilnyk and G.B. Golding. E.A. Weretilnyk and G.B. Golding supervised the analyses and writing of the manuscript.

2.2 Abstract

Background: In plants, long non-protein coding RNAs are believed to have essential roles in development and stress responses. However, relative to advances on discerning biological roles for long non-protein coding RNAs in animal systems, this RNA class in plants is largely understudied. With comparatively few validated plant long non-coding RNAs, research on this potentially critical class of RNA is hindered by a lack of appropriate prediction tools and databases. Supervised learning models trained on data sets of mostly non-validated, non-coding transcripts have been previously used to identify this enigmatic RNA class with applications largely focused on animal systems. Our approach uses a training set comprised only of empirically validated long non-protein coding RNAs from plant, animal, and viral sources to predict and rank candidate long non-protein coding gene products for future functional validation.

Results: Individual stochastic gradient boosting and random forest classifiers trained on only empirically validated long non-protein coding RNAs were constructed. In order to use the strengths of multiple classifiers, we combined multiple models into a single stacking meta-learner. This ensemble approach benefits from the diversity of several learners to effectively identify putative plant long non-coding RNAs from transcript sequence features. When the predicted genes identified by the ensemble classifier were compared to those listed in GreeNC, an established plant long non-coding RNA database, overlap for predicted genes from *Arabidopsis thaliana*, *Oryza sativa* and *Eutrema salsugineum* ranged from 51 to 83% with the highest agreement in *Eutrema salsugineum*. Most of the highest ranking predictions from *Arabidopsis thaliana* were annotated as potential natural antisense genes, pseudogenes, transposable elements, or simply computationally predicted hypothetical protein. Due to the nature of this tool, the model can be updated as new long non-protein coding transcripts are identified and functionally verified.

Conclusions: This ensemble classifier is an accurate tool that can be used to rank long non-protein coding RNA predictions for use in conjunction with gene expression studies. Selection of plant transcripts with a high potential for regulatory roles as long non-protein coding RNAs will advance research in the elucidation of long non-protein coding RNA function.

Keywords: lncRNA, classifier, machine learning, ensemble, transcript

2.3 Background

Long non-protein coding RNAs (lncRNAs) represent a diverse and functionally important class of RNAs (Kung et al. 2013), and have been classically defined as transcripts longer than 200 nucleotides with little protein-coding potential (Kapranov et al. 2007). Previously thought to be transcriptional noise, there is now evidence of their involvement in the development, disease, and stress responses of plants (Wang et al. 2017; Xu et al. 2017b); however, these transcripts are also found throughout all kingdoms of life. LncRNA transcripts often lack sequence conservation within close relatives, and the evolution of these transcripts remains poorly understood, but there exists growing evidence of positional and structural conservation that may indicate selection on transcript function (Hezroni et al. 2015).

Unlike other non-coding RNAs, the mechanisms and functions of lncRNAs can range wildly – from epigenetic regulation, as exemplified by mouse *Xist* and human X-inactive specific transcript (*Xist*) (Jeon and Lee 2011; Zhao et al. 2008), to small RNA target mimics, as seen with Induced by Phosphate Starvation 1 (*IPS1*) and *ath-miR399* in *Arabidopsis thaliana* cold assisted intronic noncoding RNA (*COLDAIR*), a lncRNA associated with flowering, functions by remodeling chromatin and alters expression of the *FLC* locus (He et al. 2013). A recent review by Ma et al. (2013) suggests that most known lncRNAs regulate transcription, both in *cis* and *trans*, while others can affect translation, splicing, post-translational regulation or are classified as "other functional mechanisms." Due to such a wide range of functionality, lncRNAs are typically classified by their position to protein coding genes as intergenic (also referred to as lincRNAs), natural antisense, or intronic (Kung et al. 2013; Ma et al. 2013).

Notably, lncRNAs can not only be functional in their long RNA form, but also act as small RNA precursors and sources of small regulatory peptides (Anderson et al. 2015; Ji et al. 2015; Juntawong et al. 2014) although extensive translation of lncRNAs has been disputed (Guttman et al. 2013). Adding to the complexity of these RNAs, some transcripts do not meet the arbitrary length cutoffs set by the classical definition for lncRNAs, such as *Brain Cytoplasmic RNA 1 (BC1)* in mice (152nt) (DeChiara and Brosius 1987). Even with recent developments in sequencing technologies, lncRNAs remain difficult to identify due to low, and condition-dependent and tissue-dependent expression levels (Derrien et al. 2012). Demonstrating minimal homology with close relatives (Hezroni et al. 2015), current research suggests these transcripts undergo fast and unclear evolution making functional predictions challenging. This lack of distinct rules for predicting and identifying lncRNAs is a likely contributor to the lack of validated plant lncRNAs.

Currently, many lncRNA prediction softwares that are available to researchers, such as PLEK (Li et al. 2014), lncRScan-SVM (Sun et al. 2015), and COME (Hu et al. 2017), use machine learning methods trained on data consisting of lncRNA transcripts yet to be empirically validated. Without empirical validation, many of these predicted lncRNA transcripts could have no regulatory function and could be produced due to spurious transcription because of the low fidelity of RNApolII (Struhl 2007). In addition, CPAT (Wang et al. 2013a) and CPC2 (Kang et al. 2017) are popular softwares used to identify non-coding transcripts. These softwares are successful at quickly predicting the protein-coding potential of mRNA sequences, but are not specific to lncRNAs and are unsuitable for identifying those lncRNAs that may code for small peptides. Additionally, since the majority of lncRNA research is on animals, software packages for lncRNAs prediction often use only animal training datasets. While the exact functions of most plant and animal lncRNAs remain poorly understood, there are known differences in biogenesis and mechanisms of other non-coding RNAs, such as miRNAs (Axtell et al. 2011). As such, ignoring the few plant lncRNA transcripts with known function could hinder the potential of future plant lncRNA predictors.

Depending on the source, lncRNA databases can also fall victim to biases toward animal systems and non-validated transcripts as they are often model organism specific with a preference for humans, and rarely differentiate between validated and predicted lncRNA transcripts. These biases can be seen in the popular lncRNA databases, LNCipedia and NONCODE (Volders et al. 2013; Zhao et al. 2016).

Outputs from lncRNA software often result in thousands of unranked predictions leaving the researcher to choose the most likely candidates for empirical validation. In combination with an RNASeq experiment that can result in tens of thousands of transcripts, filtering through thousands of lncRNA predictions can be difficult and time consuming for a researcher. Objectively ranking predictions in combination with gene expression estimates can help researchers complete functional validation of lncRNAs more efficiently.

Recently, ensemble methods have become popular for approaching difficult biological problems typically solved by machine learning (Liu et al. 2017a; You et al. 2013). Ensemble models work by combining multiple learners into a single model which helps to avoid over fitting and encourages generalization of the classifier. In addition to improved classification, ensemble methods also remove the difficulty in choosing the “best” model as all models can be used in a single classifier. Each individual classifier used in the construction of the overall ensemble model will have its own classification strengths, resulting in stronger and more accurate predictions when these classifiers are used in combination.

Here we describe a lncRNA predictor constructed using an ensemble of machine learning models developed for and tested on plant transcript sequences. We compared accuracy of this meta-learner trained on multiple machine learning models to the prediction ability of individual random forest and gradient boosting models making up the meta-learner. All models were trained on empirically validated lncRNAs to ensure only true lncRNA transcripts were used in each model’s training sets. We found the most successful method to be a stacking meta-learner constructed from eight stochastic gradient boosting models. This approach offers multiple advantages over those currently available as this machine learning approach prevents predictions from being constrained to the arbitrary classic definitions of lncRNAs, such as ignoring transcripts with high coding potential of small open reading frames (ORFs). In addition, our method numerically scores each prediction to help researchers focus their validation efforts on highly ranked lncRNA predictions. Finally, this approach uses the Diamond algorithm (Buchfink et al. 2015) that allows for efficient and fast sequence alignment in protein databases, an essential feature for lncRNA prediction.

2.4 Methods

2.4.1 Overview of classifiers

Multiple machine learning approaches to lncRNA prediction were compared to find the most accurate plant transcript classifier. Ensemble approaches were chosen due to the diversity of RNAs in the lncRNA category as these approaches are ideal for heterogeneous data. Ensemble models typically follow three main approaches: bagging, boosting, and stacking. Bagging (**b**ootstrap **agg**regat**ing**) relies on creating n models on bootstrapped training data, and averages predictions of all models for a final group prediction. This protocol is used in the random forest method. With boosting, such as in gradient boosting, one iteratively trains n learners, with each iteration attempting to reduce prediction error. The predictions are summed for a final classification. Finally, a stacking generalizer refers to training a new learner, for example by logistic regression, on the output of multiple learners. This is commonly referred to as meta-learner.

This study used all three approaches to ensemble methods, firstly by evaluating the lncRNA prediction accuracy of individual stochastic gradient boosting and random forest models. These individual models were then also combined into four ensemble classifiers explained further in the proceeding sections: 1. Arithmetic mean of scores, 2. Geometric mean of scores, 3. Majority vote, 4. Logistic regression meta-learner, and were evaluated similarly.

2.4.2 Individual stochastic gradient boosting and random forest models

Data

Positive data remained constant in each training set and consisted of a total of 436 unique, validated lncRNA sequences downloaded from two separate lncRNA databases: 1. lncRNAdb v2.0 (http://lncrnadb.org) on November 25, 2016 and 2. lncRNAdisease (http://www.cuilab.cn/lncrnadisease) on February 15, 2017. These sources for lncRNA sequences include all available validated lncRNAs, but are heavily populated by animal systems and include only six plant lncRNA sequences.

Negative data for each training set consisted of sequences from four different species: *Homo sapiens*, *A. thaliana*, *Mus musculus*, and *Oryza sativa*. *H. sapiens* and *M. musculus* sequences were included in the negative data of the training set as these species are the source for the majority of validated lncRNAs. *H. sapiens* sequences were downloaded from Ensembl (`http://www.ensembl.org`) on December 19, 2016, *A. thaliana* from Araport v11 (`https://www.araport.org`) on December 16, 2016, *M. musculus* from Ensembl on March 28, 2017 and

O. sativa from Ensembl on March 28, 2017. These data are made available in Supplemental File 2. To ensure that lncRNA, tRNAs, and rRNAs were removed from the negative training data, these types of sequences were downloaded from RNAcentral v6 (`http://rnacentral.org`) on March 28, 2017, using search terms available in Supplemental File 1 and were then removed from the dataset. Eight different training sets with different combinations of negative data from multiple species were used to construct eight different models and are described in Table 2.1. Sets denoted "A" and "B" remained constant throughout the training sets and were randomly chosen from the transcript sequences of each species. These training datasets were used in both random forest and gradient boosting methods, for a total of 16 preliminary models. The variety of training datasets was used to maximize model diversity, a requirement for the proceeding ensemble models.

Feature extraction and selection

Eleven features were chosen for use in model construction:

1. mRNA length
2. ORF length
3. GC%
4. Fickett score
5. hexamer score
6. alignment identity in SwissProt database
7. length of alignment in SwissProt database
8. proportion of alignment length and mRNA length (alignment length:mRNA length)
9. proportion of alignment length and ORF length (alignment length:ORF)
10. presence of transposable element
11. sequence percent divergence from transposable element

Features were extracted using a combination of custom Python scripts and known software (CPAT (Wang et al. 2013a) used for features 4 and 5, Diamond (Buchfink et al. 2015) used for features 6, 7, 8, 9, RepeatMasker (Smit et al. 2015) used for features 10 and 11.)

TABLE 2.1: Negative training data sets in individual models, and corresponding accuracy, sensitivity, specificity and AUC values.

Training dataset	Negative data	AUC		Accuracy		Specificity		Sensitivity	
		GB	RF	GB	RF	GB	RF	GB	RF
1	3000 *H. sapiens* (set A) 1000 *M. musculus* (set A) 3000 *O. sativa* (set A)	0.940	0.943	0.962	0.956	0.988	0.990	0.548	0.404
2	3000 *H. sapiens* (set A) 3000 *O. sativa* (set A)	0.943	0.944	0.960	0.953	0.988	0.989	0.576	0.461
3	3000 *H. sapiens* (set A) 1000 *M. musculus* (set A) 3000 *A. thaliana* (set A)	0.961	0.962	0.973	0.970	0.990	0.992	0.693	0.592
4	3000 *H. sapiens* (set A) 3000 *A. thaliana* (set A)	0.962	0.966	0.972	0.967	0.990	0.990	0.725	0.640
5	3000 *H. sapiens* (set B) 3000 *A. thaliana* (set B)	0.955	0.959	0.965	0.958	0.991	0.980	0.608	0.530
6	4500 *H. sapiens* (set A + 1500 seq) 4500 *A. thaliana* (set A + 1500 seq)	0.961	0.967	0.979	0.979	0.995	0.995	0.633	0.571
7	3000 *H. sapiens* (set A) 4500 *A. thaliana* (set A + 1500 seq)	0.963	0.967	0.976	0.971	0.993	0.992	0.700	0.603
8	2000 *H. sapiens* (2000 from set A) 1000 *M. musculus* (set A) 3000 *A. thaliana* (set A)	0.964	0.965	0.968	0.965	0.988	0.990	0.695	0.619

Training datasets of random forest (RF) and gradient boosting (GB) individual models are described. The positive training dataset, 436 validated lncRNAs, remained constant throughout all training datasets. Specificity, sensitivity, accuracy and AUC values were found using 10-fold cross validation of all training data.

CPAT model creation and application

As no publicly available plant CPAT model exists, two logit models were built using coding and non-protein coding RNA sequences from *A. thaliana* and *O. sativa*. Non-coding lncRNA, miRNA, snRNA, and snoRNA sequences from each species were downloaded from the Plant Non-coding RNA Database on September 26, 2016 (*A. thaliana*, 5062 sequences total) and July 14, 2017 (*O. sativa*, 4718 sequences total) (Yi et al. 2015). Protein coding transcript sequences from each species were downloaded from Phytozome v11 (Goodstein et al. 2012) on August 3, 2016. In order to supply a balanced training set, 5938 *A. thaliana* and 5283 *O. sativa* protein coding sequences were randomly selected for a total of 11,000 *A. thaliana* transcripts and 10,000 *O. sativa* transcripts for CPAT model construction. *A. thaliana* CPAT models were used for predictions in all species but *A. thaliana* itself, which used *O. sativa* CPAT models. Fickett and hexamer values from CPAT results were used as features in machine learning model construction.

Diamond alignment in SwissProt database

Diamond v0.8.34 (Buchfink et al. 2015) was used to quantify transcript sequence alignments to curated protein sequences in the SwissProt database (Bairoch and Apweiler 2000) downloaded February 1, 2017 from http://www.uniprot.org/downloads. We ran Diamond in “more-sensitive” mode as we aligned full transcript sequences to the SwissProt database rather than RNASeq reads. Options for each Diamond run were as follows: `-e` 0.001, `-k` 5, `--matrix` BLOSUM62, `--gapopen` 11, `--gapextend`: 1, `-f` 6 qseqid pident length qframe qstart qend sstart send evalue bitscore.

RepeatMasker

RepeatMasker (Smit et al. 2015) was used to extract information on transcription element related features. The software was run on transcript sequences using default settings, and with `-species` set to Eukaryota.

Stochastic gradient boosting and random forest model construction and hyper-parameter selection

Once features were extracted, models were constructed using Python’s scikit-learn package (Pedregosa et al. 2011). Eight separate models were constructed using both gradient boosting and random forest approaches, for a total of 16 models differing in negative training data or machine learning algorithm (Table 2.1). All transposable element related features were removed after performing recursive feature elimination as they were found to be uninformative and reduced the accuracy of models. With the 9 remaining features, a nested 4-fold cross-validation grid search was performed for 30 trials in gradient boosting hyper-parameter selection with possible hyper-parameters:

- `learning_rate`: 0.02, 0.04, 0.06, 0.08, 0.1

- `max_depth`: 4, 6, 8, 10
- `subsample`: 0.2, 0.4, 0.6, 0.8, 1
- `n_estimators`: 100, 500, 1000

Random forest hyper-parameters remained constant through all models with the only change from default parameters being `n_estimators = 5000` and `min_samples_leaf = 20`.

Models were evaluated by sensitivity, specificity, accuracy area under the curve (AUC) values using 10-fold cross validation and the `caret` R package (Jed Wing et al. 2017).

2.4.3 Ensemble learner construction

As gradient boosting and random forest models 1-8 were trained using eight different negative training sets, 3000 randomly selected *Zea mays* protein coding sequences were used as negative data in the construction and/or testing of each ensemble model for consistency through models. *Z. mays* was chosen as no training set contained sequences from this species and the genome is well annotated. *Z. mays* transcripts were downloaded from EnsemblPlants on April 27, 2017. Two separate values were used for the creation of each ensemble model – scores s_{ij} and predictions p_{ij} where i represents model number and j transcript. Scores can take any number between 0 and 1, while predictions are binary and indicate if the transcript was or was not predicted as a lncRNA. A score greater than or equal to 0.5 would indicate the transcript is predicted as a lncRNA and would have a prediction value of 1. Ensemble models were constructed for random forest and gradient boosting models separately in order to avoid potential correlation of predictions. The four ensemble approaches included both algebraic combiners and voting methods as non-trainable methods, and a stacking generalizer as a meta-learner.

The four ensemble methods are described as follows and are illustrated in Figure 2.1:

1. **Arithmetic Mean**

$$\frac{1}{n}\sum_{i=1}^{n} s_{ij} \tag{2.1}$$

   Where $n = 8$, the number of individual models combined into the ensemble approach. The ensemble decision is made from taking the arithmetic mean of each score s_{ij} from models 1-8 for each gene j. The arithmetic mean of scores will act as a new ensemble score, and prediction will be made as described previously.

2. **Geometric mean**

$$\left(\prod_{i=1}^{n} s_{ij}\right)^{\frac{1}{n}} \tag{2.2}$$

Where $n = 8$, the number of individual models combined into the ensemble approach. The ensemble decision is made from taking the geometric mean for each score s_{ij} from models 1-8 for each gene j. The geometric mean of scores will act as a new ensemble score, and prediction will be made as described previously.

3. **Majority vote**

$$\frac{1}{n}\sum_{i=1}^{n} p_{ij} \tag{2.3}$$

Where $n = 8$, the number of individual models combined into the ensemble approach. The ensemble decision depends only on final predictions and is decided on which label (0 or 1) receives the largest vote. The final prediction is made depending on the value of the majority vote score.

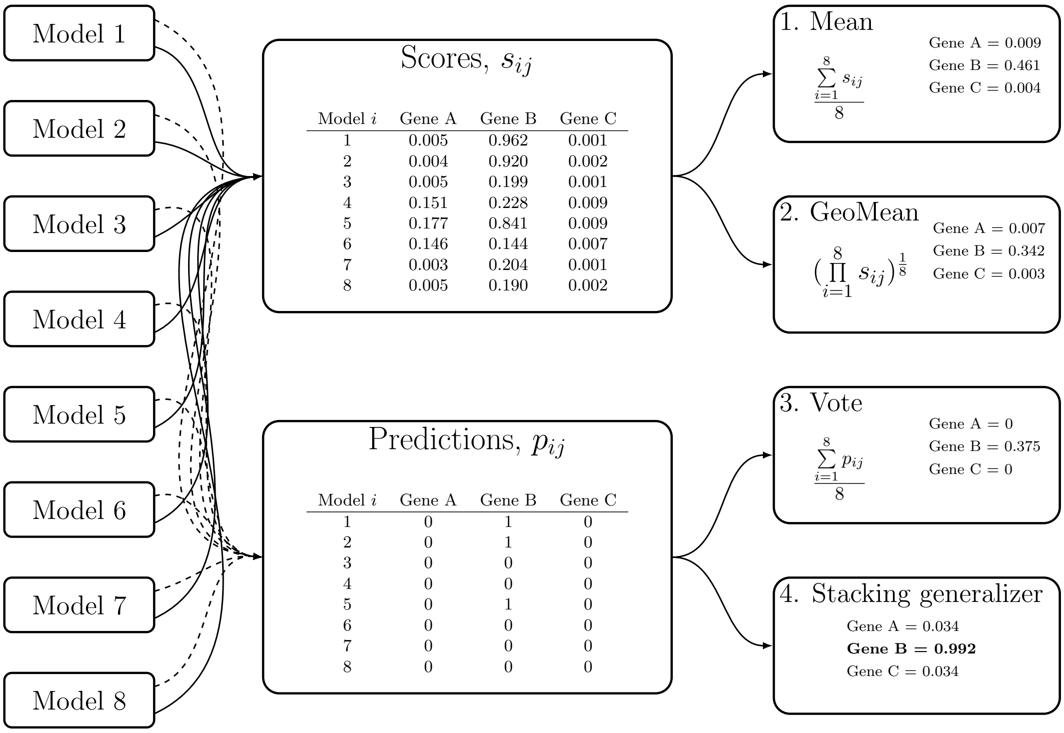
4. **Logistic regression**

This meta learner is trained on a training dataset of 3000 known *Z. mays* protein coding sequences as negative data and the 10-fold cross validation prediction outputs of known lncRNAs as positive data.

Voting, arithmetic mean, and geometric mean ensemble models were evaluated by directly comparing scores of predictions to the known outcomes of validated lncRNAs and 3000 *Z. mays* protein coding sequences. The logistic regression stacking generalizer was evaluated by 10-fold cross validation. Accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and AUC values were calculated using a custom R script and the R package `caret` (Jed Wing et al. 2017).

2.4.4 Comparison of predicted lncRNAs to GreeNC and annotation exploration

Transcript sequences of *O. sativa* and *Eutrema salsugineum* were downloaded from Phytozome v10.3 and *A. thaliana* from TAIR10 for direct comparison to GreeNC. LncRNAs predictions by GreeNC of *A. thaliana*, *O. sativa* and *E. salsugineum* were downloaded on June 19, 2017. Annotations from each species were downloaded from Phytozome v12, with extra *A. thaliana* annotation downloaded from Araport v11.



Scores, s_{ij}

Model i	Gene A	Gene B	Gene C
1	0.005	0.962	0.001
2	0.004	0.920	0.002
3	0.005	0.199	0.001
4	0.151	0.228	0.009
5	0.177	0.841	0.009
6	0.146	0.144	0.007
7	0.003	0.204	0.001
8	0.005	0.190	0.002

Predictions, p_{ij}

Model i	Gene A	Gene B	Gene C
1	0	1	0
2	0	1	0
3	0	0	0
4	0	0	0
5	0	1	0
6	0	0	0
7	0	0	0
8	0	0	0

FIGURE 2.1: **Illustration of ensemble methods.** An illustrative example of all four ensemble methods: arithmetic mean, geometric mean, majority vote and the stacking generalizer. Real examples from three different genes are given: gene A represents AT5G44470 a predicted protein, gene B represents At43G09922.1 *IPS1* a known lncRNA, and gene C represents At2G18130.1 a known protein coding gene, *AtPAP11*. Note the final stacking generalizer score of gene B compared to the individual model scores for the gene.

2.5 Results

2.5.1 Individual random forest and stochastic gradient boosting model construction

Feature selection

Researchers have proposed that specific characters in transcript sequences can be useful in lncRNA classification. For example, lncRNAs can be translated into short peptides (Anderson et al. 2015; Ji et al. 2015; Juntawong et al. 2014), however most validated lncRNAs remain functional in their RNA form with little protein coding potential. The potential for a transcript to be translated into a protein can be predicted by codon bias, often measured by Fickett score and hexamer usage bias (Wang et al. 2013a). Mammalian lncRNAs are known to have a lower GC content than protein-coding RNAs (Niazi and Valadkhan 2012), and this feature has been used as a defining feature for *A. thaliana* lncRNA prediction in the past (Di et al. 2014). Transposable elements (TEs) are also known to be sources for plant lncRNAs (Wang et al. 2017). Based on these studies, 11 features were originally chosen for use in lncRNA classification: mRNA length, ORF length, GC%, Fickett score, hexamer score, alignment identity in SwissProt database, length of alignment in SwissProt database, proportion of alignment length and mRNA length (alignment length:mRNA length), proportion of alignment length and ORF length (alignment length:ORF), presence of transposable element, and sequence percent divergence from transposable element. Using recursive feature elimination as described in Section 2.4, features that related to transposable elements were removed since inclusion of these features in classifiers decreased prediction accuracy and thus were deemed uninformative for this training data. After feature elimination, nine features were chosen for implementation in individual random forest and gradient boosting models: mRNA length, ORF length, GC%, Fickett score, hexamer score, alignment identity, length of alignment, alignment length:mRNA length, and alignment length:ORF.

Individual model configuration and model evaluation

Gradient boosting and random forest models were constructed using eight different negative training datasets for a total of sixteen models (Table 2.1). Empirically validated lncRNA transcripts were downloaded from databases as described in Section 2.4. To ensure optimal performance of each gradient boosting classifier, proper calibration of multiple hyper-parameters is required. As such, hyper-parameter tuning (`learning_rate`, `max_depth`, `subsample`, and `n_estimators`) for each gradient boosting model was completed by grid search and 30 iterations of 4-fold nested cross validation with results summarized in Table 2.2. All random forest models were constructed with the same hyper-parameters; all options were left as default other than `n_estimators=5000` and `min_samples_leaf` = 20.

TABLE 2.2: Gradient boosting hyper-parameters chosen by grid search for each model.

GB Model #	Learning rate	maxdepth	subsample	n estimators
1	0.04	10	0.6	100
2	0.04	10	0.6	100
3	0.04	10	0.6	100
4	0.02	8	0.6	100
5	0.02	10	0.6	100
6	0.02	10	0.6	100
7	0.04	10	0.6	100
8	0.04	10	0.6	100

Hyper-parameters were chosen by grid search using 30 iterations of 4-fold nested cross validation. The given hyper-parameters corresponded to models with the highest accuracy values of all given hyper-parameter combinations.

After training calibrated models, gradient boosting and random forest models were evaluated individually by 10-fold cross validation by accuracy, specificity, sensitivity and AUC measures for model validation (Table 2.1). All models performed at or above accuracy, specificity and AUC measures of 0.94, however, sensitivity values ranged from 0.40 to 0.725 (Table 2.1). Because of this wide range of sensitivity values, four alternative ensemble approaches using combined random forest and gradient boosting models were explored.

2.5.2 Ensemble classifier construction

To take advantage of the predictive strengths of each random forest and gradient boosting model, ensemble learners for all random forest and all gradient boosting models were constructed. As ensemble classifiers function by combining "diverse" learners (Brown et al. 2005), only models constructed from different training sets were used in each ensemble classifier to maintain diversity in predictors. In other words, ensemble classifiers were constructed from all eight random forest models and a separate set of ensemble classifiers were constructed from all eight gradient boosting models.

Four types of ensemble classifiers were constructed: a majority vote model, arithmetic means of scores model, geometric means of scores model, and a stacking ensemble model constructed from a logistic regression of model outputs (Figure 2.1 and Section 2.4 for details).

A final training set comprised of 3000 known *Z. mays* protein coding genes and validated lncRNAs was created. This *Z. mays* training data set was used for training the logistic regression classifier because random forest and gradient boosting models were trained on different data sets (see Section 2.4). For consistency, all four ensemble methods were also evaluated using these data. The arithmetic mean, geometric mean, and majority vote methods were evaluated by

TABLE 2.3: Evaluation of random forest (RF) and gradient boosting (GB) ensemble models

ML model type	Ensemble type	AUC	MCC	Accuracy	Sensitivity	Specificity
RF						
	Vote	0.834	0.725	0.944	0.594	0.995
	Arithmetic Mean	0.963	0.661	0.941	0.562	0.996
	Geometric Mean	0.963	0.706	0.941	0.555	0.997
	Logistic regression	0.835	0.765	0.952	0.665	0.994
GB						
	Vote	0.887	0.797	0.958	0.702	0.995
	Arithmetic Mean	0.945	0.786	0.956	0.681	0.996
	Geometric Mean	0.940	0.750	0.949	0.601	0.999
	Logistic regression	0.883	0.822	0.963	0.745	0.994

Statistics for vote, arithmetic mean, and geometric mean models were calculated using outputs of models compared to true labels. Logistic regression evaluation statistics were calculated using the scores found by 10-fold cross validation of *O. sativa* training data and validated lncRNA sequences.

comparing ensemble method outputs to true labels, and 10-fold cross validation scores were used to evaluate the logistic regression stacking model. Accuracy, specificity, and AUC values were similar for all ensemble approaches; therefore, the best performing ensemble method was largely determined by both sensitivity and MCC measures (Table 2.3). Using these values as methods of evaluation, the stacking model constructed from gradient boosting model outputs was found to be the best performing model and was used for the remainder of the study.

2.5.3 Comparison of meta-learner to GreeNC predictions

To assess the overlap of predictions to another plant lncRNA resource, the lncRNAs predicted by the stacking generalizer were compared to an established lncRNA database, GreeNC (Paytuvi-Gallart et al. 2016). This database uses a transcript filtering method, rather than a machine learning approach, where transcripts must meet the criteria of a classic lncRNA in order to be identified as putative lncRNAs. To be considered a lncRNA in the GreeNC database, the transcript must: be larger than 200nt, have an ORF smaller than 120aa, not have a hit in the SwissProt database or be considered non-coding by the Coding Potential Calculator (Kong et al. 2007), and not be already classified as another class of functional RNA as identified by Rfam.

Transcript sequences of *O. sativa*, and *E. salsugineum* were downloaded from Phytozome v10.3 and *A. thaliana* sequences from TAIR10 to enable direct comparison to the GreeNC protocol. In total, 1310, 856 and 198 lncRNAs were predicted from *A. thaliana*, *O. sativa*, and *E. salsugineum* respectively, of which 872 (66.6%), 444 (51.9%), and 164 (82.8%) have been previously predicted by GreeNC (Figure 2.2). Comparing number of predicted lncRNAs using this method to GreeNC, 1700, 4381, and 1471 fewer lncRNAs are identified in *A. thaliana*, *O. sativa* and *E. salsugineum* using the stacking method. Another 438, 412 and 34 putative lncRNAs were identified using the stacking learner that have not been predicted by GreeNC in *A. thaliana*, *O. sativa*, and *E. salsugineum*.

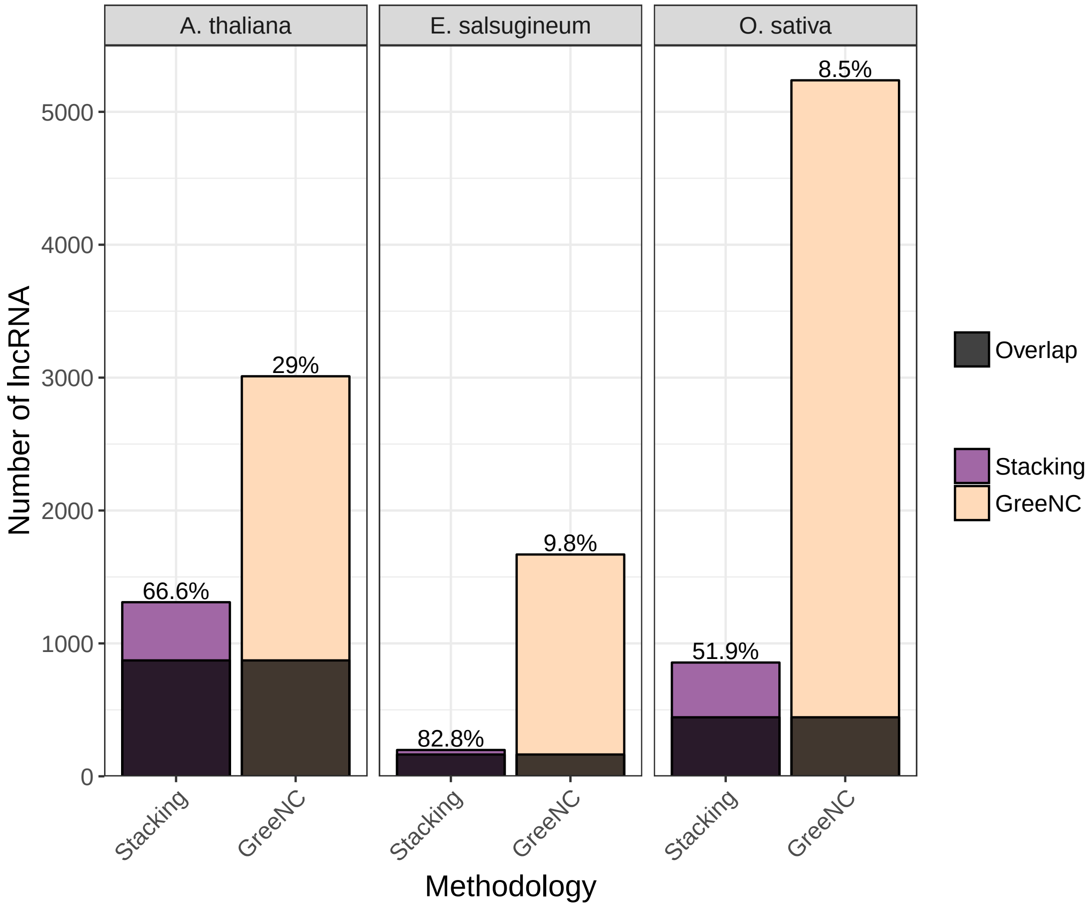


FIGURE 2.2: **Counts of predicted lncRNAs in *A. thaliana*, *E. salsugineum* and *O. sativa* from the gradient boosting stacking generalizer method and GreeNC database.** Counts of predicted lncRNAs in this work from all three species were also compared to predictions recorded in GreeNC. Overlapping predictions of the two methods are represented as shaded bars. The percentages above each bar represent the percent of the total predictions by each method that are shared

Current annotation of top ranking lncRNAs in *A. thaliana*, *E. salsugineum*, and *O. sativa*

Using the prediction scoring system of this stacking method, the current annotation of the highest ranking lncRNAs from each species was explored. Due to the nature of a logistic regression-type ensemble method, transcripts with similar features will have identical prediction scores. As such, multiple prediction score ties exist in the top ranking transcripts of each species (See A, Figure S1.1, Table S1.1 for distribution of lncRNA scores). Using a cutoff of the top three unique prediction scores, annotations of 256, 17 and 94 transcripts in *A. thaliana*, *E. salsugineum*, and *O. sativa* were identified as "top scoring" due to these multiple ties. The majority of predicted lncRNAs in *A. thaliana* were annotated by TAIR as potential natural antisense lncRNAs, pseudogenes, and transposable element related genes (Table 2.4). Only one transcript from *E. salsugineum*'s top predictions, and two transcripts from *O. sativa*'s top predictions have annotation in Phytozome v12.

TABLE 2.4: Number of transcripts in annotation categories of top ranking lncRNAs in the *A. thaliana* transcriptome.

