IDENTIFYING LONG NONCODING RNAS AND EXAMINING THEIR FUNCTIONAL ROLES IN BRASSICACEAE

by

Kyle Palos

Copyright © Kyle Palos 2022

A Dissertation Submitted to the Faculty of the

SCHOOL OF PLANT SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2022
THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Kyle Palos
titled: Identifying Long Noncoding RNAs and Examining Their Functional Roles in Brassicaceae

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Mark Beilstein
Date: Jan 6, 2022
Frans E. Tax
Date: Jan 6, 2022
Frans E. Tax
Date: Jan 7, 2022
Rebecca A Mosher
Date: Jan 6, 2022
Ravishankar Palanivelu
Date: Jan 6, 2022
Ross Buchan
Date: Jan 6, 2022
Ross Buchan

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Mark Beilstein
School of Plant Sciences

Date: Jan 6, 2022
ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Dr. Mark Beilstein, for everything that he has done for me over my graduate career. Thank you for the years of scientific discussions, presentation practice sessions, and coffee walks. You have made me a better scientist and communicator. Additionally, thank you for fostering a wonderful work environment and always being so accommodating and supportive. I would also like to thank Dr. Andrew Nelson for all the mentorship over the years. You introduced me to computational biology and have constantly provided insight and suggestions throughout my studies. Finally, I would like to acknowledge all of the academic mentorship that I received from: Dr. Ross Buchan, Dr. Becky Mosher, Dr. Ravi Palanivelu, and Dr. Frans Tax.

I would next like to thank the many graduate students who I have worked with or interacted with during my graduate career. From the Beilstein lab, thank you to Evan, Kelly, and Nick for your friendship. Thank you to Josh Trujillo, Jeff Grover, Tania Chakraborty, Cecilia Chow, Jen Noble, Emma Jong, Kelsey Pryze, Xue Pan, and Cory Keith. Additionally, thank you to the many undergraduate students who have worked in the Beilstein lab over the years for all the help with research and laughs along the way.

I would like to sincerely thank all the people closest to me. Thank you to my parents, Susan and Rick, for all the opportunities that you have afforded me and giving me the foundation to succeed. Thank you to my siblings: Alexis, Madison, and Jason for a lifetime of friendship and good times. Finally, thank you to my partner, Caroline, for being endlessly supportive and caring throughout my graduate work.
DEDICATION

To my parents, who have always been there for me.
Table of Contents

ABSTRACT .............................................................................................................................................6

INTRODUCTION ........................................................................................................................................8
  1.1 INTRODUCTION ..........................................................................................................................8
  1.2 EXPLANATION OF DISSERTATION FORMAT .............................................................................14

PRESENT STUDY ....................................................................................................................................15
  2.1 IDENTIFICATION AND FUNCTIONAL ANNOTATION OF LONG INTERGENIC NON-CODING RNAS IN BRASSICACEAE ........................................................................15
  2.2 EXPLORING THE ACTIVITY OF NUCLEAR LONG NON-CODING RNAS DURING PROLONGED SALT STRESS ...........................................................................................................17
  2.3 AN EVOLUTIONARY EXPLORATION OF THE LONG NON-CODING RNA
     HIDDEN TREASURE 1 .........................................................................................................................50

REFERENCES ............................................................................................................................................50

APPENDIX A – IDENTIFICATION AND FUNCTIONAL ANNOTATION OF LONG INTERGENIC NON-CODING RNAS IN BRASSICACEAE .................................................................88
ABSTRACT

Tremendous progress has been made over the last century characterizing the structure, function, and evolutionary dynamics of protein-coding genes. Over the last three decades, biologists have sequenced the genomes of a variety of organisms, and it quickly became apparent that a relatively small portion of most genomes encode protein-coding genes. We now know that the non-protein-coding space contains repetitive sequences along with non-coding RNAs. Many of these non-coding RNAs encode housekeeping transcripts such as: ribosomal RNAs, transfer RNAs, small nucleolar RNAs, and small nuclear RNAs. However, the remaining transcripts include regulatory RNAs that have become the focus of intense investigation over the past two decades. These include the microRNAs and the small interfering RNAs which both act to knock down gene expression.

More recently, a class of transcripts called long non-coding RNAs (lncRNAs) have been interrogated. One well described, functional lncRNA is the telomerase RNA which is necessary for genome stability via the maintenance of chromosome ends following DNA replication (Dew-Budd et al. 2020). While lncRNAs have been known for more than two decades, their functions have remained enigmatic; there are few functionally characterized lncRNAs. However, lncRNAs that have been characterized, especially in mammalian systems, appear to fine tune a plethora of developmental and regulatory processes, along with responses to environmental cues. LncRNAs are frequently expressed at low levels and in specific tissues; they are often poorly conserved among closely related species (Palos et al. 2021). As high-throughput RNA-sequencing becomes more widely available and used, researchers have employed the resulting data to identify lncRNAs en masse. For mammals, a variety of high-quality resources have been used to generate informative annotations to allow for improved hypothesis generation. These resources include GENCODE (Frankish et al. 2021), coupled with lncRNA identification efforts by Cabili et al. (2011) and Guttman et al. (2009). However, while identification and annotation of lncRNAs have been performed in a variety of plant species, they are generally focused on specific tissues or stresses with non-uniform methods for identifying lncRNAs. This inconsistency likely contributes to the paucity of functionally characterized lncRNAs in plants.
In this dissertation, I present three chapters in which I used comparative genomics and transcriptomics to increase our understanding of where lncRNAs are located inside cells, what they look like, what they are doing, and how they are conserved. First, I perform a comprehensive annotation effort across four Brassicaceae species using publicly available RNA-seq data to identify lncRNAs, study their expression patterns, epigenetic characteristics, and evolutionary relationships. This work identified thousands of candidate lncRNAs for follow up study and provided a suite of informative characteristics for many. Next, I studied how prolonged salt stress affects lncRNA expression and accumulation in purified nuclei as well as whole cells in three Brassicaceae. Finally, I performed an evolutionary analysis on a previously characterized lncRNA, *HIDDEN TREASURE 1*, showing that it has undergone extensive duplication and loss across flowering plants. In sum, my work contributes to the growing field of plant lncRNA biology and provides substantive resources for future lncRNA research.
INTRODUCTION

1.1 INTRODUCTION

Eukaryotic genomes are extensively transcribed. Reports for the human genome estimated that as much as 93% of the genome is transcribed in at least one cell type or context, even though only a small fraction of the genome is made up of protein coding genes (<2% in humans, (ENCODE Project Consortium 2012; Pertea 2012). Beyond protein coding transcripts, many of these transcriptional products have been extensively studied, including tRNAs, rRNAs, small nuclear and small nucleolar RNAs (sn/snoRNAs), as well as micro and small interfering RNAs (miRNAs, siRNAs, respectively). Most of these RNAs are transcribed at relatively high levels and perform crucial functions in the cell (Uchida and Adams 2019). Even accounting for these other types of functional RNA genes, only ~5% of the human genome is composed of the genes described above (Pertea 2012). Thus, efforts over the last decade and a half have shifted toward elucidating the enigmatic nature of the remaining portion of human genomes, and eukaryotic genomes more generally.

Some functional RNAs that do not fit within the categories described above have been characterized. One such RNA that has been the subject of intensive research is the telomerase RNA, a molecule crucial for genome stability via the maintenance of telomeres post DNA replication (Dew-Budd et al. 2020). Additionally, the X-inactive-specific transcript (XIST), is the master regulator of X-chromosome inactivation, has been studied for nearly three decades (W. Wang et al. 2021).

Many noncoding RNAs that do not fit the definition of housekeeping or regulatory RNAs are termed long noncoding RNAs (lncRNAs). LncRNAs are transcripts > 200 nucleotides (nt) long that contain no open reading frame > 100 amino acids long (Derrien et al. 2012). LncRNAs can be further characterized by their genomic context (i.e. where they are transcribed). For example, some lncRNAs overlap protein-coding genes (PCGs). With the increase in high density tiling arrays and RNA-seq data over the last 15 years, evidence that most mammalian PCGs have a sense, antisense, or intronic overlapping lncRNA has emerged (Nakaya et al. 2007; St Laurent et al. 2012; Katayama et al. 2005). This finding is likely to be consistent in plants as well, where
thousands of PCG with overlapping lncRNAs have already been identified (H. Wang et al. 2014; Ietswaart, Wu, and Dean 2012; Henz et al. 2007; Huang et al. 2021; Xu et al. 2021).

Research focusing on sense overlapping lncRNAs is sparse since this class of transcript is difficult to identify and characterize given an mRNA is encoded from the same locus. One exception in mammals is the Steroid Receptor RNA Activator (SRA) gene, which can be transcribed into two alternative isoforms (Hubé et al. 2011). One isoform is an lncRNA that acts as a transcriptional co-activator of steroid receptors while the other isoform is translated and appears to be a component of various transcription factor complexes. Another characterized sense lncRNA is encoded from the ENOD40 (early nodulin 40) gene in plants; the locus also produces two isoforms, one of which is an lncRNA while the other is an mRNA (Crespi et al. 1994; Röhrig et al. 2002). The ENOD40 lncRNA induces the proliferation of root cortical cells which ultimately leads to the formation of legume nodules, structures that house nitrogen-fixing rhizobia bacteria (Ganguly et al. 2021). Meanwhile the small peptides generated from the ENOD40 mRNA bind to a subunit of sucrose synthase in root nodules (Röhrig et al. 2002).

Antisense transcripts have been the focus of substantially more research. In mice, nearly all (~87%) coding genes have antisense transcripts (Katayama et al. 2005). This ubiquity seems to be shared in plants; Arabidopsis thaliana produces antisense transcripts from ~70% of annotated protein coding loci (H. Wang et al. 2014). In both mammals and plants, antisense transcripts appear to be expressed at much lower levels and in a more tissue-specific manner than their sense mRNA counterparts (Clark and Blackshaw 2014; H. Wang et al. 2014). However, conflicting findings regarding mRNA-antisense lncRNA expression correlation have been reported in the literature. For example, Zhao and co-authors (2018) showed that A. thaliana antisense lncRNAs tend to be positively correlated in their expression with their sense mRNA, suggesting that many of these transcripts are likely positively influencing the expression of their cognate sense mRNA or that their expression is a consequence of the expression of the mRNAs. In contrast, Henz et al. (2007) described the opposite trend in A. thaliana; they observed a significantly skewed correlation distribution of mRNA-antisense lncRNA pairs toward negative correlation. Importantly, Zhao et al. (2018) used strand specific sequencing of poly-A and ribosomal depleted RNA to identify and assess expression while Henz et al. (2007) used
microarray data of existing annotated lncRNAs, potentially explaining some of the differences reported between the two studies.

The last class of lncRNAs to consider are those that do not overlap PCGs and thus are termed long intergenic noncoding RNAs (lincRNAs). Because lincRNAs originate from loci that do not share attributes with PCGs or classical housekeeping RNAs they have long been considered “junk DNA” (Ling et al. 2015). In humans, lincRNAs make up more than half of all annotated lncRNAs (Ransohoff, Wei, and Khavari 2018). LincRNAs were first suggested to be common genomic elements in mammals from tiling arrays which observed pervasive transcription in non-genic regions (Bertone et al. 2004; Maeda et al. 2006). Support for the validity of lincRNAs was also boosted by the presence of actively transcribed epigenetic signals called “K4-K36” domains (Cabili et al. 2011).

Similar to other classes of lncRNAs, lincRNAs tend to be expressed at relatively low levels and in very tissue specific contexts (Cabili et al. 2011). Additionally, lincRNAs were initially thought to have virtually no sequence conservation and were proposed to evolve via neutral processes (Yu et al. 2004). For these reasons, some have questioned ENCODE and GENCODE’s claims that the vast majority of the human genome has biochemical function and is transcribed (ENCODE Project Consortium 2012; Palazzo and Gregory 2014). These concerns are valid, especially when other findings are considered. Namely, intergenic transcription levels are observed to increase substantially in RNA degradation mutants (Wyers et al. 2005; Thiebaut et al. 2006), suggesting that many of these transcripts are the result of erroneous transcription. Thus, the burden of proof for the validity of lincRNAs needs to move beyond basic identification from transcriptional data.

Accumulating evidence that more evolutionary selective pressure is acting on lincRNAs than initially proposed. The potential exists for lincRNAs to experience selection at the sequence, structure, and functional levels (Diederichs 2014). Multiple studies have shown that most lincRNAs have no identifiable, distantly related homologs, and recognizable sequence similarity between homologs is often limited to small, biased stretches (Hezroni et al. 2015; Nelson et al. 2016; Ulitsky et al. 2011). Interestingly, all of these studies showed that while there are not a substantial number of sequence conserved lincRNAs, there are large groups of
lncRNAs conserved at the syntenic level. This line of evidence suggests that the act of transcription at specific loci is more important than the transcriptional product, adding yet another potential level of selection beyond those that may affect the evolution of lincRNA sequences.

Broadly, lncRNAs appear to be involved in gene regulation through a variety of mechanisms. One major mechanism is the recruitment of chromatin modifying complexes such as the Polycomb repressive complexes 1 and 2 (PRC1/2) (Davidovich and Cech 2015) to epigenetically silence target genes. One particularly surprising finding is that 20% of human lncRNAs were observed to associate with PRC2 (Khalil et al. 2009). However, there is conflicting evidence regarding the specificity of PRC2-lincRNA interactions. Khalil et al. (2009) concluded that lncRNAs are the preferred RNA substrate associating with PRC2, while J. Zhao et al. (2010) concluded that PRC2-RNA binding is promiscuous (J. Zhao et al. 2010).

Other lncRNAs have been shown to bare complementarity to miRNAs and physically interact with complementary miRNA (Cesana et al. 2011; Salmena et al. 2011). These lncRNAs are sometimes called competing endogenous RNAs (ceRNA) or miRNA sponges. LncRNAs acting as miRNA sponges sequester the miRNA away from the target mRNA; the magnitude of regulatory effect appears to depend on the stoichiometric ratio between lncRNA and miRNA (Ling et al. 2015).

Hypotheses of lncRNA function can be informed by data regarding their subcellular localization. Some studies have suggested that lncRNAs are primarily exported to the cytoplasm (Zuckerman et al. 2020; Statello et al. 2021). This hypothesis is supported by the inherit features of most lncRNAs (A/U rich sequences with few exons, (Haerty and Ponting 2015; Hezroni et al. 2015) which make them a preferred target of the Nuclear RNA Export Factor (Zuckerman et al. 2020). Meanwhile, many lncRNAs appear to contain critical elements for their retention in the nucleus (Lubelsky and Ulitsky 2018; Palazzo and Lee 2018). From these studies, it seems that lncRNA transcripts are bound to be exported to the cytoplasm, unless they contain the required sequences to be retained in the nucleus.

Over the past decade and a half, a steady increase in the number of functionally characterized lncRNAs (not just lincRNAs) has occurred, though only a handful of characterized
loci are known from plants. Likely the most well-characterized plant lncRNAs are the COOLAIR/COLDAIR transcripts that are antisense and sense products transcribed from the FLOWERING LOCUS C (FLC) gene, respectively (Csorba et al. 2014; Y. Zhao et al. 2021; Kim, Xi, and Sung 2017; Tian et al. 2019). Briefly, these transcripts repress, and maintain the repression status, of FLC to initiate flowering during cold weather. FLC is a key transcriptional repressor of the transition to flowering following vegetative growth (Deng et al. 2011). Upon vernalization (prolonged cold under short day conditions) FLC is repressed through epigenetic silencing. This silencing is partially induced by COOLAIR dependent recruitment of PRC2 and subsequent H3K27me3 deposition (Tian et al. 2019), while the stable repression is mediated by the COLDAIR lncRNA (Heo and Sung 2011). Interestingly the presence and function of COOLAIR appears to be conserved in monocots (Jiao et al. 2019) while it is unclear whether COLDAIR is present or functional. Finally, it was shown that COOLAIR displays strong structural conservation across the Brassicaceae despite low overall sequence conservation (Hawkes et al. 2016), reinforcing the idea that sequence conservation is an imperfect metric to assess whether a lncRNA is functional.