Annotation category	Number of annotations
Natural antisense lncRNA	64
Pseudogene	75
Transposable element gene	10
Transposase	46
miRNA primary transcript	4
Hypothetical protein	5
Protein	8
Other	8

Novel lncRNAs identified by the stacking generalizer

Annotation of the predicted lncRNAs not previously identified by GreeNC from all three species were explored. While all of the newly predicted lncRNAs from *E. salsugineum* and *O. sativa* were annotated as homologs of *A. thaliana* genes, 10 of 34 novel lncRNAs from *E. salsugineum* and 11 of 412 novel lncRNAs from *O. sativa* were annotated specifically as proteins. Of the newly predicted lncRNAs from *A. thaliana*, 417 remain unannotated, with only seven predicted as potential proteins.

2.6 Discussion

Our approach to lncRNA prediction by stacking with logistic regression allows researchers to combine the strengths of various machine learning models without restricting predictions to arbitrary feature cutoffs of a classic lncRNA definition. The flexible nature of this lncRNA prediction tool allows the model to be updated when additional lncRNAs are validated, helping researchers focus on empirical validation of plant lncRNA transcripts. As lncRNA research has previously primarily focused on animal systems with a large emphasis on humans and mice, this tools' training sets may have a human and mouse bias that is present out of necessity. When more plant lncRNAs are added to the tool's training set, the human and mouse lncRNA bias that may be found in the model will be reduced. Acting as positive feedback, as more plant lncRNAs are added to the model, the predictions themselves will improve.

To help researchers choose the best lncRNAs for validation, the predictions are ranked. While softwares that rank lncRNA predictions, such as COME exist (Hu et al. 2017), they are trained on a majority of non-empirically validated transcripts adding a potential bias towards non functional transcripts. A combination of ranked predictions and models trained only on true lncRNAs will help ensure researchers focus on the most likely functional lncRNAs

A lower number of identified lncRNAs in comparison to other prediction methods, such as GreeNC, was expected. Using a machine learning classification method, lncRNA predictions were not constrained to arbitrary criteria for this RNA classification. Instead, the classifiers were trained on validated lncRNAs and are expected to identify only true functional lncRNA transcripts. In other words, although transcripts were subjected to less rules for lncRNA identification, the stacking method is expected to have higher accuracy. Further, this work was tested only on sequence information available from Phytozome v10.3 in order to compare predictions directly to GreeNC. Additional transcript sequences available in public repositories, or from researchers' own sequencing libraries, would add to the number of putative lncRNAs and could be used to improve accuracy. Moreover, *COOLAIR* and *COLDAIR*, known *A. thaliana* lncRNAs, are not predicted by GreeNC because the database relies on transcript sequences provided by Phytozome and these transcript sequences were not available in the database at the time of prediction. Our stacking generalizer method for lncRNA prediction is not restricted to a single data source, and allows researchers to calculate a lncRNA score from any transcript sequence, not solely those available from an online repository.

While we expect a lower number of putative lncRNAs than other protocols, of interest is the lower proportion of predicted lncRNAs in the *E. salsugineum* genome compared to *O. sativa* or *A. thaliana*. A reason for the low lncRNA discovery rate in *E. salsugineum*, could potentially be that plants were not subjected to conditions sufficient for observable lncRNA expression. For example, *IPS1* (Franco-Zorrilla et al. 2007) and *COLDAIR* (He et al. 2013), two well studied *A. thaliana* lncRNAs, are induced by phosphate or cold-related stresses respectively. This hypothesis is supported by Derrien et al. (2012) who found human lncRNA expression to be at low levels in a condition, tissue and developmental state specific manner. It is also possible that there exists natural variation in the numbers of putative lncRNAs in different species. Further investigation on the number of putative lncRNA and their relationships to plant growth conditions for transcriptome sequencing of multiple plant species is currently underway.

Although the quantity of detected lncRNAs was low in *E. salsugineum*, the quality of putative lncRNAs in all three species is high, demonstrating that this tool can accurately classify transcripts no matter size or quality of input transcript sequence data. When exploring the annotations of the top scoring predictions in *A. thaliana*, the majority of transcripts were annotated as potential natural antisense lncRNA, pseudogenes, transposable elements, small RNA primary transcripts, or remain computationally predicted as hypothetical proteins (Table 2.4). Pseudogenes remain poorly understood, however there is evidence of pseudogene derived lncRNAs regulating their parental genes (Milligan and Lipovich 2014), making pseudogene derived lncRNAs targets of potential regulatory interest. Transposable elements are another known source of lncRNAs, particularly in vertebrates (Kapusta et al. 2013) and long intergenic non-protein coding RNAs in plants (Wang et al. 2017). This study did not find evidence that features related to transposable elements were helpful at predicting plant lncRNAs as the addition of transposable related features decreased the quality of lncRNA predictions. However, exploration of the training data used for model creation indicates that only 19 of the 436 (4.4%) validated lncRNAs show evidence of transposable element association. Of this minor group of transposable element associated lncRNAs, none were from plant species. Nonetheless, the tool did not favour lncRNAs that are not associated with transposable elements, as the tool remained successful at identifying these types of transcripts. Additionally, as novel lncRNAs are validated and added to this tool, an update to the models' feature selection step may be required, and may lead to future inclusion of transposable element associated characters. However, by not including transposable element information, the computational time for data preprocessing before transcript classification is significantly reduced to minutes from days as RepeatMasker is no longer needed.

Features of secondary RNA structure have previously been used in other RNA classifiers, such as nRC (Fiannaca et al. 2017) and GraPPLE (Childs et al. 2009), that are used to classify RNAs into functional categories. These classifications include RNAs such as miRNAs, tRNAs, rRNA, ribozymes, and riboswitch domains, all of which have conserved secondary structures. Rather than using sequence homology, commonly used with protein coding genes, structural homology has previously been used in lncRNA functional prediction, and identification (Hezroni et al. 2015). However, a lack of secondary structure conservation in animal lncRNAs with conserved sequences (*e.g. HOTAIR, ncSRA and Xist*) was recently observed (Rivas et al. 2017). As structural conservation may not be as pervasive in lncRNA classification as previously thought, we did not include structural features in our ensemble learner. A lack of structural features allows the predictor to identify a wide variety of lncRNAs and does not limit the predictor to the structures of the small number of validated plant lncRNAs available. An additional test was completed to ensure our predictor, lacking structural features, did not merely distinguish non-coding transcripts from protein coding genes. By comparing the results of the ensemble learner to predicted CPAT protein coding probabilities (Wang et al. 2013a), our ensemble method was able distinguish between other CPAT-predicted non-coding transcripts and likely lncRNAs (Appendix A, Figure S1.2). A portion of putative lncRNAs in all three plant species are also predicted to be protein coding and may encode small regulatory peptides.

High quality lncRNA predictions from this method require sequences from fully processed transcripts and cannot be predicted directly from genomic sequences. Nevertheless, potential lncRNA sequences of interest are typically more accessible by transcriptome sequencing rather than complete genome sequencing, which remains technically challenging for crop plants with large and/or polyploid genomes. This tool is flexible and can be used to identify lncRNAs from all transcriptional units of an organism, or to check the lncRNA score of a single transcript. Furthermore, as mentioned in their summary, Kang et al. (2017) suggest that researchers should now consider working on uncovering the biological implications of lncRNAs rather than solely using computational tools for transcript classification. We agree that future work should centre around using software to also further knowledge on these types of transcripts. Due to the diversity of these transcripts, there is increasing need for classification of lncRNAs into categories based on mechanism and function, as well as continuation of empirical validation, particularly for plants. Once validated, not only can novel lncRNAs mechanisms be explored, but their features can be added to this tool for further improvement in lncRNA prediction.

2.7 Conclusion

For this machine learning based tool for lncRNA prediction, we have used only empirically validated lncRNAs for training. Although lncRNAs from multiple species were used, our tool identified putative plant lncRNAs with high scores. Ranking of lncRNA predictions should improve the confidence by which gene products meriting validation are selected for empirical testing. The machine learning structure and its open source availability allows for the flexible inclusion of validated lncRNAs as our knowledge of this class of RNA improves. An important consideration of this tool is that it is not constrained by preconceived rules that may or may not appropriately classify lncRNA properties. As Kung et al. (2013) suggest, setting rules for the detection of these non-conforming transcripts could be detrimental due to the diversity in functionality, structure, expression and mechanism of these transcripts. Accordingly, our stacking generalizer model based on gradient boosting models will facilitate lncRNA identification without imposing arbitrary rules for lncRNA detection.

2.8 Acknowledgements

We thank two anonymous reviewers for their helpful questions and comments.

2.9 Funding

This work was supported by grants from the Ontario Research Fund-Research Excellence and Natural Science and Engineering Research Council of Canada to EAW (RGPIN-2015-06530) and to GBG (RGPIN-2015-04477). The funding bodies supported the publication costs of this manuscript.

2.10 Availability of data and materials

The first version of code used in this work is available at www.github.com/gbgolding/crema. Random protein and lncRNA datasets used for machine learning model training are attached as supplementary material.

2.11 Authors' contributions

The idea of this article was conceived by all authors. CMAS developed the tool, acquired the datasets, obtained and analyzed the results, and prepared the manuscript. GBG and EAW supervised the analysis, and edited the manuscript. All authors have read and approved the manuscript.

2.12 Ethics approval and consent to participate

Not applicable as this study did not use plant or animal material, and instead genomic and transcriptomic data available from various databases.

2.13 Consent for publication

Not applicable.

2.14 Competing interests

The authors declare that they have no competing interests.

2.15 Supplemental Files

Supplemental file 1 .txt — Non-coding RNA search terms
Available at publication: `https://doi.org/10.1186/s12864-018-4665-2`.

Supplemental file 2 .zip — Random protein training data sets, lncRNA data sets
Fasta files of protein coding and lncRNA sequences in data sets used for training machine learning classifiers. Available at publication: `https://doi.org/10.1186/s12864-018-4665-2`.

Supplemental file 3 .pdf — Distribution of predicted lncRNA scores
Figure and table of distribution of scores (Figure S1.1 and Table S1.1) are found in Appendix A and are available at publication: `https://doi.org/10.1186/s12864-018-4665-2`.

Supplemental file 4 .pdf — Comparison of predicted lncRNAs to CPAT results
Table of results and explanation of additional test (Table S1.2) is found in Appendix A and is available at publication `https://doi.org/10.1186/s12864-018-4665-2`.

Chapter 3

Molecular traits of long non-protein coding RNAs from diverse plant species show little evidence of phylogenetic relationships

Caitlin M.A. Simopoulos, Elizabeth A. Weretilnyk, G. Brian Golding

3.1 Preface

Chapter 3 describes how the long non-protein coding RNA (lncRNA) prediction tool (as described in Chapter 2) can be applied to comparative evolutionary studies. Using RNA sequencing data gathered from the Sequence Read Archive (SRA), we predicted lncRNAs from evolutionarily diverse plant species that were grown under ideal conditions without exposure to stress. We explored the contributions of lncRNAs to the genomes of the tested species by quantifying predicted lncRNA numbers that were represented in available reference annotations. This work also contributes to knowledge on lncRNA evolution by estimating phylogenetic signal in select molecular features of lncRNAs. Chapter 3 was submitted on January 9, 2019 for consideration of publication in G3.

I made significant contributions to this study. I jointly conceived of the experiment with E.A. Weretilnyk and G.B. Golding. I gathered data, mapped reads, assembled transcripts and predicted lncRNAs and completed all consequent statistical and evolutionary analyses. I wrote the first version of the manuscript which was edited and approved by E.A. Weretilnyk and G.B. Golding. E.A. Weretilnyk and G.B. Golding supervised the analyses and writing of the manuscript.

3.2 Abstract

Long non-coding RNAs (lncRNAs) represent a diverse class of regulatory loci with roles in development and stress responses throughout all kingdoms of life. LncRNAs, however, remain under-studied in plants compared to animal systems. To address this deficiency, we applied a machine learning prediction tool, Classifying RNA by Ensemble Machine learning Algorithm (CREMA), to analyse RNAseq data from 11 plant species chosen to represent a wide range of evolutionary histories. Transcript sequences of all expressed and/or annotated loci from plants grown in unstressed (control) conditions were assembled and input into CREMA for comparative analyses. On average, 6.4% of the plant transcriptomes were identified by CREMA as encoding lncRNAs. Gene annotation associated with the transcripts showed that up to 99% of all predicted lncRNAs for *Solanum tuberosum* and *Amborella trichopoda* were missing from their reference annotations whereas the reference annotation for the genetic model plant *Arabidopsis thaliana* contains 96% of all predicted lncRNAs for this species. Thus a reliance on reference annotations for use in lncRNA research in less well-studied plants can be impeded by the near absence of annotations associated with these regulatory transcripts. Moreover, our work suggests that molecular traits of plant lncRNAs display different evolutionary patterns than all other transcripts in plants and have molecular traits that do not follow a classic evolutionary pattern as suggested by phylogenetic signal analysis. Specifically, GC content was the only tested trait of lncRNAs with significant high phylogenetic signal, contrary to high signal in all tested molecular traits for other transcripts in our tested plant species.

Keywords: lncRNA, CREMA, phylogenetic signal, molecular traits, transcriptome, RNASeq, annotation, evolution

3.3 Introduction

Long non-protein coding RNAs (lncRNAs), a heterogeneous class of regulatory transcripts, remain greatly understudied in plant species. Although these transcripts have been implicated in development and stress responses of plants, only 13 of these transcripts have been empirically functionally characterized to date (Wang and Cheksnova 2017; Nejat and Mantri 2018; Zhao et al. 2018b). While researchers often focus on computational prediction of these transcripts, particularly lncRNAs expressed under stressful conditions, biological insights on the evolution, mechanisms and function of lncRNAs remain uncertain.

Simopoulos et al. (2018) reported that the genome of *Eutrema salsugineum*, an extremophile, contains a lower proportion of putative lncRNAs in comparison to the genome of model plants *Arabidopsis thaliana* and *Oryza sativa*. A lower number of predicted lncRNAs in *E. salsugineum* is surprising due to the naturally high capacity of this species to tolerate extreme environmental conditions (Champigny et al. 2013; Kazachkova et al. 2018) and the oft-cited association between expressed lncRNAs and stress responses (Wang et al. 2017; Xu et al. 2017b). *E. salsugineum*'s unexpectedly low number of predicted lncRNAs compared to its close and more stress sensitive relative, *A. thaliana*, leads to questions of potential natural variation in lncRNA number. However, the differences in predictions of lncRNAs in these species may be due to data availability as few plant species have had their reference annotation updated regularly in genomic databases. For example, novel gene information has yet to be updated for *E. salsugineum* since the official reference genome was released in 2013 (Yang et al. 2013) although Champigny et al. (2013) presented an additional 665 transcriptional units for which the reference genome had no annotation. Recently, Yin et al. (2018) have added to the number of novel transcripts in *E. salsugineum* with evidence of expression of an additional 65 transcripts with neither update available in the reference annotation of *E. salsugineum*.

LncRNAs may be missing from genome annotations because they are difficult to identify due to their low, tissue- and condition-dependent expression (Derrien et al. 2012). Further, contrary to protein-coding genes and other non-coding loci, the evolution of lncRNAs is not well understood. Limited nucleotide conservation has been identified in mammalian lncRNAs (Hezroni et al. 2015), and structural conservation remains controversial (Rivas et al. 2017). Instead of using homology, distinguishing traits such as transcript length (Kapranov et al. 2007), open reading frame (ORF) ,or lack of, length (Kapranov et al. 2007), GC content (Niazi and Valadkhan 2012), and number of exons in a transcript (Derrien et al. 2012), are often used in lncRNA prediction studies. Detected phylogenetic signal in traits of transcripts, rather than sequence homology, can indicate that trait values follow the expected evolutionary patterns of tested species. For example, high phylogenetic signal implies traits are more similar in closely related species, whereas low phylogenetic signal suggests the opposite: less similarity in tested traits than expected in closely related species. Identifying which evolutionary process may be influencing a significant phylogenetic signal is complex and many different processes are associated with both high or low signal estimates (Revell et al. 2008).

High phylogenetic signal detected using a signal estimation method that considers evolution following a random walk, as in Brownian motion, can be observed in both natural selection and genetic drift scenarios (Revell et al. 2008). Conversely, low detected phylogenetic signal can be inferred as the lack of similarity in tested traits, as opposed to divergence of traits, and is common in adaptive radiation or other fast adaptive processes (Kamilar and Cooper 2013). Data that fit an Ornstein-Uhlenbeck process, however, suggest an adaptive process. First described by Hansen (1997), the Ornstein-Uhlenbeck process allows for a random walk, similar to Brownian motion, but also for species to evolve towards an adaptive peak or fitness optimum. Furthermore, local estimates of Moran's I, based on the concept of spatial autocorrelation, estimate phylogenetic signal throughout evolutionary time. Positive autocorrelation indicates similarity of trait values at a given phylogenetic distance, while negative autocorrelation suggests dissimilarity at a given phylogenetic distance. Diniz-Filho (2001) has shown, however, that it is the changes in local autocorrelation over phylogenetic distance beyond a significance threshold that may be influenced by evolutionary processes. A trait following an Ornstein-Uhlenbeck adaptive process would have a reduced phylogenetic distance at which I crosses the threshold of no significant autocorrelation, also called a "phylogenetic patch", compared to a trait following the Brownian motion model of evolution.

In this study, we predicted lncRNAs from transcriptomes of 11 plant species with widely different evolutionary histories. Transcripts were assembled from RNASeq data without restriction of existing reference annotation in order to obtain a representation of all expressed loci in each study without reliance on accompanying known transcriptional units. Transcript sequences were then input into CREMA (Simopoulos et al. 2018) for accurate lncRNA prediction and ranking. Following lncRNA prediction, we observed that up to 99% of predicted lncRNAs may not be present in their corresponding reference annotation. Thus, we caution that researchers should not rely only on publicly available annotation for lncRNA research. Finally, as there has been little evidence for sequence conservation in lncRNAs between species in different families (Nelson et al. 2016), a phylogenetic signal was not expected in the distinguishing molecular traits of lncRNAs, such as transcript length and GC content. However, our comparative study detected a consistently high phylogenetic signal in GC content of lncRNAs with no conservation found for the other traits tested in these regulatory transcripts. In particular, GC content differences relative to protein-coding RNA represents a trait that could help researchers distinguish putative functional lncRNAs from non-functional and spurious transcription, or fragmented protein-coding RNAs.

3.4 Results

3.4.1 Multispecies lncRNA prediction

For this work, plant species with diverse and divergent evolutionary histories were chosen for lncRNA comparisons. Included in the analysis are various Angiosperms (including both monocots and dicots), a Lycophyte, a Bryophyte and an algal species (See Table 3.1 in Methods). As novel transcripts were of importance to this work, RNASeq data from published experiments were used to assemble transcripts. After read mapping to appropriate plant genomes, transcripts were assembled using StringTie allowing for identification of novel transcripts. Sequences of assembled transcripts were input into CREMA, a lncRNA prediction tool (Simopoulos et al. 2018) and total lncRNA numbers in each plant species are described in Figure 3.1 and Table 3.2. Ranked prediction scores of all transcripts in each species are available on GitHub: `https://github.com/caitsimop/lncRNA-compGenomics`. The percentage of total transcripts predicted as lncRNAs range from 3% in *E. salsugineum* to 16.6% in *Amborella trichopoda* with a mean percentage of 6.4% ±1.1% for the 11 analyzed plant species (Table 3.2).

Table 3.1: Information of the data sources of all RNASeq libraries

Species	# high quality reads	# mapped reads	BioProject	SRA	Source of RNASeq	Genome Source
Solanum tuberosum	12,469,853	11,106,056	PRJNA311702	SRR3162008	Sprenger et al. (2016)	Sharma et al. (2013)
Solanum lycopersicum	18,624,814	18,277,415	PRJNA307656	SRR3095793	Cardenas et al. (2016)	Tomato Genome Consortium (2012)
Eutrema salsugineum	49,522,792	46,130,371	PRJNA494564	SRR7962298	This manuscript	Yang et al. (2013)
Arabidopsis thaliana	23,490,825	23,111,430	PRJNA186843	SRR2079778	Woo et al. (2016)	Cheng et al. (2017)
Zea mays	15,141,539	14,481,792	PRJNA269060	SRR1688291	Gonzalez-Munoz et al. (2015)	Schnable et al. (2009)
Oryza sativa	23,501,682	22,145,297	PRJNA301554	SRR2931278	Wilkins et al. (2016)	Ouyang et al. (2007)
Amborella trichopoda	17,913,230	17,355,462	PRJNA212863	SRR5293262	*Amborella* Genome Project (2013)	*Amborella* Genome Project (2013)
Selaginella moellendorffi	108,008,790	92,873,912	PRJNA351923	SRR4762345	James et al. (2017)	Banks et al. (2011)
Physcomitrella patens	10,520,395	8,243,406	PRJNA265205	SRR1553300	Frank and Scanlon (2015)	Lang et al. (2018)
Chlamydomonas reinhardtii	22,002,690	21,222,625	PRJNA264777	SRR1622084	Panchy et al. (2014)	Merchant et al. (2007)
Boea hygrometrica	16,972,867	15,594,598	PRJNA210992	SRR929426	Xiao et al. (2015)	Xiao et al. (2015)

To determine how reference annotation may affect lncRNA research in plants, all assembled transcripts from each species, including novel transcripts, were compared to those found in the corresponding reference annotation. Transcripts that were not found in reference annotation and also predicted as a putative lncRNA were identified and are referred to as “novel” lncRNAs throughout this manuscript. The proportion of novel lncRNA in all predicted lncRNAs ranged among species from a low 4.5% predicted in *A. thaliana* and a high 99.6% in *Solanum tuberosum* (Figure 3.1). Because *A. thaliana* is a well studied model plant with an almost fully annotated genome, we expected this species to have fewer novel transcripts assembled from the RNASeq data. Additionally, we expected that most lncRNAs predicted from the *A. thaliana* transcriptome to already be found in the reference annotation. The high percentage of lncRNAs found in the reference annotation of *A. thaliana* indicates that CREMA makes accurate predictions and suggests that the lower percentages of known lncRNAs identified in the other species are due to incomplete annotations (Figure 3.1).

Table 3.2: Number of predicted lncRNAs in each species

Species	Total # of assembled transcripts	# predicted lncRNAs	% lncRNAs
Solanum tuberosum	73,656	3,783	5.1%
Solanum lycopersicum	43,936	2,721	6.2%
Eutrema salsugineum	34,862	1,040	3.0%
Arabidopsis thaliana	61,480	2,918	4.8%
Zea mays	95,713	7,225	7.6%
Oryza sativa	66,562	3,753	5.6%
Amborella trichopoda	42,118	6,972	16.6%
Selaginella moellendorffi	33,266	2,269	6.8%
Physcomitrella patens	88,649	4,648	5.2%
Chlamydomonas reinhardtii	21,467	1,383	6.4%
Boea hygrometrica	58,531	1,796	3.0%

3.4.2 Phylogenetic signal in molecular traits of plant lncRNAs

We first considered overall trends in typical distinguishing traits of lncRNAs: ORF length, GC content, number of exons, and transcript length. All species showed a similar trend where putative lncRNAs had a lower GC%, fewer exons, and shorter ORF length compared to the other transcripts in their corresponding transcriptomes (Figure 3.2). Length of transcripts,

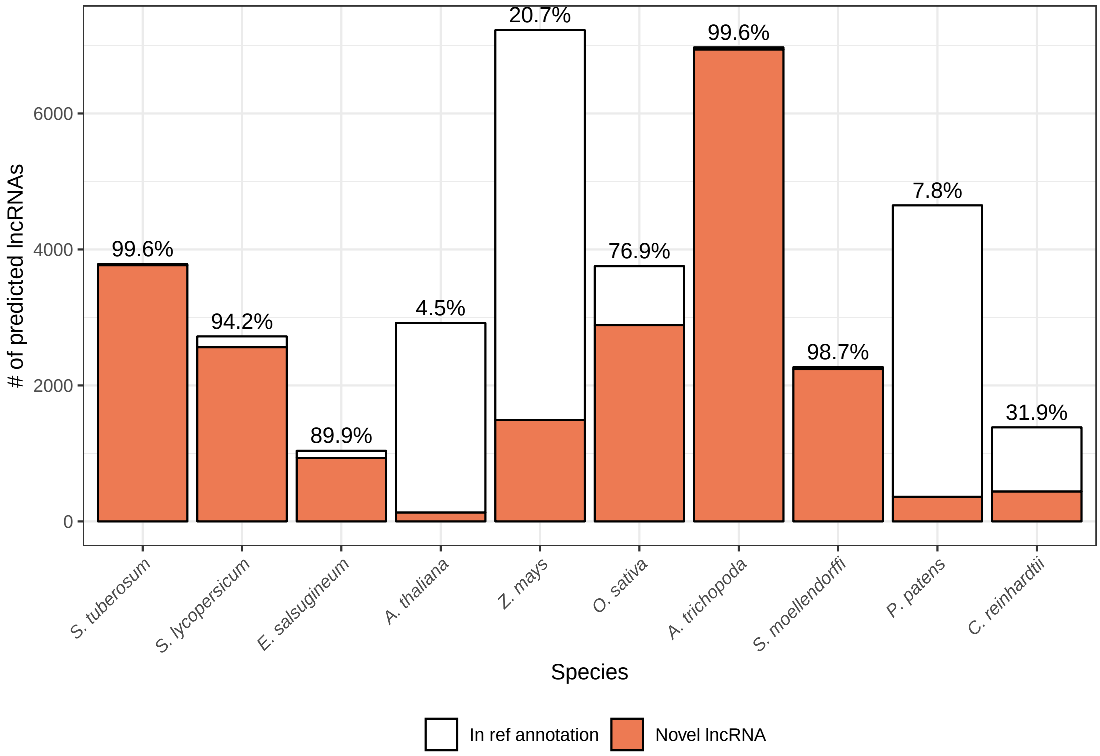


FIGURE 3.1: Total predicted lncRNAs from 10 plant species. The counts of putative lncRNAs are categorized by transcripts that appear in the reference annotation of each species (white) and novel transcripts, or those that did not appear in transcriptome annotation (coral). The percentages of novel transcripts predicted as lncRNAs appear above each bar.

however, deviated from this trend where *Zea mays*, *Selaginella moellendorffi* and *A. trichopoda* all have putative lncRNAs longer than other transcripts in their transcriptome (Figure 3.2). The deviation from the expected trend of shorter lncRNA transcripts in three of the selected species suggests that transcript length may not be a useful distinguishing trait in lncRNA prediction.

We tested for phylogenetic signal in mean trait values of the four molecular traits previously mentioned in both lncRNAs and all transcripts other than lncRNAs. Phylogenetic signal estimates were calculated using three different indices, Moran's I, Pagel's λ and Blomberg's K, that employ two different models of evolution, Brownian motion and autocorrelation. We estimated phylogenetic signal in all species but *Boea hygrometrica* due to the incomplete status of its genome annotation.

Since each phylogenetic signal estimation index is based on different concepts, all estimates cannot be interpreted the same. Firstly, Moran's I is a measure of autocorrelation (Gittleman and Kot 1990). Autocorrelation, when referring to phylogenetic signal, indicates how correlated traits are in terms of phylogenetic distance. Due to the use of 10 species, Moran's I must be compared

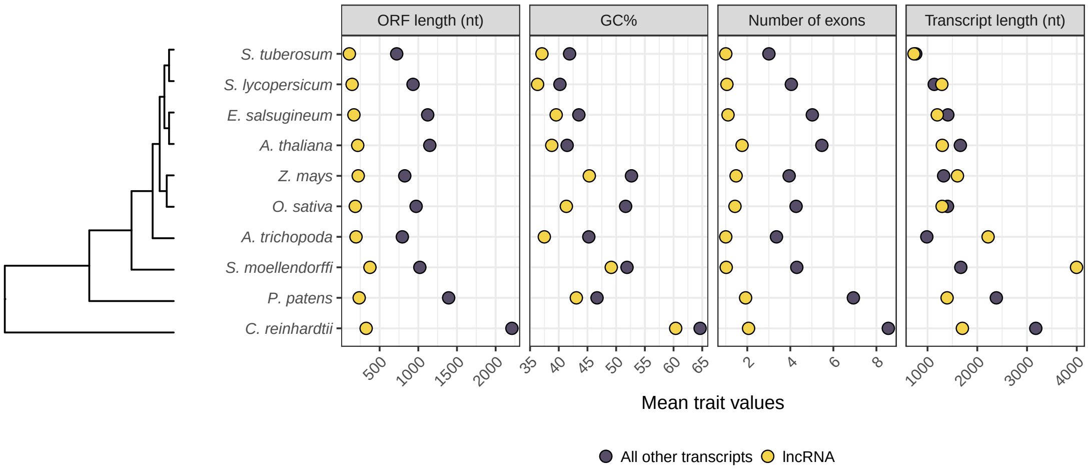


FIGURE 3.2: Mean trait values of transcripts predicted as lncRNAs (yellow) and all other assembled transcripts (purple). Species are ordered as per phylogenetic relationships.

to a calculated threshold of -0.111 (Keck et al. 2016) to determine significant autocorrelation and phylogenetic signal. A significant estimate greater than -0.111 indicates positive significant global autocorrelation and that trait values of closely related species are more similar to each other. Conversely, a significant estimate less than -0.111 suggests global significant negative autocorrelation. The Brownian motion model, originally used to describe the motion of particles suspended in fluid, is another model used to describe how traits evolve through time. In the case of phylogenetic signal, a trait following the Brownian motion model exhibits a random walk where the value of the trait can change in any direction at any time. Pagel's λ uses this Brownian motion model and can be interpreted as the transformation the phylogeny requires to explain trait distribution if the trait followed Brownian motion (Pagel 1999). Thus, a value of 1 would indicate a phylogeny as expected under Brownian motion and high phylogenetic signal, and a significant value of 0 would mean a trait distribution that does not follow Brownian motion, and consequently, low phylogenetic signal. Finally, Blomberg's K, which also uses the Brownian motion model, can be interpreted as the ratio of observed values over expected values if the trait follows the Brownian motion model (Blomberg et al. 2003). A value of $K = 1$ can be interpreted as a trait distribution following Brownian motion, and as K becomes larger than 1, a stronger signal is detected. Conversely a value of $K < 1$ indicates low phylogenetic signal, and less similarity between closely related tested species.

Table 3.3 shows phylogenetic signal estimates using all three indices for each of the four molecular traits. The mean trait values and phylogenetic relationships of the tested species are presented in Figure 3.2. In predicted lncRNAs, GC content was the only trait that demonstrates high phylogenetic signal using all phylogenetic signal detection indices (Table 3.3). While ORF length was had a significant positive global autocorrelation ($I = 0.040$; Table 3.3), a value of K

< 1 indicates less similarity than expected under Brownian motion, suggesting unclear phylogenetic signal estimation. Blomberg's K also indicates less similarity than expected in the number of exons of lncRNAs, however no other index displayed detectable significant phylogenetic signal. Transcript lengths of lncRNAs in tested species also demonstrate a moderate positive global autocorrelation. Conversely, all four traits consistently had significant phylogenetic signal estimates when all transcripts other than lncRNAs were evaluated, although λ for ORF length, number of exons and transcript length were slightly less than 1.

3.4.3 Evolutionary processes and phylogenetic signal

We examined traits with an estimated $K > 1$ with an evolutionary model that may suggest natural selection, the Ornstein-Uhlenbeck process (Hansen 1997), because high phylogenetic signal defined as $K > 1$ (Kamilar and Cooper 2013) can indicate similarity by both genetic drift and natural selection. In lncRNAs, GC content is the only trait with $K > 1$, and has significant phylogenetic signal detected using all three indices (Table 3.3). We compared the fit of a Brownian motion model versus an Ornstein-Uhlenbeck model in our data using log likelihood values and a chi square test for significance estimates. Although the Ornstein-Uhlenbeck model had the smallest log-likelihood, a chi square test indicated that there was no significant fit difference when comparing the Brownian motion and Ornstein-Uhlenbeck model (p=0.81). Because the Brownian motion model has the least number of parameters, this suggests that a Brownian motion model is the most reasonable fit for the data, and there is a lack of evidence for an adaptive process.

All four traits of all transcripts other than lncRNAs had significant high phylogenetic signal when estimated using Blomberg's K ($K > 1$), therefore we also tested for a better fit explained by the Ornstein-Uhlenbeck process. Again, the Ornstein-Uhlenbeck process was not a significantly better fit than a Brownian motion model (ORF length: p = 1, GC content: p = 1, number of exons: p = 1, transcript length: p = 0.75).

TABLE 3.3: Phylogenetic signal estimates

Feature	lncRNA			All other transcripts		
	I	λ	K	I	λ	K
ORF length	0.040*	0.975	0.621*	0.010*	0.974*	1.746*
GC content (%)	0.032*	1.027*	1.614*	0.048*	1.020*	1.038*
Number of exons	-0.053	0.620	0.336*	0.010*	0.922*	1.068*
Transcript length	-0.020*	1.007	0.642	0.038*	0.953*	1.436*

* $p < 0.05$
I = Moran's I, K = Blomberg's K, λ = Pagel's λ

We tested for local autocorrelation at 100 phylogenetic distances considering the most recent common ancestors of all species as a robust approach to an analysis on a small phylogeny. Confidence intervals were computed using 1000 bootstrapping replicates for a non-parametric significance estimate using a calculated threshold of -0.111. Figure 3.3 visualizes the local correlations of traits in both lncRNAs and all other transcripts and are limited to the phylogenetic distances of the tested phylogeny (0-1 phylogenetic distance). We detected significant positive local autocorrelation at short phylogenetic distances in ORF length and GC content of lncRNAs (Figure 3.3). This suggests that closely related species contain lncRNAs with similar ORF lengths and GC content. There was no significant autocorrelation at any short phylogenetic distances in any tested traits in all other transcripts (Figure 3.3). Detected phylogenetic patches are shorter in the ORF and transcript lengths of lncRNAs compared to all other transripts. The opposite is true in the GC content of lncRNAs, where longer phylogenetic patches are observed. Shorter phylogenetic patches suggest an adaptive process as described by an Ornstein-Uhlenbeck model, rather than genetic drift.