Another well-characterized plant lncRNA is HIDDEN TREASURE 1 (HID1). HID1 plays a role in A. thaliana photomorphogenesis during continuous red light (Y. Wang et al. 2014). It appears that HID1 acts through the direct repression of PHYTOCHROME INTERACTING FACTOR 3 (PIF3) by forming an RNA-protein complex that binds the PIF3 promoter. Impressively, mutant A. thaliana hid1 can be complemented with the rice HID1 homolog, despite the fact that the two species diverged from each other ~200 million years ago (Wolfe et al. 1989). This represents the most diverged lncRNA complementation result published to date.

There are likely multiple reasons for the disparity between the number of lncRNAs functionally characterized in mammals versus plants. One primary reason is likely the detailed and comprehensive nature of genome annotation efforts in mammalian systems. Unlike PCGs, the accurate prediction of lncRNAs using computational analyses of genome sequences alone remains challenging. For humans and mice, the GENCODE project has integrated genomic, transcriptomic, and proteomic datasets, along with manual curation, to provide accurate and detailed gene annotations for numerous classes of genes, including lncRNAs (Frankish et al. 2021). Incorporation of expression information along with accurate transcript annotation allows
for better reverse genetic discovery of candidate lncRNAs. While many lncRNA annotations have been performed in a variety of plant species (Sun et al. 2020; Liu et al. 2012; H. Wang et al. 2015), very few have incorporated the breadth of high throughput data included in ENCODE and GENCODE. Additionally, most plant annotation studies only utilize select tissues and/or stress conditions to identify lncRNAs, even though most lncRNAs are expressed at very low levels and in response to specific environmental cues (Quinn and Chang 2016). Finally, although candidate lncRNAs can be easily identified through high throughput sequencing data, additional annotation through expression, conservation, and epigenetic data is required to separate bona fide functional lncRNAs from transcriptional noise.

In this dissertation, I present three chapters which further our understanding of plant lncRNA biology. A constant theme throughout this dissertation is a motivation to better understand the trends of lncRNA sequence characteristics, expression patterns, and evolutionary history. The dissertation begins with a broad scope identification and annotation effort, and is followed by an examination of lincRNAs transcribed during salt stress while also incorporating subcellular localization information. I conclude with a focused evolutionary analysis of the HID1 lncRNA.

The first chapter, “Identification and Functional Annotation of Long Intergenic Non-coding RNAs in Brassicaceae”, is a manuscript in review at The Plant Cell. This work presents a substantial resource for the community. In it, my co-authors and I describe the identification and annotation of over 100,000 putative lincRNAs across four model Brassicaceae species. For each annotated transcript, we performed sequence, expression, and evolutionary analyses. Additionally, we provide tissue of expression and contextual environmental data as a means of generating hypotheses about the functions of lincRNAs.

The second chapter, “Exploring the Activity of Nuclear Long Non-Coding RNAs During Prolonged Salt Stress”, is a manuscript in which my co-authors and I develop a prolonged salt stress treatment in three Brassicaceae species and characterize the expression patterns of nuclear lincRNAs under salt stress and control conditions.

The third and final chapter, “An Evolutionary Exploration of the Long Noncoding RNA HIDDEN TREASURE 1”, is a manuscript in which my co-authors and I use a phylogenetic
approach to better understand the complex evolutionary history of the molecularly characterized lncRNA \textit{HID1}. Most Brassicaceae contain multiple \textit{HID1} copies, but we show that the duplicates are primarily the result of relatively recent lineage specific duplications and not an early set of Brassicaceae duplications. This work will aid in determining whether the function of \textit{HID1} is conserved within the Brassicaceae, if and how \textit{HID1} copies may have sub- or neo-functionalized, and what functions single-copy \textit{HID1} genes may be performing.

\textbf{1.2 EXPLANATION OF DISSERTATION FORMAT}

The goal of this dissertation is to improve our understanding of lncRNA biology in plants. Included is one manuscript (Appendix A), which details my efforts to annotate and characterize over 100,000 lncRNAs in four Brassicaceae species. Additionally, I included two independent chapters. The first characterizes the response of nuclear and whole cellular lncRNAs to prolonged salt stress (Section 2.2), and the second chapter uses comparative genomics and phylogenetics to characterize the duplication and loss history of a previously characterized lncRNA (Section 2.3). The manuscript presented in Appendix A has been provisionally accepted in \textit{The Plant Cell}, pending revisions. Additionally, I am the sole first author of this manuscript, however it was an extremely collaborative effort. Thus, I will outline my major contributions to this manuscript, along with the major methods, results, and findings in the ‘PRESENT STUDY’ section of this dissertation.
PRESENT STUDY

2.1 IDENTIFICATION AND FUNCTIONAL ANNOTATION OF LONG INTERGENIC NON-CODING RNAS IN BRASSICACEAE

Long intergenic non-coding RNAs (lincRNAs) are lncRNAs that do not overlap a protein-coding gene. LncRNAs that do overlap protein-coding sequences are generally thought to be playing a role in the regulation of their overlapping gene. Thus, these overlapping lncRNAs have a clear hypothesis to test for functional characterization. In contrast, the lincRNAs do not have such clear functional characteristics. Thus, researchers have relied on genomic and transcriptomic characterization of these transcripts before pursuing functional tests. Until this study, a rich resource has not existed for the wide-spread annotation and computational characterization of lincRNAs. In Appendix A, my co-authors and I process more than 37,000 RNA-sequencing datasets, assemble transcripts, and call lincRNAs in four Brassicaceae species, *Arabidopsis thaliana*, *Camelina sativa*, *Brassica rapa*, and *Eutrema salsugineum*. We identified over 100,000 distinct high-confidence lincRNAs across the four species. Our basic characterization of their sequence and expression patterns supports previous analyses of lincRNAs in bulk, they are A-T rich transcripts that are primarily mono-exonic, lowly expressed and expressed in specific tissues, and they are poorly conserved. To each lincRNA, we assign the following characteristics: tissue of maximum expression, experiments differentially expressed in, membership in co-expression modules, as well as the presence of small open reading frames, ribosome binding, and many other valuable characteristics. This study will serve as a useful resource for the future investigation of lincRNAs in the Brassicaceae.

Because Appendix A was a highly collaborative effort, I wanted to highlight my contributions to the project, as well as the contributions of others. My roles in Appendix A include the design of the project, data analyses, and preparation of the manuscript.

Prior to performing the analyses, I performed the lincRNA identification workflow in Brassica, Camelina, and Eutrema. Additionally, I processed the long-read Nanopore RNA-seq data that we generated for subsequent annotation and expression analyses. Finally, I performed the manual filtering and curation to ensure reliable and consistent lincRNA annotations for all species, which involved over 100,000 transcripts.

I performed many of the analyses that were presented in this study. This includes the basic sequence characteristics of lincRNAs along with their epigenetic trends (Appendix 1,
Figure 3), expression dynamics (Figure 4), inferring functional information (Figure 6), among
many of the supplemental analyses. Additionally, I wrote all the methods and results for the
analyses that I performed, the entire introduction, and portions of the discussion section for the
manuscript.

Many others contributed to this project as well, I have summarized their contributions
here. Anna C. Nelson Dittrich processed the plant samples and prepared the RNA-seq libraries
that were generated in this study. Li’ang Yu incorporated all of the publicly available expression
data into a publicly available interactive resource using Clustergrammer. Jordan R. Brock
generated the Camelina RNA-seq libraries from early embryos that allowed for exploration of
sub-genome expression dynamics. Hsin-Yen Larry Wu and Polly Yingshan Hsu identified small
open reading frames and ribosome association in a subset of our lincRNAs. Elewina Sokolowska
and Aleksandra Skirycz searched for lincRNA derived protein products using mass-spec data.
Andrew Nelson performed the evolutionary analyses throughout the project. Eric Lyons, Mark
Beilstein, and Andrew Nelson were involved in the development of the project, analyzing the
results, and writing or editing the manuscript.
2.2 EXPLORING THE ACTIVITY OF NUCLEAR LONG NON-CODING RNAs
during prolonged salt stress

Abstract

Soil salinity reduces yield in many crop species and thus is a major detriment to food security. While salinity stress is well studied in model plant species like Arabidopsis thaliana, there are fewer analyses of salt stress responses in stress tolerant, non-model species. Additionally, long intergenic non-coding RNAs (lincRNAs) are an emerging class of transcripts that appear to play myriad roles in genome regulation, often by affecting the transcription of protein coding genes. In this study, the nuclear and cellular transcriptomes of three Brassicaceae species: the salt sensitive Arabidopsis thaliana, the hexaploid oilseed crop Camelina sativa, and the halophytic Eutrema salsugineum were investigated under control and prolonged salt stress conditions. Camelina sativa was shown to be similarly salt tolerant to its relative Eutrema salsugineum. Additionally, conserved and species-specific suites of mRNAs and lincRNAs were identified that are differentially regulated based on multiple specific experimental comparisons. Surprisingly, ~ 10-fold more genes differentially retained or depleted from the nucleus compared to genes differentially expressed in response to the salt stress. No evidence was found for sub-genome bias in expression in the hexaploid Camelina sativa. Finally, candidate lincRNAs for each species were also identified that may regulate a neighboring salt responsive mRNA in-cis. In all, our comparative approach provides useful expression and contextual data for subsequent molecular analyses of lincRNAs mediating the salt response in two Brassicaceae.
Introduction

Soil salinization is an increasingly serious issue in agriculture and is especially prevalent in arid regions with a lack of rainfall requiring increased irrigation. It is estimated that 20% of irrigated agricultural land is affected by elevated salt levels (Shrivastava and Kumar 2015). Additionally, with increasingly severe effects of climate change, soil salinization is expected to worsen. High concentrations of salt in soil leads to detrimental effects such as physiological drought and metabolic perturbations due to intracellular sodium accumulation (Munns and Tester 2008).

The plant family Brassicaceae is ideal for studies of salt tolerance because it contains a number of model species with varying degrees of salt sensitivity (Arabidopsis, Eutrema), as well as species of agricultural importance (Brassica oleracea, B. rapa, B. napus). Moreover, there is a robust phylogeny for the family and a well-documented history of genome duplication events (Walden et al. 2020; Beilstein et al. 2010, 2008). Finally, many species in the family are amenable to genetic transformation via the floral dip method (Clough and Bent 1998) which allows for additional genetic and molecular approaches to investigate any number of biological processes, including responses to salt stress.

Molecular responses to salt stress are becoming well characterized in various model species such as Arabidopsis thaliana (hereafter Arabidopsis or A. thaliana). The SALT OVERLY SENSITIVE (SOS) pathway in A. thaliana represents one such well characterized molecular response in which the three SOS genes act to eliminate excess sodium (Zhu 2003; Yang et al. 2009; Cheng et al. 2004; Qiu et al. 2004). During salt stress, cytosolic calcium levels increase and are perceived by SOS3 in the roots and CALCINEURIN B-LIKE10 (CBL10) in the leaves. Either of these two proteins then recruits SOS2, a Serine/Threonine protein kinase, to the plasma membrane to activate SOS1. SOS1 is a plasma membrane Na+/H+ antiporter which prevents toxic levels of intracellular Na+.

Like most plant species, A. thaliana is considered a glycophyte (salt sensitive). Some species maintain their growth in higher concentrations of soil salinity (halophytes). The molecular underpinnings of salt stress responses in halophytes are beginning to emerge. Eutrema salsugineum (hereafter Eutrema or E. salsugineum) is a salt tolerant crucifer relative of Arabidopsis, belonging to lineage II of the Brassicaceae – approximately 43 million years
diverged from Arabidopsis (Beilstein et al. 2008, 2010). The basis of salt tolerance in Eutrema appears to be partially attributed to the duplication of *CBL10* in which the duplicated genes have undergone sub- or neofunctionalization (Monihan et al. 2019). While both *CBL10* paralogs contribute to salt tolerance, one of the paralogs confers enhanced activity of the *SOS* pathway.

Emerging model species in the family include *Camelina sativa* (hereafter Camelina or *C. sativa*), a hexaploid oilseed crop with a whole genome sequence (Kagale et al. 2014) and a member of lineage I, along with Arabidopsis. Because of the high levels of desirable seed oils (36-47% oil, > 90% unsaturated fat), and the ability of Camelina to grow on marginal soils (X. Li and Mupondwa 2014) there is considerable interest in developing Camelina as an agricultural staple. The physiological and molecular responses of Camelina to abiotic stresses have only recently been investigated. At least two studies examined the effects of salt stress on Camelina (Steppuhn, Falk, and Zhou 2010; Heydarian, Gruber, et al. 2018). Steppuhn et al. (2010) found that Camelina experienced severe seed yield decreases in response to salinity stress. Meanwhile, Heydarian et al. (2018) showed that Camelina root length does not change during salt stress and that critical micro-elements, such as Zn, Fe, Mn, Cu, do not change drastically in response to salt stress. Interestingly, Camelina increased specific macro-elements (Ca and Mg) while others increased only marginally (S and P). This was postulated to be an adaptation mechanism for maintaining osmotic homeostasis during physiological drought stress. This is relevant because soil salinity is known to inhibit both macro- and micro-element uptake in a variety of crops, such as barley and olive (Wu et al. 2013; Loupassaki et al. 2002) which leads to metabolic disorders. In sum, Camelina is an emerging oilseed crop which researchers are interested in better understanding its resilience and response to a variety of stresses.

Long noncoding RNAs (lncRNAs) are emerging players in the molecular responses that occur during stress. LncRNAs are RNA transcripts that lack protein-coding potential (no ORF over 100 amino acids), are arbitrarily long transcripts (200 nucleotides or longer) and are not classified as housekeeping noncoding RNAs (ribosomal, transfer, etc.). LncRNAs can be broadly generated from a variety of locations in the genome including sense and antisense overlapping mRNAs, intronic, and intergenic lncRNAs (lincRNAs). Additionally, lncRNAs often share transcript characteristics with mRNAs, this includes the presence of a 5’ cap, intron removal, and poly-adenylation (Guttman et al. 2009). LncRNAs are generally expressed at low levels relative
to protein coding genes, frequently lack sequence conservation among species, and are often restricted to specific cell types and tissues. Palos et al. (2021) developed comprehensive lincRNA annotations and performed a variety of sequence, conservation, and expression characterization analyses on lincRNAs in Arabidopsis, Camelina, and Eutrema. Comprehensive annotations are necessary prerequisites that facilitate molecular characterization of tissue specificity, sequence conservation, presence of epigenetic marks, and numerous other features of candidate lincRNAs.

Most candidate lincRNAs remain uncharacterized, although paradigms regarding their functions are emerging from studies across eukaryotic systems. One emerging paradigm is that lincRNAs often act in the nucleus. For example, the Telomerase RNA Component (TERC/TR) is one of the most important nuclear IncRNAs due its role in telomere maintenance, a key component of genome stability (Feng et al. 1995; Dew-Budd et al. 2020). Additionally, XIST and HOTAIR are well studied IncRNAs important for X-chromosome inactivation and regulating cellular signaling pathways, respectively (Brown et al. 1992; Gupta et al. 2010). In plants, COOLAIR, HID1, and SVALKA are recently characterized IncRNAs that are involved in a variety of responses to environmental cues (Csorba et al. 2014; Y. Wang et al. 2014; Kindgren et al. 2019). COOLAIR is involved in downregulating its antisense protein coding gene FLOWERING LOCUS C during the vernalization response. HIDDEN TREASURE 1 (HID1) is involved in physically repressing the transcription of PHYTOCHROME INTERACTING FACTOR 3 during red light photomorphogenesis. Finally, read-through transcription of the IncRNA SVALKA results in RNA-polymerase II collision with the neighboring CBF1 gene and subsequent downregulation. Together, these studies indicate that IncRNAs can function in the nucleus, where they promote genome stability and influence gene regulation.