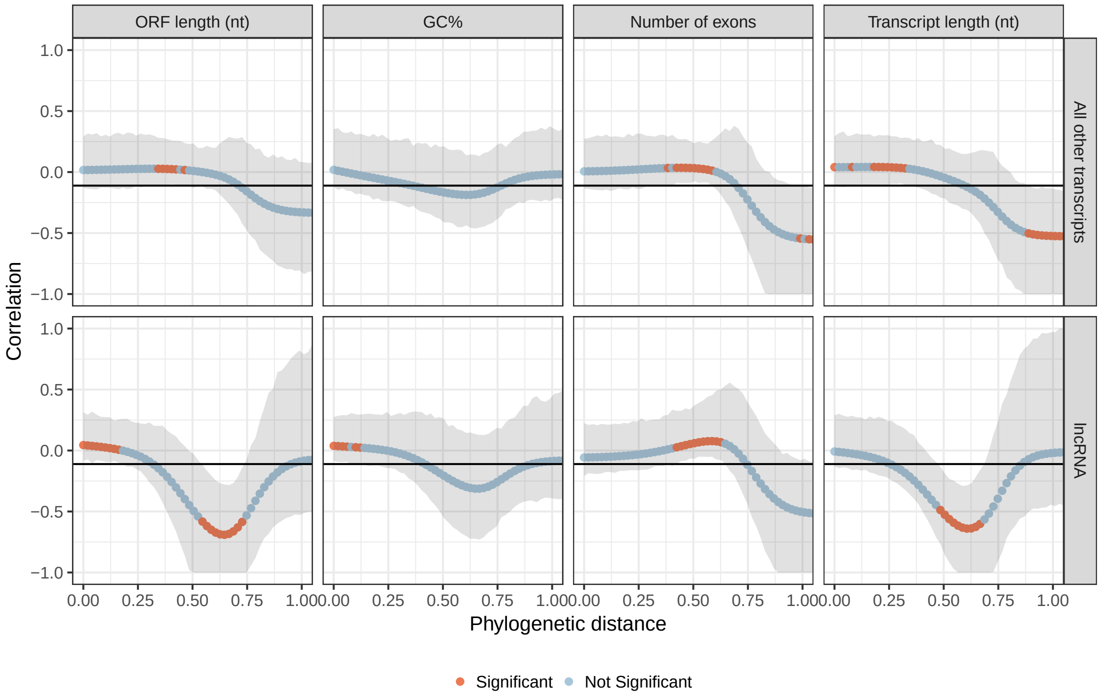


Figure 3.3: Moran's I local correlogram of mean trait values in lncRNAs and All Other Transcripts. Coral points indicate significant phylogenetic signal at a particular phylogenetic distance. The horizontal line represents a value of the null hypothesis that no phylogenetic signal is detected. The null hypothesis value is -0.111, or $-1/(n-1)$ where $n = 10$, or the number of tested species. The 95% confidence intervals, computed using bootstrapping, are also plotted and were used to identify significant values.

3.5 Discussion

We used raw RNASeq data from multiple independent studies to make inferences on the numbers of predicted lncRNAs in 11 phylogenetically divergent plant species, and to identify putative phylogenetic signal in these regulatory loci. Our data-mining approach enabled us to use the same protocols for read mapping, transcript assembly, and lncRNA prediction for each species. In performing the same read-mapping and lncRNA prediction protocols, we were able to address a concern raised by Kapusta and Feschotte (2014) that comparisons between lncRNA numbers in animals can be misleading when prediction numbers are products of meta-analyses involving different prediction methods and lncRNA criteria. We found that the percentage of transcripts predicted as lncRNAs was on average 6.4% with percentages ranging from 16.6% in *A. trichopoda* to 3% in *E. salsugineum* and *B. hygrometrica*. These estimates for lncRNA contributions to plant genomes are higher than comparable values for humans, where a review by Palazzo and Lee (2015) identified that generally less than 1% of the human genome is predicted as lncRNAs.

The review by Kapusta and Feschotte (2014) also included a meta-analysis describing variation in predicted lncRNA numbers among multiple animal species, a comparison similar to our observed prediction numbers in plants. In addition to their concern about transcriptome data arising from different methodologies, Kapusta and Feschotte (2014) also raised the issue of temporal and location specific lncRNA expression. We share a comparable concern in that plant lncRNAs have yet to be predicted in all tissue types for all developmental time points in all possible environments, so undoubtedly the number of putative lncRNAs detected in plants will increase over time. In our study, we identified 2918 putative lncRNAs in *A. thaliana* plants that were grown under conditions designed to avoid exposing plants to sources of stress. In contrast, although using different prediction methods, Zhao et al. (2018b) identified 6150 putative lncRNAs in *A. thaliana* plants undergoing cold, ABA and drought treatments. This difference in predicted lncRNAs is consistent with the expectation that lncRNAs likely play a role in stress responses and hence finding increased diversity and transcript abundance in stressed relative to unstressed plants. Interestingly, Zhao et al. (2018b) found that lncRNAs in *A. thaliana* are shorter and have fewer exons than all other transcripts, observations that agree with our study (Figure 3.2). Thus our machine learning-based methodology that was trained on only empirically characterized, functional lncRNAs and the filtering method employed by Zhao et al. (2018b) lead to similar conclusions on traits shared by lncRNAs that distinguish them from other transcripts.

The reported differences in lncRNA numbers between humans and plants described above is interesting and merits future research, but equally intriguing is our finding of the large variation in lncRNAs predicted from transcriptomes between different plant species. A question to raise is whether the number of lncRNAs predicted for *E. salsugineum*, *B. hygrometrica* and *A. trichopoda* are truly extreme examples in lncRNA prediction numbers, or are there naturally considerable ranges of lncRNA contribution to diverse plant genomes?

The genome of *A. trichopoda*, the sister taxa to all other extant angiosperms, represents a unique evolutionary history. During genome annotation, The Amborella Genome Project (2013) observed a larger number of the atypical 23 to 24nt plant miRNAs than expected as they were found in two times greater frequency than any other land plant. Additionally, eight predicted miRNA families in *A. trichopoda* have evidence of loss in more recent angiosperms (*Amborella* Genome Project 2013). The excess of miRNAs in *A. trichopoda* may reflect the high proportion of lncRNAs predicted in this study (at 16%; Table 3.2) as miRNA progenitors are considered to be lncRNAs (Saini et al. 2008).

Two plants, namely *E. salsugineum* and *B. hygrometrica*, were found to have the lowest proportion of lncRNAs in their transcriptomes (Table 3.2). *E. salsugineum* represents a plant with a halophytic life strategy and a capacity to tolerate a variety of extreme environmental conditions (Kazachkova et al. 2018). Indeed, *E. salsugineum* has been used as a model plant in stress response studies due to its naturally high tolerance to abiotic stresses such as salt (Taji et al. 2004), cold (Griffith et al. 2007), drought (MacLeod et al. 2014), and nutritional deficiencies (Velasco et al. 2016). Moreover, *E. salsugineum* shows constitutive expression of genes reported to be stress-responsive in many plants (Taji et al. 2004; Gong et al. 2005; Wong et al. 2006; Velasco et al. 2016). *B. hygrometrica*, aptly named "the resurrection plan", is also considered an extremophile by virtue of its capacity to survive desiccation (Xiao et al. 2015). However, *B. hygrometrica* shows different expression pattern changes when experiencing stress compared to *E. salsugineum*. Zhu et al. (2015) did not observe constitutively high expression of stress tolerance genes in *B. hygrometrica* during desiccation. Instead, *B. hygrometrica* seemed to require gradual dehydration priming for survival after rehydration post-desiccation (Zhu et al. 2015). *B. hygrometrica* plants that have been consequently rehydrated after this dehydration "training" have expression patterns more similar to desiccated plants than those without drought priming (Zhu et al. 2015). In other words, after experiencing a first gradual dehydration there are expression differences between well-watered *B. hygrometrica* plants and ones that experienced desiccation. The observation that *B. hygrometrica* plants can show "preparedness" among expressed genes normally responsive to a stressful condition is somewhat analogous to the constitutive nature of expressed genes in *E. salsugineum*. Specifically, *E. salsugineum* plants grown in the absence of high salt display the expression of genes reported to be salt-responsive in other plants (Taji et al. 2004; Wong et al. 2006). Interestingly, both *E. salsugineum* and *B. hygrometrica* display a low proportion of predicted lncRNAs in their transcriptome, suggestive of a possible connection between high natural stress tolerance and low lncRNA number (Table 3.2). Conceivably, with stress-related genes constantly expressed under a primed condition, a plant adapted to an extreme environment may not require the precise regulation conferred by the recruitment of diverse lncRNAs, an important role proposed for the function of lncRNAs in plant stress responses.

E. salsugineum and *A. trichopoda* have distinct evolutionary patterns and both species have few putative lncRNAs present in their reference annotations. In total, five of the ten tested plant species had less than 50% of predicted lncRNAs in their respective genome annotations. As genome annotation often relies on homology of predicted genes for functional annotation (Bolger et al. 2018), particularly homology to *A. thaliana* protein-coding genes, lncRNAs can often be left out of genome annotation. Researchers studying plant lncRNAs frequently rely on bioinformatic analyses to assemble novel transcripts for lncRNA prediction (Liu et al. 2018b; Shuai et al. 2014), indicating that missing lncRNA annotation should be taken into consideration in forthcoming genome annotation projects. Similar to our work, Jackson et al. (2018) recently described mis-annotation of lncRNAs in mammalian genomes. Gaps in annotation and the ensuing problem with lncRNA identification is exacerbated by the fact that lncRNAs do not follow classic evolutionary conservation. Instead, lncRNAs mostly depict a positional conservation pattern rather than transcript sequence conservation making functional predictions also difficult (Hezroni et al. 2015). A lack of extensive lncRNA conservation between species led to our investigation into phylogenetic signal detection in molecular traits of lncRNAs.

While few studies have compared lncRNA sequence conservation between plant and animal systems, conservation within more closely related species has been a topic of recent interest. Hezroni et al. (2015) describe little conservation of entire lncRNA sequences between vertebrate species with divergence over 50 million years ago. Instead of overall sequence conservation, the authors observed short regions of homology and syntenic conservation in vertebrate lncRNAs (Hezroni et al. 2015). Notably, when looking at specific classes of lncRNAs in animals, Ultra Conserved Regions (UCRs) and Human-Accelerated Regions (HARs), both regions found in validated lncRNAs, offer opposing ideas to lncRNA conservation. UCRs (Calin et al. 2007) describe regulatory sequences that are 100% conserved in humans, mice and rat genomes. HARs instead, found in lncRNAs *HAR1A* and *HAR1B*, describe regions in vertebrate genomes with an extremely high number of mutations in human sequences (Pollard et al. 2006). This discrepancy suggests that lncRNA homology is not completely straightforward and may vary depending on lncRNA classification. However, lncRNA classification is complex and still lacks a set of agreed-upon rules for each lncRNA type. A recent review even suggests that there exists over 50 overlapping lncRNA categories in the literature, not all of which are based on function or structure (St. Laurent et al. 2015).

In plant species, lncRNAs homology has been shown to be virtually non-existent outside of the family classification. Only 1% of predicted *A. thaliana* (Brassicaceae) lncRNAs were identified as homologous in *Tarenaya hassleriana* (Cleomaceae) (Nelson et al. 2016). Interestingly, although distantly related, human and plant lncRNAs bear similarities in number of exons, and transcript and ORF length. For example, Derrien et al. (2012) describe human lncRNAs as being shorter than protein coding genes, and by commonly having fewer exons which was also observed in our plant species analyses (Figure 3.2). However, human lncRNAs are typically spliced and most often have two exons, while our study suggests plant lncRNAs are more often unspliced and comprised of a single exon (Figure 3.2). Domination of single exon novel lncRNAs in our work may be indication of genomic contamination in source data or bioinformatic transcript assembly artefacts. Nevertheless, genomic contamination is unlikely seen in all ten independent experiments, and transcript assembly artefacts more likely result from *de novo* assembly rather than genome guided mapping. Further, Haerty and Ponting (2015) also observed a lower GC content in lncRNAs than the protein coding genes of metazoan lncRNAs, with single-exon lncRNAs having the lowest GC content of all exon-types, reminiscent of the uni-exonic plant lncRNA majority of our analyses. However, the extent to which conserved traits typify plant and animal lncRNAs is difficult to assess at present. CREMA is trained on only validated lncRNA and may be subject to a prediction bias to animal-like lncRNA sequences given that non-plant sources currently comprise the majority of validated lncRNAs (Simopoulos et al. 2018). As more plant lncRNA undergo validation, the extent of conservation among lncRNAs from diverse organisms will be easier to detect and estimate with greater precision.

Despite these concerns over bias, among the lncRNAs predicted by CREMA we found phylogenetic signal by at least one method in all four tested traits of lncRNAs: ORF length, GC content, the number of exons and transcript length (Table 3.3). Phylogenetic signal detection in traits of lncRNAs was not expected given the lack of evidence for sequence conservation in this class of RNA (Nelson et al. 2016; Hezroni et al. 2015). If the phylogenetic relationships of species are influencing the tested traits, as indicated by detectable phylogenetic signal, the species can no longer be considered independent. To demonstrate the pitfalls of using a statistical test that assumes independence when data are not independent, we statistically tested for differences in trait values of lncRNAs and all other transcripts in 10 species. Using an ANOVA and subsequent post-hoc tests without consideration of the species' phylogenetic relationships, we found that trait value differences between predicted lncRNAs and all other transcripts were significantly different in 172 of 190 post-hoc tests (Figure S2.1). The differences between the results of analyses that incorrectly assume independence from testing for phylogenetic signal and acknowledging potential non-independence of data underscore the importance of phylogenetic signal detection in data before carrying out statistical anlyses in comparative genomics studies.

While it is possible to identify phylogenetic signal in our data, it is difficult to infer evolutionary processes from phylogenetic signal estimates as many unique processes can invoke a similar signal estimation (Revell et al. 2008). However, the high K estimates in both the GC content trait values of lncRNAs and in all traits of all other transcripts suggests limited similarity of molecular traits in lncRNAs, contrary to overall high phylogenetic signal in the tested traits of all other transcripts. High phylogenetic signal, and consequently, high similarity of GC content in lncRNAs of closely related species is somewhat expected, as Haerty and Ponting (2015) identified evidence of selection on the GC content of intergenic lncRNAs in animal species. As previously mentioned, GC content in animal lncRNAs is lower than in protein-coding genes, mirroring what our work identified in plant lncRNAs. In the GC content of lncRNAs, an Ornstein-Uhlenbeck process model was not found to explain the evolutionary processes more than a Brownian motion model and suggests more evidence for genetic drift than an adaptive process. However, as discussed by Cooper et al. (2016), Ornstein-Uhlenbeck models may be prone to error when tested on small phylogenies with less than 200 species. As such, we cannot discount that a high K estimate may be indicative of selection on the GC content of lncRNAs.

A lack of similarity in ORF length, number of exons, and transcript length of lncRNAs in close relatives may be due to a variety of processes, including but not limited to: stabilizing selection with high selective strength, selection with variable strength that is bounded by phenotypic limits, punctuated divergent selection, or genetic drift of which rate of drift began low and increased towards present time (Revell et al. 2008). Because of the variety of possible complex interpretations of phylogenetic signal and process, Revell et al. (2008) do not recommend over-interpretting evolutionary processes from signal data. We have found, however, unique patterns in the phylogenetic signals of molecular traits of lncRNAs compared to all other transcripts in plant species that imply lncRNAs are not following similar evolutionary trends as most other transcripts. Moreover, the lack of similarity in three of the four tested molecular traits in lncRNAs is of interest and this observation implies that evolution of lncRNAs could be species specific, and is not be easily defined by an over-arching evolutionary process. On the other hand, it is possible that there are subclasses of lncRNAs with conserved molecular traits yet to be defined due to a lack of validated transcripts.

Because CREMA (Simopoulos et al. 2018) predicts lncRNAs using a complex ensemble machine learning model that initially uses ORF length, GC content and transcript length as features for transcript classification, it is possible that high detected phylogenetic signal in these features is a product of the lncRNA prediction tool. However, CREMA's logisic regression ensemble classifier that is used for the final lncRNA prediction does not use molecular traits as prediction features, but instead binary outputs from eight gradient boosting models. Additionally, we have identified low phylogenetic signal in two of these three molecular traits (Table 3.3) suggesting that CREMA is able to predict lncRNAs with varying ORF and transcript lengths.

In this work we show that the annotation status of plant species can affect lncRNAs prediction with up to 99% of predicted lncRNAs missing from reference annotation. While researchers may be striving to increase the volume of lncRNA research, the effort to annotate genomes with lncRNAs is not reflective of the increased interest in this RNA class. As such, we caution researchers interested in these regulatory loci to be wary of relying solely upon genome and transcriptome annotations for lncRNA identification. Additionally, our work shows that plant lncRNAs have inconsistent detectable phylogenetic signal in sequence traits, further confirming the complex evolutionary history of lncRNAs. In particular, the differences in detected phylogenetic signal in lncRNAs compared to all other transcripts suggests that lncRNAs evolve, on average, differently than other loci. Finally, a low proportion of transcripts predicted as lncRNAs in *E. salsugineum* and *B. hygrometrica*, two species highly tolerant to abiotic stress, may indicate that adaptation to extreme conditions may not be orchestrated by many, diverse lncRNAs and that indeed, the converse may be true.

3.6 Methods

3.6.1 Data collection

RNASeq data from multiple plant species were downloaded from the SRA database (See Table 3.1 for accession and SRA IDs). All plants in this analysis have a publicly available sequenced genome. All RNASeq reads except for those from *E. salsugineum* were downloaded from the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). To be considered, plants must have been grown under control conditions without being subjected to stress. For consistency, preference was given to studies that used leaf tissue from mature plants, although use of older seedlings was accepted. Additionally, only RNASeq reads from Illumina technology were considered, however both paired and single end reads were used.

E. salsugineum reads were sequenced using Illumina technology from Shandong ecotype rosette leaves grown under control, unstressed conditions as outlined in MacLeod et al. (2014). Fully-expanded leaves were used for RNA sequencing, and were collected between 8 and 10 hours into the day cycle. RNA was extracted from leaves flash-frozen using liquid nitrogen using a modified hot borate method as described Champigny et al. (2013). A quality control analysis was competed on the RNA using RNA Nano 600 chips on a Bioanalyzer 2100 and purified using three on-column purifications by Genelute mRNA miniprep kit (Cat. No. MRN10, Sigma). Finally, preparation of cDNA for sequencing was performed with the NEBNext multiplex cDNA synthesis kit for Illumina using random hexamers (Cat. No. E7335, New England Biolabs, Ipswich, MA). Cleanup of fragmented RNA was performed with Agencourt AMPure XP Beads (Cat. No. 163987, Beckman Coulter, Mississauga, ON) following the manufacturer's protocol. Raw FASTQ files were deposited to the SRA with submission ID SRR7962298 and BioProject accession PRJNA494564.

3.6.2 Transcript assembly and lncRNA prediction

Reads from all plant species were trimmed using Trimmomatic v0.36 (Bolger et al. 2014) and aligned to their corresponding genomes using STAR v2.5.2b (Dobin et al. 2013) using default settings other than `--outFilterIntronMotifs` set to `RemoveNoncanonical` and `--alignEndsType EndtoEnd`. Aligned reads were assembled into transcripts by StringTie v1.3.4d (Pertea et al. 2015). GTF files of assembled transcripts were merged with GTF files of annotated genomes and are stored on GitHub: `https://github.com/caitsimop/lncRNA-compGenomics`. Alignment quality was tested using gffcompare v0.10.4 (`https://github.com/gpertea/gffcompare`) by comparing assembled transcript GFF files with reference genome GFF files. Alignment quality metrics were used to confirm alignment quality and transcript assembly quality using accuracy and precision values File S2. Gffcompare output was also used to identify novel transcripts and to quantify transcript exon numbers in each RNASeq library.

3.6.3 Identifying lncRNAs from RNASeq data

Assembled transcript sequences were input into CREMA (https://github.com/gbgolding/crema, Simopoulos et al. 2018) for ranked lncRNA prediction. The number of lncRNAs in each species was calculated as a percentage of all transcripts (the sum of novel assembled transcripts and transcripts in reference annotation). The percentage of lncRNAs was used for normalization across all studied plant species to ensure appropriate comparisons to species with different sized transcriptomes.

3.6.4 Phylogenetic signal in lncRNA traits

Four continuous molecular traits were chosen for phylogenetic signal analysis on predicted lncRNAs: 1. Number of exons in transcript, 2. GC content of transcript, 3. Length of transcript, and 4. Length of maximal ORF. Features were extracted from transcript sequences and gffcompare outputs using a custom Python script. The `phylosignal` R package (Keck et al. 2016) was used to detect phylogenetic signal in lncRNAs, all other transcripts, and the differences between lncRNAs and all other transcripts for each trait in all species except for *B. hygrometrica*. Separate phylogenetic signal tests were completed for each trait. Although we expect there to be correlation between transcript length and ORF length, we did not observe a correlation, particularly in lncRNAs. Phylogenetic signal of the mean value of the four traits was calculated using three separate methods: Moran's I (Moran 1948; Gittleman and Kot 1990), Blomberg's K (Blomberg et al. 2003) and Pagel's λ (Pagel 1999). Local autocorrelation estimates at 100 phylogenetic distance points were also computed using Moran's I and `phylosignal` to identify the location and sign of the detected autocorrelation. To identify significant autocorrelation estimates, 1000 bootstrap replicates were used for 95% confidence interval calculation. Autocorrelation estimates were considered significant if 95% confidence intervals did not overlap the null hypothesis threshold of -0.111. The null hypothesis that there is no detectable phylogenetic signal, or, autocorrelation, was a threshold of $-1/(n-1)$ where $n = 10$, or the number tested species, as suggested by Keck et al. (2016). Because branch lengths were required by the `phylosignal` package, branch lengths were estimated from a MAFFT v7.205 (Katoh et al. 2002) alignment of *rps16*, *atp2*, *18s*, *26s* and *SMC1* (File S3) using the `dnaml` program in PHYLIP (Felsenstein 1993) (phylogeny with branch lengths is available in Figure S2.2). *B. hygrometrica* was not included in phylogenetic signal analysis due to the limited percentage of annotated loci in genome annotation. The tree topology of land plants as reported by the *Amborella* Genome Project (2013) was used, and branch lengths were estimated from this topology. Branch lengths representing site changes were converted to relative age of branches using the R package `ape`

(Paradis and Schliep 2018). A lambda value of 0 was chosen from 0, 0.1 and 1 after testing for the lowest log likelihood of lambda options.

Trait values with high K values (K >1) were chosen for further testing for better fit to models that consider an Ornstein-Uhlenbeck process and may indicate selection on traits. Traits values were fit with macroevolutionary models using the `geiger` (Harmon et al. 2008) R package and the `fitContinuous` function. Both "BM" (Brownian motion) and "OU" Ornstein-Uhlenbeck models were considered. Fit was tested for using the log liklihood estimate while considering the number of parameters in each model.

3.6.5 Data availability

Raw FASTQ RNA sequencing data is available in the SRA with submission ID SRR7962298 and BioProject accession PRJNA494564. Ranked lncRNA prediction scores and GFF files of assembled transcripts are available on the author's GitHub `https://github.com/caitsimop/lncRNA-compGenomics`. Results of post-hoc t-tests that do not consider phylogenetic relationships are described in Figure S2.1. Quality of transcriptome assemblies are available in File S2. The FASTA files of genes used in the estimation of branch lengths is available in File S3. The phylogenetic tree branch lengths adjusted relative to time are found in Figure S2.2.

3.7 Acknowledgements

We thank the Ontario Research Fund-Research Excellence and Natural Science and Engineering Research Council of Canada for our funding support (EAW: RGPIN-2015-06530 and GBG: RGPIN-2015-04477). We also thank Dr Mitchell MacLeod for growing *E. salsugineum* plant material and for preparing the RNASeq library. Finally, we thank Dr Benjamin Furman for his invaluable comments on manuscript preparation. The authors confirm there are no conflicts of interest.

3.8 Supplemental files

File S1: Results of uncorrected ANOVA (Appendix B; Figure S2.1)
File S2: Transcriptome assembly qualities
File S3: Fasta file with gene sequences and IDs for branch length estimation
File S4: Phylogenetic tree with branch lengths (Appendix B; Figure S2.2)

Chapter 4

RNA-Seq shows evidence of transcriptional reprogramming that distinguishes two *Eutrema salsugineum* ecotypes undergoing a progressive drought treatment

Caitlin M.A. Simopoulos*, Mitchell J.R. MacLeod*, Solmaz Irani, Wilson W.L. Sung, Marc J. Champigny, Peter Summers, G. Brian Golding, Elizabeth A. Weretilnyk

* Authors contributed equally

Formatted for submission to BMC Genomics

4.1 Preface

Chapter 4 describes how the long non-protein coding RNA (lncRNA) prediction tool (as described in Chapter 2) can be used in conjunction with transcriptomic studies. In this work, we describe the molecular responses of two natural accessions of *Eutrema salsugineum* subjected to a two-stage progressive drought. Previous work by MacLeod et al. (2014) demonstrated that the *E. salsugineum* ecotypes displayed similar physiological responses following a first drought, but differed significantly following a second drought. Our work tested the molecular contributions, including lncRNAs, to these drought responses using RNA sequencing and computational analyses to identify differentially expressed genes and co-expressed gene clusters.

Chapter 4 is formatted for submission to BMC Genomics. The original experiment was designed by M. MacLeod and E. A. Weretilnyk. M. MacLeod grew the *Eutrema salsugineum* plants used in the experiment, extracted RNA from 16 plants for RNA sequencing and contributed to the original manuscript draft. S. Irani extracted RNA from 15 more plants for RNA sequencing and performed the RT-qPCR experiment. W. Sung and M. Champigny performed preliminary data analyses that I advanced with additional data and broadened computational tools. I completed all bioinformatic analyses described in the manuscript including read mapping, transcript assembly, novel transcript identification, lncRNA prediction, multivariate analyses and network analysis. E. A. Weretilnyk and I wrote the manuscript. P. Summers contributed significantly to the original experimental design and initial manuscript draft. G. B. Golding and E. A. Weretilnyk supervised the analyses and revision of the manuscript.

4.2 Abstract

Background: The extremophile crucifer, *Eutrema salsugineum*, is a halophyte and hence highly tolerant of osmotic stress. Previously, we developed a two-stage, progressive drought treatment that delineates the drought response and recovery from water deficit for two *E. salsugineum* ecotypes that originate from Yukon, Canada and Shandong, China. Few physiological traits discriminate the ecotypes during a first exposure to water deficit although Yukon plants have a heightened capacity to accumulate solutes and delay turgor loss during a second drought treatment relative to Shandong plants. In this study we compared 31 leaf transcriptomes corresponding to plants undergoing the progressive water deficit protocol.

Results: The first water deficit exposure led to the differential expression of almost 1100 genes for the Yukon ecotype whereas only 63 genes were differentially expressed for Shandong *E. salsugineum*. Transcriptomes from plants undergoing the second drought treatment provided a different outcome in that almost 5000 genes were differentially expressed in Shandong plants compared to about 1,900 genes in Yukon plants. Only 13 genes showed similar drought-responsive patterns for both ecotypes. About 300 (2%) of the differentially expressed genes were predicted as long non-protein coding RNAs (lncRNAs) with only 14 drought-responsive lncRNAs found to overlap between the ecotypes. Co-expression network analysis of the transcriptomes produced eight gene clusters containing over half of the differentially expressed genes. While gene clusters were correlated to drought treatments, few clusters correlated similarly to drought for both ecotypes.

Conclusions: Yukon and Shandong *E. salsugineum* plants are not equally drought tolerant. Relative to Yukon plants, Shandong plants displayed a weak transcriptional response following initial drought treatment and yet displayed a strong response during the second drought treatment. This ecotype-specific transcriptomic response would have escaped notice had we used a single exposure to a water deficit. Notably, the comparatively robust, early transcriptional response shown by Yukon plants is associated with an improved capacity to withstand a second drought exposure. The capacity to improve tolerance and grow after a single drought episode represents an important adaptive trait for a plant that thrives under semi-arid Yukon conditions and may be similarly advantageous for crop species experiencing stresses attributed to climate change.

4.3 Introduction

Crop losses due to limited soil water availability brought on by periods of drought exceed losses attributed to all other abiotic and biotic stressors (Boyer 1982). Scientists predict that climate change will likely exacerbate these losses in the near future and there is evidence that this process has already reduced plant productivity globally (Zhao and Running 2010; Knapp et al. 2017).

Understanding how plants respond to, and recover from, drought is vital to not only maintaining, but also improving global crop yields (Mittler and Blumwald 2010; Boyer et al. 2013).

Plant responses to drought are complex and variable, but our understanding of this subject has advanced nonetheless, in part through the benefits accrued from using different experimental approaches. For example, using tissues from plants with documented physiological responses to an imposed stress allows for drawing correlative associations between the physiological and molecular responses to drought (Harb et al. 2010). Meyer et al. (2014) used a correlative approach with switchgrass to show that some genes only respond to drought-treatment exposures that extended beyond critical physiological thresholds (*e.g.* water potential and photochemical quenching measurements). Sequential drought treatments can also produce plants that display altered responses to subsequent exposures to water deficits (Wang et al. 2014a).

The transcriptional response to repeated drought exposures has been shown to be distinct from the response to a single water deficit (Ding et al. 2012). When *Arabidopsis thaliana* seedlings grown on media plates were exposed to repeated cycles of dehydration, the relative expression of several drought-responsive genes showed evidence of “training”, a phenomenon also referred to as “drought memory”. A genome-wide RNA-Seq approach helped resolve four distinct classes of drought memory genes in *A. thaliana* that reflect their broad strategic roles in protecting plants from the deleterious aspects of drought (Ding et al. 2013).

In this report we describe the transcriptional responses of the extremophile crucifer *Eutrema salsugineum* (synonymous with *Thellungiella salsuginea*), to water deficits. The geographic range where *E. salsugineum* is found is broad and extends across the Asian and North American continents (Wang et al. 2015) and so, not surprisingly, across very different climatic conditions. In the semi-arid, sub-arctic Yukon, Canada, *E. salsugineum* experiences periods with little precipitation in parts of its natural range (Guevara et al. 2012). In contrast, an accession originating in Shandong, China, is found in a temperate region that is subject to higher precipitation (Inan et al. 2004). Importantly, both the Yukon and Shandong accessions are halophytes and consequently equipped with a strong capacity for coping with high osmotic stress, and thrive when exposed to concentrations of NaCl exceeding 300 mM (Kazachkova et al. 2013; Lee et al. 2016b). Despite this unusually high tolerance to osmotic stress, MacLeod et al. (2014) reported that the Yukon and Shandong *E. salsugineum* accessions respond differently to a drought treatment that includes two periods of water deficit separated by a brief recovery period. Plants of the Yukon accession accumulate solutes in response to an initial water deficit and during a second drought treatment, the plants retain water content longer and maintain leaf expansion. Conversely, plants of the Shandong accession show no obvious benefit from the initial drought exposure. These physiological responses are consistent with Yukon plants showing drought tolerance and Shandong plants displaying drought avoidance. Notably, the first drought exposure treatment did little to distinguish the drought-responsive phenotypes that characterize the two accessions.

An indication that the initial drought exposure elicits different responses at the molecular level between the Yukon and Shandong accessions was given by measures of gene expression for four genes classically found to be drought-responsive in many species namely *RAB18*, *RD29A*, *ERD1* and *RD22* (Yamaguchi-Shinozaki and Shinozaki 1994; Ding et al. 2012; Panchbhai et al. 2017). Thus we undertook this comparative RNA-Seq study to provide a more complete understanding of how differently these two accessions respond to water deprivation. In this comparison we also evaluated the contribution of predicted long non-protein coding RNAs (lncRNAs), an interest prompted by their perceived and growing role as gene expression regulators during plant development and in response to stress, including water deficits (Bastow et al. 2004; Franco-Zorrilla et al. 2007; Bardou et al. 2014; Qin et al. 2017). Based on the RT-qPCR analysis of *RAB18*, *RD29A*, *ERD1* and *RD22* reported by MacLeod et al. (2014), we hypothesized that the ecotypes would undergo very different patterns of transcriptional re-programming during water deficits and that ecotype-specific lncRNAs may be implicated in their differential responses. In this work, we show that this prediction was borne out by comparative transcriptome analyses showing substantive differences in gene expression patterns of both protein-coding loci and lncRNAs that distinguish Yukon and Shandong *E. salsugineum* plants with respect to their response to reduced water availability.

4.4 Results

4.4.1 RNA-Seq of *E. salsugineum* accessions following drought and recovery

We prepared and analysed leaf transcriptomes of Yukon and Shandong *E. salsugineum* plants subjected to a progressive, two-stage drought treatment protocol as described by MacLeod et al. (2014). A total of 31 cDNA libraries were prepared with 15 and 16 plants from the Yukon and Shandong genotypes, respectively. Libraries corresponded to plants harvested at various fraction of transpirable soil water (FTSW) percentages: WW1 (100% FTSW), severe drought at D1 (10% FTSW), following re-watering and recovering from drought at WW2 (100% FTSW), and a second severe drought at D2 (10% FTSW). This experiment included two different RNA-Seq library preparation protocols (See Section 4.7). Table 4.1 shows that a comparable number of genes were detected in each of the cDNA libraries when considering both genotype as well as the two different, albeit similar, library preparation methods. To confirm that sequencing timing did not interfere with gene expression detection, two previously prepared and sequenced cDNA libraries (SD2.2 and YD2.1) were resequenced. Using principal component analysis (PCA), few differences were observed between the two sequencing time points by way of library overlap visualized in a PCA biplot (Fig. S3.1). However, as the pairs represented technical replicates, data from the resequenced libraries were not used in further bioinformatic analyses. Additionally, we assessed the capacity of our transcriptomic database to discern differentially expressed genes, particularly drought and/or accession-specific genes. To do so, we compared $\log_2$-fold change values derived by RNA-Seq to expression data derived by an independent approach using RT-qPCR (Fig. S3.2). We chose four genes for relative abundance determinations (*EsRAB18*, *EsRD22*, *EsRD29a* and *EsERD1*) as these four dehydrin-related genes were previously shown by RT-qPCR to distinguish the responses displayed by Shandong and Yukon ecotypes at various stages of the progressive drought protocol (MacLeod et al. 2014). We found excellent agreement between RNA-Seq and RT-qPCR results for these four genes at the three stages tested (D1, WW2 and D2) relative to their levels of expression under control, WW1 conditions (Fig. S3.2).

On average, approximately 17,400 genes were detected in each library with the lowest number of genes identified in the YWW2.1 library at 16,860. Using a minimum threshold for detection of 1 fragment per kilobase per million mapped reads (FPKM), we found read support for 20,841 genes, or 79% of the 26,531 genes comprising the predicted coding capacity of the JGI *E. salsugineum* v1.0 genome (Yang et al. 2013). A number of genes (11%) were expressed only in Shandong (1268 genes) or Yukon leaves (1023 genes). Thus, for each accession, less than 5% of the total protein-encoding capacity of the genome was expressed in an accession-specific manner.

TABLE 4.1: Number of detected loci by RNA-Seq in all 31 RNA-Seq libraries at a threshold of > 1 FPKM

	All loci		Putative lncRNA	
Library	Yukon	Shandong	Yukon	Shandong
WW1.1	17268	17749	446	498
WW1.2	17303	17440	462	520
WW1.3	17106	17595	442	480
WW1.4	17320	17423	475	489
D1.1	17608	17512	509	489
D1.2	17587	17479	522	449
D1.3	17170	17532	459	534
WW2.1	16860	17118	445	484
WW2.2	17545	17424	514	510
WW2.3	17217	17866	462	548
WW2.4	17037	17758	437	524
D2.1	17484	17589	517	531
D2.2	17493	17516	507	534
D2.3	17367	17277	477	520
D2.4	17312	17075	470	483
D2.5	NA	17357	NA	509
Mean	17311	17481	486	506

We did not restrict our analyses to reads mapping to annotated regions in the genome, but instead used a conservative approach to identify novel transcripts that are expressed but remain without annotation. In addition, we looked for expression of other transcripts previously described by Champigny et al. (2013) and Yin et al. (2018). Of the 411 transcripts previously identified by Champigny et al. (2013), 383 were expressed in at least one genotype and condition during the progressive drought (Table 4.2). An additional 1608 previously unidentified transcripts, referred to as DLOCs in this work, were expressed at one point during the experiment, of which 24 overlapped in genomic location with those described by Yin et al. (2018). In total, we detected expression of 919 putative lncRNAs, of which only 71 (7.7%) were present in the *E. salsugineum* reference annotation.

TABLE 4.2: Number of detected unannotated genes in all 31 RNA-Seq libraries at a threshold > 1 FPKM

	MacLeod et al. (2014)		Yin et al. (2018)		DLOC	
Library	Yukon	Shandong	Yukon	Shandong	Yukon	Shandong
WW1.1	308	287	15	23	650	722
WW1.2	316	281	15	23	668	763
WW1.3	298	274	13	23	643	697
WW1.4	312	272	14	22	695	711
D1.1	316	286	14	23	767	717
D1.2	322	292	14	22	798	639
D1.3	302	280	14	21	671	794
WW2.1	307	287	14	22	654	675
WW2.2	310	285	14	22	760	731
WW2.3	307	281	14	20	670	793
WW2.4	300	278	14	20	637	771
D2.1	316	296	13	20	776	756
D2.2	319	301	13	22	757	776
D2.3	307	281	14	22	729	764
D2.4	305	272	13	22	706	687
D2.5	NA	274	NA	22	NA	727
Mean	310	283	14	22	705	724

Counts only considered Yin et al. (2018) genes not previously identified by MacLeod et al. (2014). DLOC loci are those identified as novel by this study.

4.4.2 Identifying differentially expressed genes

PCA was used to explore sources of variance in transcript abundance among the 31 sequenced leaf cDNA libraries. PCA provides factor loading scores for each library with each score representing the extent to which the abundances of transcripts from each library contribute to a given principal component. PC1 accounted for 94.2% of the variance but did not distinguish the libraries on the basis of genotype or treatment, a feature reported by Champigny et al. (2013) to correspond to gene expression levels (Fig. S3.1). In contrast, PC2, PC3 and PC4 accounted for far less variance than PC1 (2.1%, 0.9%, 0.7%, respectively) but offered more meaningful biological insights into genotype and treatment-specific differences between the transcriptomes. By way of example, Fig. 4.1 is a biplot of PC2 and PC4 and it displays the variability due to ecotype, and to a lesser extent, variation due to treatment. Specifically, PC2 only explains 2.1% of the variance in the data but it clearly distinguishes the scores for Yukon transcriptomes from those of Shandong plants along the horizontal axis. For Yukon transcriptomes, PC4 discerned drought-treated from well-watered, including re-watered plants. The scores for cDNA libraries of drought-treated Yukon plants have positive loadings along PC4 (YD1, YD2) whereas more negative scores are associated with plants that have either not experienced a water deficit (YWW1) or have been re-watered and allowed to recover following a drought treatment (YWW2). In contrast, the scores for Shandong libraries produced from well-watered plants (SWW1, SWW2) cluster with plants experiencing drought (SD1, SD2) and re-watered plants (SWW2). Thus, PC4 appears to describe a source of variance that is related to water deficit for Yukon plants, with a far less clear distinction for the response to water deficits given by transcriptomes of Shandong plants.

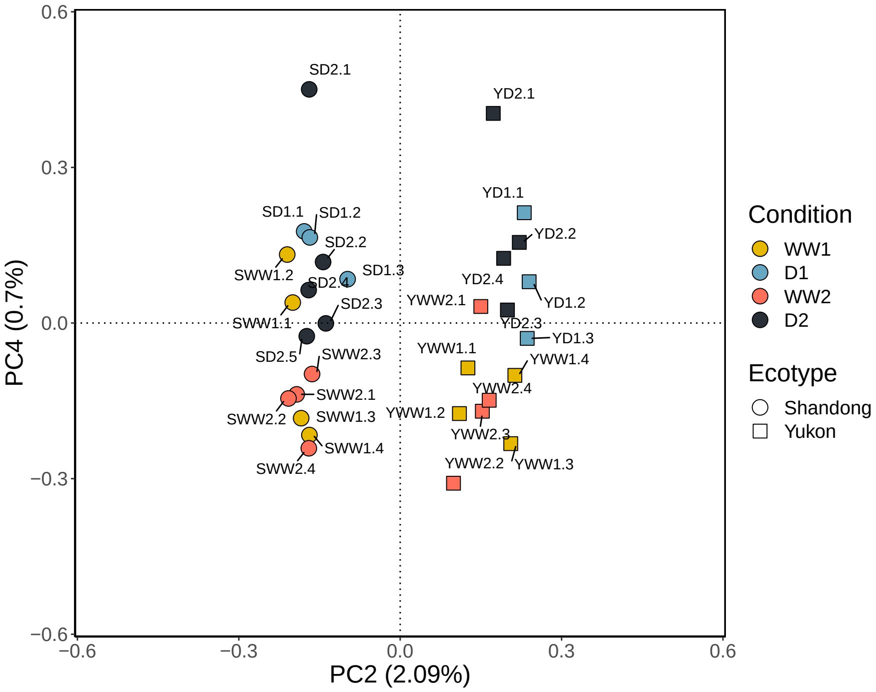


Figure 4.1: Principal component analysis of transcript abundances of Yukon and Shandong *E. salsugineum* plants undergoing stages of a progressive drought treatment protocol.

Analysis using `DESeq2` yielded 4650 and 2454 drought-responsive genes that were differentially expressed only in either Shandong or Yukon plants, respectively, while 1599 differentially expressed genes (DEGs) were found in transcriptomes for both accessions (Additional file 2). Fig 4.2 provides an overview of DEG numbers identified in comparisons of the 31 transcriptomes over the course of the progressive drought protocol for each genotype separately (Fig. 4.2A,B) and as a summary of overlapping DEGs (Fig. 4.2C). The transition from a well-watered condition to D1 provides a striking impression. In Shandong plants, only 63 DEGs were identified as undergoing significant changes in expression after the first drought exposure, whereas 1085 DEGs were detected in Yukon plants (Fig. 4.2A,B). A mere 29 DEGs were common between the two ecotypes. Fig 4.2 also provides the estimated contribution of DEGs predicted to be lncRNAs at each stage of the protocol for both natural accessions. Notably, none of the DEGs identified in Shandong plants at D1 were predicted as lncRNAs whereas 2.7% of the DEGs detected in Yukon plants at D1 were predicted as being lncRNAs. During the recovery from the initial drought (D1) to the re-watered and recovery stage (WW2), the two ecotypes again show different gene expression responses. Of the total DEGs identified in each genotype, over 82% and 77% were unique to Yukon and Shandong plants, respectively.

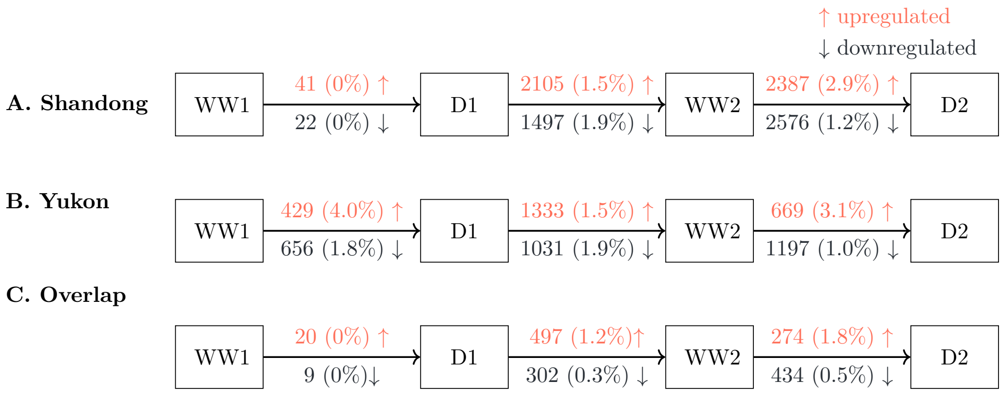


Figure 4.2: Number of DEGs detected in each *E. salsugineum* ecotype and overlap between DEGs at each stage of the progressive drought treatment. The number of upregulated DEGs are described in coral above the transition arrow. The number of downregulated DEGs are given in blue below the transition arrow. Numbers in brackets refer to the percentage of DEGs predicted by CREMA as encoding lncRNAs.

The overall impression is that Shandong and Yukon plants undergo different transcriptional reprogramming during both stages of the progressive drought protocol. The DEG complement they share, based on empirical evidence, can be as low as 2% for a comparison between WW1 → D1 and up to 14% in D1 → WW2 plants. The overlap of DEGs predicted as lncRNAs is negligible as only 14 unique, putative lncRNAs were identified among the DEGs of both *E. salsugineum* genotypes (Fig. 4.2C; Additional file 2). We also tested whether any genes showed a similar pattern of drought-responsive expression during the progressive drought protocol. Figure 4.3 shows that only 13 DEGs displayed the same expression patterns, with eight showing increased transcript abundance following water deficit and decreased abundance under watered/re-watered conditions (Fig. 4.3A,B) while five DEGs showed the inverse response (Fig 4.3C,D). For the eight DEGs undergoing increased transcript abundance with water deficit, the transcript levels in well-watered Yukon plants were typically already higher relative to those detected in well-watered Shandong plants, notwithstanding the drought-responsive increases found for both ecotypes (Fig. 4.3A,B).

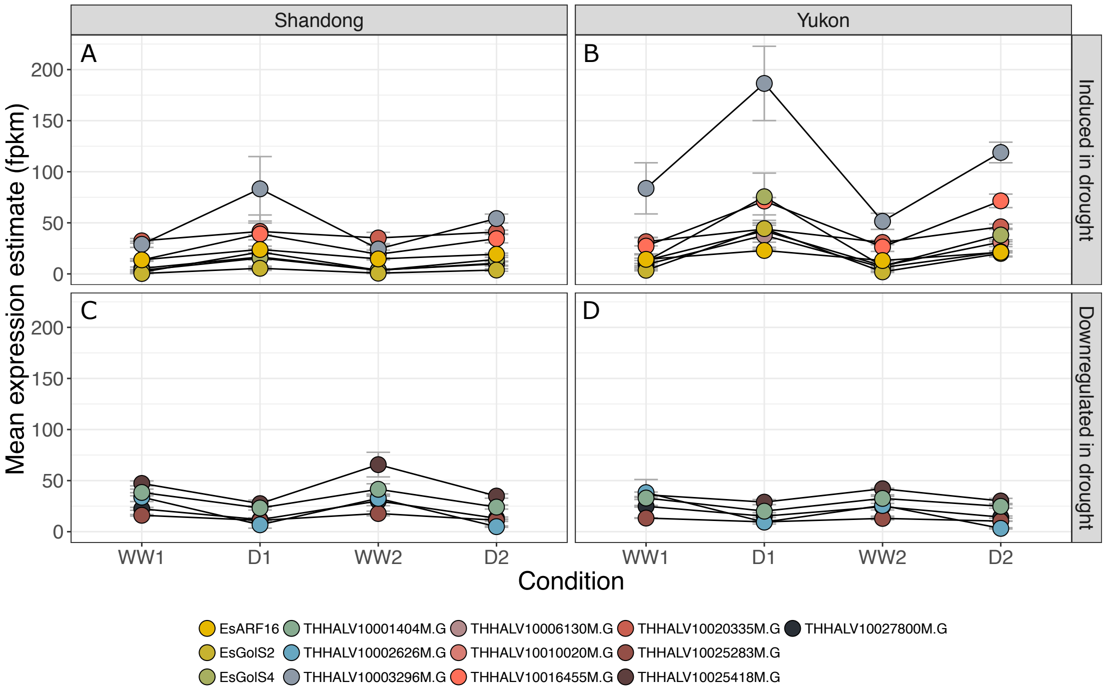


Figure 4.3: Average estimated FPKM values of DEGs identified by DESeq2 that follow the same direction of fold change in both *E. salsugineum* ecotypes. Standard error of the expression values are represented by grey error bars.

4.4.3 Correlating altered gene expression with biological responses to water deficit

MacLeod et al. (2014) reported that the initial exposure to water deficit (D1) altered the way that Yukon plants responded to the second water deficit (D2). For example, both leaf water content and leaf ψ_s were different for Yukon plants during D2 compared to D1. Conversely, Shandong plants responded similarly to the two drought exposures with no discernible changes in leaf water content or leaf ψ_s. Moreover, Yukon leaves took longer to wilt relative to leaves of Shandong plants during D2 relative to D1. Longer times to wilt following an initial drought episode suggest that Yukon plants underwent changes during D1 that improved their water holding capacity during subsequent water deficits. Thus we hypothesized that the transcriptional response of Yukon plants would be distinctive between D1 and D2, whereas D1-associated changes in Shandong plants would likely be repeated during D2. This prediction, however, was not consistent with the DEGs identified in Figure 4.2. Rather, we found that only about half of the genes differentially expressed in D1 in Shandong plants were also differentially expressed in D2, despite the fact that many more genes (over 78-fold more) were differentially expressed in D2 compared to D1. Conversely, only a two-fold increase was found for DEGs in leaves of Yukon plants when comparing D2 to D1 with over half (630 or almost 60%) of the DEGs showing drought-responsive expression during both stages of the progressive drought treatment.

We considered that differential gene expression analysis is limited to pairwise comparisons and hence would not identify groups of co-expressed genes that may contribute to insights into the unique drought responses of the Yukon and Shandong *E. salsugineum* genotypes. As such, a weighted gene co-expression network analysis (WGCNA) was used to cluster genes using estimated transcript abundances during the progressive drought treatment conditions. WGCNA is a systems biology analysis method that assumes co-expressed genes that belong to the same cluster have a similar function. This "guilt-by-association" approach can help predict functionality of un-annotated protein-coding or lncRNA-coding loci with information on the directionality (up- or down-regulated) of their possible roles during each treatment condition. For this analysis, expression estimates from all genes as opposed to only DEGs were used to allow for an unbiased, unsupervised clustering method with a summary of the results shown in Figure 4.4. Eigengene values, summary statistics calculated using a dimensionality reduction method similar to PCA, were used to quantify the "average" gene expression values of each cluster. Using these cluster eigengene values, we correlated each gene cluster to drought treatment and ecotype (Additional file 3) and then selected clusters with 50% or more DEGs for gene ontology (GO) term enrichment analysis (Table S3.1). A reduced list of highly significant biological processes in selected clusters was produced using REVIGO (Supek et al. 2011) and the results are summarized in Additional file 4.

The heat map of cluster eigengene correlations to drought treatment (Fig. 4.4) shows correlations of ecotype and eigengenes grouping separately, suggestive of distinct responses to the

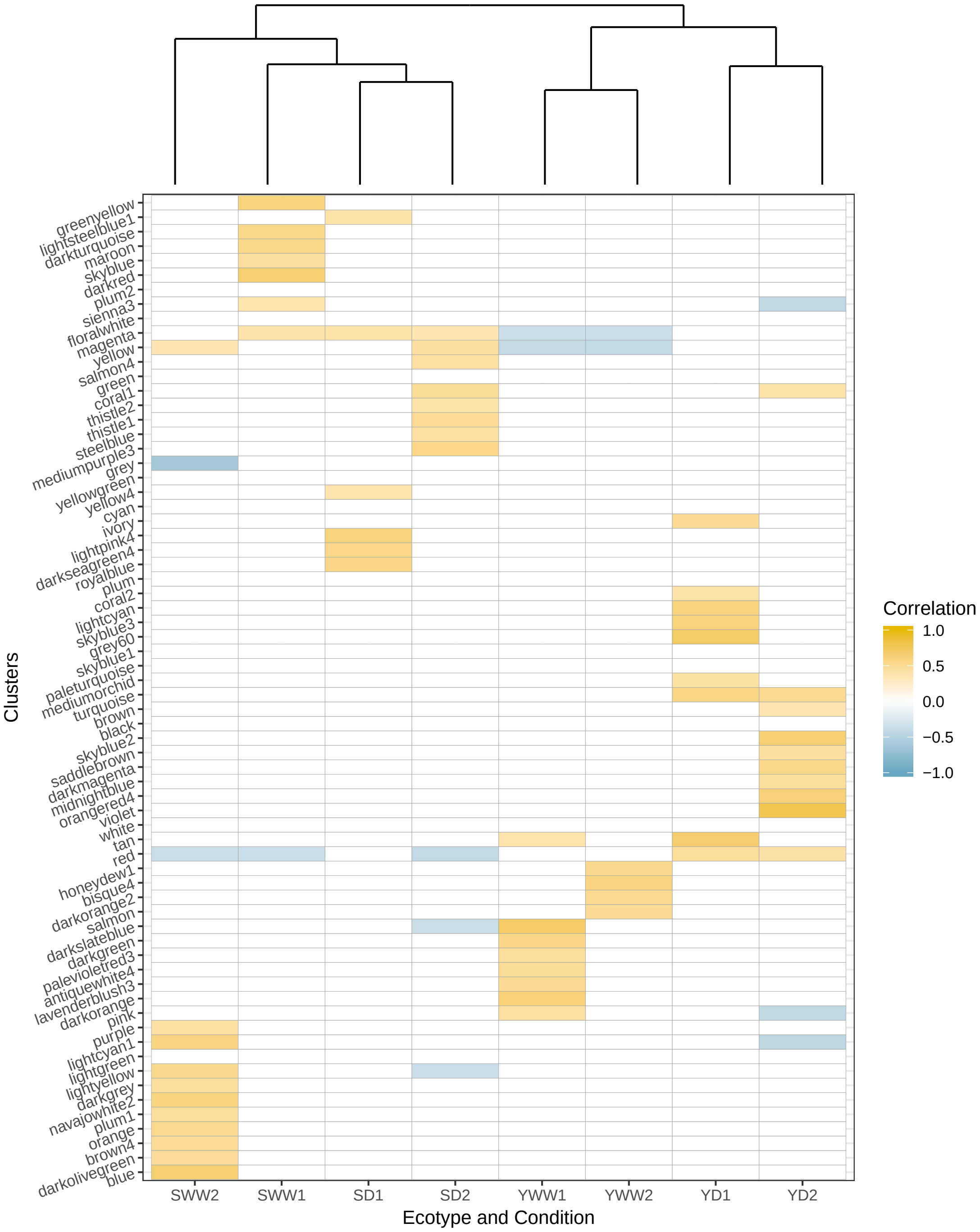


FIGURE 4.4: WGCNA cluster heatmap illustrating correlations of cluster eigengenes (Y axis) to ecotype and progressive drought conditions (X axis). Positive correlations are represented in yellow, while negative correlations are represented in blue. Only significant correlations are displayed (p < 0.05 after false discovery rate (FDR) adjustment). Non-significant correlations are coloured in white.

progressive drought treatment by Shandong and Yukon plants. Moreover, the heat map also shows the correlated data for drought (D1, D2) and watered (WW1, WW2) treatments being grouped separately for Yukon plants whereas for Shandong plants the WGCNA results grouped both drought treatments together with the well-watered control (WW1) plants. Consistent with different drought response strategies for the two ecotypes, only one cluster containing at least 50% DEGs, "coral1", also showed significant correlation to the same stage of drought treatment (D2) for both ecotypes (Fig 4.4; Table S3.1). The "coral1" gene cluster is significantly enriched in GO terms relating to sulfur assimilation and sulfur utilization which infers a common connection between sulfur nutrition and a more prolonged exposure to drought stress (Additional file 4). Identifying only one ecotype-overlapping cluster suggests that groups of co-expressed genes are more highly correlated to a single ecotype and not shared by both ecotypes, an interpretation consistent with Shandong and Yukon plants expressing genes with different functions during drought. By way of example, genes of the "lightyellow" cluster are only correlated for the drought response of Shandong plants and interestingly, the clustered genes are negatively correlated with D2 but positively correlated with WW2 (Fig. 4.4; Table S3.1). This associated set of correlated differences is particularly relevant given the DEGs summarized in Figure 4.2A. Shandong plants did not significantly alter their gene expression during the initial drought exposure (D1) but a re-watering treatment following D1 and subsequent drought (D1→WW2→D2) triggered major transcriptional changes. The "lightyellow" cluster is composed of genes associated with metabolic processes, with the most significant GO terms associated with lipid biosynthetic processes, and ketone and carbohydrate metabolic processes (Fig 4.4; Table S3.1; Additional file 4). The directionality of the correlated transcriptional changes suggests that biosynthetic pathways promoted by re-watering were subsequently reversed by the second drought (D2).

For Yukon plants, co-expressed genes positively correlated with both D1 and D2 are grouped in the "turquoise" cluster (Fig 4.4; Table S3.1). Containing 3415 co-expressed transcripts, "turquoise" is the largest identified cluster and is enriched in genes associated with water deprivation, peptide transport, and cellular lipid catabolic processes (Additional file 4). The category "lightcyan1" offers a different type of response in being populated by genes negatively correlated to D2 in Yukon plants but also positively correlated to a re-watering (WW2) response in Shandong plants (Fig 4.4; Table S3.1). The "lightcyan1" cluster is comprised of genes associated with proteolysis and negative regulation of catalytic activity, offering an indication of related functions elicited by the recovery of Shandong plants from water deficit as compared to Yukon plants (Additional file 4). In contrast, both "purple" and "blue" clusters are only positively correlated to the re-watering (WW2) response for Shandong plants (Fig 4.4; Table S3.1). The "purple" cluster is highly enriched in genes associated with translation and DNA packaging and replication while genes in the "pink" cluster are related to cell cycle regulation, vitamin biosynthetic processes, nucleoside-related biosynthetic processes, and photosynthesis (Additonal file 4). Thus the broad functionality of the Shandong-specific recovery associated clusters reflects the large transcriptional response of Shandong plants transitioning from D1 through WW2 (Fig. 4.2, Additional file 4). On the other hand, there is a notable lack of significant correlation of Yukon re-watered plants to any clusters primarily encompassed by DEGs (Fig. 4.4; Table S3.1) although Yukon plants, like Shandong plants, continued to grow during the entire progressive drought treatment (MacLeod et al. 2014).

4.5 Discussion

In this work, we hypothesized that *E. salsugineum* ecotypes would have different global expression responses to a progressive drought, hinted at by different expression patterns of select dehydrins during the same progressive drought treatment (MacLeod et al. 2014). MacLeod et al. (2014) also reported overall differences in the physiological responses of *E. salsugineum* ecotypes during the progressive drought protocol. For example, Yukon plants grown in controlled environment chambers were found to respond to an initial drought by a 46% reduction in stomatal conductance and 25% reduction in rosette water loss relative to unstressed control plants, evidence of drought avoidance to conserve water (MacLeod et al. 2014). Upon wilting, Yukon plants re-established turgor at significantly lower leaf solute potentials than the level for consistently well-watered Yukon plants which suggests osmotic adjustment. In contrast, while Shandong plants also showed signs of undergoing drought avoidance, the leaf solute potentials in re-watered Shandong plants returned to pre-drought levels after re-watering. Thus while the physiological responses during D1 seemed similar for both ecotypes, their distinct responses are more clearly seen during subsequent exposure to drought where Yukon plants take longer before turgor is lost relative to Shandong plants. In this regard, the very different transcriptional responses that we observed for the plants of two ecotypes during a first drought (D1) is particularly notable. That is, with the initial drought exposure, about 1100 differentially expressed genes were detected in leaves of Yukon plants compared to only 63 in Shandong plants (Fig. 4.2A,B). Following recovery with re-watering (WW2) and subsequent exposure to a second drought (D2), 1866 genes were differentially expressed in leaves of Yukon plants while Shandong plants now underwent a much larger transcriptional response with almost 5000 genes showing differential expression. While there is overlap between transcriptional changes of both ecotypes, most of the DEGs are unique to each genotype (Fig. 4.2C). The approach of using a progressive drought regime was valuable in that, if it had not been used, we would not have appreciated the large differences in gene expression changes observed in the Shandong ecotype in comparison to Yukon plants upon a second exposure to a water deficit.

Progressive drought protocols have been used to study transcriptional reprogramming of alfalfa (Kang et al. 2011) and switchgrass (Meyer et al. 2014). These studies both show that fewer genes are differentially expressed during recovery from drought than during a severe water deficit. This is not consistent with our finding of a higher number of DEGs identified in both Yukon and Shandong plants that were re-watered (WW2) relative to plants undergoing an initial drought (D1) (Fig 4.2A,B). This difference between the earlier work on alfalfa and switchgrass and our research with *E. salsugineum* is especially evident in Shandong plants, where the 63 DEGs identified during D1 are followed by over 3600 DEGs in plants that were re-watered and allowed to recover. One potential explanation that could explain the strong differential response by Shandong plants may relate to the extent by which these plants perceive the initial drought. Specifically, MacLeod et al. (2014) found that physiological measurements of Shandong plants (including cut rosette water loss, static leaf water content, and specific leaf area) during D1 were not different from the same measurements of well-watered, control Shandong plants. This negligible physiological response suggests that Shandong plants were either not stressed or did not sense the severe water stress imposed at D1 at 10% FTSW and hence few DEGs were associated with the SWW1 → SD1 transition (Fig. 4.2B). The comparatively stronger transcriptional responses of Yukon plants during D1 and upon re-watering implies that what happens during drought is largely reversed during recovery, an inference that is largely borne out by the data. For example, 62% of the 429 genes up-regulated and 75% of the 656 down-regulated genes in Yukon plants identified during D1 showed altered expression in the opposite direction during re-watering. This observation led us to question why more genes show changed expression during D2 compared to D1 in Yukon plants (Fig. 4.2). We originally predicted that the changed expression patterns of many genes may not return to pre-stress levels, and indeed 43% of the 429 genes up-regulated during D1 were also up-regulated during D2. However, when looking specifically at genes associated with drought for both ecotypes (Fig. 4.3) we see that genes induced during D1, although still differentially expressed in the second drought, show levels of expression that are lower in Yukon plants during D2. This behaviour is exemplified by two of the genes encoding dehydrins selected for RT-qPCR analysis, namely *EsRAB18* and *EsRD29A*, where drought-responsive changes in transcript abundance for Yukon plants are lower at D2 relative to D1 (Fig. S3.2). Conceivably, whereas re-watering returns plants to the same water status achieved before drought, the transcriptional reprogramming during D1 has an enduring impact that may benefit Yukon plants during subsequent stress exposures.

As discussed earlier, MacLeod et al. (2014) reported that Yukon plants tolerate repeated drought exposure better than Shandong plants with benefits seen in solute accumulation and a longer time taken before turgor loss. However, the stress protective effect is not specific to drought. Exposure of Yukon plants to an initial drought treatment improves the freezing tolerance of Yukon plants from -19C to -21C with a shortened cold acclimation period (Griffith et al. 2007; Khanal et al. 2017). By not fully reverting to pre-stress levels, the constitutive expression of stress-responsive genes may enable the plant to retain a complement of gene products that serve as a "molecular buffer" for prolonged stress protection, products that may promote a greater coping capacity should the stress return. This further implies that Yukon plants, once stressed by exposure to water deficits, are no longer "naive" to stress and that their tolerance to other sources of adverse abiotic or biotic conditions can be improved. By way of contrast, the expression patterns for *EsRAB18*, *EsRD29A*, and *EsRD22* were very different in Shandong plants compared to Yukon plants (Fig. S3.2). The ecotype-specific expression changes are particularly evident in the expression of *EsRAB18* and *EsRD29A* where their relative transcript levels remain high in Shandong re-watered plants (WW2), but are downregulated Yukon plants experiencing the same re-watering treatment. This pattern of expression appears to be shared by a large number of drought-responsive genes given our finding of a large increase in gene expression changes in Shandong plants at WW2 and D2 (Fig. 4.2B). This different transcriptional response suggests that Shandong plants, unlike Yukon plants, may not be appropriately "primed" by the water deficit stress during D1 and, by consequence, are less able to cope with stress during the D2 treatment relative to Yukon plants.

We used WGCNA and the "guilt-by-association" approach to address the global transcriptomes in order to identify genes undergoing significant changes in expression during the progressive drought protocol for insight into their predicted functionality. By way of examples, cuticular waxes have been shown to be altered in a drought-responsive manner in a comparative study using Shandong and Yukon *E. salsugineum* plants (Xu et al. 2014). Xu et al. (2014) reported that the two ecotypes alter the composition and amount of cuticular waxes between non-stressed and drought-stress conditions with Yukon plants exhibiting a 4.6-fold increase in leaf wax content, although both ecotypes showed increases in the total amount of wax. MacLeod et al. (2014) did not measure cuticular waxes to an enhanced drought tolerance in Yukon plants but rather focused on a variety of physiological changes including differences in accumulated solutes with drought exposure. Our WGCNA distinguished clusters of co-expressed genes relevant to the studies just described. The "lightyellow" cluster is negatively correlated to SD2 and positively correlated to SWW2 (Table S3.1) and contained genes enriched in functions associated with ketone metabolic processes (Additional file 4). Indeed, the DEGs found in the "lightyellow" cluster were significantly increased in abundance in re-watered conditions and decreased in drought conditions in both Shandong and Yukon plants. While these transcriptional changes in the direction of expression cuticular of wax-related genes seems counter intuitive, studies exploring the regulation of wax biosynthesis in rice describe DROUGHT HYPERSENSITIVE (*DHS*), encoding a RING-type E3 ligase, as a negative regulator of wax biosynthesis (Wang et al. 2018e) and hence its overexpression reduces drought tolerance in transgenic rice lines. DECREASE WAX BIOSYNTHESIS (*DEWAX*) is a transcriptional repressor of wax production with over-expression also reducing wax deposition (Go et al. 2014). Thus the enriched status of the "lightyellow" cluster by putative wax-related gene products may indicate that plants recovering from drought-stress are better positioned with respect to their capacity to alter cuticular wax composition and/or content, rather than a reflection of changes in the activity of wax-related biosynthetic processes themselves. Plants like Yukon *E. salsugineum* that are adapted to dry environments are classically known to develop thick cuticular waxes but the regulatory mechanisms responsible remain a topic of considerable interest and are likely very complex as suggested in a recent review by Xue et al. (2017).

Among the transcripts clustered by WGCNA were lncRNAs. LncRNAs are proposed to function as gene expression regulators, particularly in organisms experiencing stress (Xu et al. 2017b). In this study the two *E. salsugineum* ecotypes display different transcriptional responses to water deficits. Hence, we predicted that there should be differences in lncRNA expression in the plants undergoing the progressive drought treatment. Unexpectedly, we found an almost complete lack of overlap in the drought-associated lncRNAs expressed in Yukon and Shandong plants. This finding of negligible overlap among lncRNAs is perhaps not surprising given their fast evolution (Hezroni et al. 2015) and extreme conditions that have led to the local adaptation of *E. salsugineum* ecotypes to different natural environments. We particularly focused our analysis on genes associated with the "turquoise" cluster, a group positively correlated to D1 and D2 drought treatments in Yukon plants. The "turquoise" cluster was functionally enriched in genes with GO terms associated with plant response to water deprivation, including responses to abscisic acid (ABA). The "turquoise" cluster contains 36 differentially expressed putative lncRNAs, 30 of which are up-regulated during both drought treatments. Eight of the drought-induced lncRNAs are only differentially expressed in Yukon plants whereas 18 are specific to Shandong plants indicating that both ecotypes deploy distinct lncRNAs in response to the same stress treatment protocol. The "turquoise" cluster was enriched in GO terms with functions similar to a previously identified lncRNA, drought induced lncRNA (*DRIR*), first identified in *A. thaliana* (Qin et al. 2017). We did not find evidence of genes in this cluster with sequence homology to DRIR indicating the gene products we described are most likely previously unidentified water deficit stress-associated lncRNA transcripts.

Interestingly, we found only 13 drought-responsive genes that display similar expression patterns in both Shandong and Yukon plants. Of the eight genes that display a positive response to drought (Fig. 4.3A,B), all but one, Thhalv10020335m.g, are found in the "turquoise" cluster that is enriched in drought-related genes. We explored the functions of the overlapping genes as we hypothesized that these products may be part of a conserved drought response for *Eutrema* and likely other plants. Thhalv10024122m.g, is homologous to the *A. thaliana* gene AT2G38800.1 and encodes a plant calmodulin-binding protein that has previously been characterized by Lovell et al. (2015) as a quantitative trait locus (QTL) associated with drought in *A. thaliana*. Thhalv10003296m.g (AT5G43150) is a predicted mitochondiral protein with no known function, however, this gene is expressed under a variety of abiotic stresses in both *A. thaliana* and *Oryza sativa*, consistent with a role in a conserved stress response throughout plants (Narsai et al. 2010). Using a combined expression ranking and co-expression analysis, Ransbotyn et al. (2015) also identified AT3G57540 (Thhalv10006130m.g) to be stress responsive, and found its expression to cluster with other ABA-responsive genes, a finding similar to this work. Thhalv10023585m.g (AT1G60470) and Thhalv10024122m.g (AT5G43150) are both annotated as encoding galactinol synthases, enzymes known to act in the biosynthesis of raffinose family oligosaccharides, known osmoprotectants in plants (Nishizawa et al. 2008) and hence likely playing a similar role in osmoprotection for *E. salsugineum* experiencing drought. Galactinol and raffinose accumulate during stress treatments in leaves of *E. salsugineum* and these metabolites are detected in *E. salsugineum* plants collected at a highly saline Yukon field site (Guevara et al. 2012). Rasheed et al. (2016) identified AT1G34060 (Thhalv10010020), a tryptophan aminotransferase, to be upregulated during drought, as well as other auxin-related genes, similar to Thhalv10024601 (AT4G30080), an auxin response factor. Thus the comparatively small group of drought-induced genes shared by both ecotypes are well-known to be associated with osmotic stress. Of additional interest for this group of drought-responsive genes is the differences in their expression levels between the two ecotypes with the comparatively muted transcriptional changes detected for Shandong plants relative to Yukon plants with drought stress (Fig. 4.3).

4.6 Conclusion

Although Yukon and Shandong *E. salsugineum* plants are both halophytes, several studies show that they do not respond similarly to abiotic and biotic stress. For example, these *E. salsugineum* ecotypes modulate their photosynthetic responses to light and temperature differently (Khanal et al. 2017) and, as discussed in this work, they respond to water deficits by divergent mechanisms as shown by differential alterations in wax composition and water use (Xu et al. 2014; MacLeod et al. 2014). There is also recent evidence that Shandong *E. salsugineum* is not as well equipped for drought tolerance relative to Yukon *E. salsugineum*, and even *A. thaliana* (MacLeod et al. 2014; Xu et al. 2014; Pinheiu et al. 2019). Similarly, Shandong plants show greater constitutive resistance to infection by *Pseudomonas syringae* than Yukon plants indicating that these ecotypes also diverge with respect to their responses to biotic stress (Yeo et al. 2015). Moreover, the potential for Yukon *E. salsugineum* to retain a “molecular memory” conferred by an early exposure to stress appears to be a distinguishing feature of this ecotype. This priming response would have an advantageous role for an extremophyte that likely already invests heavily to survive the extreme conditions that are endemic features of its semi-arid and subartic natural habitat.