Moreover, a longstanding observation in molecular biology is the substantial fraction of RNA in the makeup of chromatin (Huang and Bonner 1965; Holmes et al. 1972). The wealth of high throughput sequencing data-sets has provided extensive evidence that IncRNAs are critical components of chromatin (Guh, Hsieh, and Chu 2020; Nickerson et al. 1989; Caudron-Herger and Rippe 2012; Kung et al. 2015). Due to the low expression levels of most IncRNAs, targeted nuclear analyses for IncRNA focused studies may prove more informative to adequately characterize IncRNA expression dynamics for nuclear-acting lincRNAs.
In this study we performed a prolonged and comparative salt stress experiment using Arabidopsis, Camelina, and Eutrema to analyze the response of the nuclear transcriptome as well as whole cellular transcriptomes. Because each of these species are amenable to genetic transformation, we stably transformed a biotin labeled nuclear membrane protein to facilitate the immunoprecipitation of pure nuclei using the INTACT system (Deal and Henikoff 2010). We show that the emerging biofuel crop, Camelina, is a salt tolerant Brassicaceae, tolerating salt treatments as high as Eutrema. We uncovered species specific as well as conserved responses to the prolonged salt stress, including distinct repertoires of differentially regulated lincRNAs and mRNAs. Surprisingly, we identified ~ 10-fold more genes differentially expressed when comparing the nuclear transcriptomes to whole cellular transcriptomes compared to stress vs control comparisons. Additionally, by sampling both the nuclear and whole cellular transcriptomes we were able to define which cohorts of mRNAs and lincRNAs appear to be nuclear acting versus those performing their role outside the nucleus. We additionally did not find evidence for sub-genome bias in differential expression patterns in the hexaploid Camelina. Finally, we searched for candidate functional lincRNAs that may be regulating relevant salt tolerant mRNAs in-cis. In total, we performed the first comparative transcriptomic study in the Brassicaceae examining the effects of prolonged salt stress and identified conserved, as well as lineage-specific, responses to the prolonged stress.

**Results**

**Camelina is a Salt Tolerant Relative of Arabidopsis and Eutrema**

We wanted to grow plants of our three focal species (*Arabidopsis thaliana, Camelina sativa, and Eutrema salsugineum*) under similar levels of prolonged salt stress to characterize and compare their transcriptional responses. We used previously described salt concentrations from experiments that compared Arabidopsis (150 mM NaCl) and Eutrema (300 mM NaCl) salt tolerance (Monihan et al. 2019). These concentrations were chosen as they were not lethal when applied gradually. Additionally, these concentrations also reduce overall growth by approximately 50% (Supplemental Figure 1). In brief, plants were germinated on soil and once the seedlings developed first true leaves, salt treatments for the experimental plants began. Salt treatments started by watering with 50 mM NaCl, and the concentration of salt was increased in 50 mM increments up to maximal concentration and then continued at this level for 14 days (see
Materials and Methods). We used the same watering regimen for Camelina plants, but because the salt concentration that reduced overall growth by 50% was not known, we first determined this concentration by growing Camelina plants under a range of different final salt concentrations. In these experiments, we measured shoot height, shoot fresh weight, shoot dry weight, and the weight of ten seeds across final NaCl concentrations that differed by 50mM increments (from 0 to 300mM, Figure 1). Shoot height was significantly reduced at all concentrations (66.32% median decrease at 300 mM NaCl, Figure 1B, all 50 mM NaCl increments are significantly different from 0 mM \( P < 7.3 \times 10^{-7} \) Dunnett’s Test), while fresh and dry weight were significantly reduced at all salt concentrations except 50 mM (78.55% and 57.14% median decreases at 300 mM, respectively, Figures 1C and 1D, 50 mM to 0 mM comparisons \( P = .0549 \) and \( .0527 \) for fresh and dry weight, all other comparisons \( P < 1 \times 10^{-7} \)). Ten seed weight for 100 mM and 150 mM NaCl were significantly reduced compared to 0 mM (42.86% median decrease at 150 mM, Figure 1E, \( P < 4.7 \times 10^{-5} \)); salt concentrations above 150 mM resulted in Camelina plants that failed to produce seed. Because Camelina was vegetatively viable after the prolonged 300 mM NaCl treatment, but could not set seed, and thus we classified Camelina as salt tolerant with equivalent responses to Eutrema and proceeded with applying 300 mM NaCl to Camelina for subsequent analyses.
Figure 1: *Camelina sativa* is a salt tolerant Brassicaceae species. (A) Representative growth of Camelina shoots after two weeks of treatment where the salt concentration listed is the maximum concentration reached through the two-week salt treatment. (B) Height of the shoot in centimeters after the two-week salt treatment. (C) Fresh weight of shoot tissue after the two-week salt treatment. (D) Dry weight (approximately two weeks of drying after harvesting) of the shoot tissue after the two-week salt treatment. (E) Ten seed weight after the two-week salt treatment. Plants treated above 150 mM NaCl did not set seed. Significance was calculated using Dunnett’s multiple comparison test comparing treatment against control *** $P < 4.7e-5$.

**RNA sequencing of Nuclei and Whole Cellular Transcriptomes Yields Distinct data-sets**

Because most characterized plant IncRNAs are nuclear-acting (Qin et al. 2017; Y. Wang et al. 2014; D.-H. Kim, Xi, and Sung 2017), are up-regulated in response to stress, and expressed at relatively low levels, we performed RNA-seq on purified nuclei and whole cellular extracts
from leaves following prolonged salt stress. These experimental comparisons allowed us to
detect nuclear vs cytoplasmic lincRNAs, as well as salt stress responsive lincRNAs. Each of our
three focal species were stably transformed with the previously characterized INTACT construct
(Isolation of Nuclei TAgged in specific Cell Types, pUBQ10::NTF/pACT2::BirA, (Deal and
Henikoff 2010) which targets a biotinylated protein to the outer nuclear envelope to allow for the
immuno-purification of whole nuclei. After three weeks of salt treatment at the maximum NaCl
concentration determined for each species or no-salt control conditions, nuclei from leaf tissue
were immuno-purified from all three focal species. In addition, we purified RNA and constructed
libraries from cellular extracts of the same INTACT genotype used for control and salt stress
experiments. In total, we generated 36 RNA-sequencing (RNA-seq) libraries.

To understand the similarities and differences among the RNA-seq libraries we
generated, we performed a principal component analysis (PCA) for data sets from each focal
species (Arabidopsis Figure 2A, Camelina Figure 2C, and Eutrema Figure 2D). In Arabidopsis
we identified three data-sets that did not cluster with those from similar experimental conditions
or cellular compartment, but rather clustered closely with data sets generated from other
conditions or cellular compartments. This result suggested to us that the samples collected were
mislabeled, and thus we excluded them from further analysis. Following their removal, we
performed a second PCA on the remaining Arabidopsis data sets (Figure 2B). As a result, our
conclusions for Arabidopsis are based on fewer data sets than similar analyses performed for
Camelina and Eutrema. After removing the three Arabidopsis data sets, the PCA for libraries
from each species exhibited similar patterns. 1) Prolonged salt stress differentiated libraries
along the second axis of the principal component. 2) Whole cellular transcriptomes were tightly
clustered suggesting a highly consistent response across our experimental replicates. 3) Libraries
from immuno-purified nuclei under both control and prolonged salt stress cluster moderately
closely, albeit less than whole cellular data-sets. 4) We observed the most variation along the
primary axis, which separates nuclear and whole cellular transcriptome data-sets, indicating the
distinct transcriptomic makeup of nuclei relative to whole cellular extracts.
Figure 2: RNA-seq libraries are primarily distinguished based on nuclear purified vs whole cellular transcriptomes. (A-D) Principal component analysis plots of RNA-seq libraries normalized using a variance stabilized transformation. *Arabidopsis thaliana* with all libraries (A), *Arabidopsis thaliana* with non-clustering libraries removed (B), *Camelina sativa* with all libraries (C), and *Eutrema salsugineum* with all libraries (D).

Various studies have shown that nuclear transcriptomes are highly correlated with cytosolic or whole cellular transcriptomes (Barthelson et al. 2007; Grindberg et al. 2013). We were interested in the level of gene expression correlation observed between our nuclear samples and whole cellular transcriptome samples (see materials and methods). Arabidopsis showed the lowest level of nuclear to whole cellular transcriptome correlation (Pearson’s correlation coefficient $r = 0.667$), while Eutrema had the highest correlation level of the three species ($r = 0.852$, supplemental figure 2). This is lower expression correlation than previously reported, however the methods and species examined in the three studies differ. These results are consistent with
the PCAs above and suggest that nuclear gene expression in our species has a distinct and consistent profile relative to whole cellular transcriptomes.

**Nuclear and Whole Transcriptome Sampling Identifies Unique and Shared LincRNA Expression Patterns**

LincRNAs are generally expressed at relatively low levels, frequently stress responsive, and often nuclear localized. Thus, we searched for lincRNAs that were preferentially detected in salt stress experiments and/or from purified nuclei. In Arabidopsis and Camelina, the whole cellular transcriptome data-sets contain more transcribed lincRNAs (*i.e.*, expressed > 0.1 TPM, see materials and methods) when compared to nuclear transcriptome data-sets in these species regardless of whether they were grown under control or salt stress conditions (Figure 3): Arabidopsis 656 vs. 475 & Camelina 5,481 vs. 4,243 (these numbers represent the most expressed lincRNAs identified in whole cellular data-sets and nuclear data-sets, respectively). In contrast, the nuclear libraries of Eutrema contain more expressed lincRNAs than the whole cellular transcriptome data-sets (Figure 3; 2,912 vs. 2,364). In sum, Arabidopsis and Camelina transcriptomes from purified nuclei contained fewer (~45% for both) lincRNAs relative to transcriptomes of whole cellular extracts. Thus, either Arabidopsis and Camelina lack lincRNAs that are primarily nuclear localized, or whole cell RNA-seq is sensitive enough to identify most nuclear lincRNAs in these species, or there are a notable number of lincRNAs that are primarily cytoplasmic that are absent from purified nuclei, or some combination of these hypotheses.
Figure 3: Whole cellular transcriptomes identify more lincRNAs than purified nuclei in *A. thaliana* and *C. sativa* while more nuclear lincRNAs are identified in *E. salsugineum*. Expressed lincRNAs were defined as those expressed above 0.1 transcripts per million.
To address whether lincRNAs are expressed in nuclear, cytoplasmic, or both compartments, and whether lincRNAs are expressed under salt stress, control, or both conditions, we analyzed the common and unique lincRNAs between each compartment and condition. In Arabidopsis, there were 1,215 distinct lincRNAs identified across all libraries (Figure 4, sum of 15 columns), and 148 of these were expressed in libraries from all four experimental conditions. The majority of lincRNAs (701, Figure 4) were present in libraries from only a single experimental condition. We inferred 600 exclusively cytoplasmic lincRNAs, while 296 were exclusively nuclear.

Meanwhile, in Camelina, we observe 7,812 distinctly expressed lincRNAs, with 2,844 of those assigned to the cytoplasm and 1,306 being assigned to the nucleus. Eutrema has 4,260 distinctly expressed lincRNAs, with 767 assigned to the cytoplasm and 1,401 assigned to the nucleus. In sum, we detected what we infer to be housekeeping lincRNAs due to their presence in libraries from all experimental conditions. For Camelina and Eutrema, these housekeeping lincRNAs represented the greatest proportion of lincRNAs. For each species we were also able to infer exclusively nuclear lincRNAs, suggesting these transcripts could promote genomic stability or regulate the transcription of other genes. Finally, we inferred a number of exclusively cytoplasmic lincRNAs, suggesting additional modes of action outside the paradigm of genomic regulation.
Figure 4: LincRNAs are often expressed in a single data-set in *A. thaliana* while they tend to be expressed across all data-sets in *C. sativa* and *E. salsugineum*. Set size corresponds to the total number of lincRNAs expressed in those data-sets while intersection size corresponds to the number of expressed lincRNAs shared between the compared data-sets on the x-axis.
Conservation and Uniqueness of Differentially Regulated mRNAs

The global transcriptomic responses to prolonged salt stress are under-interrogated, especially from a comparative standpoint. To better understand the extent to which conserved transcriptional responses occur when plants experience salt stress, we searched for orthologous mRNAs across our three focal species that were differentially expressed in the same direction (i.e., up-regulated or down-regulated in the same experiments) in all three species (basic mRNA differential expression characteristics are described in Table 1). We identified > 300 orthologous mRNAs that were differentially expressed under the same experimental conditions and in the same direction in all three species (Table 2). To broadly assess the function of these shared differentially regulated genes, we searched for enriched GO terms based on the orthologous Arabidopsis gene function of these conserved differentially expressed orthogroups.

<table>
<thead>
<tr>
<th></th>
<th>Arabidopsis</th>
<th>Camelina</th>
<th>Eutrema</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upregulated</td>
<td>Downregulated</td>
<td>Upregulated</td>
</tr>
<tr>
<td>Nuclei salt vs nuclei control</td>
<td>659</td>
<td>511</td>
<td>3552</td>
</tr>
<tr>
<td>Nuclei control vs whole control</td>
<td>1424</td>
<td>1203</td>
<td>4266</td>
</tr>
<tr>
<td>Nuclei salt vs whole salt</td>
<td>1134</td>
<td>1069</td>
<td>4784</td>
</tr>
<tr>
<td>Whole salt vs whole control</td>
<td>105</td>
<td>8</td>
<td>297</td>
</tr>
</tbody>
</table>

Table 1: Differential expression statistics of mRNAs in the three focal species. Experimental comparisons within a species are not mutually exclusive.

<table>
<thead>
<tr>
<th></th>
<th>AT</th>
<th>CS</th>
<th>ES</th>
<th>AT-CS</th>
<th>AT-ES</th>
<th>AT-CS-ES</th>
<th>% Upregulated in AT-CS-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>whole salt vs whole control</td>
<td>113</td>
<td>743</td>
<td>131</td>
<td>23</td>
<td>6</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>nuclear salt vs nuclear control</td>
<td>1170</td>
<td>10815</td>
<td>1542</td>
<td>359</td>
<td>97</td>
<td>32</td>
<td>69</td>
</tr>
<tr>
<td>nuclear control vs. whole control</td>
<td>2627</td>
<td>11616</td>
<td>3731</td>
<td>923</td>
<td>555</td>
<td>323</td>
<td>61</td>
</tr>
<tr>
<td>nuclear salt vs. whole salt</td>
<td>2203</td>
<td>17796</td>
<td>4417</td>
<td>1028</td>
<td>572</td>
<td>393</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 2: Combinations of differentially expressed mRNA orthologs in our three focal species. AT represents *A. thaliana*, CS represents *C. sativa*, and ES represents *E. salsugineum*. Hyphenated abbreviations represent the number of shared differentially expressed orthologs.
(e.g., AT-CS = 23 means there are 23 conserved mRNAs which are differentially expressed in *A. thaliana* and *C. sativa*).

Table 3: Conserved differentially expressed mRNAs in the whole salt vs whole control experimental comparison.