As more is known about extremophyte species in general, it may become clearer whether other plants are similarly equipped to use drought exposure to significantly augment their tolerance to the same or different stress exposures, a capacity that would be advantageous for plants such as *Anastatica hierochuntica*, an extremophyte species that displays considerable cross-tolerance to salt though it naturally grows in a desert (Eshel et al. 2016). Certainly, as recently reviewed by Sork (2018), “-omics” technologies are now offering deeper insights into mechanisms underlying local adaptation and they can be readily applied to both model and non-model species. Our comparison between the drought responses of Yukon and Shandong plants shows that even extremophyte plants adapted to extreme environments display elements of local adaptation and that ongoing studies of their overlapping and contrasting responses could help discern the underlying mechanisms responsible.

4.7 Materials and Methods

4.7.1 Plant growth conditions and drought simulation assay

Shandong and Yukon *E. salsugineum* plants were grown in a controlled environment growth chamber and subjected to a drought simulation assay consisting of two periods of water deficit separated by a two-day recovery period (MacLeod et al. 2014). Water was withheld from four-week-old plants to initiate the first drought treatment (D1). The progress of the drought treatment was monitored gravimetrically and the FTSW was determined. FTSW was maintained at approximately 100% for well-watered, control plants. Plants undergoing drought treatment were water deprived until FTSW reached 0% and the plants visibly wilted. The re-watering treatment was started on the day a plant wilted and FTSW was restored to 100% within 48 h (WW2). After 48 h, water was again withheld from plants to begin the second drought treatment (D2). A set of plants of each accession was watered daily over the course of the entire experiment and served as well water control plants.

4.7.2 Selection of plant tissue for transcriptome profiling

Only fully-expanded rosette leaves were harvested from both Yukon and Shandong plants. Leaf samples used for RNA extraction were collected between 8 and 10 h into the day cycle under our cabinet conditions. Once harvested, the leaf tissue was flash-frozen in liquid N and then transferred to a freezer for long term storage at -80C. Leaves for RNA extraction were collected at three stages of the water deficit protocol. Plants were harvested during the initial water deficit (D1-10% FTSW, during recovery (WW2-100% FTSW), and during the second water deficit (D2-10% FTSW). In addition, leaf tissue was harvested from four well-watered control plants with one plant harvested when drought-treated plants had reached D1-10% FTSW and the final three control plants at the WW2-100% stage of the drought protocol. The difference in age between these control plants was 5 days.

4.7.3 RNA extraction cDNA library construction and transcriptome assembly

Total RNA was extracted from frozen leaves using a modified hot borate method (Wan and Wilkins 1994) as described in Champigny et al. (2013). RNA quantity and integrity was assessed using RNA Nano 6000 chips on a Bioanalyzer 2100 instrument. Two mRNA purification protocols were performed depending on sequencing date: A. Three successive on-column purifications using the Genelute mRNA miniprep kit (Cat. No. MRN10, Sigma) or B. NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490). Both mRNA purification protocols were followed by the

NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760). Preparation of cDNA libraries was performed with the NEBNext multiplex cDNA synthesis kit for Illumina using random hexamers (Cat. No. E7335, New England Biolabs, Ipswich, MA). The cleanup of fragmented RNA was performed with Agencourt AMPure XP Beads (Cat. No. A63987, Beckman Coulter, Mississauga, ON) following the manufacturer's protocol.

Quality control, amplification, and sequencing of the 31 cDNA libraries from cabinet-grown plants was conducted at the sequencing facility of the Farncombe Family Digestive Health Research Institute (McMaster University, ON, Canada). A combination of two high-output and three rapid paired-end sequencing runs using a 100 or 150 bp length were performed using the Illumina Hi-Seq 1500 platform. The libraries are identified by accession (Y or S), drought stage (WW1, D1, WW2, D2), and plant number (1, 2, 3, or 4).

Following sequencing, the reads obtained were trimmed of Illumina adaptor sequences and low quality reads with default settings of Trimmomatic v0.34 (Bolger et al. 2014). Only reads $\geq$ 36 bp after trimming were mapped to the JGI *E. salsugineum* genome assembly downloaded from Phytozome v12.1 (Goodstein et al. 2012; Yang et al. 2013) using STAR v2.5.2b (Dobin et al. 2013). We used StringTie 1.3.4d (Pertea et al. 2015) for transcript assembly on each individual RNA-Seq library to identify transcripts missing from the *E. salsugineum* reference annotation. Library-specific transcripts called by StringTie were merged with each other and the reference annotation using StringTie `--merge` default settings.

Assembled transcripts that were not found in the *E. salsugineum* reference annotation, or were not previously identified by Champigny et al. (2013), were considered novel transcripts and have the locus identifier-prefix "DLOC". The merged GTF annotation file that included all novel transcripts, those identified by Champigny et al. (2013), and those found in the reference annotation was created and used for consequent transcript abundance estimates and is available in Additional file 5. The merged GTF file was compared to the reference annotation using `gffcompare` and only those transcripts classified as "unknown" and "intergenic" were retained for further analysis to reduce errors caused by mapping, assembling, or sequencing of unprocessed transcripts.

4.7.4 Determination of transcript abundance

Gene expression estimates for transcripts that met our classification criteria (*i.e.* in reference annotation, "unknown" or "intergenic") were calculated using RSEM v1.2.31 (Li and Dewey 2011) and an internal call for mapping to the transcriptome using RNA-Seq aligner STAR v2.5.2 (Dobin et al. 2013) to accommodate the ambiguity of multi-mapping reads. Gene level transcript abundance is reported as the number FPKM, a determination accounting for both mRNA length and library size (Trapnell et al. 2010). FPKM estimates for each gene were calculated using expected counts and the median length of each transcript considering all RNA-Seq libraries.

4.7.5 lncRNA prediction

Single nucleotide polymorphisms (SNPs) were also called separately for both *E. salsugineum* ecotypes following GATK's best practices (VanderAuwera et al. 2013) for RNA-Seq data (Accessed September 1, 2018). Reads were mapped to the *E. salsugineum* genome, rather than transcriptome, with STAR v.2.5.2b (Dobin et al. 2013) using splice site junctions identified by the first read mapping step as suggested GATK's best practices. SNPs were called individually for each library and were merged into a genotype-specific single variant calling file for downstream analyses. SNPs were filtered using GATK's `VariantFiltration` software to flag: clusters of three of more SNPs in 35 base pair windows, QualByDepth (QD) <2 and FisherStrand (FS) >30. Using a custom Python script, only homozygous SNPs that were not flagged and found in the majority of each genotype's cDNA libraries were retained. VCFtools (Danecek et al. 2011) was used to create new genome consensus files for each genotype containing the consensus filtered SNPs. Transcript sequences for each ecotype were extracted using the individual genotype genome files and merged annotation files. Each transcript was then input into CREMA (https://github.com/gbgolding/crema) (Simopoulos et al. 2018) for lncRNA prediction. Because CREMA uses a scoring system for lncRNA prediction, only those transcripts with a prediction score > 0.5 were considered putative lncRNAs.

4.7.6 Multivariate analysis

Statistical analyses on gene expression data was performed using R v3.5.1 (R Core Team 2018) using FPKM values shifted by a constant of 1 to allow the data to be $\log_2$ transformed. Normalization was used to to account for the disparity in transcript abundance for genes with very low or very high expression. PCA was performed on the covariance matrix for all genes detected across the 31 RNA-Seq libraries of both *E. salsugineum* ecotypes subjected to the progressive drought treatment in order to explore variation within and between the transcriptomes with regards to gene expression estimates. Log_2 transformed FPKM values for transcripts associated with 28,712 genes were treated as variables while each of the 31 cDNA libraries were treated as observations.

4.7.7 Detection of differentially expression genes

DEGs were called using the `DESeq2` Bioconductor package (Love et al. 2014) using a FDR of 0.05 (Champigny et al. 2013). To control for a potential batch effect due to differences in library preparation protocols, library preparation type was added to the `DESeq2` regression formula. In addition, a threshold was set for differentially expressed genes to reduce predictive error that may arise with biological variance. In a differential expression test between Condition A vs Condition B, all genes identified as "upregulated" must have gene expression estimates

$>$ 1 FPKM in Condition A. Similarly, all genes identified as "downregulated" must have gene expression estimates $>$ 1 FPKM in Condition B. DEGs were identified in all biologically relevant drought progression transitions (*i.e* WW1 vs D1, D1 vs WW2, WW2 vs D2).

4.7.8 Weighted gene co-expression network analysis and GO term enrichment

A gene co-expression network was inferred from untransformed FPKM values of all expressed transcripts using the `WGCNA` R package (Langfelder and Horvath 2008). A "signed hybrid" network was constructed in a blockwise manner using a maximum block size of 10,000 genes, a soft threshold power of 9, minimum module size of 30, and a merge cut height value of 0.25. Gene expression clusters were summarized using an eigengene value equal to the first principal component of gene expression values contained in each cluster. Cluster eigengene values were correlated to each ecotype's progessive drought treatment status in order to identify genes associated with an ecotype and/or drought treatment. Correlation of treatments was also clustered using hierarchical clustering. Eight clusters composed of over 50% of the previously identified DEGs were chosen for further investigation (Table S3.1).

The DEG-containing clusters were used to identify significantly enriched GO terms based on custom GO term annotation. Because homology with *A. thaliana* genes was used for GO term assignment, we used a reciprocal best BLAST hit approach to annotate novel transcripts identified in our cDNA libraries with *A. thaliana* loci. The most recent *A. thaliana* GO terms were downloaded from TAIR on November 12, 2018 (`https://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/GO_and_PO_Annotations`). *E. salsugineum* genes were annotated with *A. thaliana* GO terms using *A. thaliana* loci available from Phytozome v12.1.5, annotation provided by Champigny et al. (2013) and the annotation of novel transcripts by reciprocal best BLAST hit. GO term enrichment of each of the eight clusters was called using the `topGO` R package (Alexa A 2018) using the Benjamini and Hochberg (1995) FDR set at a 0.05 significance threshold. Redundancy of enriched GO terms was reduced using the Revigo webserver (Supek et al. 2011), using the *A. thaliana* GO term database size and an allowed SlimRel maximum measure of 0.4. FDR adjusted p-values of the enriched GO terms were also used in the GO term summary process.

4.8 Acknowledgements

Yukon *E. salsugineum* plants and seeds were collected under a Yukon Scientists and Explorers Act License.

4.9 Funding

This work was supported by grants from the Ontario Research Fund Research Excellence (RE03-043) and Natural Science and Engineering Research Council of Canada to EAW (RGPIN-2015-06530) and to GBG (RGPIN-2015-04477). MJRM was a recipient of a NSERC Graduate Scholarship.

4.10 Availability of data and materials

Raw FASTQ RNA sequencing data is available in the SRA with submission ID SRR7962298 (SWW1.1) and BioProject accession PRJNA494564. All other RNA sequencing libraries are available upon request and will be uploaded to the SRA for publication.

4.11 Authors' contributions

The original experiment was designed by MJRM and EAW. MJRM grew the *E. salsugineum* plants used in the experiment and extracted RNA from 16 plants for RNA sequencing. SI extracted RNA from 15 plants for RNA sequencing and performed the RT-qPCR experiment. WWLS and MJC performed preliminary data analyses. CMAS performed all bioinformatic analyses. EAW and CMAS wrote the manuscript. PS contributed to experimental design and manuscript writing and revision. GBG and EAW supervised the analyses and revised the manuscript.

4.12 Ethics approval and consent to participate

Not applicable.

4.13 Consent for publication

Not applicable.

4.14 Competing interests

The authors declare that they have no competing interests.

4.14.1 Additional files

Additional file 1: Figures and tables found in Appendix C.

Additional file 2: .xlsx, significant DEGs identified at all biologically relevant drought progression stages.

Additional file 3: .xlsx, DEG composition of all clusters.

Additional file 4: .xls, GO enrichment of selected clusters.

Additional file 5: .gtf, New *E. salsugineum* genome annotation containing XLOCs identified by Champigny et al. (2013) and novel DLOCs.

Chapter 5

Conclusion

Caitlin M. A. Simopoulos

5.1 Thesis summary

Long non-protein coding RNAs (lncRNAs) are essential players in development and stress responses, yet they remain understudied in plant systems compared to animal systems. Research on this RNA classification is difficult partly due to their fast evolution and thus lack of overall sequence homology observed in these transcripts (Derrien et al. 2012). To acknowledge and attempt to remedy the discrepancy in lncRNA knowledge between plant and animal species, I first introduced a lncRNA prediction tool that uses an improved machine learning methodology. The tool was trained only on empirically validated lncRNA sequences from all species, including the nine lncRNAs of plant-origin available at time of publishing. The final stacking generalizer used by the prediction tool was evaluated using 10-fold cross validation of the training dataset and resulted in over 96% accuracy (Table 2.3). LncRNAs from *Arabidopsis thaliana*, *Oryza sativa*, and *Eutrema salsugineum* were predicted using the tool and were compared to predictions available from GreeNC, an established plant lncRNA database (Paytuvi-Gallart et al. 2016). Unlike our machine learning tool, GreeNC uses a transcript filtering approach to lncRNA prediction and considers all transcripts that meet an arbitrary feature cutoff as a lncRNA. Thus we expected our tool to predict fewer lncRNAs since the stacking generalizer was trained on known, functional lncRNA sequences and was not imposed to classification "rules" *a priori.* Consistent with this expectation, we predicted a smaller number of lncRNAs than GreeNC, although extensive overlap of the predictions was observed (Figure 2.2).

The lncRNA prediction tool, herein referred to as Classifying RNA by Ensemble Machine learning Algorithm (CREMA), was applied to RNA sequencing data to answer three main questions:

1. If lncRNAs are not conserved by nucleotide sequence, is there evidence of phylogenetic signal in the molecular traits of lncRNAs?
2. Are lncRNAs adequately represented in the reference annotation of plant species?
3. What are the transcriptional contributions, both coding and non-coding, to two unique drought tolerance strategies of two natural accessions of *E. salsugineum*, an extremophile plant species?

To answer these questions, CREMA was first applied to RNA sequencing data of plant species with diverse evolutionary histories. Novel transcripts assembled from RNA sequencing data and transcripts available in reference annotation were used in the analysis. We predicted lncRNAs using both data sources to quantify the number of lncRNAs that are present and missing from the annotation of a subset of plant species. In our tested plant species, on average, 6.4% of all assembled and annotated transcripts were predicted as lncRNAs, a larger percentage than the estimated 1% in human genomes (Kapusta and Feschotte 2014). However, the lncRNA contributions in plant genomes ranged greatly, from 3% in *E. salsugineum* to almost 17% in

Amborella trichopoda (Table 3.2). The percentages of lncRNAs that were not present in their appropriate species genome annotations also ranged significantly, from 4.5% novel lncRNAs in *A. thaliana* to over 99% in *Solanum tuberosum* (Figure 3.1). Because lncRNA evolution is not well understood, and lncRNAs typically display less sequence conservation than protein coding genes, we looked for phylogenetic patterns in molecular traits on lncRNAs rather than nucleotide sequence. Specifically, phylogenetic signal estimates were calculated to determine if phylogenetic relationships can explain the variances observed in transcript length, open reading frame (ORF) length, GC content and exon numbers in predicted lncRNAs and all other transcripts in the assembled transcriptomes of the tested species. The results suggested inconsistent and unclear phylogenetic signal patterns of lncRNA trait values compared to the consistently high and significant phylogenetic signals detected in all other transcripts. In addition, this comparative study introduces the possibility that extremophyte species, such as *E. salsugineum* and *Boea hygrometrica*, require fewer lncRNAs than the average plant species. Both species are naturally tolerant to abiotic stresses (Kazachkova et al. 2018; Xiao et al. 2015) and have the smallest percentage of transcripts predicted as lncRNAs (Table 3.2). This is unexpected as there are associations between stress responses and lncRNA expression (Xu et al. 2017b).

To further explore the relationship between predicted lncRNA numbers in *E. salsugineum* and a stress response, we sequenced the RNA of two natural accessions of *E. salsugineum* subjected to a progressive two-stage drought treatment previously described by MacLeod et al. (2014). While *E. salsugineum* genotypes are considered to be naturally tolerant to drought, the two studied ecotypes, Yukon and Shandong, have unique physiological responses to water stress that are most evident after a second drought treatment. In fact, MacLeod et al. (2014) suggested that the two tested ecotypes display different drought response strategies, with Yukon plants displaying drought tolerance and Shandong drought avoidance. Using RNA sequencing, we detected 919 lncRNAs expressed at one point during the progressive drought treatment, of which just under 8% are present in the *E. salsugineum* reference annotation. Differentially expressed genes (DEGs) were identified at biologically relevant conditions (*i.e* WW1 vs. D1, D1 vs. WW2, WW2 vs. D2). In keeping with the physiological responses found for the two ecotypes, the DEGs profiles of Yukon and Shandong plants did not generally display overlap with each other. The differences regarding DEGs are particularly evident in the responses to a D1 where Yukon plants displayed 17-fold more DEGs compared to Shandong plants. Only 14% of DEGs detected at any drought transition stage were shared by both ecotypes and only 14 of these genes were predicted to be lncRNAs. This low overlap of lncRNAs indicates that different lncRNAs were induced by each ecotype during the drought treatment. A weighted gene co-expression network analysis (WGCNA) was also constructed to identify clusters of genes that co-express and correlate to the drought responses of each ecotype. A single gene cluster correlated to the same condition in both ecotypes in the same direction, again suggesting that each natural accession also induced unique groups of genes during the responses to both drought treatments. Finally, functional enrichment of the identified gene clusters allowed for functional prediction of lncRNAs involved in each ecotype's drought response that can be used in future lncRNA validation studies.

5.2 Insights on current and future lncRNA research

The lncRNAs described in this dissertation remain computationally predicted, however lncRNAs that have been empirically characterized by other studies have displayed similar molecular characteristics as our predictions. In fact, although a general definition of lncRNAs exists, there remains few set rules or agreed-upon classifications for long non-protein coding transcript identification. For example, St. Laurent et al. (2015) described over 50 overlapping classifications of lncRNAs that group lncRNAs by genome location, function, and size. As a whole, lncRNAs are transcripts that exhibit little sequence homology between species, have seemingly no consistent distinguishing molecular feature cutoff values, and have multiple non-specific classifications. However, the heterogeneity of lncRNAs may exist due to the presence of lncRNA families or subclassifications that have yet to be identified, which may stem from a lack of empirically validated lncRNAs.

5.2.1 Putative lncRNA subclasses

Unlike small non-protein coding RNAs (ncRNAs), there are few described subclasses of lncRNAs. However, the lncRNAs that encompass the small number of distinct subclasses typically exhibit structural and functional homology. While there is controversy whether extensive lncRNA structure is conserved (Rivas et al. 2017), as discussed in a proceeding section, RNA structure in general has long been considered a feature of functionality (Blythe et al. 2016). Fitting all criteria of a lncRNA, riboswitches are a class of RNA commonly found in bacteria that regulate gene expression in *cis* using their unique secondary structures (Nudler and Mironov 2004). Their complex structure is categorized into two regions: the adaptamer domain and the expression platform (Tucker and Breaker 2005). Typically forming multiple hairpin structures, the adaptamer domain binds to a ligand, usually a metabolite, and the expression platform transitions into an active, or inactive structure for expression induction or repression (Nudler and Mironov 2004). Riboswitches are found extensively throughout eubacterial genomes and are classified into families according to secondary structure, ligand-type, and function (Montange and Batey 2008).

Circular RNAs (circRNAs) represent another unique non-polyadenylated subclass of lncRNA. While the exact mechanisms of biogenesis remain uncertain (Quan and Li 2018), circRNAs are covalently linked RNAs that lack both a 5‘ cap and a poly-A tail due to their circular shape. Similar to lncRNAs, circRNAs have a large range of validated functions from acting as endogenous microRNA (miRNA) sponges (Hansen et al. 2013), to containing functional small ORFs (Chekulaeva and Rajewsky 2018). Similarly, topological anchor point RNAs (tapRNAs) are another class of lncRNA that exhibit similarity of molecular features. The tapRNA group was originally defined as positionally conserved lncRNAs with transcription start sites preferentially located within chromatin loops and CCCTC-binding factor anchors (Amaral et al. 2018). However, tapRNAs were found to not only be syntenic and structurally conserved, but also demonstrated nucleotide sequence homology with conserved functional motifs. Similarly, enhancer RNAs (eRNAs), transcribed from enhancer regions of protein coding genes, are another subclass of lncRNA. eRNAs are typically shorter than an average lncRNA, do not undergo post-transcriptional processing, and are proposed to induce transcription of their partner gene (Kim et al. 2015).

As lncRNA research progresses, novel classes or groups of lncRNAs are emerging. Conservation of molecular features within each class of lncRNA has been observed in riboswitches, circRNAs, tapRNAs and eRNAs. While there is evidence that lncRNAs in general lack extensive conservation in nucleotide sequence or molecular traits, our poor understanding of lncRNA evolution may merely be a product of a lack of characterised lncRNAs and/or RNA subgroups. As such, specific and non-arbitrary criteria for defining lncRNAs should be identified, and lncRNA classes should ideally have specific, non-overlapping definitions. However, creating definitions for lncRNA subgroups is not trivial and will require the continuation of empirical validation and functional characterization. In addition, it is possible that lncRNAs should be categorized according to functionality rather than structure. For example, circRNAs and lncRNAs have the same range of functionally validated molecular mechanisms, but are considered different classes of RNA (Hansen et al. 2013; Chekulaeva and Rajewsky 2018). We believe that the delivery of scores on lncRNA predictions by CREMA offer a means for researchers to prioritize validation experiments. As novel lncRNAs are characterized, they can be added to CREMA's training dataset which in turn will improve the model's prediction accuracy.

5.2.2 Current reference genome annotations and lncRNAs

When we predicted lncRNAs in multiple plant species, we observed that the reference annotations of less commonly studied plants typically did not contain the majority of predicted lncRNAs. For example, over 99% of the lncRNAs predicted in *S. tuberosum* were not found in its reference annotation (Figure 3.1). Conversely, the reference annotation of *A. thaliana*, a well studied model plant, contained 95.5% of the lncRNAs identified in the study. Although most predicted lncRNAs in *A. thaliana* are found in the reference genome, they are often functionally annotated as transposable elements or pseudogenes rather than functional non-coding transcripts (Table 2.4). While this may be somewhat expected, as lncRNAs can evolve from transposable elements (TEs) (Kapusta et al. 2013), lncRNAs that are incorrectly functionally annotated may be contributing to difficulties in lncRNA research. Further, the lncRNAs contained in the reference annotations of less commonly-studied plant species, like *E. salsugineum* and *O. sativa*, typically lack any functional annotation (described in Chapter 2). The issue of inadequate lncRNA annotation is not unique to plant systems. As discussed in Chapter 3, Jackson et al. (2018) have also identified misannotation of human lncRNAs. A lack of lncRNAs available in reference annotations, or lncRNAs incorrectly functionally annotated, impedes the progress of lncRNA research.

Although not a trivial solution, reference annotations of species with inadequate lncRNA information should be updated on a regular basis. However, re-annotation would require a set definition for lncRNAs that may require criteria that define lncRNA subclasses. Alternatively, including lncRNA prediction statuses will benefit research until lncRNA subclasses have been identified. In addition, although computational prediction is an important step towards empirical validation, functional characterisation of lncRNA must continue, particularly to identify groups of lncRNAs that share common characteristics.

5.2.3 The evolution of lncRNAs with conserved nucleotide sequences

It is accepted that lncRNAs display little sequence conservation, however, there are long non-coding transcripts that exhibit at least small regions of homology. *COOLAIR*, a group of cold-induced natural antisense transcript lncRNAs (NAT lncRNAs) involved with the regulation of FLOWERING LOCUS C (*FLC*), were first described in *A. thaliana* and remain some of the most well studied plant lncRNAs (Swiezewski et al. 2009). *COOLAIR* transcripts are involved in the vernalization response and are found on the opposite strand of the 3′ end of *FLC*, their target gene. Interestingly, *FLC* is a MADS box transcription factor. There are two classes of *COOLAIR* transcripts that are dependent on alternative splicing (Hawkes et al. 2016). Class I transcripts are approximately 450nt, whereas Class II transcripts are longer at around 750nt. The complexity of the evolutionary studies on *COOLAIR* transcripts within Brassicaceae species are great examples of the unclear evolution of plant lncRNAs. COOLAIR transcript sequences are conserved within Brassicaceae species, particularly in *COOLAIR*'s first exon that interacts with an R-loop involved in regulation of the lncRNAs (Castaings et al. 2014). The 150bp region of sequence conservation within exon 1 was found to be at the 3′ end of *FLC*, and was more conserved than expected when considering other developmentally associated MADS box transcription factors that lack NAT lncRNAs. This also indicates that the promoter involved in the regulation of *COOLAIR* and its homologs in other Brassicaceae species is highly conserved. COOLAIR homologs in *Arabidopsis lyrata* and *Arabis alpina* were cold-inducible and displayed similar expression patterns as observed in *A. thaliana*.

The secondary structures of *COOLAIR* transcripts, determined by shotgun secondary structure determination, demonstrate conserved secondary structures within Brassicaceae species even within regions of diverged nucleotide sequence (Hawkes et al. 2016). Interestingly, the predicted secondary structures of *COOLAIR* identified in *A. thaliana*, *A. lyrata*, *Capsella rubella*, *A. alpina*, *Brassica rapa* and *E. salsugineum* were all similar, however, *B. rapa* contains a unique central domain as it does not contain a conserved helix. Hawkes et al. (2016) suggests that the conserved structure of *COOLAIR* transcripts supports the hypothesis that regions of conserved structural similarity are involved in regulating *FLC*, and the variable length of the H4 helix is due to adaptation to natural environments. Although the sequences and secondary structures of *COOLAIR* transcripts in multiple Brassicaceae species have been studied, *A. thaliana* is the only species with transcripts deposited in NCBI. In fact, although Hawkes et al. (2016) amplified the expression of *COOLAIR* in four species, the evolutionary studies of the transcripts were based on *FLC* sequences and the sequences of *COOLAIR* were not released. Additionally, Araport11 (Cheng et al. 2017), an updated source for *A. thaliana* gene annotations, does not contain annotation for *COOLAIR*, although a NAT lncRNA (AT5G01675) much longer than *COOLAIR* (6230nt vs the 750nt of *COOLAIR*) is annotated near the transcription start site of *COOLAIR*. A lack of publicly available *COOLAIR* transcript sequences leaves researchers dependent on either RNA sequencing data or *FLC* sequences for current evolutionary studies of this NAT lncRNA. The annotation status of *COOLAIR* also corroborates our conclusions in Chapter 3 that plant reference genomes are often missing lncRNAs.

HOX transcript anitisense RNA (HOTAIR), like *COOLAIR*, is a well studied lncRNA however it is primarily found in mammalians. Expressed from the *HoxC* cluster, HOTAIR regulates *HoxD* genes by Polycomb Repressive Complex 2 (PRC2) recruitment (Wu et al. 2013) and is involved in limb development by skin fibroblasts in humans (Schorderet and Duboule 2011). The human *HOTAIR* transcript is poorly conserved in mice, exemplified by a loss of two exons in the mouse *Hotair*. In terms of functionality, the mouse *Hotair* does not seem to play essential roles in development as observed in humans, and may have lost functionality during sequence divergence. Evolutionary studies of *Hotair* in marsupials indicates that *Hotair*, and other lncRNAs found in HOX clusters, are ancient and may have evolved over 160 million years ago before the divergence of marsupials and eutherians (Yu et al. 2012). Additionally, phylogenetic analysis of *HOTAIR* in 10 mammalian species suggests that exons 1 and 6 of the six exon *HOTAIR* transcript evolve more quickly in primates than other animals but all exons evolve more quickly than their surrounding protein coding genes (He et al. 2011). Using a subset of available *HOTAIR* transcripts, it is evident that the transcripts differ in molecular traits (Table 5.1). For example, transcript length ranges from 325 to 2370nt and there exists differences in the numbers of exons. GC content, however, remains relatively stable across the tested species, other than in mice with a lower than average GC%.

Table 5.1: Molecular traits of *HOTAIR*

Species	Transcript Info	NCBI ID	Length	Num. of exons	GC content (%)	Citation
Human	Variant 1	NR_047517	2370	6	48.6	Woo and Kingston (2007)
Mouse		NR_047528	2222	2	41.5	He et al. (2011)
Tammar		NA	325	3	46.2	Yu et al. (2012)
Dog		NR_131937	619	5	47.4	NA
Chimpanzee		NR_131936	2242	4	47.3	NA

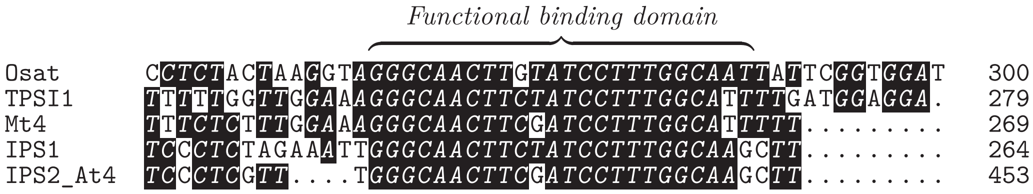
Similarly to *COOLAIR*, Somarowthu et al. (2015) demonstrated that *HOTAIR* forms into an intricate secondary structure that the authors believe may be involved in its binding functionality with PRC2, however, its structure was computationally predicted. Phylogenetic analysis using *Hotair* sequences from other mammalian species confirmed *Hotair*'s predicted secondary structure where putative protein binding domains contained either conserved or co-varying bases. However, recent analysis by Rivas et al. (2017) argues that most structural analyses of lncRNAs, like the analysis by Somarowthu et al. (2015), use invalid software that overly bias results towards "compensatory base-pair substitutions" and ignores mutations that disrupt conserved structure. Rivas et al. (2017) argue that *Hotair* transcripts do not show conserved structure, suggesting that work on conservation of lncRNA secondary structure should consider using more appropriate tools. Although a conservative approach to secondary structure prediction was taken, this suggestion also includes the analysis of *COOLAIR* transcripts by Hawkes et al. (2016) as the authors did not use a method similar to the one suggested by Rivas et al. (2017).

5.2.4 Are lncRNAs truly species-specific?

Understanding how lncRNAs evolve, in combination with functional characterization, will help researchers determine how often lncRNAs are truly species-specific. For example, lncRNAs often do not share sequence homology in two different species, but may have the same function or act on the same target gene. Induced by Phosphate Starvation 1 (*IPS1*), an endogenous miRNA sponge first identified in *A. thaliana* (Franco-Zorrilla et al. 2007), has also been identified in tomato (Liu et al. 1997), barrel clover (Burleigh and Harrison 1997), soybean (Burleigh and Harrison 1999), and rice and grape (Franco-Zorrilla et al. 2007). These *IPS1* homologs, however, differ greatly in nucleotide sequence except for a small 23nt region of homology (Figure 5.1). This small motif is the functional binding domain where *IPS1* sequesters its target gene miR-399. Thus, *IPS1*-homologs all function as miR-399 sponges, but do not display overall sequence conservation. Functional domains of other functional lncRNAs are often more obscure than motifs with homology to target genes. For example, tapRNAs are mostly defined by positional conservation and location within promoters and chromatin loops, and typically display limited sequence homology (Amaral et al. 2018). Unlike protein-coding genes whose functions can be predicted by sequence similarity due to shared evolutionary relationships (Pearson 2013), a lack of sequence homology in lncRNAs prevents researchers from using protein-coding loci based functional prediction protocols. A lack of sequence homology, however, does not necessarily indicate species-specificity in lncRNAs. Future work should not only focus on identifying subclasses of lncRNAs, but also to use criteria that define lncRNA subclasses to determine if any functionally conserved lncRNAs exist between species.

5.2.5 Do extremophytes have fewer lncRNAs than the average plant?

In Chapter 4, we discuss the potential for plants with natural abiotic stress tolerance to have fewer lncRNAs than the average we predicted in our tested plant species. In our study, we



```
Osat      CCTCTACTAAGGTAGGGCAACTTGTATCCTTTGGCAATTATTCGGTGGAT   300
TPSI1     TTTTTGGTTGGAAAGGGCAACTTCTATCCTTTGGCATTTTGATGGAGGA.   279
Mt4       TTTCTCTTTGGAAAGGGCAACTTCGATCCTTTGGCATTTTT.........   269
IPS1      TCCCTCTAGAAATTGGGCAACTTCTATCCTTTGGCAAGCTT.........   264
IPS2_At4  TCCCTCGTT....TGGGCAACTTCGATCCTTTGGCAAGCTT.........   453
```

FIGURE 5.1: Sequence alignment of the 23nt functional domain in *IPS1* in rice, tomato, barrel clover and two in *A. thaliana*. Alignment does not show entire *IPS1* sequence and is instead localised to the functional domain which is a target mimic for miR-399. Nucleotides which are found in at least 50% of the species are shaded. Gaps are represented by ".". Gene sequences are accessions BU673244 (Osat, *O. sativa*), V34808 (TPSI1, *S. lycopersicum*), U76742 (Mt4, *Medicago truncatula*), AF236376 (IPS1, *A. thaliana*), AY536062 (IPS2_At4, *A. thaliana*). Sequences previously used for a sequence alignment by (Franco-Zorrilla et al. 2007) were re-aligned using Muscle v3.8.425 (Edgar 2004). Sequence alignment is visualised using TEXshade (Beitz 2000).

observed 3% of all assembled transcripts from unstressed plants of *E. salsugineum* and *B. hygrometrica* as predicted lncRNAs, less than the average of 6.4% found in our tested plant species (Table 3.2). It is interesting that lncRNA prediction numbers were not proportional to genome size, as *E. salsugineum*'s genome is almost two-fold the size of the *A. thaliana* genome (241 Mb vs 135 Mb) and contained a smaller percentage of predicted lncRNAs (3.0% vs 4.3%; Table 3.2). In Chapter 3, we proposed that "priming", or constitutive expression of stress-associated genes due to previous stress encounters, may indicate less of a need for lncRNAs in stress tolerant plants. This hypothesis was suggested because stress-associated gene regulation is not as pervasive in our tested extremophytes compared to stress-sensitive species. However, it is also possible that extremophiles induce the expression of lncRNAs with molecular features that CREMA did not include in its training datasets and therefore, could not predict. This hypothesis is unlikely as CREMA was able to identify lncRNAs with molecular features that varied and did not follow the estimated phylogenetic relationships in tested species (Figure 3.1; Table 3.2). Additionally, over 90% of the top ranking lncRNAs predicted by CREMA with annotations were functionally characterized as putative lncRNAs or lncRNA progenitors (Table 2.4). GC content of lncRNAs is the only tested molecular trait that displayed consistently high phylogenetic signal. This finding was corroborated by a previous study by Haerty and Ponting (2015) which identified evidence of selection on GC content of predicted lncRNAs in animal species. It may be possible that CREMA identifies lncRNAs with certain GC percentage ranges leaving lncRNAs with unusual GC contents undetected. Thus, we also hypothesize that the genomes of *E. salsugineum* and *B. hygrometrica* may contain lncRNAs with unexpected features which remain undetected and are influencing the abnormally low number of detected lncRNAs in these species.

As lncRNA research moves forward, particularly with empirical validation and functional characterisation, it is essential to continue updating the training datasets of CREMA and other machine learning-based tools. In addition, novel molecular features that distinguish lncRNAs should be included in future iterations of prediction tools for more accurate results. Finally, while CREMA was originally intended to be used even when reference genomes were unavailable, genomic position in relation to protein coding genes, promoters or enhancer regions may be useful in identifying additional lncRNAs as these features have been used in other lncRNA subclass criteria (Amaral et al. 2018).

5.3 Machine learning in biology

In this thesis, we describe a machine learning classifier, CREMA, used for putative lncRNA identification that was accurate even when trained on small datasets and few distinguishing features (Chapter 2). We created this open-source computational tool to guide lncRNA validation studies and included a prediction scoring system to help focus research on likely functional lncRNA candidates. As discussed in Chapter 1, machine learning has been used extensively in biological contexts, from secondary protein structure prediction (Wang et al. 2016) to DNA-protein interaction identification (Kelley et al. 2016). LncRNA research was an excellent choice for the application of an ensemble learning algorithm due to the heterogeneous nature of the non-coding transcripts. In particular, a logistic regression meta-learner, or stacking generaliser, was the most suitable ensemble algorithm tested for accurate prediction of lncRNAs from very diverse species compared to other ensemble methods, such as majority vote or mean scores (Table 2.3).

5.3.1 Pitfalls of ensemble methods

Ensemble algorithms represent a machine learning framework that combines multiple models into a single classifier and are typically more accurate than a single model (Baba et al. 2015). Although ensemble machine learning methods are a popular choice for classification studies, there are occasions when ensemble models are not the optimal choices for predictions (Wang 2008). As Baba et al. (2015) discuss, diversity of the individual base classifiers is essential for accurate predictions and can be implemented by using different machine learning algorithms, re-sampling of training data, or using different training datasets as used in the construction of CREMA (Chapter 2). In addition, feature selection and parameter tuning are essential for the success of the ensemble model, two approaches also used in the construction of CREMA. Wang (2008) determined that "voting" and "average" ensemble models with fewer base classifiers (< 7 classifiers) may be accurate on training data, but do not typically perform well on test data. Conversely, Baba et al. (2015) suggest that fewer base classifiers are preferred for unbalanced training data, such as the data used to train CREMA. Nonetheless, ensemble methods are especially useful when classifying heterogeneous data, such as lncRNAs.

5.3.2 Stacking generalizers

Stacking refers to an ensemble-type algorithm that combines multiple models by training a meta-learner on the outputs of said base models. Stacking is not as commonly used as other ensemble model approaches, like boosting or bagging, due to extensive computational requirements. However, due to computational advancements, stacking has recently been used to solve complex problems such as remote sensing of forest change (Healey et al. 2018), prediction of ski injuries from metrological data (Delibašić et al. 2018), miRNA-mRNA target prediction (Van Peer et al. 2017), ageing biomarkers (Putin et al. 2016), and lncRNA subcellular localization (Cao et al. 2018). Although stacking can be an appropriate approach to classification problems not solvable by a single model, running multiple models requires more computational resources than a single algorithm (Healey et al. 2018). In Chapter 2, we used a stacking approach to combine custom models into a single classifier. It is also possible to train a stacking generalizer on the outputs of multiple, independent, previously-published models as Healey et al. (2018) proposed for predicting forest change. While including other lncRNA prediction tools would increase the heterogeneity of our stacking learner, a benefit to this ensemble approach, combining independent models would most likely require cloud-computing for software and data compatibility.

Machine learning models, including stacking classifiers, are also prone to a lack of interpretability where researchers tend to choose models with the highest model evaluation scores rather than "actionable insights" (Krause et al. 2016). For example, although CREMA is able to predict lncRNAs with over 96% accuracy (Table 2.3), the complexity of the individual models that make up the stacking generalizer prevent easy interpretability of the model. In other words, it is difficult to determine how a change in a feature, for example the Fickett score, changes the final lncRNA prediction. This is a disadvantage to using "black box" machine learning approaches for lncRNA prediction, especially as researchers are currently lacking criteria for lncRNA identification and cannot use machine learning tools to identify important distinguishing features.

Currently, there are efforts to help researchers visualize and understand feature contributions of "black box" machine learning models for increased interpretability (Krause et al. 2016). In Chapter 2, we used recursive feature elimination to determine that TE-related features were not informative in our training data or predictive models. While this approach does not quantify the amount each feature contributed to lncRNA predictions, in our case, it allowed us to remove a computationally intensive preprocessing step and to make inferences on our remaining training dataset. Namely, our positive training dataset of empirically validated lncRNAs showed little evidence of TE origin, contrary to the study by Kapusta et al. (2013). Krause et al. (2016) use a different approach to increase model interpretability, where partial dependence plots can be visualized for each feature demonstrating how changes in feature value affect a prediction score. Their tool, Prospector, is a web-based interactive tool applicable for any Python-based machine learning model that encourages data and feature exploration and increased interpretability of complex predictors. Future work should consider using tools like Prospector in concert with lncRNA prediction tools to identify important molecular features of lncRNAs.

5.3.3 Deep learning

The affordability of sequencing has contributed to a large increase in molecular data, and an estimate of one Zetta-basepairs (10^{21}) of DNA will be sequenced by 2025 (Stephens et al. 2015). Such large amounts of data require efficient computational processes for storage and analysis. In particular, efficient computational algorithms are needed to identify patterns in largely heterogeneous data. As Zou et al. (2019) discuss, deep learning algorithms can offer novel approaches to molecular analyses and "big data". The term "deep learning" refers to a group of machine learning algorithms commonly associated with artificial neural networks (Ching et al. 2018). Artificial neural networks connect input and output layers (data to be classified and prediction outcomes) to "hidden layers" consisting of features that are connected to other layers via edges. Deep learning is complex and, depending on the algorithm, can allow for one hidden layer to influence the arrangement of other layers for accurate predictions.

Deep learning displays a similar lack of interpretability as other machine learning models due to the complexity of hidden layer construction during training. However, unlike other supervised machine learning methods, deep learning does not always require feature selection and, therefore, can require little domain knowledge for model construction. For example, deepTarget is composed of deep recurrent neural network-based autoencoders that can predict miRNA precursors and their corresponding target genes from nucleotide sequences (Lee et al. 2016a). The autoencoders, in the case of deepTarget, are first used in an unsupervised way to identify sequence features essential to miRNA-mRNA target prediction. Autoencoders represent a classification of neural networks, where the input and output are the same, thus the model can identify low-level features of input data for the projection of output data. In our work, we could have used a deep learning approach to feature selection in Chapter 2 for lncRNA prediction to identify seemingly cryptic features that distinguish lncRNAs. Additionally, rather than manual feature selection from the literature, as presented in Chapter 2, deep learning can be used in automatic feature selection for use in a supervised model. A lncRNA subcellular localisation tool, lncLocator, also uses autoencoders to identify features for lncRNA cellular localisation. The features identified by autoencoders were then used in combination with manually curated features for classification by a stacked generalizer approach (Cao et al. 2018). We could also use this method to improve CREMA and use both autoencoder-found features and important features identified from previous research for lncRNA prediction. However, deep learning typically requires large amounts of data, and may not be suitable for use with our small training datasets.

5.4 Applying the drought tolerance strategies of *Eutrema salsugineum*

In Chapter 4, we applied CREMA to RNA sequencing data of two *E. salsugineum* genotypes experiencing a progressive drought treatment to better understand the molecular strategies behind *E. salsugineum*'s drought tolerance. *E. salsugineum* is a plant species with genotypes native to Canada. The natural environment of *E. salsugineum*'s Yukon ecotype is typically dry and has been classified as semi-arid due to the minimal amount of precipitation the region receives, making the Yukon genotype optimal for studying drought tolerance (Guevara et al. 2012). Although the natural region of Yukon plants is typically dry, other areas of Canada are not immune to abnormally dry conditions that subject plants to drought stress, the abiotic stress that contributes the most to crop yield loss (Boyer 1982) . As of September 30, 2017, The Ministry of Agriculture and Agri-Food Canada (2017) identified areas of moderate to exceptional drought in areas of Western Canada (Figure 5.2). A combination of heat and drought stress contributed to smaller than average yields in the 2017 growing season, particularly in canola plants, however soil moisture reserves helped moderate devastating reductions in yield in the Southern prairies. As climate change may contribute to reduced soil moisture (Schlaepfer et al. 2017), understanding the molecular mechanisms for drought tolerance is imperative for future crop improvement and ensuring future food security. Typically, transcription factors that regulate the expression of multiple stress-associated genes are targets for crop improvement by genetic engineering. This often requires an understanding of molecular signalling pathways and networks in combination with physiological outcomes of changed gene expression. A systems biology approach to stress tolerance is holistic and aims to piece together how the organism of interest responds to stressors. In Chapter 4, we presented a systems biology-based analysis, a gene co-expression network constructed from RNA sequencing data of our progressive drought treatment, to identify groups of genes associated with a drought response.

5.4.1 Co-expression networks and stress

Large amounts of gene expression data, whether by microarray or RNA sequencing, allows for more network-based approaches to understand stress signalling pathways. A review by Gehan et al. (2015) suggests that network-based statistical approaches should be used to identify conserved stress-associated pathways throughout tissues, plant genotypes and/or species. In Chapter 4 we demonstrated that Yukon and Shandong *E. salsugineum* ecotypes have different gene expression changes when exposed to a progressive drought treatment. When considering differential gene expression analysis, of the 5730 unique DEGs identified in the experiment, only 799 were common to both ecotypes. A gene co-expression network confirmed the observation that the ecotypes respond differently to stress where clusters of co-expressed genes tended to correlate to either

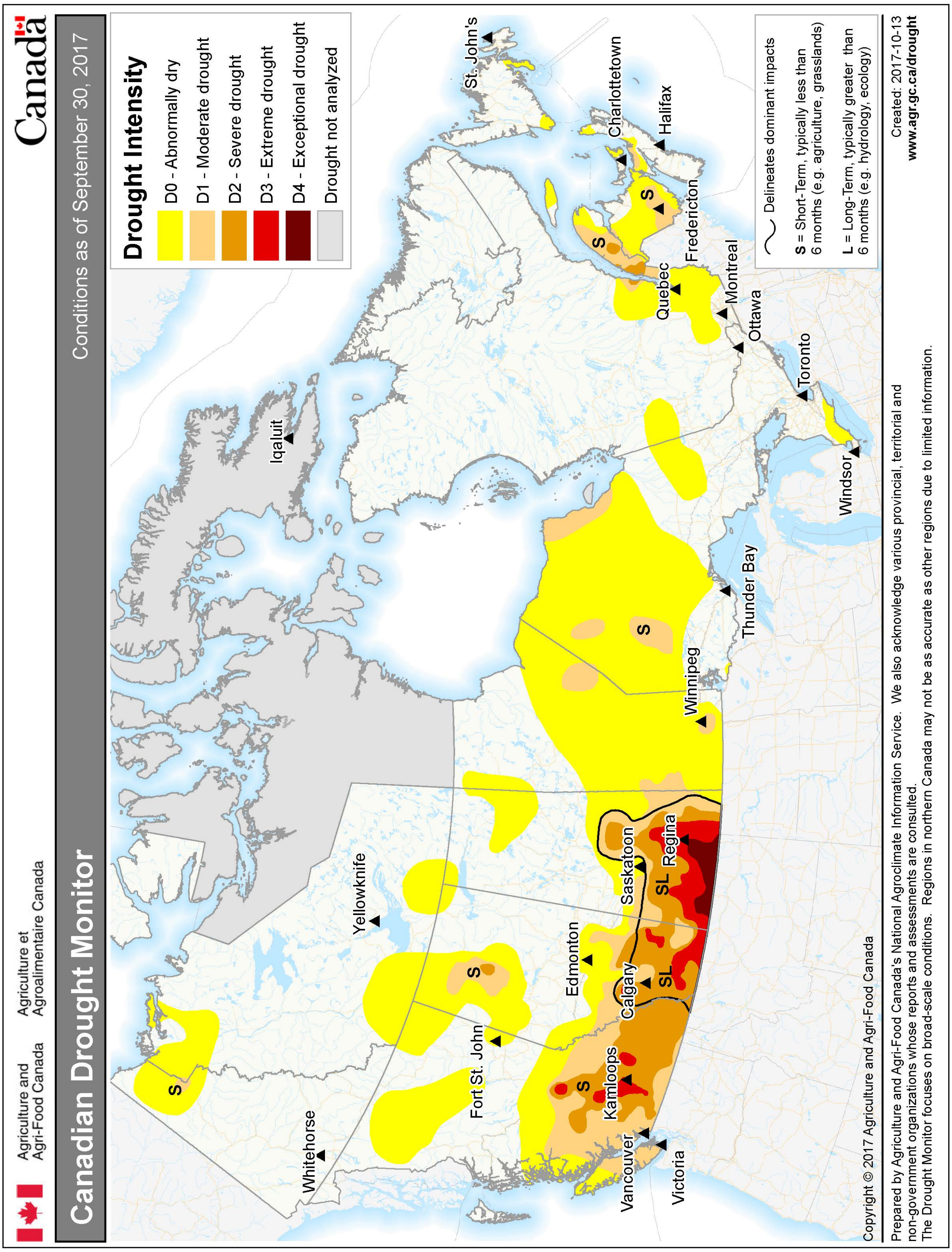

Agriculture and Agri-Food Canada
Agriculture et Agroalimentaire Canada
Canada
Canadian Drought Monitor
Conditions as of September 30, 2017
Drought Intensity
D0 - Abnormally dry
D1 - Moderate drought
D2 - Severe drought
D3 - Extreme drought
D4 - Exceptional drought
Drought not analyzed
Delineates dominant impacts
S = Short-Term, typically less than 6 months (e.g. agriculture, grasslands)
L = Long-Term, typically greater than 6 months (e.g. hydrology, ecology)
Whitehorse
Yellowknife
Iqaluit
Fort St. John
Edmonton
Calgary
Kamloops
Vancouver
Victoria
Saskatoon
Regina
Winnipeg
Thunder Bay
Windsor
Toronto
Ottawa
Montreal
Quebec
Fredericton
Halifax
Charlottetown
St. John's
S
SL
Copyright © 2017 Agriculture and Agri-Food Canada
Prepared by Agriculture and Agri-Food Canada's National Agroclimate Information Service. We also acknowledge various provincial, territorial and non-government organizations whose reports and assessments are consulted.
The Drought Monitor focuses on broad-scale conditions. Regions in northern Canada may not be as accurate as other regions due to limited information.
Created: 2017-10-13
www.agr.gc.ca/drought

FIGURE 5.2: A Canada Drought Monitor map of droughts occurring across Canada the month of September 2017. The legend includes information on drought classifications: "D0 (Abnormally Dry) an event that occurs once every 3-5 years, D1 (Moderate Drought) an event that occurs every 5-10 years, D2 (Severe Drought) an event that occurs every 10-20 years, D3 (Extreme Drought) an event that occurs every 20-25 years, D4 (Exceptional Drought) an event that occurs every 50 years" (The Ministry of Agriculture and Agri-Food Canada 2017)

ecotype, but rarely both (Figure 4.4; Table S3.1). This gene co-expression network may help researchers identify groups of genes associated with the unique drought response strategies of each ecotype where, for example, Yukon display drought tolerance and Shandong plants drought avoidance. In particular, regulatory loci like lncRNAs or transcription factors contained within gene clusters of interest may play roles in gene expression regulation of genes within clusters. A lack of overlap in differentially expressed lncRNAs in Yukon and Shandong plants experiencing the progressive drought corroborates our suggestions that both ecotypes respond differently to drought. Additionally, using functional enrichment, we were able to make functional predictions for novel lncRNAs associated with drought treatment and ecotype. For example, a large cluster of 3415 co-expressed genes was positively correlated with both of Yukon plants' drought responses. This cluster was enriched in genes associated to a dehydration response, peptide transport, and cellular lipid catabolic processes which are functions that have also been associated with drought stress in *A. thaliana* (Gigon et al. 2004).

Co-expression network analysis can also be combined with metabolite profiling (Coneva et al. 2014). Coneva et al. (2014) analysed microarray data from rice plants undergoing four different conditions of nitrogen availability. Functional enrichment of clusters identified by network analysis substantiated metabolite analysis. For example, Coneva et al. (2014) identified increased purine metabolism compounds in plants transplanted from high to low nitrogen conditions using metabolomics. Similarly, a gene expression cluster enriched with purine metabolism-related genes was positively correlated to the same nitrogen condition, indicating that co-expression network analysis is a valid approach to transcriptomic studies. Coneva et al. (2014) also discuss the merit of targeting the master regulators of clusters of interest in future studies to better understand the molecular mechanisms and regulation of genes deemed important by the analysis. Similarly, in our work, the putative lncRNAs identified by CREMA that belong to drought associated clusters should also be explored for their contributions to *E. salsugineum*'s drought tolerance.

5.4.2 Transgenic plants, crop improvement, and lncRNAs

As research on the molecular stress responses of plants continues, so does progress on applying this information to crop improvement. In addition to traditional and marker assisted breeding programs, genetic engineering by creating transgenic plants is a method that has been used in crop improvement work (Ronald 2014). An internationally successful example of transgenic crop improvement is the introduction of Bt crops that endogenously express insecticidal proteins originally isolated from *Bacillus thuringiensis* (Bt) (Shelton et al. 2002). In a long-term study describing the effects of Bt cotton in China, Lu et al. (2012) describe a reduction in pesticide use and re-introduction of more environmentally-friendly biocontrol methods for insect control. Additionally, beneficial insect generalist predator population numbers were not negatively impacted by the introduction of Bt cotton. Thus Bt crops may enable pest-free and sustainable farming practices with little negative environmental impact. The success of Bt crops suggest that transgenic crops are important to the advancement of stress-tolerant crops.

When we consider drought tolerance however, crop improvement is not straight forward. As Nuccio et al. (2018) discuss, there are few drought tolerant biotechnology products available to date. The progression from identifying a gene essential to drought tolerance to product development is time consuming, and can cost up to tens of millions of dollars. Because a plant's response to abiotic stress is metabolically expensive, transgenes inserted into a plant's genome are most commonly transiently expressed using stress-inducible promoters. However, single-stress responsive promoters are not as pervasive as those responsive to multiple stresses complicating drought-induced transgene expression attempts (Jeong and Jung 2015). Finally, public opinion on so-called "genetically modified organisms", or GMOs, is not positive (Hundleby and Harwood 2018). The European Union has recently introduced strict regulations on genetic modification and gene editing programs halting many crop improvement efforts in most of Europe. Nonetheless, lncRNAs may offer a new perspective to abiotic stress tolerance advancement through novel gene targets.

Although not stress related, Wang et al. (2018d) were successful in overexpressing LRK Antisense Intergenic RNA (*LAIR*) in *O. sativa* to increase crop yield. First identified as a yield-QTL in an *O. sativa* genotype, *LAIR* is contained inside the *LRK* gene cluster and induces the expression of other *LRK* genes. In the study, the authors transformed *LAIR* into rice and observed a significant increase in yield and larger and more panicles per plant. Cited applications of lncRNAs in crop improvement are limited, however, the study by Wang et al. (2018d) confirms that lncRNAs can be candidate genes for crop improvement. While over-expressing lncRNAs that influence drought tolerance may not improve the public's perception of genetically engineered crops, lncRNAs as gene targets for stress tolerance may solve problems stemming from a lack of drought-specific promoters. LncRNAs are functional even in low levels, thus potentially enabling researchers to improve stress tolerance without stress responsive promoters and instead with low constitutive lncRNAs expression. When considering crop improvement by drought tolerance and lncRNAs, there is a single empirically validated drought-associated lncRNA, drought induced lncRNA (*DRIR*) (Qin et al. 2017). Qin et al. (2017) also were successful in overexpressing *DRIR* in *A. thaliana* and observed a dosage dependant improvement in both drought and salt stress tolerance. However, because *DRIR* does not display sequence homology with other plants, it is unknown if this lncRNA is functional in other species, but its connection to abscisic acid (ABA)-mediated response pathways may indicate conserved functionality that warrants further research.

5.4.3 Importance of field studies in stress tolerance research

Stress experiments in growth cabinets give researchers complete control of experimental conditions but cannot mimic the natural environments of stress tolerant species (Champigny et al.

2013; Bechtold 2018). If crop improvement is the ultimate goal in drought tolerance research, plants experiencing drought in their natural environment may offer experiments with more realistic interpretations of drought tolerance. Additionally, constitutive expression of stress tolerance-related genes can have negative, unintended effects on plant growth and yield (Nakashima et al. 2007; Yang et al. 2010), two characteristics essential for the commercial success of traits that can only be fully tested during field trials.

Nuccio et al. (2018) suggest that a "disconnect" between experiments in controlled environments and field trials may be contributing to difficulties in bringing novel drought tolerant crop plants to market. However, stress experiments in controlled environments may be essential for understanding the molecular pathways induced by individual stressors. Our research group has previously considered gene expression estimates of Yukon *E. salsugineum* plants sampled from their natural environments alongside cabinet-grown plants (Guevara et al. 2012; Champigny et al. 2013). Using RNA sequencing, Champigny et al. (2013) demonstrated that field-sampled Yukon plants express genes related to photosynthesis, metabolism and stress differently compared to cabinet-grown Yukon plants. It is not feasible to carry out experiments in the remote Yukon locations where *E. salsugineum* is found, but our alternative was to devise a prolonged progressive drought treatment presented in Chapter 4. The progressive drought, however, cannot mirror the variety of other abiotic and biotic stresses plants experience daily when grown in the field. Comparisons between field-grown plants with plants stressed in growth cabinets is one approach for future research that could help identify pathways that are responsive to water deficits independent of where the plants were stressed.

5.5 RNA sequencing and lncRNA research

As discussed in Chapter 3, the majority of lncRNAs are not found in transcriptome annotations of plant species. Because of this discrepancy, lncRNA researchers cannot rely on transcriptome annotations alone, thus requiring RNA sequencing for efficient research. While CREMA, presented in Chapter 2, is able to predict lncRNAs from assembled transcripts from RNA sequencing data, the predictions rely on RNA sequencing quality. The leads to questions on how RNA sequencing parameters, such as library preparation protocols and sequencing technology, can affect lncRNA prediction and gene expression estimates.

5.5.1 PolyA+ selection is sufficient for lncRNA detection

RNA sequencing typically involves a pre-processing step where rRNA is removed from samples for more effective gene expression capture (Zhao et al. 2018a). This step is essential for accurate expression estimates because rRNA is the most abundant class of RNA and can obscure the expression results of lowly expressed genes. PolyA+ selection before sequencing is a cost-effective method for rRNA removal that results in transcriptomes enriched in polyadenylated transcripts. Conversely, it is also possible to deplete the rRNA in samples of interest, which will remove most rRNA transcripts but also retain other non-polyadenylated transcripts.

In Chapter 4, we used data from RNA sequencing protocols that used polyA+ selection methods rather than rRNA- depletion. LncRNAs are often transcribed by RNApolII and post-transcriptionally modified with 5′ caps and poly-A tails (Derrien et al. 2012). While we were able to identify lncRNA expression using a polyA+ selection approach, our sequencing data limited our lncRNA predictions to those transcripts that are polyadenylated. Zhao et al. (2018a) explored the differences in gene expression estimation results obtained by polyA+ selection compared to rRNA-depletion protocols. The authors found that a large fraction of transcript sequences from rRNA-depleted libraries were immature, and potentially non-functional, suggesting that rRNA-depleted libraries may lead to expression over-estimation. Although rRNA-depleted libraries were able to capture additional lncRNAs, particularly those without polyA tails, the polyA+ libraries still contained many lncRNAs of interest. Due to a reduced cost and the fact that it targets processed mRNA, polyA+ selection methods for RNA sequencing are an appropriate choice for lncRNA research. However, future work should consider using both polyA+ selection and rRNA-depletion protocols in combination to identify additional ncRNA transcripts.

5.5.2 Different library preparation protocols for single experiment

A challenge that may be encountered in large-scale gene expression studies are potential batch effects arising from, for example, multiple researchers, experiments completed on different days, or different sequencing protocols. In Chapter 4, the RNA used in our progressive drought treatment was extracted at two different times. The 16 original RNA sequencing libraries were sequenced in 2013 with an additional 15 libraries sequenced in 2018 from tissue that had been frozen for five years. A principal component analysis (PCA) completed on all 31 RNA sequencing libraries displayed batch effects that explained 0.9% of the variation in our data (Figure S3.1A). Wang et al. (2018b) explored the effect of variations in RNA sequencing library protocols and detected little expression differences in samples prepared using different methods. When Wang et al. (2018b) specifically tested for gene expression changes at cells stored at -80°for three years, reminiscent to our RNA sequencing experiment, the authors found only 90 genes differentially expressed between the original fresh samples and the samples that were cryoperserved. However, Wang et al. (2018b) observed a greater perturbance in lncRNA expression after cryopreservation compared to protein-coding genes. Due to our the visual confirmation of batch effects using PCA (Figure S3.1A), and the potential for gene expression differences due to cryopreservation rather than biological variation, we accounted for batch effect using the `DESeq2` R package (Love et al. 2014) during differential gene expression analysis. `DESeq2` accounts for batch effect by adding a “batch” variable to the generalised linear model in DEG identification removing the potential effects of different library preparation protocols from differential gene expression analysis.

5.6 Proposed studies

5.6.1 Prediction of lncRNAs induced by other abiotic stresses in the Yukon *E. salsugineum* genotype

In Chapter 4 we used CREMA in combination with RNA sequencing data to identify lncRNAs induced in two *E. salsugineum* genotypes during a progressive drought. Previous work has shown that the Yukon and Shandong ecotypes of *E. salsugineum* display unique molecular and physiological strategies to drought tolerance (MacLeod et al. 2014). Additionally, MacLeod et al. (2014) identified that Yukon plants experience "priming" and were more prepared than Shandong plants to tolerate a subsequent water stress while maintaining growth. Drought, however, is not the only abiotic stress where *E. salsugineum* displays innate abiotic stress tolerance. *E. salsugineum* is able to tolerate cold (Wong et al. 2005), salt stress (Gong et al. 2005) and nutritional deficiencies (Velasco et al. 2016) demonstrating characteristics of an extremophile (Kazachkova et al. 2018). I propose that transcripts not currently annotated in the *E. salsugineum* reference annotation are likely expressed in response to other abiotic stresses and may be contributing to the Yukon ecotype's superior stress tolerance. I also hypothesize that many of these "novel transcripts" will be predicted as lncRNAs, similar to the results of Chapter 4 that indicated 42% of novel transcripts were predicted as lncRNAs.

Our research group has RNA sequencing data from two unique nutrient deprivation studies. The first experiment consists of sequencing data from both leaves and roots of Yukon plants where phosphate was withheld (0mM and 2.5mM added phosphate). In the second experiment, Yukon plants were exposed to a combination of both sulfur and phosphate deprivation (0mM or 2.5mM added phosphate, and 0mM or 5000ppm calcium sulfate). The bioinformatics pipeline for transcript assembly and lncRNA prediction by CREMA that was presented in Chapter 4 can be applied to the previously described RNA sequencing libraries. By estimating the expression of all detected transcripts in Yukon plants undergoing multiple stresses, it will be possible to identify lncRNAs associated with a single stress rather than a general stress response. Using a single network created by WGCNA, clusters of genes with expression patterns associated with certain stresses or tissues can be identified. Alternatively, networks for each experiment can be constructed, and conserved or differential clusters can be identified using software such as MODA (Li et al. 2016a).

5.6.2 Yukon genome assembly

The reference genome and corresponding annotation for *E. salsugineum* that is currently available was compiled using sequence data from the Shandong ecotype (Yang et al. 2013). However, we found over 60,000 single nucleotide polymorphisms (SNPs) in Yukon plants compared to the reference genome in expressed regions using RNA sequencing data. This analysis did not consider InDels or mutations within non-coding regions which may be contributing to even more genomic variation between the two ecotypes. Additionally, in Chapter 4 we identified that the expression of 1023 and 1268 transcripts were unique to Yukon and Shandong ecotypes respectively. Of those 2291 transcripts, 169 and 237 were lncRNAs detected only in Yukon or Shandong plants. Ecotype-specific expression and known mutations in coding-regions may indicate that using Shandong's genome sequence in Yukon-focused studies may be leading to results biased by the Shandong ecotype.

I propose that an improved Yukon genome assembly should be completed and will be beneficial to any future experiments, either computational or empirical, on *E. salsugineum* plants of the Yukon ecotype. This could be done using available genomic data from an in-house, paired-end Illumina sequencing run. A genome-guided approach to expanding and connecting the 639 scaffolds of the *E. salsugineum* genome assembled by Yang et al. (2013) using our DNA sequencing reads may be an appropriate choice for the Yukon genome assembly. For example, By Adaptive Unique Mapping (BAUM) is a software that has been recently released to improve genome assemblies using an iterative approach (Wang et al. 2018a). Using an overlap-layout-consensus method, rather than a k-mer based assembly method, BAUM is less prone to assembly errors due to repetitive sequences. The iterative approach of BAUM also reduces the computational load of a typical overlap-layout-consensus assembly method. Guided by the *O. sativa* genome, BAUM has previously been used to assemble the *Oryza longistaminata* genome, a demonstration of success with plant genomes.

Conversely, a *de novo* genome assembly may be preferred, especially to identify previously un-sequenced genomic regions. Single molecule sequencing reads produced by technology from Pacific BioSciences (Eid et al. 2009) or Oxford Nanopore (Jain et al. 2015) may be more useful than short reads for a *de novo* genome assembly. It is difficult, however, to assemble single molecule reads due to a large computational requirement. A recent tool, MECAT, aims to reduce the computational need by single molecule *de novo* aligners by using pseudoalignment methods (Xiao et al. 2017).

5.6.3 The evolutionary rate of lncRNAs in Brassicaceae species

Sequence homology and secondary structure conservation of *COOLAIR* has been studied in multiple Brassicaceae species (Castaings et al. 2014; Hawkes et al. 2016). Similarly, the conserved and functional domain of *IPS1* and Induced by Phosphate Starvation 2 (*IPS2*) has been identified in *A. thaliana*, tomato, barrel clover and alfalfa. However, the rate at which both lncRNAs evolve has not yet been explored. Previous research on the evolutionary relationships of animal-derived lncRNAs has relied on sequence similarity for homolog identification (Necsulea et al. 2014), a method which may exclude lncRNA homologs with very fast evolution and bias results to a slower evolutionary rate than what is true. For example, lncRNAs such as *IPS1* and its homologs would have been missed in this analysis as the genes only contain a small 23nt region of homology. It is accepted that lncRNAs evolve more quickly than protein coding genes (Ulitsky 2016), however the rate at which they evolve has not been fully explored in plants. Due to the lack of validated lncRNAs in plants, large scale studies that quantify evolutionary rates of all predicted lncRNAs are difficult. Thus, I propose a study focused on known, functional lncRNAs that vary in the levels of sequence conservation in Brassica species. The study should quantify the evolutionary rate of both exons and introns, if applicable, of *COOLAIR* and *IPS1*/*IPS2* in *A. thaliana*, *A. lyrata*, *C. rubella*, *A. alpina*, *E. salsugineum* and *B. rapa*, mirroring the species chosen by Hawkes et al. (2016) in their study that identified conserved secondary structures of *COOLAIR*. A similar analysis should be completed on the target genes of each lncRNA as *FLC* has been proposed a target of *COOLAIR* and miR-399 is a known target of *IPS1* and *IPS2*. This work will quantify the evolutionary rates of known, functional lncRNAs and will compare these rates to conserved target genes.

5.7 Conclusion

In this work we presented a novel lncRNA prediction tool that addresses the gaps in knowledge between plant- and animal-derived lncRNAs. The accurate ensemble machine learning tool, CREMA, was then applied to RNA sequencing data of evolutionarily diverse plant species to identify conserved molecular traits in lncRNAs. The phylogenetic signal analysis supported previous research that suggested an unclear evolution of lncRNAs as signal estimates of molecular traits other than GC content were inconsistent or not significant. However, phylogenetic signal estimates were different in lncRNAs compared to all other assembled transcripts indicating that lncRNAs follow different evolutionary patterns than most transcripts. This analysis also highlighted a lack of lncRNAs contained in the annotations of plant genomes and cautioned researchers to not rely on genome annotations for lncRNA research. Finally, we used RNA sequencing to identify the molecular mechanisms behind the drought tolerance strategies of two *E. salsugineum* ecotypes. Differential gene expression analysis revealed little overlap between the gene regulation responses of the two genotypes. We also predicted lncRNAs from RNA sequencing data and identified only 14 lncRNAs differentially expressed in both ecotypes. Co-expression network analysis provides evidence that both ecotypes invoke unique molecular pathways when responding to drought treatment. Functional enrichment of clusters identified by co-expression network analysis was used to make functional predictions for lncRNAs expressed during drought.

Our novel lncRNA prediction tool contributes to lncRNA research by offering a tool specifically created to identify functional lncRNAs, and has been tested on multiple plant species. The tool can be applied to transcriptome data and can be used on species without reference genomes with the goal of being used in concert with gene expression studies to focus on empirical validation of lncRNAs. Our phylogenetic analysis of lncRNA molecular traits used a novel approach to evolutionary studies of lncRNAs, and corroborated the uncertainty of an evolutionary path given by nucleotide sequence-based methods. Finally, by identifying different lncRNAs expressed in Yukon and Shandong *E. salsugineum* ecotypes, we have identified potential gene expression regulators that may be contributing to drought tolerance and avoidance strategies. *E. salsugineum*'s drought-induced lncRNAs offer potential gene leads for improved drought tolerance of crops, and may be suitable targets without an accompanying need for single-stress associated promoters.

Bibliography

Akhade, V. S., P, D., and Kanduri, C. (2017). Long Noncoding RNA: Genome Organization and Mechanism of Action. In: *Long Non Coding RNA Biology*. Ed. by M. R. S. Rao. Springer Nature, 47–74.

Alexa A, R. J. (2018). *topGO: Enrichment Analysis for Gene Ontology. R package version 2.34.0.*

Amaral, P., Leonardi, T., Han, N., Vire, E., Gascoigne, D., Arias-Carrasco, R., Buscher, M., Pandolfini, L., Zhang, A., Pluchino, S., et al. (2018). Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol* 19, 32.

Amborella Genome Project (2013). The *Amborella* genome and the evolution of flowering plants. *Science* 342, 1241089.

Amtmann, A. (2009). Learning from evolution: Thellungiella generates new knowledge on essential and critical components of abiotic stress tolerance in plants. *Mol Plant* 2, 3–12.

Anderson, D., Anderson, K., Chang, C., Makarewich, C., Nelson, B., McAnally, J., Kasaragod, P., Shelton, J., Liou, J., Bassel-Duby, R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160, 595–606.

Angermueller, C., Parnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol Syst Biol* 12, 878.

Axtell, M., Westholm, J., and Lai, E. (2011). Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol* 12, 221.

Baba, N. M., Makhtar, M., Fadzli, S. A., and Awang, M. K. (2015). Current issues in ensemble methods and its applications. *Journal of Theoretical & Applied Information Technology* 81(2).