<table>
<thead>
<tr>
<th>Arabidopsis ID</th>
<th>AT log2FC</th>
<th>Camelina ID</th>
<th>CS Subgenome</th>
<th>CS log2FC</th>
<th>Extrema ID</th>
<th>ES log2FC</th>
<th>Arabidopsis gene description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT2G47770</td>
<td>7.661929456</td>
<td>Csa06g054300</td>
<td>CS_2</td>
<td>4.259406</td>
<td>EUTRA_v100101647mg</td>
<td>4.562557736</td>
<td>Outer membrane Tryptophan-rich sensory protein; induced in vegetative tissues by cosmetic or salt stress or abiotic acid treatment</td>
</tr>
<tr>
<td>AT5G52300</td>
<td>8.203029145</td>
<td>Csa1lg085080</td>
<td>CS_1</td>
<td>4.703317</td>
<td>EUTRA_v10013021mg</td>
<td>4.201440335</td>
<td>Low-temperature-induced 65; induced during water deprivation</td>
</tr>
<tr>
<td>AT5G52300</td>
<td>8.203029145</td>
<td>Csa18g023200</td>
<td>CS_2</td>
<td>5.251025</td>
<td>EUTSA_v10013021mg</td>
<td>4.201440335</td>
<td>Low-temperature-induced 65; induced during water deprivation</td>
</tr>
<tr>
<td>AT5G52300</td>
<td>8.203029145</td>
<td>Csa02g045840</td>
<td>CS_3</td>
<td>5.465585</td>
<td>EUTSA_v10013021mg</td>
<td>4.201440335</td>
<td>Low-temperature-induced 65; induced during water deprivation</td>
</tr>
<tr>
<td>AT5G64400</td>
<td>6.118941306</td>
<td>Csa02q076390</td>
<td>CS_2</td>
<td>1.836096</td>
<td>EUTRA_v10004006mg</td>
<td>3.203655019</td>
<td>Drought induced 8, dehydrin protein family</td>
</tr>
<tr>
<td>AT5G36780</td>
<td>5.093061093</td>
<td>Csa16p008430</td>
<td>CS_2</td>
<td>5.383004</td>
<td>EUTSA_v10024697mg</td>
<td>7.905528415</td>
<td>RadC-like cupin superfamily protein</td>
</tr>
</tbody>
</table>

We started by identifying conserved, differentially expressed orthogroups in the libraries from whole cell extract from salt treated plants vs. whole cell extract control plants (Table 3). We did not compute enriched GO terms for this comparison because only four conserved genes were differentially expressed, likely because of the reduced number of differentially expressed genes identified in Arabidopsis in general. Three of the four conserved, differentially expressed genes had gene descriptions consistent with stress response (e.g., low-temperature-induced; induced during water deprivation).

There were considerably more conserved, differentially expressed genes among libraries from the other experimental conditions, and thus we summarized the top 10 most significantly enriched ontology terms of up-regulated mRNAs from these conditions (Table 4). The significant GO terms for conserved, up-regulated mRNAs from the comparison of nuclear salt vs. nuclear control libraries included response to water deprivation and response to abiotic stimulus. For the comparison of conserved, up-regulated mRNAs from libraries of nuclear control vs. whole cellular extract control, significant GO terms were associated with nuclear processes such as DNA binding, RNA processing, and helicase activity. We considered
transcripts that were significantly up-regulated in nuclear vs. whole cellular extract libraries as a putatively “nuclear retained” transcript. We were unable to determine what class of transcripts may be preferentially retained in nuclei from previous studies beyond aberrant transcripts with retained introns or other processing defects (Palazzo and Lee 2018). Additionally, we looked at the nuclear depleted ontology terms (i.e., those genes that are down-regulated in nuclear vs. whole cellular samples) in Table 5. Because our list of orthologs between the three focal species does not contain mitochondria or chloroplast genes, we also had no a priori expectations as to which transcripts may preferentially exist in the cytoplasm relative to the nucleus. We observed many significant ontology terms related to various metabolism processes such as: photosystem II light harvesting, pigment metabolism, and carboxylic acid metabolism. These nuclear vs. whole cellular comparisons should serve as an additional layer of context to the complex modes of gene regulation. To our knowledge, this is the first comparative study of nuclear transcriptome profiling which reports the enriched and depleted RNA species in nuclei and cytosol.
Table 4: Enriched GO-terms of up-regulated conserved mRNAs across the three focal species in different experimental conditions.

<table>
<thead>
<tr>
<th>Term</th>
<th>pvalue</th>
<th>FDR</th>
<th>experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>response to abscisic acid stimulus</td>
<td>2E-11</td>
<td>9.9E-10</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to water deprivation</td>
<td>2.7E-09</td>
<td>0.000000039</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to hormone stimulus</td>
<td>3.6E-09</td>
<td>0.000000059</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to endogenous stimulus</td>
<td>0.000000032</td>
<td>0.0000004</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to chemical stimulus</td>
<td>0.000000062</td>
<td>0.00000051</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to organic substance</td>
<td>0.000000035</td>
<td>0.0000025</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to abiotic stimulus</td>
<td>0.00000099</td>
<td>0.000062</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to stress</td>
<td>0.000002</td>
<td>0.00011</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>response to stimulus</td>
<td>0.000024</td>
<td>0.00012</td>
<td>nuclear salt vs nuclear control</td>
</tr>
<tr>
<td>helicase activity</td>
<td>1.2E-09</td>
<td>0.00000012</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>binding</td>
<td>6.7E-10</td>
<td>0.00000012</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>nucleic acid binding</td>
<td>0.000000014</td>
<td>0.00000099</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>nucleus</td>
<td>8.6E-09</td>
<td>0.0000011</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>purine NTP-dependent helicase activity</td>
<td>0.0000000086</td>
<td>0.0000036</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>ATP-dependent helicase activity</td>
<td>0.0000000086</td>
<td>0.0000036</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>RNA metabolic process</td>
<td>0.000000091</td>
<td>0.0005</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>DNA binding</td>
<td>0.000015</td>
<td>0.00054</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>post-embryonic development</td>
<td>0.0000034</td>
<td>0.00063</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>RNA processing</td>
<td>0.0000026</td>
<td>0.00063</td>
<td>nuclear control vs whole control</td>
</tr>
<tr>
<td>binding</td>
<td>3.4E-14</td>
<td>6.90E-12</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>nucleus</td>
<td>1.60E-12</td>
<td>2.20E-10</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>nucleic acid binding</td>
<td>5.00E-12</td>
<td>5.00E-10</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>RNA metabolic process</td>
<td>6.90E-10</td>
<td>4.00E-07</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>DNA binding</td>
<td>1.50E-08</td>
<td>9.90E-07</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>nucleobase, nucleoside, nucleotide and nucleic acid metabolic process</td>
<td>9.40E-08</td>
<td>2.70E-05</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>cellular macromolecule metabolic process</td>
<td>7.40E-07</td>
<td>0.00014</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>RNA processing</td>
<td>1.10E-06</td>
<td>0.00015</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>protein binding</td>
<td>8.20E-06</td>
<td>0.00042</td>
<td>nuclear salt vs whole salt</td>
</tr>
<tr>
<td>nitrogen compound metabolic process</td>
<td>5.00E-06</td>
<td>0.00058</td>
<td>nuclear salt vs whole salt</td>
</tr>
</tbody>
</table>
Table 5: Enriched GO terms of conserved, differentially expressed, and nuclear depleted mRNAs (i.e., those mRNAs that are down regulated in nuclear vs whole comparisons and in all three focal species)

<table>
<thead>
<tr>
<th>Term</th>
<th>pvalue</th>
<th>FDR</th>
<th>experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSII associated light-harvesting complex II catabolic process</td>
<td>2.6E-10</td>
<td>0.00000013</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>oxoacid metabolic process</td>
<td>0.000000021</td>
<td>0.0000026</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>organic acid metabolic process</td>
<td>0.00000026</td>
<td>0.0000026</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>carboxylic acid metabolic process</td>
<td>0.00000026</td>
<td>0.0000026</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>cellular ketone metabolic process</td>
<td>0.0000003</td>
<td>0.0000029</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>ATP-dependent peptidase activity</td>
<td>0.0000015</td>
<td>0.000012</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>0.00000041</td>
<td>0.000023</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>plastid envelope</td>
<td>0.0000044</td>
<td>0.00003</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>cellular nitrogen compound metabolic process</td>
<td>0.0000044</td>
<td>0.000036</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>metallopeptidase activity</td>
<td>0.0000014</td>
<td>0.000059</td>
<td>nuclei control vs. whole control</td>
</tr>
<tr>
<td>cellular macromolecule metabolic process</td>
<td>0.0028</td>
<td>0.035</td>
<td>nuclear salt vs. nuclear control</td>
</tr>
<tr>
<td>macromolecule metabolic process</td>
<td>0.0047</td>
<td>0.035</td>
<td>nuclear salt vs. nuclear control</td>
</tr>
<tr>
<td>primary metabolic process</td>
<td>0.016</td>
<td>0.046</td>
<td>nuclear salt vs. nuclear control</td>
</tr>
<tr>
<td>cellular process</td>
<td>0.013</td>
<td>0.046</td>
<td>nuclear salt vs. nuclear control</td>
</tr>
<tr>
<td>cellular metabolic process</td>
<td>0.013</td>
<td>0.046</td>
<td>nuclear salt vs. nuclear control</td>
</tr>
<tr>
<td>PSII associated light-harvesting complex II catabolic process</td>
<td>1.3E-09</td>
<td>0.00000089</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>monosaccharide metabolic process</td>
<td>0.00003</td>
<td>0.01</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>pigment metabolic process</td>
<td>0.000076</td>
<td>0.013</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>cellular nitrogen compound metabolic process</td>
<td>0.000063</td>
<td>0.013</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>cellular macromolecule catabolic process</td>
<td>0.00015</td>
<td>0.017</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>metabolic process</td>
<td>0.00013</td>
<td>0.017</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>cellular catabolic process</td>
<td>0.00045</td>
<td>0.02</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>cellular carbohydrate metabolic process</td>
<td>0.00031</td>
<td>0.02</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>cellular membrane organization</td>
<td>0.00034</td>
<td>0.02</td>
<td>nuclear salt vs. whole salt</td>
</tr>
<tr>
<td>cofactor biosynthetic process</td>
<td>0.00048</td>
<td>0.02</td>
<td>nuclear salt vs. whole salt</td>
</tr>
</tbody>
</table>

Differential Expression Status of LincRNAs in Nuclear and Whole Cellular Salt Stress

We identified lincRNAs that were differentially expressed under salt stress conditions to understand the extent to which lincRNA expression is influenced by salt stress. In total, we found 100 Arabidopsis, 1843 Camelina, and 736 Eutrema lincRNAs that were differentially expressed across four experimental comparisons (nuclei salt vs. nuclei control, nuclei salt vs. whole control, nuclei salt vs. whole salt, whole salt vs. whole control; Figure 5 & Table 6). Most differentially expressed lincRNAs were identified in comparisons of libraries generated from purified nuclei with libraries from whole cell extract. This suggests that libraries from purified
nuclei better capture the expression dynamics of lincRNAs and/or lincRNAs in the nucleus are more likely to be differentially regulated, a conclusion consistent with the idea that lincRNAs often act in the nucleus.

In Arabidopsis and Eutrema a disproportionate number of lincRNAs were up-regulated in libraries from purified nuclei vs. whole cells, for both control and salt conditions (Table 6). In contrast, in Camelina there were more down-regulated lincRNAs for the same comparisons. To determine if this opposing trend is due to the selective repression of lincRNAs from one of the three sub-genomes that comprise Camelina, we sorted differentially expressed transcripts (i.e., both mRNAs and lincRNAs) by sub-genome, and looked for evidence that transcripts residing on a specific sub-genome were more often down-regulated. We found no evidence that lincRNAs or mRNAs are selectively up- or down-regulated based on their sub-genome location (Supplemental Figure 3, multiple Chi-squared tests for given probability, Benjamini-Hochberg corrected, all adj- \( P \) values > 0.624). Interestingly, our observation contradicts other findings for salt stress in Camelina (Heydarian, Yu, et al. 2018), which showed a ~10% increase in differentially expressed genes from sub-genome II. Importantly, aspects of the experimental design differed between our study and Heydarian et al. (2018) including the accession of Camelina used, salt stress protocol employed, and the tissues sampled.

To determine whether differentially expressed lincRNAs are part of a conserved response, or are species specific, we searched for homologs among differentially expressed lincRNAs from the three focal species. We did not identify any conserved and differentially expressed lincRNAs that were shared among all three species and in the same experimental comparison. However, we identified three homologous lincRNAs from Arabidopsis and Eutrema that were differentially expressed in comparisons of purified nuclei vs. whole cell, for both control and salt conditions, (Table 7). We also identified six homologous lincRNAs from Arabidopsis and Camelina that were differentially expressed in comparisons of purified nuclei control vs. whole control experimental comparison. Thus, while nuclear lincRNAs are abundant in all three species, those that are differentially expressed are, with only a few exceptions, unique to the species where they were detected.
Figure 5: Differentially expressed lincRNAs tend to be shared in purified nuclear vs whole cellular comparisons, regardless of control or salt conditions. Set size corresponds to the total number of lincRNAs expressed in those data-sets while intersection size corresponds to the number of expressed lincRNAs shared between the compared data-sets on the x-axis.
Table 6: Differentially expression statistics of lincRNAs in our three focal species. Experimental comparisons within a species are not mutually exclusive.

<table>
<thead>
<tr>
<th></th>
<th>Arabidopsis</th>
<th></th>
<th>Camelina</th>
<th></th>
<th>Eutrema</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upregulated</td>
<td>Downregulated</td>
<td>Upregulated</td>
<td>Downregulated</td>
<td>Upregulated</td>
<td>Downregulated</td>
</tr>
<tr>
<td>Nuclei salt vs nuclei control</td>
<td>8</td>
<td>13</td>
<td>114</td>
<td>398</td>
<td>68</td>
<td>59</td>
</tr>
<tr>
<td>Nuclei control vs whole control</td>
<td>79</td>
<td>4</td>
<td>418</td>
<td>732</td>
<td>452</td>
<td>69</td>
</tr>
<tr>
<td>Nuclei salt vs whole salt</td>
<td>53</td>
<td>1</td>
<td>433</td>
<td>989</td>
<td>478</td>
<td>79</td>
</tr>
<tr>
<td>Whole salt vs whole control</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>14</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7: Combinations of differentially expressed lincRNA orthologs in our three focal species. AT represents A. thaliana, CS represents C. sativa, and ES represents E. salsugineum. Hyphenated abbreviations represent the number of shared differentially expressed orthologs.

<table>
<thead>
<tr>
<th></th>
<th>AT</th>
<th>CS</th>
<th>ES</th>
<th>AT-CS</th>
<th>AT-ES</th>
<th>AT-CS-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>whole salt vs. whole control</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nuclear salt vs nuclear control</td>
<td>21</td>
<td>512</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nuclear control vs. whole control</td>
<td>83</td>
<td>1150</td>
<td>410</td>
<td>6</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>nuclear salt vs. whole salt</td>
<td>54</td>
<td>1422</td>
<td>442</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Identification of Putative Cis-Regulating lincRNAs

A commonly described molecular function of lincRNAs is the regulation of nearby mRNAs (Qin et al. 2017; Kindgren et al. 2019). A hallmark of cis regulation is the strong positive or negative expression correlation between nearby lincRNAs and mRNAs. We used our expression data from each species to calculate pairwise Pearson correlation coefficients to identify any pairs of adjacent or nearby genes that show a strong correlation, either positive or negative. Our expectation for the distribution of adjacent lincRNA-mRNA gene expression correlations is an approximately normal distribution centered on zero, indicating no correlation...
between the expression levels of the two loci. The distribution of Pearson correlation coefficients followed this expectation for both Arabidopsis and Eutrema lincRNA-mRNA pairs (Figure 6). However, in Camelina there was a strong positive skew because there was a disproportionately high frequency of lincRNA-mRNA pairs whose expression was positively correlated (Figure 6). As of now it is unclear whether the correlation skew observed in Camelina is an artifact of our sequencing and sampling, or whether there are certain characteristics of Camelina gene expression that leads to this observation. For Arabidopsis and Eutrema, the majority of lincRNAs did not show evidence of cis-regulation, but in Camelina it is possible that a substantial fraction of the lincRNAs that we detect are playing a role in the positive regulation of adjacent genes.