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28, 45–48.

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Jr., G. W., Kundaje, A., Gunawardena, H., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* 22, 1646–1657.

Banks, J., Nishiyama, T., Hasebe, M., Bowman, J., Gribskov, M., dePamphilis, C., Albert, V., Aono, N., Aoyama, T., Ambrose, B., et al. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332, 960–963.

Bardou, F., Ariel, F., Simpson, C., Romero-Barrios, N., Laporte, P., Balzergue, S., Brown, J., and Crespi, M. (2014). Long noncoding RNA modulates alternative splicing regulators in *Arabidopsis*. *Dev Cell* 30, 166–176.

Barsyte-Lovejoy, D., Lau, S., Boutros, P., Khosravi, F., Jurisica, I., Andrulis, I., Tsao, M., and Penn, L. (2006). The c-Myc oncogene directly induces the *H19* noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res* 66, 5330–5337.

Bartolomei, M., Zemel, S., and Tilghman, S. (1991). Parental imprinting of the mouse *H19* gene. *Nature* 351, 153–155.

Bastow, R., Mylne, J., Lister, C., Lippman, Z., Martienssen, R., and Dean, C. (2004). Vernalization requires epigenetic silencing of FLC by histone methylation. *Nature* 427, 164–167.

Bazin, J., Baerenfaller, K., Gosai, S., Gregory, B., Crespi, M., and Bailey-Serres, J. (2017). Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc Natl Acad Sci U S A* 114, E10018–E10027.

Bechtold, U. (2018). Plant life in extreme environments: How do you improve drought tolerance? *Front Plant Sci* 9, 543.

Beitz, E (2000). TEXshade: shading and labeling of multiple sequence alignments using $\LaTeX\,2_\varepsilon$. *Bioinformatics* 16(2), 135–9.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothsis testing. *J R Statist Soc B* 57, 289–300.

Bhan, A. and Mandal, S. (2015). LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. *Biochim Biophys Acta* 1856, 151–164.

Blomberg, S., GarlandT, J., and Ives, A. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57, 717–745.

Blythe, A., Fox, A., and Bond, C. (2016). The ins and outs of lncRNA structure: How, why and what comes next? *Biochim Biophys Acta* 1859, 46–58.

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.

Bolger, M., Arsova, B., and Usadel, B. (2018). Plant genome and transcriptome annotations: from misconceptions to simple solutions. *Brief Bioinform* 19, 437–449.

Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J.-K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* 123(7), 1279–91.

Boyer, J. (1982). Plant productivity and environment. *Science* 218, 443–448.

Boyer, J., Byrne, P, Cassman, K., Cooper, M, Delmer, D, Greene, T, Gruis, F, Habben, J, Hausmann, N, Kenny, N, et al. (2013). The US drought of 2012 in perspective: A call to action. *Global Food Security* 2(3), 139–143.

Brannan, C., Dees, E., Ingram, R., and Tilghman, S. (1990). The product of the *H19* gene may function as an RNA. *Mol Cell Biol* 10, 28–36.

Brown, C., Hendrich, B., Rupert, J., Lafreniere, R., Xing, Y., Lawrence, J., and Willard, H. (1992). The human *XIST* gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527–542.

Brown, G., Wyatt, J., Harris, R., and Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion.*

Buchfink, B., Xie, C., and Huson, D. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12, 59–60.

Burleigh, S. and Harrison, M. (1997). A novel gene whose expression in *Medicago truncatula* roots is suppressed in response to colonization by vesicular-arbuscular mycorrhizal (VAM) fungi and to phosphate nutrition. *Plant Mol Biol* 34, 199–208.

Burleigh, S. and Harrison, M. (1999). The down-regulation of Mt4-like genes by phosphate fertilization occurs systemically and involves phosphate translocation to the shoots. *Plant Physiol* 119, 241–248.

Calin, G., Liu, C., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E., Wojcik, S., et al. (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12, 215–229.

Campalans, A., Kondorosi, A., and Crespi, M. (2004). Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. *Plant Cell* 16, 1047–1059.

Cao, Z., Pan, X., Yang, Y., Huang, Y., and Shen, H. (2018). The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics* 34, 2185–2194.

Cardenas, P., Sonawane, P., Pollier, J., VandenBossche, R., Dewangan, V., Weithorn, E., Tal, L., Meir, S., Rogachev, I., Malitsky, S., et al. (2016). GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nat Commun* 7, 10654.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 491, 454–457.

Castaings, L., Bergonzi, S., Albani, M., Kemi, U., Savolainen, O., and Coupland, G. (2014). Evolutionary conservation of cold-induced antisense RNAs of FLOWERING LOCUS C in *Arabidopsis thaliana* perennial relatives. *Nat Commun* 5, 4457.

Champigny, M., Sung, W., Catana, V., Salwan, R., Summers, P., Dudley, S., Provart, N., Cameron, R., Golding, G., and Weretilnyk, E. (2013). RNA-Seq effectively monitors gene expression in *Eutrema salsugineum* plants growing in an extreme natural habitat and in controlled growth cabinet conditions. *BMC Genomics* 14, 578.

Chekulaeva, M. and Rajewsky, N. (2018). Roles of Long Noncoding RNAs and Circular RNAs in Translation. *Cold Spring Harb Perspect Biol.*

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 41, D983–6.

Cheng, C., Krishnakumar, V., Chan, A., Thibaud-Nissen, F., Schobel, S., and Town, C. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J* 89, 789–804.

Childs, L., Nikoloski, Z., May, P., and Walther, D. (2009). Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Res* 37, e66.

Ching, T., Himmelstein, D., Beaulieu-Jones, B., Kalinin, A., Do, B., Way, G., Ferrero, E., Agapow, P., Zietz, M., Hoffman, M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 15.

Clemson, C., McNeil, J., Willard, H., and Lawrence, J. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol* 132, 259–275.

Coneva, V., Simopoulos, C., Casaretto, J., El-Kereamy, A., Guevara, D., Cohn, J., Zhu, T., Guo, L., Alexander, D., Bi, Y., et al. (2014). Metabolic and co-expression network-based analyses associated with nitrate response in rice. *BMC Genomics* 15, 1056.

Conley, A., Miller, W., and Jordan, I. (2008). Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet* 24, 53–56.

Cooper, N., Thomas, G., Venditti, C., Meade, A., and Freckleton, R. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol J Linn Soc Lond* 118, 64–77.

Costa, M., Farrant, J., Oliver, M., Ligterink, W., Buitink, J., and Hilhorst, H. (2016). Key genes involved in desiccation tolerance and dormancy across life forms. *Plant Sci* 251, 162–168.

Csorba, T., Questa, J., Sun, Q., and Dean, C. (2014). Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proc Natl Acad Sci U S A* 111, 16160–16165.

Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.

Danquah, A., deZelicourt, A., Colcombet, J., and Hirt, H. (2014). The role of ABA and MAPK signaling pathways in plant abiotic stress responses. *Biotechnol Adv* 32, 40–52.

DeChiara, T. and Brosius, J. (1987). Neural BC1 RNA: cDNA clones reveal nonrepetitive sequence content. *Proc Natl Acad Sci U S A* 84, 2624–2628.

Delibašić, B., Radovanović, S., Jovanović, M., Bohanec, M., and Suknović, M. (2018). Integrating knowledge from DEX hierarchies into a logistic regression stacking model for predicting ski injuries. *Journal of Decision Systems*, 1–8.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22, 1775–1789.

Di, C., Yuan, J., Wu, Y., Li, J., Lin, H., Hu, L., Zhang, T., Qi, Y., Gerstein, M., Guo, Y., et al. (2014). Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. *Plant J* 80, 848–861.

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends Genet* 30, 121–123.

Ding, Y., Fromm, M., and Avramova, Z. (2012). Multiple exposures to drought 'train' transcriptional responses in *Arabidopsis*. *Nat Commun* 3, 740.

Ding, Y., Liu, N., Virlouvet, L., Riethoven, J., Fromm, M., and Avramova, Z. (2013). Four distinct types of dehydration stress memory genes in Arabidopsis thaliana. *BMC Plant Biol* 13, 229.

Diniz-Filho, J. (2001). Phylogenetic autocorrelation under distinct evolutionary processes. *Evolution* 55, 1104–1109.

Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5), 1792–7.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138.

Eshel, G., Shaked, R., Kazachkova, Y., Khan, A., Eppel, A., Cisneros, A., Acuna, T., Gutterman, Y., Tel-Zur, N., Rachmilevitch, S., et al. (2016). *Anastatica hierochuntica*, an *Arabidopsis* Desert Relative, Is Tolerant to Multiple Abiotic Stresses and Exhibits Species-Specific and Common Stress Tolerance Strategies with Its Halophytic Relative, *Eutrema (Thellungiella) salsugineum*. *Front Plant Sci* 7, 1992.

Faghihi, M. and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* 10, 637–643.

Fang, M., Zhang, P., Zhao, Y., and Liu, X. (2017). Bioinformatics and co-expression network analysis of differentially expressed lncRNAs and mRNAs in hippocampus of APP/PS1 transgenic mice with Alzheimer disease. *Am J Transl Res* 9, 1381–1391.

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res* 46, D308–D314.

Fang, Y. and Xiong, L. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. *Cell Mol Life Sci* 72, 673–689.

Felsenstein, J. (1993). *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein.

Fiannaca, A., LaRosa, M., LaPaglia, L., Rizzo, R., and Urso, A. (2017). nRC: non-coding RNA Classifier based on structural features. *BioData Min* 10, 27.

Franco-Zorrilla, J., Valli, A., Todesco, M., Mateos, I., Puga, M., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* 39, 1033–1037.

Frank, M. and Scanlon, M. (2015). Transcriptomic evidence for the evolution of shoot meristem function in sporophyte-dominant land plants through concerted selection of ancestral gametophytic and sporophytic genetic programs. *Mol Biol Evol* 32, 355–367.

Gehan, M., Greenham, K., Mockler, T., and McClung, C. (2015). Transcriptional networks-crops, clocks, and abiotic stress. *Curr Opin Plant Biol* 24, 39–46.

German, D. and Koch, M. (2017). *Eutrema salsugineum* (*Cruciferae*) new to Mexico: a surprising generic record for the flora of Middle America. *PhytoKeys*, 13–21.

Gigon, A., Matos, A., Laffray, D., Zuily-Fodil, Y., and Pham-Thi, A. (2004). Effect of drought stress on lipid metabolism in the leaves of *Arabidopsis thaliana* (ecotype Columbia). *Ann Bot* 94, 345–351.

Gittleman, J. L. and Kot, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology* 39(3), 227–241.

Go, Y. S., Kim, H., Kim, H. J., and Suh, M. C. (2014). *Arabidopsis* Cuticular Wax Biosynthesis Is Negatively Regulated by the *DEWAX* Gene Encoding an AP2/ERF-Type Transcription Factor. *Plant Cell* 26, 1666–1680.

Gomes, C., Spencer, H., Ford, K., Michel, L., Baker, A., Emanueli, C., Balligand, J., and Devaux, Y. (2017). The Function and Therapeutic Potential of Long Non-coding RNAs in Cardiovascular Development and Disease. *Mol Ther Nucleic Acids* 8, 494–507.

Gong, Q., Li, P., Ma, S., InduRupassara, S., and Bohnert, H. (2005). Salinity stress adaptation competence in the extremophile *Thellungiella halophila* in comparison with its relative *Arabidopsis thaliana*. *Plant J* 44, 826–839.

Gonzalez-Munoz, E., Avendano-Vazquez, A., Montes, R., deFolter, S., Andres-Hernandez, L., Abreu-Goodger, C., and Sawers, R. (2015). The maize (*Zea mays* ssp. mays var. B73) genome encodes 33 members of the purple acid phosphatase family. *Front Plant Sci* 6, 341.

Goodstein, D., Shu, S., Howson, R., Neupane, R., Hayes, R., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40, D1178–86.

Gramates, L., Marygold, S., Santos, G., Urbano, J., Antonazzo, G., Matthews, B., Rey, A., Tabone, C., Crosby, M., Emmert, D., et al. (2017). FlyBase at 25: looking to the future. *Nucleic Acids Res* 45, D663–D671.

Griffith, M., Timonin, M., Wong, A., Gray, G., Akhter, S., Saldanha, M., Rogers, M., Weretilnyk, E., and Moffatt, B. (2007). *Thellungiella*: an *Arabidopsis*-related model plant adapted to cold temperatures. *Plant Cell Environ* 30, 529–538.

Grondin, A., Rodrigues, O., Verdoucq, L., Merlot, S., Leonhardt, N., and Maurel, C. (2015). Aquaporins Contribute to ABA-Triggered Stomatal Closure through OST1-Mediated Phosphorylation. *Plant Cell* 27, 1945–1954.

Gudenas, B. and Wang, L. (2018). Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features. *Scientic Reports* 8, 16385.

Guevara, D., Champigny, M., Tattersall, A., Dedrick, J., Wong, C., Li, Y., Labbe, A., Ping, C., Wang, Y., Nuin, P., et al. (2012). Transcriptomic and metabolomic analysis of Yukon *Thellungiella* plants grown in cabinets and their natural habitat show phenotypic plasticity. *BMC Plant Biol* 12, 175.

Guttman, M., Russell, P., Ingolia, N., Weissman, J., and Lander, E. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251.

Haerty, W. and Ponting, C. (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* 21, 333–346.

Hangauer, M., Vaughn, I., and McManus, M. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* 9, e1003569.

Hansen, T., Jensen, T., Clausen, B., Bramsen, J., Finsen, B., Damgaard, C., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388.

Hansen, T. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51, 1341–1351.

Harb, A., Krishnan, A., Ambavaram, M., and Pereira, A. (2010). Molecular and physiological analysis of drought stress in Arabidopsis reveals early responses leading to acclimation in plant growth. *Plant Physiol* 154, 1254–1271.

Harmon, L., Weir, J., Brock, C., Glor, R., and Challenger, W (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics* 24, 129–131.

Hawkes, E., Hennelly, S., Novikova, I., Irwin, J., Dean, C., and Sanbonmatsu, K. (2016). COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures. *Cell Rep* 16, 3087–3096.

He, C., Huang, H., and Xu, L. (2013). Mechanisms guiding Polycomb activities during gene silencing in *Arabidopsis thaliana*. *Front Plant Sci* 4, 454.

He, S., Liu, S., and Zhu, H. (2011). The sequence, structure and evolutionary features of *HOTAIR* in mammals. *BMC Evol Biol* 11, 102.

Healey, S. P., Cohen, W. B., Yang, Z., Brewer, C. K., Brooks, E. B., Gorelick, N., Hernandez, A. J., Huang, C., Hughes, M. J., Kennedy, R. E., et al. (2018). Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment* 204, 717–728.

Heo, J. and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* 331, 76–79.

Hezroni, H., Koppstein, D., Schwartz, M., Avrutin, A., Bartel, D., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep* 11, 1110–1122.

Hu, L., Xu, Z., Hu, B., and Lu, Z. (2017). COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res* 45, e2.

Huang, L., Dong, H., Zhou, D., Li, M., Liu, Y., Zhang, F., Feng, Y., Yu, D., Lin, S., and Cao, J. (2018). Systematic identification of long non-coding RNAs during pollen development and fertilization in *Brassica rapa*. *The Plant Journal* 96, 203–222.

Huang, X., Liu, J., and Chen, X. (2010). Overexpression of PtrABF gene, a bZIP transcription factor isolated from *Poncirus trifoliata*, enhances dehydration and drought tolerance in tobacco via scavenging ROS and modulating expression of stress-responsive genes. *BMC Plant Biol* 10, 230.

Hundleby, P. A. and Harwood, W. A. (2018). Impacts of the EU GMO regulatory framework for plant genome editing. *Food and Energy Security*, e00161.

Hüttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? *TRENDS in Genetics* 21, 289–297.

Inan, G., Zhang, Q., Li, P., Wang, Z., Cao, Z., Zhang, H., Zhang, C., Quist, T. M., Goodwin, S. M., Zhu, J., et al. (2004). Salt cress. A halophyte and cryophyte Arabidopsis relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant physiology* 135(3), 1718–1737.

Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A., Khan, O., Brewer, J., Skadow, M., Duizer, C., et al. (2018). The translation of non-canonical open reading frames controls mucosal immunity. *Nature*.

Jain, M., Fiddes, I., Miga, K., Olsen, H., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 12, 351–356.

James, A., Jayasena, A., Zhang, J., Berkowitz, O., Secco, D., Knott, G., Whelan, J., Bond, C., and Mylne, J. (2017). Evidence for Ancient Origins of Bowman-Birk Inhibitors from *Selaginella moellendorffii*. *Plant Cell* 29, 461–473.

Jed Wing, M. K. C. from, Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, the, Benesty, M., et al. (2017). *caret: Classification and Regression Training.*

Jeon, Y. and Lee, J. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* 146, 119–133.

Jeong, H. and Jung, K. (2015). Rice tissue-specific promoters and condition-dependent promoters for effective translational application. *J Integr Plant Biol* 57, 913–924.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* 4, e08890.

Johnson, R. and Guigo, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* 20, 959–976.

Juntawong, P., Girke, T., Bazin, J., and Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in *Arabidopsis*. *Proc Natl Acad Sci USA* 111, E203–12.

Kamilar, J. and Cooper, N. (2013). Phylogenetic signal in primate behaviour, ecology and life history. *Philos Trans R Soc Lond B Biol Sci* 368.

Kang, Y., Han, Y., Torres-Jerez, I., Wang, M., Tang, Y., Monteros, M., and Udvardi, M. (2011). System responses to long-term drought and re-watering of two contrasting alfalfa varieties. *Plant J* 68, 871–889.

Kang, Y., Yang, D., Kong, L., Hou, M., Meng, Y., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

Kapusta, A. and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet* 30, 439–452.

Kapusta, A., Kronenberg, Z., Lynch, V., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 9, e1003470.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30, 3059–3066.

Kazachkova, Y., Batushansky, A., Cisneros, A., Tel-Zur, N., Fait, A., and Barak, S. (2013). Growth platform-dependent and -independent phenotypic and metabolic responses of *Arabidopsis* and its halophytic relative, *Eutrema salsugineum*, to salt stress. *Plant Physiol* 162, 1583–1598.

Kazachkova, Y., Eshel, G., Pantha, P., Cheeseman, J., Dassanayake, M., and Barak, S. (2018). Halophytism: What Have We Learnt From *Arabidopsis thaliana* Relative Model Systems? *Plant Physiol* 178, 972–988.

Keck, F., Rimet, F., Bouchez, A., and Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol* 6, 2774–2780.

Kelley, D., Snoek, J., and Rinn, J. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 26, 990–999.

Khanal, N., Bray, G., Grisnich, A., Moffatt, B., and Gray, G. (2017). Differential Mechanisms of Photosynthetic Acclimation to Light and Low Temperature in Arabidopsis and the Extremophile *Eutrema salsugineum. Plants* 6.

Kiegle, E., Garden, A., Lacchini, E., and Kater, M. (2018). A Genomic View of Alternative Splicing of Long Non-coding RNAs during Rice Seed Development Reveals Extensive Splicing and lncRNA Gene Families. *Front Plant Sci* 9, 115.

Kim, D. and Sung, S. (2017). Vernalization-Triggered Intragenic Chromatin Loop Formation by Long Noncoding RNAs. *Dev Cell* 40, 302–312.e4.

Kim, T., Hemberg, M., and Gray, J. (2015). Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb Perspect Biol* 7, a018622.

Knapp, A., Ciais, P., and Smith, M. (2017). Reconciling inconsistencies in precipitation-productivity relationships: implications for climate change. *New Phytol* 214, 41–47.

Kong, L., Zhang, Y., Ye, Z., Liu, X., Zhao, S., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 35, W345–9.

Krause, J., Perer, A., and Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In: *ACM Conf Hum Factors Comput Syst.* ACM, 5686–5697.

Kung, J., Colognori, D., and Lee, J. (2013). Long noncoding RNAs: past, present, and future. *Genetics* 193, 651–669.

Lamesch, P., Berardini, T., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D., Garcia-Hernandez, M., et al. (2011). The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40, D1202–10.

Lang, D., Ullrich, K., Murat, F., Fuchs, J., Jenkins, J., Haas, F., Piednoel, M., Gundlach, H., VanBel, M., Meyberg, R., et al. (2018). The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J* 93, 515–533.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559.

Latge, G., Poulet, C., Bours, V., Josse, C., and Jerusalem, G. (2018). Natural Antisense Transcripts: Molecular Mechanisms and Implications in Breast Cancers. *Int J Mol Sci* 19.

Lauressergues, D., Couzigou, J., Clemente, H., Martinez, Y., Dunand, C., Becard, G., and Combier, J. (2015). Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520, 90–93.

Lee, B., Baek, J., Park, S., and Yoon, S. (2016a). deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks, 434–442.

Lee, S., Kang, J., Park, H., Kim, M., Bae, M., Choi, H., and Kim, S. (2010). DREB2C interacts with ABF2, a bZIP protein regulating abscisic acid-responsive gene expression, and its overexpression affects abscisic acid sensitivity. *Plant Physiol* 153, 716–727.

Lee, Y. P., Funk, C., Erban, A., Kopka, J., Köhl, K. I., Zuther, E., and Hincha, D. K. (2016b). Salt stress responses in a geographically diverse collection of *Eutrema/Thellungiella* spp. accessions. *Functional Plant Biology* 43(7), 590–606.

LeProvost, G., Domergue, F., Lalanne, C., RamosCampos, P., Grosbois, A., Bert, D., Meredieu, C., Danjon, F., Plomion, C., and Gion, J. (2013). Soil water stress affects both cuticular wax content and cuticle-related gene expression in young saplings of maritime pine (*Pinus pinaster Ait*). *BMC Plant Biol* 13, 95.

Lesk, C., Rowhani, P., and Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature* 529, 84–87.

Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* 15, 311.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1), 323.

Li, D., Brown, J., Orsini, L., Pan, Z., Hu, G., and He, S. (2016a). MODA: MOdule Differential Analysis for weighted gene co-expression network. *bioRxiv*.

Li, Y., Wang, Z., Wang, Y., Zhao, Z., Zhang, J., Lu, J., Xu, J., and Li, X. (2016b). Identification and characterization of lncRNA mediated transcriptional dysregulation dictates lncRNA roles in glioblastoma. *Oncotarget* 7, 45027–45041.

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D., Zhao, H., et al. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Research* 39, 3864–3878.

Liu, B., Wang, S., Long, R., and Chou, K. (2017a). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 33, 35–41.

Liu, C., Muchhal, U., and Raghothama, K. (1997). Differential expression of *TPSI1*, a phosphate starvation-induced gene in tomato. *Plant Mol Biol* 33, 867–874.

Liu, L., Lu, Y., Wei, L., Yu, H., Cao, Y., Li, Y., Yang, N., Song, Y., Liang, C., and Wang, T. (2018a). Transcriptomics analyses reveal the molecular roadmap and long non-coding RNA landscape of sperm cell lineage development. *The Plant Journal* 96, 421–437.

Liu, S., Horlbeck, M., Cho, S., Birk, H., Malatesta, M., He, D., Attenello, F., Villalta, J., Cho, M., Chen, Y., et al. (2017b). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* 355.

Liu, W., Cheng, C., Lin, Y., XuHan, X., and Lai, Z. (2018b). Genome-wide identification and characterization of mRNAs and lncRNAs involved in cold stress in the wild banana (*Musa itinerans*). *PLoS One* 13, e0200002.

Lobell, D., Roberts, M., Schlenker, W., Braun, N., Little, B., Rejesus, R., and Hammer, G. (2014). Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. *Science* 344, 516–519.

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550.

Lovell, J., Mullen, J., Lowry, D., Awole, K., Richards, J., Sen, S., Verslues, P., Juenger, T., and McKay, J. (2015). Exploiting Differential Gene Expression and Epistasis to Discover Candidate Genes for Drought-Associated QTLs in *Arabidopsis thaliana*. *Plant Cell* 27, 969–983.

Lu, Y., Wu, K., Jiang, Y., Guo, Y., and Desneux, N. (2012). Widespread adoption of Bt cotton and insecticide decrease promotes biocontrol services. *Nature* 487, 362–365.

Ma, L., Bajic, V., and Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biol* 10, 925–933.

Ma, L., Li, A., Zou, D., Xu, X., Xia, L., Yu, J., Bajic, V., and Zhang, Z. (2015). LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res* 43, D187–92.

MacLeod, M. J., Dedrick, J., Ashton, C., Sung, W. W., Champigny, M. J., and Weretilnyk, E. A. (2014). Exposure of two *Eutrema salsugineum* (*Thellungiella salsuginea*) accessions to water deficits reveals different coping strategies in response to drought. *Physiologia plantarum* 155(3), 267–280.

Merchant, S., Prochnik, S., Vallon, O., Harris, E., Karpowicz, S., Witman, G., Terry, A., Salamov, A., Fritz-Laylin, L., Marechal-Drouard, L., et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318, 245–250.

Meyer, E., Aspinwall, M., Lowry, D., Palacio-Mejia, J., Logan, T., Fay, P., and Juenger, T. (2014). Integrating transcriptional, metabolomic, and physiological responses to drought stress and recovery in switchgrass (*Panicum virgatum L.*) *BMC Genomics* 15, 527.

Millar, A. and Waterhouse, P. (2005). Plant and animal microRNAs: similarities and differences. *Funct Integr Genomics* 5, 129–135.

Milligan, M. and Lipovich, L. (2014). Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front Genet* 5, 476.

Mitra, J. (2001). Genetics and genetic improvement of drought resistance in crop plants. *Current Science*, 758–763.

Mittler, R. and Blumwald, E. (2010). Genetic engineering for modern agriculture: challenges and perspectives. *Annu Rev Plant Biol* 61, 443–462.

Mohammadin, S., Edger, P., Pires, J., and Schranz, M. (2015). Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the *Brassicaceae* and *Cleomaceae*. *BMC Plant Biol* 15, 217.

Monfort, A., DiMinin, G., Postlmayr, A., Freimann, R., Arieti, F., Thore, S., and Wutz, A. (2015). Identification of Spen as a Crucial Factor for *Xist* Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. *Cell Rep* 12, 554–561.

Montange, R. and Batey, R. (2008). Riboswitches: emerging themes in RNA structure and function. *Annu Rev Biophys* 37, 117–133.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10(2), 243–251.

Muniz-Garcia, M., Giammaria, V., Grandellis, C., Tellez-Inon, M., Ulloa, R., and Capiati, D. (2012). Characterization of StABF1, a stress-responsive bZIP transcription factor from *Solanum tuberosum L.* that is phosphorylated by StCDPK2 in vitro. *Planta* 235, 761–778.

Munschauer, M., Nguyen, C., Sirokman, K., Hartigan, C., Hogstrom, L., Engreitz, J., Ulirsch, J., Fulco, C., Subramanian, V., Chen, J., et al. (2018). The *NORAD* lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* 561, 132–136.

Nakashima, K., Tran, L., VanNguyen, D., Fujita, M., Maruyama, K., Todaka, D., Ito, Y., Hayashi, N., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2007). Functional analysis of a NAC-type transcription factor OsNAC6 involved in abiotic and biotic stress-responsive gene expression in rice. *Plant J* 51, 617–630.

Narsai, R., Castleden, I., and Whelan, J. (2010). Common and distinct organ and stress responsive transcriptomic patterns in *Oryza sativa* and *Arabidopsis thaliana*. *BMC Plant Biol* 10, 262.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640.

Nejat, N. and Mantri, N. (2018). Emerging roles of long non-coding RNAs in plant response to biotic and abiotic stresses. *Crit Rev Biotechnol* 38, 93–105.

Nelson, A., Forsythe, E., Devisetty, U., Clausen, D., Haug-Batzell, A., Meldrum, A., Frank, M., Lyons, E., and Beilstein, M. (2016). A Genomic Analysis of Factors Driving lincRNA Diversification: Lessons from Plants. *G3* 6, 2881–2891.

Niazi, F. and Valadkhan, S. (2012). Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA* 18, 825–843.

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L., et al. (2016). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res* 44, D980–5.

Nishizawa, A., Yabuta, Y., and Shigeoka, S. (2008). Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiol* 147, 1251–1263.

Nuccio, M., Paul, M., Bate, N., Cohn, J., and Cutler, S. (2018). Where are the drought tolerant crops? An assessment of more than two decades of plant biotechnology effort in crop improvement. *Plant Sci* 273, 110–119.

Nudler, E. and Mironov, A. (2004). The riboswitch control of bacterial metabolism. *Trends Biochem Sci* 29, 11–17.

Ogburn, R. M. and Edwards, E. J. (2010). The ecological water-use strategies of succulent plants. In: *Advances in Botanical Research*. Ed. by J.-C. Kader and M. Delseny. Vol. 55. Elsevier, 179–225.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573.

O'Leary, N., Wright, M., Brister, J., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44, D733–45.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R., Lee, Y., Zheng, L., et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* 35, D883–7.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* 401, 877–884.

Palazzo, A. and Lee, E. (2015). Non-coding RNA: what is functional and what is junk? *Front Genet* 6, 2.

Panchbhai, A., Char, B., and Kharat, A. S. (2017). The ALDH7 promoter of *Acacia nilotica L.* is a moisture stress inducible promoter. *Plant Gene* 10, 1–7.

Panchy, N., Wu, G., Newton, L., Tsai, C., Chen, J., Benning, C., Farre, E., and Shiu, S. (2014). Prevalence, evolution, and cis-regulation of diel transcription in *Chlamydomonas reinhardtii. G3 (Bethesda)* 4, 2461–2471.

Pang, K., Frith, M., and Mattick, J. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* 22, 1–5.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics.*

Parna, S., Verma, S., Pathak, R. U., and Mishra, R. K. (2017). Long Noncoding RNAs in Mammalian Development and Diseases. In: *Long Non Coding RNA Biology.* Ed. by M. R. S. Rao. Springer Nature.

Paytuvi-Gallart, A., Hermoso-Pulido, A., Lagran, I. Anzar-Martinez de, Sanseverino, W., and Aiese-Cigliano, R. (2016). GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Res* 44, D1161–6.

Pearson, W. (2013). An introduction to sequence similarity (“homology”) searching. *Curr Protoc Bioinformatics* Chapter 3, Unit3.1.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

Pellegrina, D., Severino, P., Barbeiro, H., deSouza, H., Machado, M., Silva, F. Pinheiro-da, and Reis, E. (2017). Insights into the Function of Long Noncoding RNAs in Sepsis Revealed by Gene Co-Expression Network Analysis. *Noncoding RNA* 3.

Pertea, M., Pertea, G., Antonescu, C., Chang, T., Mendell, J., and Salzberg, S. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33, 290–295.

Pinheiu, C., Dickinson, E., Marriott, A., Ribeiro, I. C., Pintó-Marijuan, M., António, C., Zarrouk, O., Chaves, M. M., Dodd, I. C., Munné-Bosch, S., et al. (2019). Distinctive phytohormonal and metabolic profiles of *Arabidopsis thaliana* and *Eutrema salsugineum* under similar soil drying. *Planta.*

Pintacuda, G., Young, A., and Cerase, A. (2017). Function by Structure: Spotlights on *Xist* Long Non-coding RNA. *Front Mol Biosci* 4, 90.

Polikar, R. (2012). Ensemble Learning. In: *Ensemble Machine Learning: Methods and Applications.* Ed. by C. Zhang and Y. Ma. Boston, MA: Springer US, 1–34.

Pollard, K., Salama, S., Lambert, N., Lambot, M., Coppens, S., Pedersen, J., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167–172.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research* 17(5), 000–000.

Prensner, J., Chen, W., Iyer, M., Cao, Q., Ma, T., Han, S., Sahu, A., Malik, R., Wilder-Romans, K., Navone, N., et al. (2014). *PCAT-1*, a long noncoding RNA, regulates *BRCA2* and controls homologous recombination in cancer. *Cancer Res* 74, 1651–1660.

Provart, N., Alonso, J., Assmann, S., Bergmann, D., Brady, S., Brkljacic, J., Browse, J., Chapple, C., Colot, V., Cutler, S., et al. (2016). 50 years of *Arabidopsis* research: highlights and future directions. *New Phytol* 209, 921–944.

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., Ostrovskiy, A., Cantor, C., Vijg, J., and Zhavoronkov, A. (2016). Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging* 8, 1021–1033.

Qin, T., Zhao, H., Cui, P., Albesher, N., and Xiong, L. (2017). A nucleus-localized long non-coding RNA enhances drought and salt stress tolerance. *Plant Physiol* 175, 1321–1336.

Quan, G. and Li, J. (2018). Circular RNAs: biogenesis, expression and their potential roles in reproduction. *J Ovarian Res* 11, 9.

Quek, X., Thomson, D., Maag, J., Bartonicek, N., Signal, B., Clark, M., Gloss, B., and Dinger, M. (2015). lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* 43, D168–73.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundaion for Statistical Computing. Vienna, Austria.

Ransbotyn, V., Yeger-Lotem, E., Basha, O., Acuna, T., Verduyn, C., Gordon, M., Chalifa-Caspi, V., Hannah, M., and Barak, S. (2015). A combination of gene expression ranking and co-expression network analysis increases discovery rate in large-scale mutant screens for novel *Arabidopsis thaliana* abiotic stress genes. *Plant Biotechnol J* 13, 501–513.

Rasheed, S., Bashir, K., Matsui, A., Tanaka, M., and Seki, M. (2016). Transcriptomic Analysis of Soil-Grown *Arabidopsis thaliana* Roots and Shoots in Response to a Drought Stress. *Front Plant Sci* 7, 180.

Renganathan, A and Felley-Bosco, E. (2017). Long noncoding RNAs in Cancer and Therapeutic Potential. In: *Long Non Coding RNA Biology.* Ed. by M. R. S. Rao. Springer Nature.

Revell, L., Harmon, L., and Collar, D. (2008). Phylogenetic signal, evolutionary process, and rate. *Syst Biol* 57, 591–601.

Rivas, E., Clements, J., and Eddy, S. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods* 14, 45–48.

Rocha, S. T. da and Heard, E. (2017). Novel players in X inactivation: insights into *Xist*-mediated gene silencing and chromosome conformation. *Nat Struct Mol Biol* 24, 197–204.

Rohrig, H., Schmidt, J., Miklashevichs, E., Schell, J., and John, M. (2002). Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc Natl Acad Sci U S A* 99, 1915–1920.

Romito, A. and Rougeulle, C. (2011). Origin and evolution of the long non-coding genes in the X-inactivation center. *Biochimie* 93, 1935–1942.

Ronald, P. (2014). Lab to farm: applying research on plant genetics and genomics to crop improvement. *PLoS Biol* 12, e1001878.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6), 386.

Saini, H., Enright, A., and Griffiths-Jones, S. (2008). Annotation of mammalian primary microRNAs. *BMC Genomics* 9, 564.

Schlaepfer, D., Bradford, J., Lauenroth, W., Munson, S., Tietjen, B., Hall, S., Wilson, S., Duniway, M., Jia, G., Pyke, D., et al. (2017). Climate change reduces extent of temperate drylands and intensifies drought in deep soils. *Nat Commun* 8, 14196.

Schnable, P., Ware, D., Fulton, R., Stein, J., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115.

Schorderet, P. and Duboule, D. (2011). Structural and functional differences in the long noncoding RNA hotair in mouse and human. *PLoS Genet* 7, e1002071.

Sharma, S., Bolser, D., deBoer, J., Sonderkaer, M., Amoros, W., Carboni, M., D'Ambrosio, J., delaCruz, G., DiGenova, A., Douches, D., et al. (2013). Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3* 3, 2031–2047.

Shelton, A., Zhao, J., and Roush, R. (2002). Economic, ecological, food safety, and social consequences of the deployment of Bt transgenic plants. *Annu Rev Entomol* 47, 845–881.

Shin, J. and Chekanova, J. (2014). Arabidopsis RRP6L1 and RRP6L2 function in FLOWERING LOCUS C silencing via regulation of antisense RNA synthesis. *PLoS Genet* 10, e1004612.

Shuai, P., Liang, D., Tang, S., Zhang, Z., Ye, C., Su, Y., Xia, X., and Yin, W. (2014). Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in *Populus trichocarpa*. *J Exp Bot* 65, 4975–4983.

Simopoulos, C., Weretilnyk, E., and Golding, G. (2018). Prediction of plant lncRNA by ensemble machine learning classifiers. *BMC Genomics* 19, 316.

Singh, U., Khemka, N., Rajkumar, M., Garg, R., and Jain, M. (2017). PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Res* 45, e183.

Smit, A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. *http://www.repeatmasker.org*.

Somarowthu, S., Legiewicz, M., Chillon, I., Marcia, M., Liu, F., and Pyle, A. (2015). HOTAIR forms an intricate and modular secondary structure. *Mol Cell* 58, 353–361.

Sork, V. L. (2018). Genomic Studies of Local Adaptation in Natural Plant Populations. *Journal of Heredity* 109(1), 3–15.

Sprenger, H., Kurowsky, C., Horn, R., Erban, A., Seddig, S., Rudack, K., Fischer, A., Walther, D., Zuther, E., Kohl, K., et al. (2016). The drought response of potato reference cultivars with contrasting tolerance. *Plant Cell Environ* 39, 2370–2389.

St. Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The Landscape of long noncoding RNA classification. *Trends in Genetics* 31, 239–251.

Stephens, Z., Lee, S., Faghri, F., Campbell, R., Zhai, C., Efron, M., Iyer, R., Schatz, M., Sinha, S., and Robinson, G. (2015). Big Data: Astronomical or Genomical? *PLoS Biol* 13, e1002195.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research* 10(9), 2997–3011.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* 14, 103–105.

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., and Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res* 41, e166.

Sun, L., Liu, H., Zhang, L., and Meng, J. (2015). lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine. *PLoS One* 10, e0139654.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* 6(7), e21800.

Swiezewski, S., Liu, F., Magusin, A., and Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* 462, 799–802.

Szczesniak, M., Rosikiewicz, W., and Makalowska, I. (2016). CANTATAdb: A Collection of Plant Long Non-Coding RNAs. *Plant Cell Physiol* 57, e8.

Taji, T., Seki, M., Satou, M., Sakurai, T., Kobayashi, M., Ishiyama, K., Narusaka, Y., Narusaka, M., Zhu, J., and Shinozaki, K. (2004). Comparative genomics in salt tolerance between *Arabidopsis* and *Arabidopsis*-related halophyte salt cress using *Arabidopsis* microarray. *Plant Physiol* 135, 1697–1709.

Tavares, R. C., Pyle, A. M., and Somarowthu, S. (2018). Covariation analysis with improved parameters reveals conservation in lncRNA structures. *bioRxiv*, 364109.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

The Ministry of Agriculture and Agri-Food Canada (2017). *2017 Annual Report of Agroclimate Conditions Across Canada.*

Tichon, A., Gil, N., Lubelsky, Y., HavkinSolomon, T., Lemze, D., Itzkovitz, S., Stern-Ginossar, N., and Ulitsky, I. (2016). A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nat Commun* 7, 12209.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., vanBaren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28, 511–515.

Tucker, B. and Breaker, R. (2005). Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* 15, 342–348.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* 17, 601–614.

Umezawa, T., Sugiyama, N., Takahashi, F., Anderson, J., Ishihama, Y., Peck, S., and Shinozaki, K. (2013). Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in *Arabidopsis thaliana*. *Sci Signal* 6, rs8.

Van Peer, G., De Paepe, A., Stock, M., Anckaert, J., Volders, P.-J., Vandesompele, J., De Baets, B., and Waegeman, W. (2017). miSTAR: miRNA target prediction through modeling quantitative and qualitative miRNA binding site information in a stacked model structure. *Nucleic Acids Research* 45, e51–e51.

VanderAuwera, G., Carneiro, M., Hartl, C., Poplin, R., DelAngel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43, 11.10.1–33.

Velasco, V., Mansbridge, J., Bremner, S., Carruthers, K., Summers, P., Sung, W., Champigny, M., and Weretilnyk, E. (2016). Acclimation of the crucifer *Eutrema salsugineum* to phosphate limitation is associated with constitutively high expression of phosphate-starvation genes. *Plant Cell Environ* 39, 1818–1834.

Volders, P., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J., and Mestdagh, P. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* 41, D246–51.

Wan, C. and Wilkins, T. (1994). A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum L.*) *Anal Biochem* 223, 7–12.

Wang, A., Wang, Z., Li, Z., and Li, L. (2018a). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics* 34, 2019–2028.

Wang, D., Qu, Z., Yang, L., Zhang, Q., Liu, Z., Do, T., Adelson, D., Wang, Z., Searle, I., and Zhu, J. (2017). Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. *Plant J* 90, 133–146.

Wang, H., Iacoangeli, A., Lin, D., Williams, K., Denman, R., Hellen, C., and Tiedge, H. (2005). Dendritic BC1 RNA in translational control mechanisms. *J Cell Biol* 171, 811–821.

Wang, H. V. and Cheksnova, J. A. (2017). Long Noncoding RNAs in Plants. In: *Long Non Coding RNA Biology.* Ed. by M. R. S. Rao. Springer Nature.

Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F., and Fan, Q. (2010). CREB up-regulates long non-coding RNA, *HULC* expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* 38, 5366–5383.

Wang, L., Park, H., Dasari, S., Wang, S., Kocher, J., and Li, W. (2013a). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 41, e74.

Wang, L., Felts, S., VanKeulen, V., Pease, L., and Zhang, Y. (2018b). Exploring the effect of library preparation on RNA sequencing experiments. *Genomics.*

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* 6, 18962.

Wang, W. (2008). Some fundimental issues in ensemble methods, 2244–2251.

Wang, X., Vignjevic, M., Jiang, D., Jacobsen, S., and Wollenweber, B. (2014a). Improved tolerance to drought stress after anthesis due to priming before anthesis in wheat (*Triticum aestivum L.*) var. Vinjett. *J Exp Bot* 65, 6441–6456.

Wang, X., Shi, D., Wang, X., Wang, J., Sun, Y., and Liu, J. (2015). Evolutionary Migration of the Disjunct Salt Cress *Eutrema salsugineum* (= *Thellungiella salsuginea*, Brassicaceae) between Asia and North America. *PLoS One* 10, e0124010.

Wang, X., Hu, Q., Guo, X., Wang, K., Ru, D., German, D., Weretilnyk, E., Abbott, R., Lascoux, M., and Liu, J. (2018c). Demographic expansion and genetic load of the halophyte model plant *Eutrema salsugineum. Mol Ecol* 27, 2943–2955.

Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., Hu, Y., Xu, J., Yi, L., Yang, J., et al. (2013b). Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis* 4, e765.

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., and Deng, X. (2014b). *Arabidopsis* noncoding RNA mediates control of photomorphogenesis by red light. *PNAS* 111, 10359–10364.

Wang, Y., Luo, X., Sun, F., Hu, J., Zha, X., Su, W., and Yang, J. (2018d). Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. *Nat Commun* 9, 3516.

Wang, Z., Tian, X., Zhao, Q., Liu, Z., Li, X., Ren, Y., Tang, J., Fang, J., Xu, Q., and Bu, Q. (2018e). The E3 Ligase DROUGHT HYPERSENSITIVE Negatively Regulates Cuticular Wax Biosynthesis by Promoting the Degradation of Transcription Factor ROC4 in Rice. *Plant Cell* 30, 228–244.

Wierzbicki, A., Haag, J., and Pikaard, C. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* 135, 635–648.

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M., Pham, G., Nicotra, A., Gregorio, G., Jagadish, S., Septiningsih, E., Bonneau, R., et al. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. *Plant Cell* 28, 2365–2384.

Wong, C., Li, Y., Whitty, B., Diaz-Camino, C., Akhter, S., Brandle, J., Golding, G., Weretilnyk, E., Moffatt, B., and Griffith, M. (2005). Expressed sequence tags from the Yukon ecotype of *Thellungiella* reveal that gene expression in response to cold, drought and salinity shows little overlap. *Plant Mol Biol* 58, 561–574.

Wong, C., Li, Y., Labbe, A., Guevara, D., Nuin, P., Whitty, B., Diaz, C., Golding, G., Gray, G., Weretilnyk, E., et al. (2006). Transcriptional profiling implicates novel interactions between abiotic stress and hormonal responses in *Thellungiella*, a close relative of *Arabidopsis. Plant Physiol* 140, 1437–1450.

Woo, C. and Kingston, R. (2007). *HOTAIR* lifts noncoding RNAs to new levels. *Cell* 129, 1257–1259.

Woo, H., Koo, H., Kim, J., Jeong, H., Yang, J., Lee, I., Jun, J., Choi, S., Park, S., Kang, B., et al. (2016). Programming of Plant Leaf Senescence with Temporal and Inter-Organellar Coordination of Transcriptome in *Arabidopsis*. *Plant Physiol* 171, 452–467.

Wu, L., Murat, P., Matak-Vinkovic, D., Murrell, A., and Balasubramanian, S. (2013). Binding interactions between long noncoding RNA *HOTAIR* and PRC2 proteins. *Biochemistry* 52, 9519–9527.

Xiang, J., Yin, Q., Chen, T., Zhang, Y., Zhang, X., Wu, Z., Zhang, S., Wang, H., Ge, J., Lu, X., et al. (2014). Human colorectal cancer-specific *CCAT1-L* lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res* 24, 513–531.

Xiao, C., Chen, Y., Xie, S., Chen, K., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* 14, 1072–1074.

Xiao, L., Yang, G., Zhang, L., Yang, X., Zhao, S., Ji, Z., Zhou, Q., Hu, M., Wang, Y., Chen, M., et al. (2015). The resurrection genome of *Boea hygrometrica*: A blueprint for survival of dehydration. *PNAS* 112, 5833–5837.

Xiong, L. and Zhu, J. (2003). Regulation of abscisic acid biosynthesis. *Plant Physiol* 133, 29–36.

Xu, J., Wang, Q., Freeling, M., Zhang, X., Xu, Y., Mao, Y., Tang, X., Wu, F., Lan, H., Cao, M., et al. (2017a). Natural antisense transcripts are significantly involved in regulation of drought stress in maize. *Nucleic Acids Res* 45(9), 5126–5141.

Xu, Q., Song, Z., Zhu, C., Tao, C., Kang, L., Liu, W., He, F., Yan, J., and Sang, T. (2017b). Systematic comparison of lncRNAs with protein coding mRNAs in population expression and their response to environmental change. *BMC Plant Biol* 17, 42.

Xu, X., Feng, J., Lu, S., Lohrey, G., An, H., Zhou, Y., and Jenks, M. (2014). Leaf cuticular lipids on the Shandong and Yukon ecotypes of saltwater cress, Eutrema salsugineum, and their response to water deficiency and impact on cuticle permeability. *Physiol Plant* 151, 446–458.

Xue, D., Zhang, X., Lu, X., Chen, G., and Chen, Z.-H. (2017). Molecular and Evolutionary Mechanisms of Cuticular Wax for Plant Drought Tolerance. *Front Plant Sci* 8, 621.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (1994). A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *The Plant Cell* 6(2), 251–264.

Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M., and Zhu, H. (2018). LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics* 34, 3825–3834.

Yang, R., Jarvis, D., Chen, H., Beilstein, M., Grimwood, J., Jenkins, J., Shu, S., Prochnik, S., Xin, M., Ma, C., et al. (2013). The Reference Genome of the Halophytic Plant *Eutrema salsugineum*. *Front Plant Sci* 4, 46.

Yang, S., Vanderbeld, B., Wan, J., and Huang, Y. (2010). Narrowing down the targets: towards successful genetic engineering of drought-tolerant crops. *Mol Plant* 3, 469–490.

Yeo, M., Carella, P, Fletcher, J, Champigny, M., Weretilnyk, E., and Cameron, R. (2015). Development of a *Pseudomonas syringae-Eutrema salsugineum* pathosystem to investigate disease resistance in a stress tolerant extremophile model plant. *Plant Pathology* 64(2), 297–306.

Yi, X., Zhang, Z., Ling, Y., Xu, W., and Su, Z. (2015). PNRD: a plant non-coding RNA database. *Nucleic Acids Res* 43, D982–9.

Yin, J., Gosney, M. J., Dilkes, B. P., and Mickelbart, M. V. (2018). Dark period transcriptomic and metabolic profiling of two diverse *Eutrema salsugineum* accessions. *Plant Direct* 2(2), e00032.

Yin, Y., Zhao, Y., Wang, J., Liu, C., Chen, S., Chen, R., and Zhao, H. (2007). antiCODE: a natural sense-antisense transcripts database. *BMC Bioinformatics* 8, 319.

You, Z., Lei, Y., Zhu, L., Xia, J., and Wang, B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics* 14 Suppl 8, S10.

Yu, H., Lindsay, J., Feng, Z., Frankenberg, S., Hu, Y., Carone, D., Shaw, G., Pask, A., O'Neill, R., Papenfuss, A., et al. (2012). Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC Genomics* 13, 251.

Zerbino, D., Achuthan, P., Akanni, W., Amode, M., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C., et al. (2018). Ensembl 2018. *Nucleic Acids Res* 46, D754–D761.

Zhang, C., Tang, G., Peng, X., Sun, F., Liu, S., and Xi, Y. (2018). Long non-coding RNAs of switchgrass (Panicum virgatum L.) in multiple dehydration stresses. *BMC Plant Biol* 18, 79.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution* 18(6), 292–298.

Zhang, W., Han, Z., Guo, Q., Liu, Y., Zheng, Y., Wu, F., and Jin, W. (2014). Identification of maize long non-coding RNAs responsive to drought stress. *PLoS One* 9, e98958.

Zhao, J., Sun, B., Erwin, J., Song, J., and Lee, J. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322, 750–756.

Zhao, M. and Running, S. (2010). Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. *Science* 329, 940–943.

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and vonSchack, D. (2018a). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep* 8, 4781.

Zhao, X., Li, J., Lian, B., Gu, H., Li, Y., and Qi, Y. (2018b). Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. *Nat Commun* 9, 5056.

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M., et al. (2016). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* 44, D203–8.

Zhu, L., Guo, J., Zhu, J., and Zhou, C. (2014). Enhanced expression of EsWAX1 improves drought tolerance with increased accumulation of cuticular wax and ascorbic acid in transgenic Arabidopsis. *Plant Physiol Biochem* 75, 24–35.

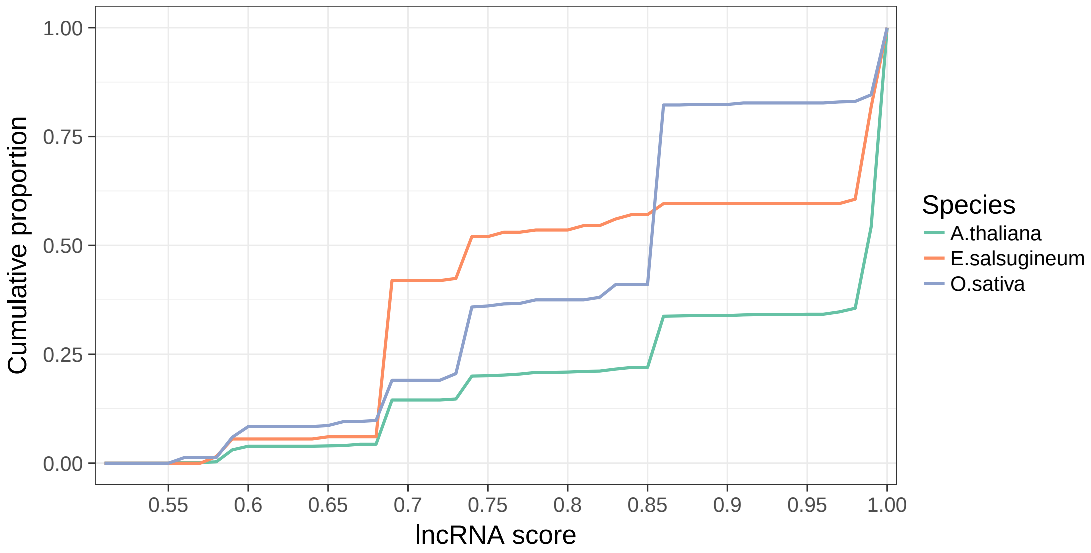
Zhu, Y., Wang, B., Phillips, J., Zhang, Z., Du, H., Xu, T., Huang, L., Zhang, X., Xu, G., Li, W., et al. (2015). Global Transcriptome Analysis Reveals Acclimation-Primed Processes Involved in the Acquisition of Desiccation Tolerance in *Boea hygrometrica*. *Plant Cell Physiol* 56, 1429–1441.

Zinad, H., Natasya, I., and Werner, A. (2017). Natural Antisense Transcripts at the Interface between Host Genome and Mobile Genetic Elements. *Front Microbiol* 8, 2292.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nat Genet* 51, 12–18.

Appendix A

Chapter 2 Supplementary Files



Supplemental Figure S1.1: **Cumulative proportions of lncRNA scores in *A. thaliana*, *E. salsuginiem*, and *O. sativa* found using the gradient boosting stacking generalizer**. The figure depicts the proportions of lncRNAs that are predicted as a lncRNA equal to or less than a particular score.

Supplemental Table S1.1: The distribution of lncRNA prediction scores in *A. thaliana E. salsugineum*, and *O .sativa*

Score	*A. thaliana*	*E. salsugineum*	*O. sativa*
0.50-0.55	0	0	0
0.55-0.60	51	11	72
0.60-0.65	1	1	2
0.65-0.70	138	71	89
0.70-0.75	73	20	146
0.75-0.80	11	3	12
0.80-0.85	14	7	30
0.85-0.90	156	5	354
0.90+	866	80	151
Total	1310	80	148

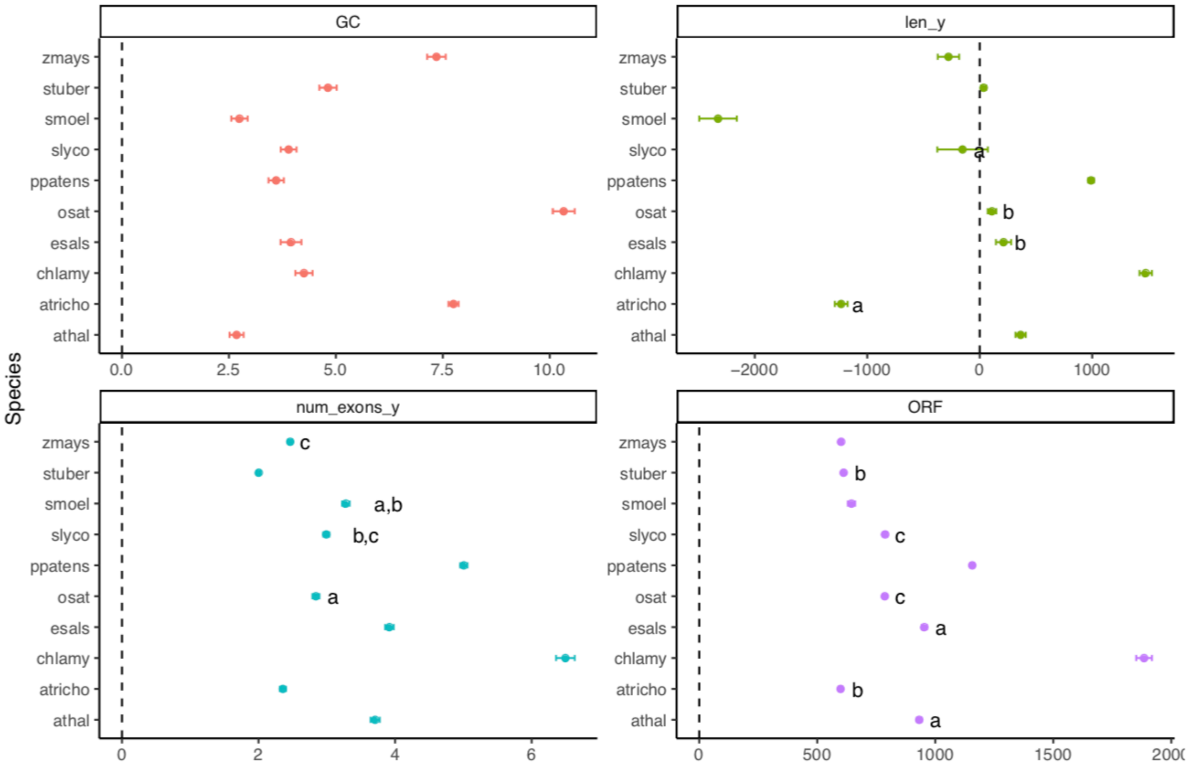
Supplemental Table S1.2: Ensemble predictor has no preference for coding or noncoding sequences.

	CPAT prediction:			
	Coding		Noncoding	
Stacking prediction	lncRNA positive	lncRNA negative	lncRNA positive	lncRNA negative
Species				
A. thaliana	243	37279	1067	2314
O. sativa	133	49325	723	1762
E. salsugineum	44	28079	154	1161

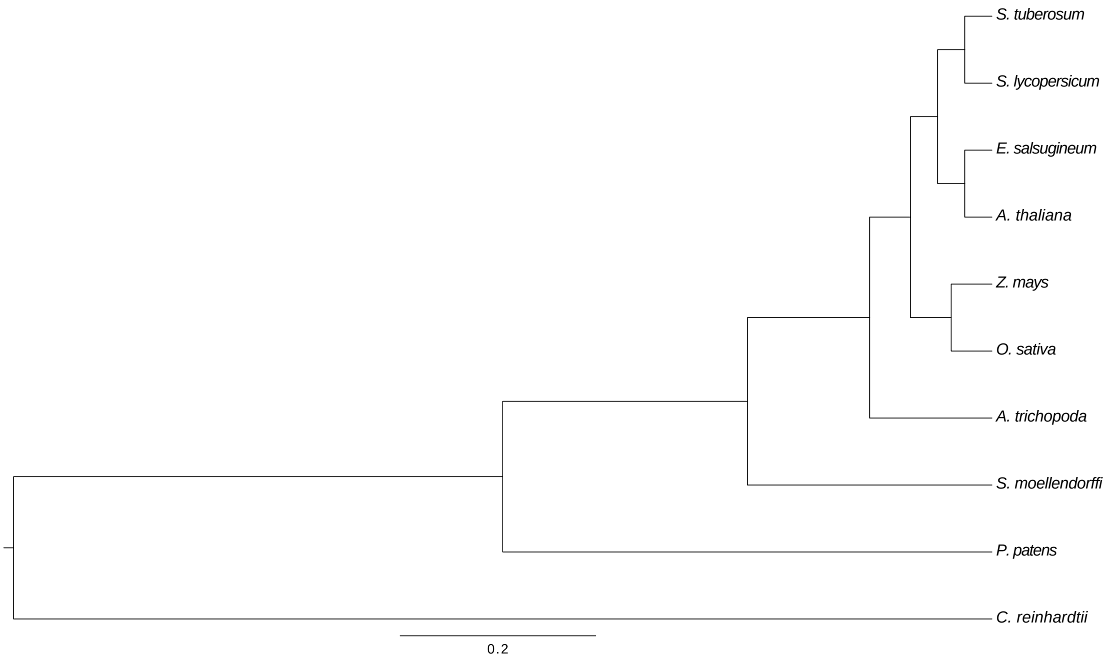
Transcripts were identified as either protein coding or non-coding using the CPAT software. Coding probability cutoffs were calculated by intersect of sensitivity and specificity via 10-fold cross validation. A coding probability cut off of 0.38 was used for *A. thaliana* and 0.52 for both *O. sativa* and *E. salsugineum*.

Appendix B

Chapter 3 Supplementary Files



Supplemental Figure S2.1: Post-hoc pairwise t-test results of uncorrected mean comparisons. The trait value differences between all other transcripts and lncRNAs are plotted. Pairs without significant mean differences are indicated with corresponding letters.



Supplemental Figure S2.2: Phylogenetic tree visualizing the calculated branch lengths used in phylogenetic signal detection. Branch lengths were estimated from a MAFFT v7.205 alignment of *rps16, atp2, 18s, 26s* and *SMC1* (FASTA file of sequences available in File S3) using the dnaml program in PHYLIP. The tree topology reported by the *Amborella* Genome Project (2013) was used. Branch lengths representing site changes were converted to relative age of branches using the R package ape.

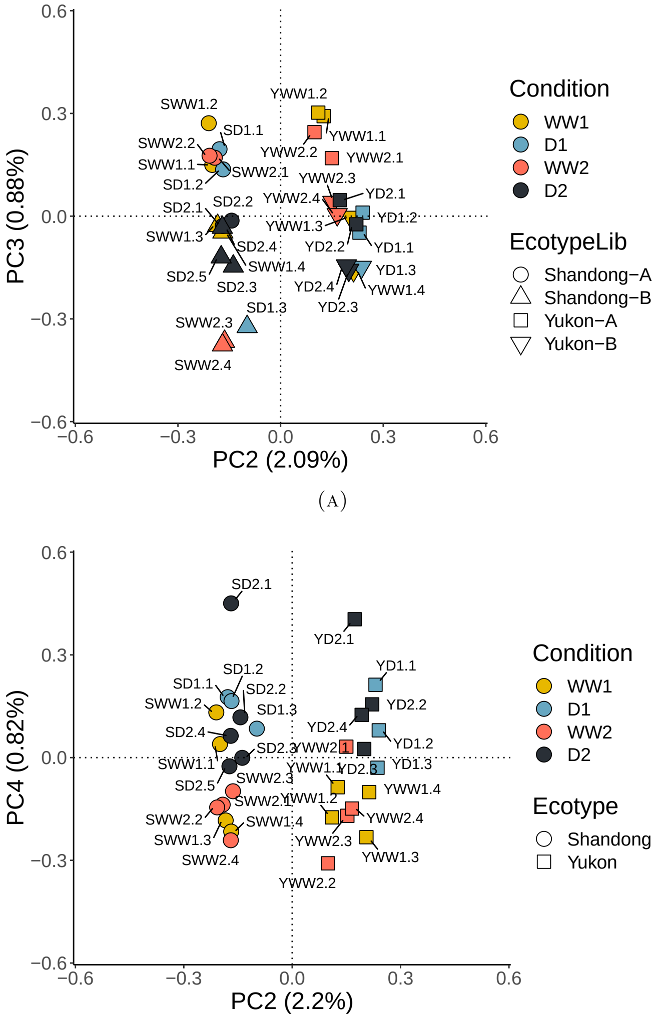
Appendix C

Chapter 4 Supplementary Files

Supplemental Table S3.1: Correlation of select cluster eigengenes to genotype and drought treatment

Cluster	Number of genes			Correlation to Condition							
	DEGs		Total	SD1	SD2	SWW1	SWW2	YD1	YD2	YWW1	YWW2
lightcyan1	121	(89%)	136	ns	ns	ns	0.56	ns	-0.45	ns	ns
lightyellow	184	(76%)	241	ns	-0.36	ns	0.53	ns	ns	ns	ns
purple	393	(75%)	525	ns	ns	ns	0.43	ns	ns	ns	ns
pink	444	(66%)	677	ns	ns	ns	ns	ns	-0.43	0.43	ns
blue	2008	(61%)	3311	ns	ns	ns	0.66	ns	ns	ns	ns
darkslateblue	63	(55%)	115	ns	-0.36	ns	ns	ns	ns	0.70	ns
turquoise	1756	(51%)	3415	ns	ns	ns	ns	0.57	0.51	ns	ns
coral1	42	(51%)	83	ns	0.48	ns	ns	ns	0.40	ns	ns

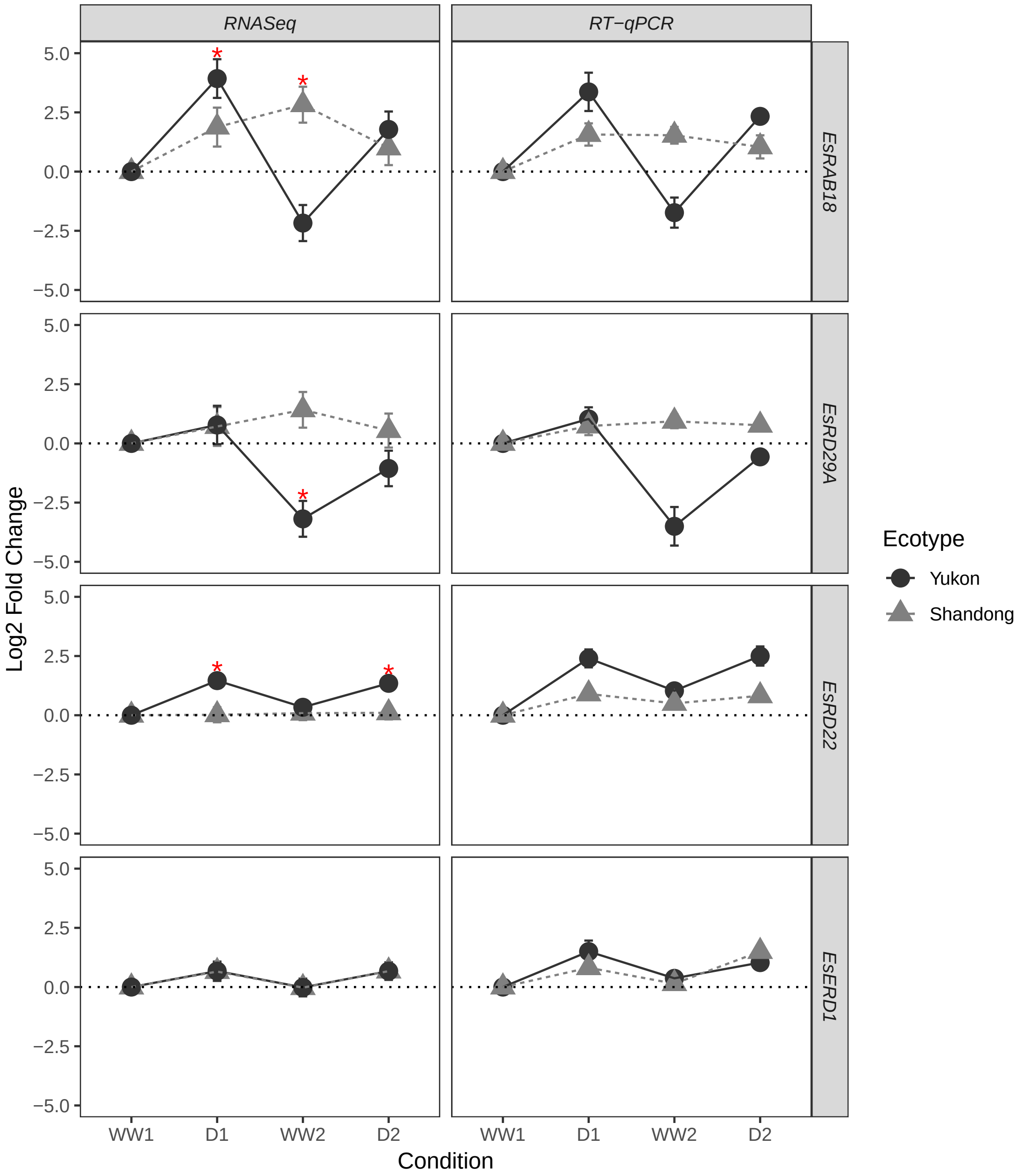
Clusters for GO term enrichment were chosen if at least 50% of genes in each cluster were identified as a DEG at one progressive drought condition progression. Only significant correlations are displayed ($p<0.05$ after FDR adjustment). NS indicates that the correlation was not significant.



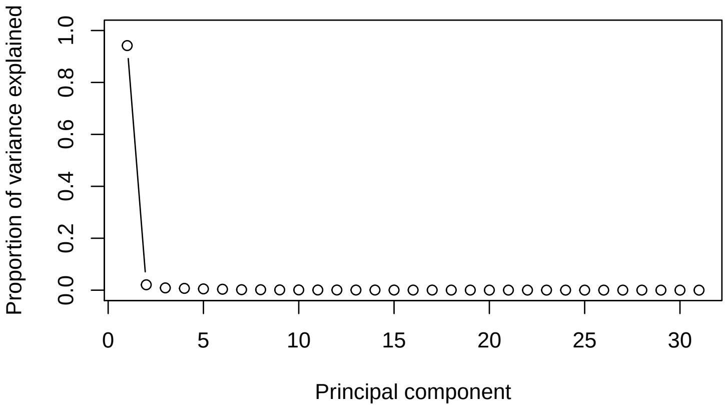
(a)

(b)

Supplemental Figure S3.1: PCA biplots used to identify possible batch effect caused by different cDNA library preparation protocols. A. PC2 and PC4 biplot that shows clustering of libraries sequenced at different times and prepared using different library preparation protocols. cDNA libraries prepared by protocol A are shown in circles (Shandong) and squares (Yukon). cDNA libraries prepared by protocol B are shown in upwards (Shandong) and downwards (Yukon) facing triangles. Library preparation A can be found clustering positively on PC3, while library preparation B is negatively loading on PC3. The clustering was used for batch effect detection. Batch effect was considered in the differentially expressed gene (DEG) analysis using DESeq2. B. PC2 and PC3 biplot shows overlapping technical replicates of resequenced cDNA libraries YD2.1 and SD2.2 suggesting that it is library preparation methods, not sequencing technologies, that are causing a putative batch effect.



SUPPLEMENTAL FIGURE S3.2: Log_2 fold change of the dehydrins described by MacLeod et al. (2014). All $\log_2$ estimates are relative to WW1, or control, conditions for each ecotype. Significant fold changes are described by a red asterisk (*). Error bars represent the standard error of the $\log_2$ fold change. Log_2 fold change results of RNASeq data from all 31 libraries were identified using DESeq2 and an FDR adjust p-value threshold of 0.05. Log_2 fold change results of RT qPCR data of three biological replicates from each condition were identified using a t-test and an FDR adjusted p-value threshold of 0.05



Supplemental Figure S3.3: Screeplot describing proportion of variances explained by each principal component of the PCA completed on the estimated expression abundances of *E. salsugineum* ecotypes subjected to a progressive drought. Expression estimates were calculated from RNASeq data.