To recover high confidence lincRNA candidates that may act as cis-regulators, we filtered the lincRNA-mRNA pairs above to exclude pairs where the probability of such regulation was unlikely. First, we removed lincRNA-mRNA pairs that were separated by less than 100 base pairs to avoid misannotated, untranslated regions associated with the mRNA. Next, we omitted lincRNA-mRNA pairs in which the mRNA member was not differentially expressed; and finally, we only computed correlations between lincRNA-mRNA pairs in which the lincRNA member showed variable expression across the libraries (see materials and methods). Our goal in using the latter two filters was to exclude lincRNA-mRNA pairs for which we lacked evidence that lincRNA expression could have altered the expression of the adjacent mRNA. In sum, we used RNA-seq data to identify putative cis-regulatory lincRNA-mRNA candidate pairs and present 12 pairs across our focal species which likely act in salt stress responses (Table 8). Four co-expression patterns are possible for each species: significantly up-regulated mRNA with either a highly correlated (pattern 1), or anti-correlated (pattern 2) lincRNA; a significantly down-regulated mRNA with a highly correlated (pattern 3) or anti-correlated (pattern 4) lincRNA. Many of the mRNAs that we identify have functions related to stress response, calcium signaling, and transcriptional regulation (Table 8), suggesting we may have recovered lincRNA that modulate the expression of adjacent protein coding genes during the salt response.
Figure 6: Neighboring mRNA-lincRNA expression correlation coefficients are non-normally distributed in *C. sativa*. The density of Pearson correlation coefficient values (*r*) between the expression values of neighboring mRNA-lincRNA pairs. Only mRNAs which were differentially expressed in at least one experimental comparison were included for correlation calculations.
Table 8: Summary of candidate lincRNA-mRNA pairs identified based on their expression correlation or anti-correlation.

<table>
<thead>
<tr>
<th>Species</th>
<th>mRNA Log2TC</th>
<th>Nuclear Salt stress</th>
<th>lincRNA ID</th>
<th>r</th>
<th>mRNA ID</th>
<th>Genomic distance (bp)</th>
<th>Arabidopsis Ortholog ID</th>
<th>Arabidopsis Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>7.81912</td>
<td>XLOC_011741</td>
<td>-4.9824057</td>
<td>-0.1718</td>
<td>NA</td>
<td>-1718</td>
<td>NA</td>
<td>Defense responsive, nuclear stress-responsive</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>2.08395</td>
<td>XLOC_065397</td>
<td>0.67661575</td>
<td>-1042</td>
<td>NA</td>
<td>1042</td>
<td>NA</td>
<td>Phenolic differentiation</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>-3.47276</td>
<td>XLOC_000105</td>
<td>0.90317885</td>
<td>5113</td>
<td>NA</td>
<td>5113</td>
<td>NA</td>
<td>MATH domain containing family</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>-1.78873</td>
<td>YLOC_001918</td>
<td>-0.49909264</td>
<td>245</td>
<td>NA</td>
<td>245</td>
<td>NA</td>
<td>G-box</td>
</tr>
<tr>
<td>Cicer arietinum</td>
<td>4.25169</td>
<td>XLOC_000656</td>
<td>0.897548615</td>
<td>-648</td>
<td>AT5G17210</td>
<td>-648</td>
<td>AT5G17210</td>
<td>Inhibitors of protein phosphatases (DAP) respond to salt stress</td>
</tr>
<tr>
<td>Cicer arietinum</td>
<td>1.03272</td>
<td>XLOC_015441</td>
<td>-0.928472594</td>
<td>513</td>
<td>AT5G06140</td>
<td>513</td>
<td>AT5G06140</td>
<td>20S1 chromatin regulator, transcriptional regulator</td>
</tr>
<tr>
<td>Cicer arietinum</td>
<td>-1.417</td>
<td>XLOC_019346</td>
<td>-0.77357418</td>
<td>-2431</td>
<td>AT5G05420</td>
<td>-2431</td>
<td>AT5G05420</td>
<td>G protein binding protein, responds to water deprivation</td>
</tr>
<tr>
<td>Cicer arietinum</td>
<td>-5.197115</td>
<td>XLOC_007082</td>
<td>-0.97752278</td>
<td>-3201</td>
<td>AT5G23630</td>
<td>-3201</td>
<td>AT5G23630</td>
<td>ATPase-glycosyltransferase transmembrane transporter</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>2.05367</td>
<td>XLOC_016940</td>
<td>-0.320481043</td>
<td>-973</td>
<td>AT5G21340</td>
<td>-973</td>
<td>AT5G21340</td>
<td>EF1A-like transcription factor, responds to water deprivation</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>1.82663</td>
<td>XLOC_011883</td>
<td>-0.89400517</td>
<td>1242</td>
<td>AT5G22350</td>
<td>1242</td>
<td>AT5G22350</td>
<td>RACK1-like, Hsp90 family</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>-1.849916</td>
<td>XLOC_0193014</td>
<td>-0.48311986</td>
<td>-940</td>
<td>AT5G07320</td>
<td>-940</td>
<td>AT5G07320</td>
<td>18S rRNA, methionine-dependent ribosome</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>-3.636089</td>
<td>XLOC_016935</td>
<td>0.77794648</td>
<td>908</td>
<td>AT5G47480</td>
<td>908</td>
<td>AT5G47480</td>
<td>CALMOSCIN-LIKE 4, Calmodulin-like EF-hand family</td>
</tr>
</tbody>
</table>

Discussion:

In this study, we used nuclear and whole cellular transcriptome profiling to identify lincRNAs and mRNAs that are expressed in response to systemic salt stress in three Brassicaceae species. This work serves as an important first step in uncovering lincRNAs that mediate salt stress responses in Brassicaceae.

One unexpected finding from our study is that we observed the fewest number of differentially expressed mRNAs and lincRNAs in our whole cell salt vs. whole cell control libraries (Table 1). Libraries generated from purified nuclei yielded 1 order of magnitude more differentially expressed loci. This is also apparent in our principal component analyses which show a substantial amount of variation due to nuclear vs whole cell sampling (Figure 2). Recent studies that have examined nuclear and total RNA-seq show varying levels of correlation among these types of data-sets in plants (Reynoso et al. 2018; Sunaga-Franze et al. 2021). Reynoso et al. (2018) found that both gene expression and transcript identity were loosely correlated in libraries generated from INTACT purified nuclei from rice root tips vs. poly-A RNA from whole root tips (Pearson’s correlation coefficient r = 0.298). Importantly, the authors sequenced ribosomal-
depleted nuclear RNA, whereas we sequenced poly-A nuclear RNA. Sunaga-Franze et al. (2021) sequenced poly-A RNA from single nuclei and showed that those data-sets had a stronger Pearson correlation coefficient with bulk whole cellular RNA-seq data-sets ($r = 0.74$), which agrees with our observations of poly-A nuclear RNA-seq (Supplemental Figure 1). Some studies of nuclear transcriptomic profiling in various mammalian cell lines show that there is much less variation between nuclear and total or cytoplasmic transcriptomic profiles (Barthelson et al. 2007; Grindberg et al. 2013). Meanwhile, Zaghlool et al. (2021) identified over 10,000 transcripts with differential abundance between nuclear and cytosolic compartments in human brain tissue (Zaghlool et al. 2021). In addition, the authors showed that there are approximately 10-fold more lincRNAs expressed in nuclei relative to cytoplasm, which fits the developing paradigm of lincRNAs performing regulatory roles in the nucleus. Additional experiments should be performed in all species to better understand which cohorts of lincRNAs are specifically retained in the nucleus and which are shuttled to the cytoplasm.

We observed approximately 10-fold fewer lincRNAs differentially expressed in our experimental comparisons compared to mRNAs. This observation is consistent with previous reports (Q. Li et al. 2019). Published lincRNA characteristics for RNA-seq experiments include: low levels of expression, high variance in expression among samples and replicates. In addition, the function of some lincRNAs may be independent of the RNA product (i.e., the act of transcription at the locus is necessary while the actual RNA product may be non-functional (Tuck et al. 2018)). Taken together, these attributes make transcript quantification of lincRNAs difficult. Moreover, our finding that whole cell salt vs. whole cell control library comparisons contained the fewest differentially expressed mRNAs and lincRNAs, is not without precedent. For example, heat stress in Arabidopsis, mammalian cell lines, and microorganisms is associated with a rapid and global transcriptional reprogramming, followed by an adaptive response to prolonged stress which increases the levels of proteins important for protein folding and degradation (Lukoszek, Feist, and Ignatova 2016; Shalgi et al. 2013; Liu, Han, and Qian 2013; Richter, Haslbeck, and Buchner 2010). We applied salt stress over ten treatments, which covers nearly the entire vegetative life stage of our focal species. We were unable to identify an equivalent transcriptomic study in any model organism, so further work will need to be performed to understand how our prolonged salt stress temporally reprograms the transcriptome and whether the number of differentially regulated genes gradually reduces over time.
We observed that mRNAs and lincRNAs in Camelina were predominantly down-regulated in purified nuclei relative to whole cell libraries, while Arabidopsis and Eutrema had a much more balanced ratio of up- and down-regulated genes. Transcripts induced in the nucleus relative to the whole cell libraries represent “nuclear retained” transcripts. There are emerging features for nuclear retained transcripts known from studies in mammals. These include transcripts with retained introns and extended poly-A tails, as well as transcripts that are putative cis-acting elements (Wegener and Müller-McNicoll 2018). The molecular mechanisms and outcomes of these nuclear retained transcripts are widely varied. Nuclear retained mRNAs are often incompletely spliced mRNAs, which stockpile in nuclear speckles during stressful conditions where they are stored before processing is completed, after which they are released for translation or other functions (Ninomiya, Kataoka, and Hagiwara 2011). We found that nuclear retained mRNAs encoded proteins whose GO terms indicated a nuclear function such as DNA binding, RNA processing, and helicase activity (Table 4). Conversely, GO terms for transcripts that were depleted in the nucleus were enriched for metabolic processes, suggesting that these mRNAs are not targets for nuclear retention (Table 5). This study is the first, to our knowledge, to identify nuclear retained mRNA transcripts in plants. Moreover, this is the first study to recover mRNAs and lincRNAs that respond to prolonged salt stress in plants.

We recovered a small subset of lincRNAs that may act as cis-regulatory elements by identifying strongly correlated or anti-correlated lincRNA-mRNA pairs within close physical proximity. One particularly promising lincRNA may regulate the expression of its neighbor, \textit{AT1G32350} (Table 8), which encodes an alternative oxidase that localizes to mitochondria and is responsible for scavenging reactive oxygen species during osmotic stress, among other stresses (D. S. Kim et al. 2011; D. Wang et al. 2021). Without additional information about RNA-chromatin physical interactions, or RNA-protein interactions during systemic salt stress, it is difficult to generate detailed molecular hypotheses concerning the functions of other nuclear lincRNAs. Most characterized lincRNAs are nuclear acting, and many of these regulate a gene in cis, assessing neighboring localized gene expression correlation should be a routine analysis in transcriptomic studies of lincRNAs.

Finally, our study provides additional context for the types of activities that lincRNAs may be engaging in during an agriculturally relevant abiotic stress. Our findings are consistent
with previous observations that the majority of lincRNAs act in the nucleus. Importantly, we
identified a variety of candidate mRNAs and lincRNAs that may be involved in responding to
prolonged salt stress, or that may perform regulatory roles in the nucleus. Another important take
away from this work is that we did not recover a suite of lincRNAs that were differentially
regulated in response to prolonged salt stress, suggesting that lincRNA responses may be highly
species specific. Future work will need to be performed to answer a variety of outstanding
questions: how do the whole cellular transcriptomes of our focal species have relatively few
differentially expressed genes given the drastic phenotypic changes resulting from the
treatments; does cytoplasmic RNA-seq during prolonged salt stress recapitulate the general
whole cellular comparisons that we make here; and finally are the lincRNA candidates that are
differentially expressed and show strong proximal expression correlation to mRNAs bona fide
regulators of gene expression?

Materials and Methods

Plant Material, Growth Conditions and Transformation

The following species and genotypes were used for experiments in this study: the Arabidopsis
thaliana accession Columbia (Col-0), the Eutrema salsugineum accession Shandong (kindly
provided by Dr. Karen Schumaker), and the Camelina sativa accession Ames (kindly provided
by Dr. John McKay). All species were surface sterilized in 30% bleach/1% TWEEN 20 then
washed 3 times with sterile water before being plated on either on ½ MS agar plates
(Arabidopsis and Camelina) or on 1/10 MS agar plates for Eutrema, both plates contained 1%
(w/v) sucrose and were adjusted to pH 5.8. Plates were stratified at 4°C for 1 week before being
transferred to a growth chamber at long day conditions (16 hours light 22°C, 8 hours dark 20°C).
Seedlings were transplanted to soil 7-10 days after germination once true leaves formed.

To generate a constitutively expressed nuclear targeting factor (NTF) to perform INTACT, the
Arabidopsis Ubiquitin 10 gene (AT4G05320.1) promoter was introduced upstream of the NTF.

Plant transformations were carried out by Agrobacterium-mediated floral dip transformation as
described by Clough and Bent
**Salt Stress Treatments**

A modified Hoagland’s nutrient solution (Hoagland and Arnon 1938) was used to administer the salt, with minor alterations. First, cobalt nitrate Co(NO₃)₂ was replaced with cobalt chloride CoCl₂. Second, the solution was made to be ¼ strength to optimize growth conditions for the species analyzed.

Salt stress began when plants developed their first true leaves. Plants were watered every Monday, Wednesday, and Friday with the Hoagland’s solution and supplemented NaCl. Salt concentrations increased in increments of 50 mM NaCl until the maximum concentration is reached. Plants are salt stressed until the plants reach reproductive development, this usually consists of approximately 10 salt treatments. Control plants are watered with the Hoagland’s solution without supplemented NaCl until salt stressed plants are ready to be harvested.

The different growth forms of each species necessitated different harvesting methods. Arabidopsis and Eutrema had their rosettes removed from the base of the soil, inflorescences removed if applicable, and placed in a 50 mL falcon tube until the mass reached 1.5 grams and flash frozen. Camelina and Brassica had their vegetative leaves removed from the central stems and placed in a 50 mL falcon tube until the mass reached 1.5 grams and flash frozen.

**Nuclei Purification**

The INTACT method was used to purify nuclei from Arabidopsis, Camelina, and Eutrema using a modified protocol from Wang and Deal, 2015. Briefly, three grams of frozen tissue was ground in liquid nitrogen and resuspended in 21 mL of NPB (20 mM MOPS, 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 0.5 mM Spermidine, 0.2 mM Spermine, pH 7.0) containing one tablet of cOmplete UTRA EDTA-free protease inhibitors (Sigma-Aldrich SKU 5892953001) per 50 mL of NPB. Lysate was filtered through a 70 μm cell strainer then split into two 10.5 mL aliquots to be spun at 1,200 x G for seven minutes at 4° C. Supernatant was decanted and nuclei pellets were gently resuspended in 1 mL of NPB. 25 μL of washed M-280 streptavidin Dynabeads (ThermoFisher 11205D) are added to the resuspended nuclei pellet and incubated on a nutator at 4 ° C for 30 minutes. The nuclei-bead mixture was diluted to 14 mL with NP8t
(NPB supplemented with 0.1% (v/v) Triton X-100) and placed on a magnet for 7 minutes. The unbound lysate was aspirated and the beads were washed twice more with 12 mL of NPBt before being pooled.

**Total and Nuclear RNA Purification and Sequencing Library Preparation**

Purified nuclei or ground tissue was resuspended in an appropriate volume of TRIzol (ThermoFisher 15596026) depending on the quantity of input sample. The TRIzol user guide instructions were followed through the phase-separation step. The aqueous phase was then mixed with an equal volume of pure ethanol before being processed through the Direct-zol RNA Microprep kit (Zymo R2062). Purified RNA was further processed by Amaryllis Nucleics (Oakland, California) which includes mRNA isolation, QC, and library preparation using the YourSeq FT v1.5 kit (Amaryllis Nucleics). Libraries were sequenced using an Illumina (Illumina, San Diego, CA, USA) NextSeq 500 producing single-end 80 bp reads (Arabidopsis) or paired-end 80 bp reads (Camelina and Eutrema).

**Bioinformatic Processing**

Raw RNA-seq reads were quantified against lincRNA annotations generated from Palos et al. 2021 using Salmon version 1.5.1 (Patro et al. 2017). We generated a decoy-aware transcriptome following https://combine-lab.github.io/alevin-tutorial/2019/selective-alignment/ before quantifying the RNA-seq data using the --validateMappings and --gcBias parameters. Salmon output files were further processed using the tximport R package (Soneson, Love, and Robinson 2015; Team and Others 2013) to get TPM values. Differential expression analyses were conducted using the R package DESeq2 (version 1.32.0 (Love, Huber, and Anders 2014). The design matrix specified for each species before running the “DESeq” function was: “~ Test + Condition + Test:Condition” where test specified whether the data-set was whole or nuclear, and condition specified whether the data-set was control or salt stressed. We analyzed four differential expression experimental comparisons: whole salt vs whole control, nuclear salt vs nuclear control, nuclear salt vs whole salt, and nuclear control vs whole control which were generated using the “results” command and the “contrast” argument. Genes were considered
differentially expressed if they had a log2 fold change value greater or less than 1 or -1, respectively, as well as a FDR of 0.05 or lower.

Expression correlation was calculated for each species using the variance stabilized transformation (vst) function from DESeq2. Prior to calculating pairwise correlations, genes were filtered so that only the top 50% of genes based on expression interquartile range values were kept. Pearson correlation coefficients were calculated in a pairwise fashion between all of the retained variable genes. Neighboring lincRNA-mRNA gene pairs were then filtered.

**Comparative Genomic Analyses**

Orthologous mRNAs relative to Arabidopsis were gathered from Phytozome (Goodstein et al. 2012) and Ensembl Plants (Bolser et al. 2017). LincRNA sequence homologs and syntelogs were generated using Evoline II (Nelson et al. 2017).
Supplemental Figure 1: *A. thaliana* is a salt sensitive model organism while *E. salsugineum* is a salt tolerant model species in Brassicaceae. Each species was allowed to germinate and develop true leaves before receiving two-weeks of gradually increasing salt treatment (50 mM increments) up to the maximum concentration shown.
Supplemental Figure 2: Nuclear and whole cellular transcriptomes are relatively similar. The expression levels of mRNAs and lincRNAs was compared between averaged nuclear samples and averaged whole cellular samples. These averaged values were \( \log_2 + 1 \) transformed before being plotted. \( r \) represents Pearson’s correlation coefficient between the nuclear and whole cellular samples along with the 95% confidence interval (95% CI).
Supplemental Figure 3: Differentially expressed genes are not sub-genome biased in *Camelina sativa*. Differentially expressed genes were separated based on sub-genome as well as up or down-regulated. Sub-genome bias is not significantly different in any comparison (e.g., sub-genome 1 vs sub-genome 2 for nuclei control vs whole control in lincRNAs) adj-\(P > 0.624\) Chi-square tests for given probability, Benjamini-Hochberg multiple testing correction.
2.3 AN EVOLUTIONARY EXPLORATION OF THE LONG NON-CODING RNA
HIDDEN TREASURE 1

Abstract

Responses to environmental cues are a critical part of plant life cycles as well as their adaptation to changing environments. Light cues regulate many aspects of development from germination to flowering. HIDDEN TREASURE 1 (HID1) is a previously described long non-coding RNA (lncRNA) involved in red light mediated photomorphogenesis in Arabidopsis. We searched for paralogs and orthologs of HID1 and used the recovered sequences to infer phylogeny. Our results revealed an extensive history of duplication across flowering plants. Moreover, HID1 diversity in Brassicaceae is largely the result of relatively recent, lineage specific tandem duplication events. In addition, our comparison of aligned HID1 homologs indicated that one of the four stem-loops identified in Arabidopsis HID1 is highly conserved in HID1 loci from other species. Interestingly, we also recovered an Arabidopsis HID1 paralog (HID1B) with high sequence similarity, but with distinct domains of expression that includes roots, suggesting that HID1B may not be exclusively involved in red light mediated photomorphogenesis. Our findings provide an evolutionary and comparative framework for future investigations of Arabidopsis HID1 and its homologs across angiosperms.
Introduction

Plants have a variety of photoreceptors capable of perceiving specific wavelengths of light, and which control a wide range of developmental and physiological traits (Legris, Ince, and Fankhauser 2019). Well-characterized plant photoreceptors include the blue light sensing phototropins and cryptochromes, the UV-B absorbing UVR8, and the red/far-red light perceiving phytochromes (Mawphlang and Kharshiing 2017; Christie et al. 2015; Chen and Chory 2011; Rizzini et al. 2011). Germination is one of the most critical light regulated processes in plants. During germination, desiccated seeds imbibe water and activated photoreceptors promote germination along with early seedling growth. Early seedling growth is characterized by de-etiolation (greening), chlorophyll synthesis, and chloroplast development (Franklin and Quail 2010; Mawphlang and Kharshiing 2017).

In Arabidopsis, phytochromes are the dominant photoreceptor mediating germination and early seedling growth (Franklin and Quail 2010; Chung and Paek 2003), processes that collectively are termed photomorphogenesis. In brief, phytochromes are activated when they absorb red light which leads to their translocation to the nucleus. When in the nucleus, phytochromes interact with a variety of proteins to induce transcriptional changes that promote germination and de-etiolation (Pham, Kathare, and Huq 2018). One family of direct interaction partners for phytochromes are the Phytochrome Interacting Factors (PIFs). PIFs encode basic helix-loop-helix transcription factors that suppress de-etiolation and photomorphogenesis. PIFs are targeted for rapid degradation upon interaction with nuclear-localized phytochromes (Pham, Kathare, and Huq 2018; Leivar and Monte 2014). Thus, PIFs have a critical role in repressing seedling photomorphogenesis in the dark, ensuring developmental processes occur in the proper light environment.

An interesting and unusual insight into the regulation of photomorphogenesis was the characterization of the long non-coding RNA (lncRNA) *HIDDEN TREASURE1* in Arabidopsis (*AtHIDI*) (Y. Wang et al. 2014). *AtHIDI* is the first of four units of a highly expressed,
polycistronic non-coding RNA transcript. Wang et al. (2014) analyzed the phenotype of Arabidopsis seedlings with a T-DNA insertion at the 3’ end of the polycistronic transcript. HID1 transcript was detected, albeit at lower levels, in the Athid1 insertion mutant lines. Under continuous red light, Athid1 mutants displayed signs of incomplete photomorphogenesis characterized by elongated hypocotyls relative to wild type plants with normal AtHID1 expression. Genetic and molecular studies of AtHID1 showed it forms an RNA-protein complex that physically interacts with the promoter of PIF3 during continuous and intense (60 μmol m$^{-2}$ s$^{-1}$) red light exposure, resulting in transcriptional repression of PIF3. Using computational approaches, the authors predicted that the AtHID1 secondary structure is composed of 4 stem-loops (SL1–4). Through sequence deletion analysis, they demonstrated that SL2 and SL4 are critical for AtHID1 function. Finally, in a surprising display of functional conservation, the Oryza sativa (rice) HID1 promoter and homologous sequence (OsHID1) complemented the Athid1 mutant phenotype despite that fact that Arabidopsis and rice last shared a common ancestor ~200 million years (Wolfe et al. 1989).

To date, AtHID1 represents one of the few well characterized plant lncRNAs. The growing consensus from lncRNA research is that these transcripts fine tune a plethora of developmental and regulatory processes (Datta and Paul 2019), and thus mutant lncRNA phenotypes are likely to be subtle, or context specific. Additionally, in comparison to protein coding loci, lncRNAs exhibit less sequence conservation across even closely related species and are often expressed in only one or a few tissues at low levels (Palos et al. 2021). Taken together, these factors have hindered the functional characterization of lncRNAs.

In this study we explored the evolutionary history of HID1 in Brassicaceae and across other angiosperm lineages. Our phylogenetic analyses of HID1 across angiosperms show that HID1 duplication has been a common occurrence across most plant lineages we investigated. We identified a highly similar paralog to AtHID1 in Arabidopsis that likely originated as the result of a tandem duplication event – AtHID1B. Using publicly available RNA-seq data, we determined the expression domains of both paralogs. AtHID1 is most highly expressed in seedlings and seed, while AtHID1B is induced in seedlings and root. Additionally, we found that the expression of AtHID1 and AtHID1B are anti-correlated with PIF3 expression in seedling meristems, suggesting that AtHID1B could also contribute to the regulation of PIF3. Finally, we show that
SL4 drives the conservation of HID1 across angiosperms, resulting in the maintenance of a multi
stem-loop structure known to be important for AtHID1 function. Our results provide a
comparative framework for future functional characterization of AtHID1 and its homologs across
angiosperms.

Results

AtHID1 Overlaps an Orphan Protein Coding Gene

We wanted to better understand the genomic context of AtHID1 and the nature of the
polycistronic transcript from which it is derived. Wang et al. (2014) described AtHID1 as the
first unit of a polycistronic cluster that is processed into four non-coding RNAs (Figure 1A).
Upon further investigation of this genomic region, we observed that all four non-coding loci are
annotated as small-nucleolar RNAs (snoRNAs), including AtHID1 – AtHID1 is not annotated as
a snoRNA in Rfam, but is on TAIR and Araport11 (Lamesch et al. 2012; C.-Y. Cheng et al.
2017; Griffiths-Jones et al. 2003). Additionally, in the Arabidopsis annotation databases, AtHID1
is annotated as AT2G356747, however this locus corresponds to the second unit of the
polycistronic cluster, an 80 nucleotide (nt) snoRNA. The accurate gene ID for AtHID1 is
AT2G08770 (Figure 1A).

Both AtHID1 and the second unit of the polycistronic RNA (AT2G356747) entirely overlap an
annotated sense protein coding gene of unknown function (AT2G35750) (Figure 1A). The TAIR
annotation indicates this locus produces a 61 amino acid transmembrane protein that localizes to
the mitochondria. This overlapping protein coding gene was not mentioned by Wang et al.
(2014), and we could find no other examples of loci encoding both mRNAs and snoRNAs, and
thus we wondered about the validity of these seemingly conflicting annotations. To resolve these
conflicts, we searched for sequences similar to the cDNA sequence of AT2G35750 in 6
representative Brassicaceae genomes (Arabidopsis thaliana, Arabidopsis lyrata, Capsella
rubella, Brassica rapa, Eutrema salsugineum, and Aethionema arabicum). Our homology search
returned multiple sequences in each species except Eutrema salsugineum, where only one
homolog was identified. To determine if a conserved open reading frame is present in these
homologs, we retrieved the surrounding genomic regions (see methods) and aligned these
regions to AT2G35750. None of the returned sequence homologs from other Brassicaceae
species contained an open reading frame capable of producing the protein predicted from \textit{AT2G35750} (Alignment: Supplemental Figure 1). Additionally, we searched four independent peptide atlases (Castellana et al. 2008; van Wijk et al. 2021; Baerenfaller et al. 2008; Bairoch et al. 2005) as well as ribosome profiling experiments (Lee and Bailey-Serres 2019) and found no evidence for the presence of translation or protein products from \textit{AT2G35750}. Thus, we conclude that \textit{AT2G35750} is likely not a bona fide protein coding gene, or its protein products are enigmatic. Regardless, the sequence conservation observed for \textit{AtHID1} is independent of \textit{AT2G35750} since the latter gene appears to be a recently emerged orphan that is not present in other species of the family.

Figure 1: \textit{HID1} overlaps an annotated sense mRNA and the paralogous \textit{HID1B} non-coding RNA cluster is upstream. (A) The genomic region of \textit{HID1} with annotated genes. The T-DNA insertion is salk\_017318. (B) The \textit{HID1B} lncRNA cluster is approximately 2,000 base pairs upstream of \textit{HID1} and has a similar makeup of non-coding RNA clusters as \textit{HID1}.
**AtHID1B is a Nearby Paralog of AtHID1 that Resulted from a Tandem Duplication**

While searching for the overlapping annotated protein coding (AT2G35750) sequence across Brassicaceae, we discovered that Arabidopsis contained a highly sequence similar homolog less than 2,000 basepairs (bps) away from the AtHID1 polycistronic cluster (Figure 1B). This nearby homolog (AT2G35743) is annotated as a pseudogene of the protein coding gene AT2G35750. Because HID1 is entirely overlapped by AT2G35750, we wondered whether an AtHID1 paralog also existed within the pseudogene region. We searched for a sequence similar to AtHID1 and recovered a sequence with 90.7% identity that lies partially within the annotated pseudogene. We referred to this paralog as AtHID1B (AT2G08760).

The predicted structure of AtHID1 has four stem-loops (SL1–4), and SL2 and SL4 are critical for HID1 function (Wang et al. 2014). We generated a pairwise alignment of AtHID1 and AtHID1B and found that SL1, SL2, and SL4 all had nucleotide identities > 94%, while SL3 only shared 78% nucleotide identity. Not surprisingly, the structure predicted for AtHID1B (Figure 2B) contains four stem-loops suggesting that AtHID1A and AtHID1B could function redundantly.

We wondered if the two paralogs resulted from a tandem duplication. We analyzed the organization of genes from both genomic regions and found that the two are nearly identical in gene order. AtHID1B is the first unit of a non-coding RNA cluster with two annotated snoRNAs adjacent to AtHIDB, as well as a 445 nt novel transcribed region at the end of the cluster (AT2G00880, Figure 2A). This transcribed region contains an open reading frame that could encode a polypeptide of 69 amino acids. However, like AT2G35750, conservation analyses, peptide atlases, and ribosome profiling datasets did not yield evidence of the presence of a conserved open reading frame, polypeptide, or interaction with ribosomes. Overall, our pairwise alignment of the two regions showed that they are remarkably similar in sequence and gene structure (997 bp alignment length, 59.7% pairwise identity; Figure 2A). The major difference between the two regions is the presence of a “novel transcribed region” (AT2G00880) in the AtHID1B region while the AtHID1 region has a final snoRNA gene. Unsurprisingly, this aligned region has a much lower identity (38% pairwise identity) relative to the rest (72%).
Figure 2: HID1 has tandemly duplicated to form a highly conserved polycistronic non-coding RNA cluster. (A) Alignment of the HID1 polycistronic region with the HID1B region. The top identity bar represents the sequence similarity between the two regions. Green means identical while white gaps mean differences. Sequence bars are colored according to differences between the regions. SL1-4 represents the predicted stem-loops from Wang et al. (2014). (B) Predicted structures using UNAFold of HID1 (left) and HID1B (right). Stem loops are colored for clarity and named according to Wang and colleagues. (C) Long read RNA-seq reads aligned to HID1 (top) and HID1B (bottom).

We finally wondered whether the AtHID1B region generated a polycistronic transcript, as Wang et al. (2014) described for the AtHID1 containing region. We interrogated our own previously published long read Nanopore cDNA sequencing (PRJNA765684) to determine whether we could find evidence that AtHID1 and/or AtHID1B are contained within polycistronic transcripts. We found that each IncRNA is part of a longer transcript which included the overlapping mRNA/pseudogene, the smaller snoRNA clusters, and the novel transcribed region in the case of AtHID1B (Figure 2C). Curiously, our long-read RNA-seq transcript assembly suggested that the locus containing AtHID generates a transcript ~400 nucleotides longer than the presumed last unit of the polycistron. In contrast, the assembled transcripts for the AtHID1B locus ended within the last annotated unit of the polycistron. We conclude from the long-read data along with the alignment and gene structure results that AtHID1 and AtHID1B are the result of a tandem duplication event that yielded two copies of the IncRNA as well as associated snoRNAs.


tAtHID1 and AtHID1B have Distinct Expression Patterns

We characterized the expression dynamics of AtHID1 and AtHID1B, as well as the direct target of AtHID1, PIF3, by re-processing two sets of Arabidopsis RNA-seq data. The first set we used was derived from the expression atlas of Klepikova et al. (2016), which contains 69 different tissues and developmental stages for Arabidopsis, and thus represents one of the most thorough expression atlases in plants. The second set we examined was the RNA-seq dataset
(GSE57806) was generated by Wang et al. (2014) during their characterization of *AtHID1*; they extracted RNA from five-day old wild-type and *Athid1* seedlings grown under red light.

Across tissues and developmental stages, *AtHID1* and *AtHID1B* are relatively similar in their expression patterns (Figure 3A, Pearson correlation coefficient, *r* = 0.52). *AtHID1* was expressed at its highest levels in seeds and seedlings (1.07 and 1.46 standard deviations [SDs or Z-score] above their average, respectively), while *AtHID1B* was expressed highest in roots and seedlings (1.49 and 1.69 SDs above average, respectively). Hence, both paralogs were expressed in seedlings, but each paralog had a unique domain of expression. These results suggest that the expression patterns of the paralogs have diverged from each other post-duplication, consistent with the processes of either sub- or neofunctionalization. *PIF3* achieved its highest expression levels in stem, leaf, and seedling, and appeared to be repressed in roots, stamens, and anthers. The expression pattern of neither paralog was correlated with the expression pattern of *PIF3* (*AtHID1*-*PIF3*, *r* = 0.009; *AtHID1B*-*PIF3* *r* = -0.03).

The regulatory relationship between *AtHID1* and *PIF3* was detected in 5-day old seedlings grown under continuous red light (Wang et al. 2014), and thus we focused more detailed analyses of expression on these tissues and conditions. Additionally, because seedlings have numerous diverse cell types and tissues, we analyzed four datasets from one-day old white-light grown seedlings (hypocotyl, cotyledon, shoot apical meristem, and root) (Klepikova et al. 2016) to determine in which tissue type these genes may be acting. We detected low levels of expression for *AtHID1* in the *Athid1* mutant (Figure 3B), consistent with the findings of Wang et al. (2014). In addition, we detected low levels in the cotyledons of one-day old white light grown seedlings (Figure 3B). The maximum expression level for both paralogs was in seedling meristem, where *PIF3* expression was low (Figure 3B). Interestingly, the expression of *AtHID1B* did not appear to be affected by the T-DNA insertion adjacent to the *AtHID1* polycistron encoding region in *Athid1* seedlings, suggesting that the transcription of the two regions is independently controlled. Taken together, our results indicate partially overlapping and distinct expression domains for *AtHID1* and *AtHID1B*, and that the expression of both loci is anti-correlated with the expression of *PIF3* in seedling meristems.
Figure 3: *HID1B* is expressed in distinct tissues compared to *HID1* and both transcripts are anti-correlated with *PIF3* in the meristem. (A) A heatmap of gene expression for *HID1*, *HID1B*, and *PIF3* in broad tissue categories from Klepikova et al. (2016). Female floral organs correspond to stigmas, ovules, and carpels. Male floral organs correspond to anthers and stamens. Z-score represents the number of standard deviations expression is from the mean. Expression was normalized using a variance stabilized transformation. (B) A line plot showing the direction and magnitude of expression change between various seedling tissues and treatments. The first four categories on the x-axis are datasets from Klepikova et al. (2016), the last two are datasets from Wang et al. (2014) in the *HID1* study. Dots represent the variance stabilized transformation value averaged between replicates.
**HID1 has Experienced Extensive Duplication and Loss Across Brassicaceae**

We identified highly similar paralogs of *HID1* in Arabidopsis likely resulting from a tandem duplication event, and thus we searched 18 publicly available Brassicaceae genomes for *HID1* homologs. We aligned the sequences recovered in our similarity searches (alignment length: 266 bp, pairwise identity: 80.9%) and subjected them to phylogenetic analysis (Figure 4A, all species relationships mentioned in this study can be found in Supplemental Figure 2). Our analyses revealed a complex history of duplication and loss across the Brassicaceae. We identified at least two *HID1* homologs in every sampled species from lineage I of the family (Beilstein et al. 2010). In the clade the included *Capsella* species and *Camelina sativa* we observed two well-supported sub-clades. One sub-clade contained *HID1* paralog from each of the three *Capsella* species with the four copies identified in *Camelina sativa*, indicating that the duplication giving rise to these copies occurred before the divergence of *Capsella* and *Camelina* (i.e. there was likely one *HID1* copy prior to the divergence of the two lineages ~10 million years ago (Mandáková et al. 2017)) and that *Camelina sativa* subsequently lost one of those duplicates. Meanwhile, each *Capsella* species underwent at least one tandem duplication event generating additional copies in these species (Figure 4A and B).

In the *Arabidopsis* clade of *HID1* sequences, we observed *Arabidopsis lyrata* and *A. halleri* (which are more closely related to each other than either is to *A. thaliana*, supplemental Figure 2) experienced a duplication event prior to their divergence yielding two copies in each species. These four *HID1* homologs were sister to *AtHID1*, while *AtHID1B* fell outside this clade. This suggests that *HID1* likely duplicated early in the evolution of the genus *Arabidopsis*, but the *AtHID1B* paralog was subsequently lost in *A. lyrata* and *A. halleri*. Taken together these results indicate that the common ancestor of lineage I had a single *HID1* locus, and tandem duplications in different clades of lineage I produced the diversity of *HID1* loci currently present in these species. This includes recent lineage specific duplications and older duplications.

The duplication and loss history of lineage II is less clear. *Brassica rapa* and *B. oleracea* each have at least three copies of *HID1*, meanwhile we identified at least five copies of *HID1* in *B. nigra*. These species of *Brassica* were specifically chosen they are the three major diploid progenitors of the many vegetable and oilseed crops in the genus (corner species of the Triangle...
of U (Nagaharu 1935; F. Cheng, Wu, and Wang 2014)). All Brassica species share a whole genome triplication event that occurred ~16–8 million years ago (Kagale et al. 2014; Lysak et al. 2005). While each of the aforementioned Brassica species returned to a diploid status, they retain three syntenic sub-genomes, and thus a higher gene copy number than other Brassicaceae that did not undergo the triplication event (S. Liu et al. 2014; Waminal et al. 2016; Park et al. 2021; Perumal et al. 2020). In B. rapa, we identified three HID1 paralogs, each on different chromosomes. Hence, the paralogs present in these species did not result from tandem duplication events but were more likely the result of past polyploidy events. Both B. oleracea and B. nigra also showed evidence of tandem duplications for at least one set of HID1 paralogs since some HID1 copies are present together on the same chromosome (Figure 4A and B).

Regarding other lineage II species, we identified a single HID1 copy in Eutrema salsugineum and Schrenkiella parvula and two copies in Thlaspi arvense, Caulanthus amplexicaulis, and Sisymbrium irio. For these latter three lineage II species, the paralogs are located on different chromosomes, suggesting that they were not the result of tandem duplication. Similar to lineage I, we can confidently conclude that a single HID1 copy was present in the common ancestor of lineage II, and that the diversity currently observed was generated by subsequent duplication events during the expansion of the lineage.
**Figure 4:** *HID1* has duplicated extensively across the Brassicaceae. (A) Maximum-likelihood phylogenetic tree of *HID1* sequences identified in various Brassicaceae. Lineage I is colored in blue, and lineage II is colored in orange. (B) Diagram showing which chromosomes the *HID1* sequences identified in (A) are located on. *HIDs* on the same chromosome are likely to have resulted from tandem duplication. Distances between *HIDs* is arbitrary.

**HID1 is Maintained Across Angiosperms with Strong Selection for Inferred Structural Features**

We also searched for *HID1* homologs in genomes of select monocot and eudicot species. These major lineages within angiosperms diverged from each other ~200 million years ago (Wolfe et al. 1989). Our sequence similarity searches returned 44 sequences from 18 species. We aligned these sequences with six *HID1* sequences representing the diversity present in Brassicaceae. We aligned the sequences recovered in our similarity searches (alignment length: 284 bp, pairwise identity: 58.9%) and subjected them to phylogenetic analysis (Figure 5A). Across monocot and eudicot *HID1* homologs, we observed the greatest sequence identity in the 3’ region of *HID1* homologs relative to the rest of the gene; this region encodes SL4, (SL4
pairwise identity: 79.4%, pairwise identity outside SL4: 49.1%). Sequences from eudicot species fell within three clades (Figure 5A). Two of these clades were modestly well-supported: Brassicaceae + Cleome + Lotus + Medicago + Glycine (bootstrap support [bs] = 66%), and Manihot + Populus (bs = 68%). In addition, a third poorly supported clade (bs < 50%) contained sequences from Lotus, Medicago, and Glycine. All of the sequences we recovered from monocots formed a modestly supported monophyletic group (bs = 62%). Within monocots, sequences from Poaceae (grasses) were strongly monophyletic (bs = 98%), and we recovered evidence that \textit{HID1} duplicated early in the evolutionary history of the family, with subsequent duplications yielding additional copies of \textit{HID1}. In particular, relatively recent lineage specific duplications occurred in Zea mays (maize). In sum, our phylogenetic analysis of angiosperm \textit{HID1} sequences revealed a dynamic and complex evolutionary history characterized by both ancient and more recent duplication events. Strikingly, all the retained duplicates from the genomes we sampled occurred after the divergence of monocots and eudicots, suggesting that loss may be more common than retention for \textit{HID1}.

Using our angiosperm alignment of \textit{HID1} (Supplemental Figure 3), we generated consensus RNA structures to infer conserved structural elements that may have been present in the common ancestor of extant Brassicaceae, eudicots, and monocots (Figure 5). Overall, we observed structures that confirm the results from our alignment; SL4 is a driving factor in the sequence and structural conservation of \textit{HID1}. Every consensus structure possessed a 3’ stem-loop which corresponds to the critical structural element in \textit{AtHID1}. By analyzing the evolutionary history of \textit{HID1} across various levels of flowering plants, we uncovered a history of duplication and loss of the \textit{HID1} locus with apparently strong selective pressure for maintenance of a potentially critical stem-loop structure.
Figure 5: *HID1* has been duplicated and lost across flowering plants with selective pressure to maintain a 3’ stem-loop. (A) Maximum-likelihood phylogenetic tree of *HID1* in select flowering plants. (B) Predicted RNA structure based on the *Arabidopsis thaliana* *HID1* sequence alone (top left), the sequence alignment of Brassicaceae *HID1*s (top right), the sequence alignment of all dicot *HID1*s (bottom left), and the sequence alignment of all *HID1*s in this tree representing angiosperms (bottom right).

Discussion

In this study, we provided pertinent genomic, transcriptomic, and comparative context for future analyses of the *HID1* lncRNA along with its duplicates. *HID1* is the first characterized lncRNA with a role in photomorphogenesis through physical interaction with the *PIF3* promoter (*in trans*). Successful repression of *PIF3* under continuous red light represents another aspect of plant biology mediated by lncRNAs.

We uncovered a paralog of *AtHID1* in *A. thaliana* (*AtHID1B*) that is remarkably sequence and structurally conserved. This paralog has a semi-distinct expression pattern from *HID1*, with higher expression in the roots, possibly signifying retention due to tissue specific sub- or neofunctionalization. *PIF3* exhibits its lowest expression level in roots, and thus one intriguing
hypothesis is that AtHID1B may contribute to PIF3 repression in roots. It seems likely that AtHID1B would also have the ability to physically interact with the promoter of PIF3 due to the nearly identical sequence and predicted structures of AtHID1 and AtHID1B.

Our homology searches and phylogenetic analyses of HID1 in Brassicaceae and across flowering plants yielded multiple, highly similar, HID1 homologs for nearly all the species we examined. These paralogs were the result of largely independent, species-specific duplication events. The sequence encoding SL4 of HID1 is more highly conserved compared to other regions of the gene, potentially underscoring its functional significance. Inferred consensus structures representing nodes in our tree indicated that the common ancestor of major angiosperm clades all could have formed SL4. It is currently unclear what aspects of HID1 function can be attributed to SL4. Because SL4 shares little homology with the PIF3 promoter, it may serve as a binding platform or scaffold for the recruitment of repressive machinery. Regardless, these questions concerning AtHID1 and AtHID1B are likely to provide important insights into the mechanisms of IncRNA function.

One outstanding question remaining is whether AtHID1 and AtHID1B are acting as snoRNAs, or whether they are IncRNAs that were mis-annotated as snoRNAs. SnoRNAs are functional non-coding RNAs that are usually ~50-300 nt in length and usually guide chemical modifications of rRNAs (Dieci et al. 2009; Kufel and Grzechnik 2019). Multiple lines of evidence suggest that AtHID1 and AtHID1B are in fact snoRNAs. The first is that they are part of polycistronic snoRNA clusters. Polycistronic clusters of snoRNAs are the predominant makeup of snoRNA genes in Arabidopsis and Oryza sativa (Dieci, Preti, and Montanini 2009). Additionally, plants have substantially more near-identical snoRNA copies in their genomes relative to Drosophila (Dieci, Preti, and Montanini 2009). While snoRNA duplicates may have arisen from whole genome duplications, the group which characterized AtHID1 noted in a previous study that over 50% of snoRNA duplicates were generated through tandem duplication in rice (T.-T. Liu et al. 2013). The AtHID1 and AtHID1B cluster of snoRNAs seem to be perfect examples of these past findings. Additionally, very few IncRNAs are known to have strong sequence conservation across flowering plants. Even the best characterized plant IncRNAs, the telomerase RNA, as well as COOLAIR and COLDAIR have relatively little sequence conservation across flowering plants, even though they are performing similar or identical functions (Fajkus et al. 2019; Jiao et al. 2019).
2019; Hawkes et al. 2016). Meanwhile we identified numerous HID1 homologs across angiosperms.

There are two main lines of evidence that AtHID1 (and by extension, AtHID1B) is a bona fide lncRNA. First, Rfam does not return any known family of snoRNA, or other housekeeping RNA, when querying any of the Brassicaceae HID1 homologs that we identified in this study. However, we can identify the characteristic motifs of H/ACA snoRNAs in all our HID1 sequences. These snoRNAs are characterized by having the H box (ANANNA, N = A/T/C/G) and the ACA sequence (Falaleeva and Stamm 2013). Albeit the H/ACA motifs are likely to appear in many biological sequences by chance. The second line of evidence that AtHID1 is acting as a bona fide lncRNA comes from its described physical interaction with the PIF3 promoter (Wang et al. 2014). To our knowledge, there are no known examples of snoRNAs involved in the direct transcriptional repression of a gene. Wang et al. (2014) provided extensive genetic evidence, as well as molecular evidence for the physical HID1-PIF3 promoter interaction, and further that HID1 repressed the expression of PIF3 under continuous red light.

Future studies of HID1 should take the evolutionary history of the gene, particularly its proclivity to duplicate. Additionally, as genome editing becomes more efficient, precise, and approachable with CRISPR, mutational studies interrogating the function of AtHID1B and other homologs could yield important clues about the function of HID1. Wang et al. (2014) utilized a T-DNA insertion mutant which inserted ~365 base pairs downstream of HID1, into the last snoRNA of the polycistronic transcript (Figure 1A). This insertion merely knocked down the expression of HID1 (Wang et al. 2014). Analysis of null mutants could recover additional functions distinct from PIF3 repression. These future studies will help to elucidate the functional relationship between the two paralogous transcripts and provide a novel example of the fates of lncRNAs post-duplication.

Materials and Methods

Sequence Identification

Sequence homologs of the Arabidopsis thaliana HID1 IncRNA were identified using CoGe BLAST (Lyons et al. 2008) with the following parameters: E-Value < 1e-10, Word size: 8, Gap
Costs: Existence: 5 Extension: 2. All Brassicaceae genomes available on CoGe and Phytozome (Goodstein et al. 2012) were queried. Two distinct groups of sequences were returned from the BLAST analysis, hits that had a quality score ranging from 66.1%-100% and hits with a quality score ranging from 19.5%-30.1%, both groups of hits seemed to have representation from all sampled Brassicaceae. Only sequences from the higher quality score group were kept for phylogenetic analyses. All sequences were padded on the 5’ and 3’ ends with 600 base pairs of surrounding genomic sequence.

**Phylogenetic Analysis**

The sequences generated above were imported into Geneious Prime (Geneious Prime 2021.2.2) and used in the Map to Reference tool against the *Arabidopsis thaliana* *HID1* sequence with the Geneious mapper, highest sensitivity setting, and the fine-tuning setting: move many gaps. This alignment was manually edited to trim the excess sequences that aligned outside of the *Arabidopsis thaliana* *HID1* region. This edited alignment was re-aligned using MUSCLE (Edgar 2004) with the default parameters as a Geneious plugin. This alignment was used as input for RAxML Next Generation v 1.0.3 (Kozlov et al. 2019) specifying the -all, -model GTR+G, and -bs-trees 500 options and parameters.

**Pseudo-Species Tree Generation**

To aid the reader’s understanding of organismal relationships investigated in this study, we generated a gene tree of the largest subunit of Nuclear RNA Polymerase 2 (*Arabidopsis thaliana* *NRPB1; AT4G35800*). The *AT4G35800* cDNA sequence was used as input for CoGe BLAST as described above. The highest quality hit of each species shown in Figure 5 was padded with 600 base pairs of sequence on the 5’ and 3’ ends. Sequences were mapped, aligned, and used to generate a phylogeny as above.

**Transcriptional Analysis**

RNA-seq processing and analysis was performed as in (Palos et al. 2021).
Supplemental Figure 1: The annotated mRNA overlapping Arabidopsis thaliana HID1 does not contain a conserved open reading frame. The start codon of AT2G35750 is boxed in blue on the previous page. The start codon is boxed in red on this page.

Supplemental Figure 2: An approximate organismal phylogeny of species analyzed in this study for relationship reference.

Supplemental Figure 3: Alignment of HID1s across flowering plants to generate the phylogenetic tree in Figure 5. Stem-loops are annotated on A. thaliana HID1. The identity bar height and color correspond to the level of sequence conservation. Nucleotides within the alignment are shaded according to similarity.
REFERENCES


Cabili, Moran N., Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv


Dieci, Giorgio, Milena Preti, and Barbara Montanini. 2009. “Eukaryotic snoRNAs: A Paradigm


Heydarian, Zohreh, Margaret Gruber, Bernard R. Glick, and Dwayne D. Hegedus. 2018. “Gene Expression Patterns in Roots of Camelina Sativa With Enhanced Salinity Tolerance Arising From Inoculation of Soil With Plant Growth Promoting Bacteria Producing 1-Aminocyclopropane-1-Carboxylate Deaminase or Expression the Corresponding acdS


Kung, Johnny T., Barry Kesner, Jee Young An, Janice Y. Ahn, Catherine Cifuentes-Rojas, David Colognori, Yesu Jeon, et al. 2015. “Locus-Specific Targeting to the X Chromosome Revealed by the RNA Interactome of CTCF.” Molecular Cell 57 (2): 361–75.


OF SALT STRESS ON CONCENTRATION OF NITROGEN, PHOSPHORUS, POTASSIUM, CALCIUM, MAGNESIUM, AND SODIUM IN LEAVES, SHOOTS, AND ROOTS OF SIX OLIVE CULTIVARS.” *Journal of Plant Nutrition* 25 (11): 2457–82.


Xu, Yong-Chao, Jie Zhang, Dong-Yan Zhang, Ying-Hui Nan, Song Ge, and Ya-Long Guo. 2021. “Identification of Long Noncoding Natural Antisense Transcripts (lncNATs) Correlated


APPENDIX A

IDENTIFICATION AND FUNCTIONAL ANNOTATION OF LONG INTERGENIC NON-CODING RNAS IN BRASSICACEAE
Title:
Identification and Functional Annotation of Long Intergenic Non-coding RNAs in Brassicaceae

Kyle Palos¹, Anna C. Nelson Dittrich², Li’ang Yu², Jordan R. Brock³, Hsin-Yen Larry Wu⁴, Ewelina Sokolowska⁵, Aleksandra Skirycz², Polly Yingshan Hsu⁴, Eric Lyons¹, Mark A. Beilstein¹, and Andrew D. L. Nelson²,⁶

¹The School of Plant Sciences, University of Arizona, Tucson, AZ.
²The Boyce Thompson Institute, Cornell University, Ithaca, NY.
³Department of Horticulture, Michigan State University, East Lansing, MI.
⁴Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI.
⁵Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany.

⁶Address correspondence to:
Dr. Andrew D. L. Nelson
The Boyce Thompson Institute
533 Tower Road, Ithaca, NY, 14850.
Phone (979-220-0641)
an425@cornell.edu

Current email address: KP: kpalos@email.arizona.edu, ACND: an426@cornell.edu, LY: ly76@cornell.edu, JRB: brockjor@msu.edu LW: larrywu@msu.edu, ES: ESokolowska@mpimp-golm.mpg.de, AS: as4258@cornell.edu, PH: pollyhsu@msu.edu, EL: elyons.uoa@gmail.com, MAB: mbeilstein@arizona.edu

Running title: LincRNAs in Brassicaceae
Abstract:

Long intergenic noncoding RNAs (lincRNAs) are a large yet enigmatic class of eukaryotic transcripts that can have critical biological functions. The wealth of RNA-seq data available for plants provides the opportunity to implement a harmonized identification and annotation effort for lincRNAs that enables cross-species functional and genomic comparisons as well as prioritization of functional candidates. In this study we processed >24 Tbp of RNA-seq data from >16,000 experiments to identify ~130,000 lincRNAs in four Brassicaceae: Arabidopsis thaliana, Camelina sativa, Brassica rapa, and Eutrema salsugineum. We used Nanopore RNA-seq, transcriptome-wide structural information, peptide data, and epigenomic data to characterize these lincRNAs and identify functional motifs. We then used comparative genomic and transcriptomic approaches to highlight lincRNAs in our dataset with sequence or transcriptional conservation, including lincRNAs transcribed adjacent to orthologous genes that display little sequence similarity though they may function as transcriptional regulators. Finally, we used guilt-by-association analyses to assign putative functions to lincRNAs within our dataset based on similar expression patterns to protein coding genes of known function. LincRNAs with Brassicaceae-conserved putative miRNA binding motifs, sORFs, or abiotic-stress modulated expression are a few of the annotations that will prioritize and guide functional analyses into this cryptic portion of the transcriptome.

Key words: RNA-sequencing, long intergenic noncoding RNAs, comparative genomics, transcriptomics, RNA evolution, genome annotation, RNA-based transcriptional regulation, RNA structure
Introduction:

As genomic and transcriptomic analyses have become more prevalent, it has become clear that genomes are not solely composed of protein-coding genes, housekeeping RNAs, and transposable elements. One particularly important set of findings came from the Human ENCODE (ENCODE Project Consortium 2012) project where it was discovered that over 60% of the human genome is transcribed at some point in development into long non-coding RNAs (lncRNAs). The term “lncRNA” refers broadly to a class of transcripts united by two key features: a length > 200 nt and poor protein coding potential (i.e., low likelihood of being translated). The term lncRNA is further subdivided into transcripts that are natural antisense (NAT-lncRNAs), intergenic (lincRNAs), sense overlapping (SOT-lncRNAs), and intronic (int-lncRNAs). Each of these classes of lncRNAs appears in analyses of RNA-seq data because they share features with mRNAs (e.g., they are capped, polyadenylated, and often multi-exonic) (Guttman et al., 2009). Most lncRNAs were missed or ignored in earlier expressed sequence tag (EST)-based screens because of their low or tissue-specific expression and lack of open reading frames. However, RNA sequencing data from more than 37,000 experiments reflecting ~60 different tissues under different experimental and developmental conditions led to the identification of > 100,000 high confidence lncRNAs in humans (Volders et al., 2013; Zhao et al., 2019; Zhao et al., 2020).

In contrast to proteins, which were the focus of study long before the genomes from which they are encoded were sequenced, an appreciation for the abundance and varied roles of lncRNAs has primarily emerged along with the accumulation of sequencing data. As a result, the catalog of functionally characterized lncRNAs is limited, both in number and in diversity of organisms where they have been annotated (Statello et al., 2020; Seifuddin et al., 2020; Chekanova JA, 2021). Moreover, the extent to which functionally characterized lncRNAs are archetypal across plants, animals, and fungi is unknown. Not surprisingly, lncRNA identification and functional characterization lags far behind similar efforts for proteins, representing a fundamental gap in our understanding of how genomes operate.

Findings from across eukaryotes serve to illustrate the importance of lncRNAs to genome stability and regulation. Prominent mammalian examples include the telomerase RNA component (TERC), a scaffolding RNA that is crucial for chromosome maintenance (Feng et al. 1995); XIST, a guide RNA responsible for X chromosome inactivation (Brown et al. 1992); and HOTAIR, a developmentally-linked signaling RNA (31). In Arabidopsis, TERC has been characterized, with
sequence and structural homologs present across the plant lineage, highlighting the potential for lncRNA conservation over long evolutionary timescales (Dew-Budd et al. 2020); (Fajkus et al. 2019); (Song et al. 2019). Most other lncRNAs functionally characterized in plants, such as COOLAIR, ELENA1, SVALKA, MAS, APOLO, and HID1, change expression or function in response to environmental cues, and can thus be classified as environmental sensors (Csorba et al. 2014); (Seo et al. 2017); (Kindgren et al. 2019); (Zhao et al. 2018; Ariel et al. 2020; Y. Wang et al. 2014). These examples reflect the myriad of different mechanisms by which lncRNAs play important biological roles in plants, and also likely represent only a small subset of the mechanisms and modes of action of lncRNAs.

One critical factor behind the paucity of functionally described lncRNAs in plants relative to mammalian systems is the lack of annotated candidate lncRNAs available for interrogation. Moreover, across studies where lncRNAs have been annotated there are disparities in the types of transcriptional data analyzed as well as the criteria used for their classification. For example, in Arabidopsis the bulk of annotated lincRNAs are derived from two studies (Amor et al. 2009); (Liu et al. 2012), although other genome-wide examinations have been performed (Moghe et al. 2013); (Y. Wang et al. 2014). The former examined full length cDNA libraries for lack of coding potential, whereas the latter utilized TILING arrays to infer gene structure and transcriptional status. In both cases the maximum allowable open reading frame (ORF) was 100 amino acids (AA) or less. Other lincRNA identification efforts in select angiosperms with genomic data (e.g., GREENC; Gallart et al, 2016), used official genome annotations generated by MAKER (Cantarel et al. 2008) without direct transcriptional evidence, and maximum allowable ORFs of 120 AA. Yet in other plant systems, lincRNA identification efforts are limited to a few tissues or developmental stages (Qi et al. 2013; Moghe et al. 2013; L. Li et al. 2014; Shuai et al. 2014). In sum, the disparity in identification schemes and discordant developmental stages and environmental conditions makes it difficult to make sequence or structural based comparisons within and across species, as is typically done for protein-coding genes.

Here we present a comprehensive and unified annotation of lincRNAs, using criteria established for mammals, across four model or agriculturally important Brassicaceae: Arabidopsis thaliana, Camelina sativa, Brassica rapa, and Eutrema salsugineum. We reprocessed more than 16,000 different publicly available RNA-seq experiments (> 24 Tbp of raw data) and generated our own Oxford Nanopore (ONT) and Illumina RNA-seq data, to identify lncRNAs in each of
these species. We focus primarily on the intergenic class of IncRNAs for evolutionary and technical reasons: the evolution of NAT and SOT- IncRNAs is expected to be heavily influenced by the protein-coding genes they overlap, and the unstranded nature of much of the publicly available RNA-seq data makes confident strand assignment of single exon transcripts difficult. Using transcriptomic, proteomic, epigenetic, and genome-wide RNA-protein interaction data, we examined our lincRNA catalog for features that separate and define lincRNAs from other transcriptional units. We used evolutionary and comparative genomic approaches, leveraging the unique strength of plant polyploidy, to identify conserved lincRNAs among the four species and the rest of the Brassicaceae as well as to identify conserved motifs for functional testing. Finally, we used all of these contextual clues, as well as guilt-by-association techniques, to assign putative function to lincRNAs within our catalog.

Results:
Identification of lincRNAs in four species of Brassicaceae:

We processed all RNA-seq data deposited to the Sequence Read Archive (SRA) at the NCBI [accessed December, 2018] for Arabidopsis thaliana, Brassica rapa, Camelina sativa, and Eutrema salsugineum (hereafter Arabidopsis, Brassica, Camelina, Eutrema) with the goal of detecting the full suite of lincRNAs, including those with low-expression and/or tissue/environmental specificity. We excluded SRAs with epigenetic mutants, degradome experiments (GMUCT and PARE), small RNA-sequencing, and experiments with low sequencing depth (fewer than 1 million quantified/mapped reads; Figure 1). In addition to publicly available short read RNA-seq, we also performed Oxford Nanopore Technology (ONT) PCR-free cDNA sequencing on three tissues (10-day seedlings, 4-week mature rosettes, and open flowers) for all four species. We used previously developed workflows (Peri et al. 2019) utilizing the CyVerse computational infrastructure (Merchant et al. 2016) to map, in high throughput, ~24 terabases of RNA-sequencing data associated with 16,076 experiments (listed in Supplemental File 1). We then identified initial candidate lincRNAs using the Evolinc computational pipeline (Nelson et al. 2017). We filtered these candidates based on a set of heuristic filtering steps similar to those used by Cabili et al, 2011 to identify the “gold standard” set of human lincRNAs. Transcripts were included among our high confidence (HC) lincRNAs if they met the criteria for one of the filters.
Figure 1: Basic identification and characterization of lincRNAs in each of the four target Brassicaceae. A) Number of experiments processed for each experiment using the RMTA and Evolinc pipelines in CyVerse’s cloud computing infrastructure. B) The metrics used for lincRNA additional hierarchical filtering. Note, lincRNAs only had to pass one additional filter to be considered a high confidence lincRNA. C) The number of identified lincRNAs

The first filter sorted as HC any lincRNAs identified by ONT cDNA sequencing because of the potential for capturing full-length lincRNA transcripts using ONT technology (Seki et al. 2019). ONT cDNA sequencing from across the three tissues yielded 200 unannotated (i.e., not present in the Araport 11 annotation) lincRNAs in Arabidopsis, 945 in Brassica, 1,669 in Camelina, and 563 in Eutrema. Our next filter retained lincRNAs as HC if they were multi-exonic. This filter selects for transcripts that are less likely to be artifacts of transcript assembly algorithms (Cabili et al. 2011). By this criterion, 678 Arabidopsis, 12,422 Brassica, 6,200 Camelina, and 1,812
Eutrema multi-exonic lincRNAs were identified and annotated as HC (Figure 1C). Some previously identified and characterized lincRNAs are mono-exonic (Lorenzi et al. 2021; Sun and Ma 2019; West et al. 2014). We designated as HC mono-exonic lincRNAs that met one of the following filtering criteria: 1) the transcript was conserved in sequence and syntenic at least one of 10 Brassicaceae genomes examined (See Materials and Methods); 2) the transcript length was > 500 nts; or 3) the transcript was expressed above 3 transcripts per million (TPM) in at least 10 RNA-seq experiments. All Evolinc candidate lincRNAs that did not pass these filters were retained within our dataset as low confidence (LC) lincRNAs, since there is greater potential for these transcripts to be artifactual. In total, we identified 9,244 Arabidopsis, 58,155 Brassica, 13,163 Camelina, and 20,744 Eutrema high confidence lincRNAs (HC-lincRNAs) (Figure 1B), while 8,867, 11,977, 7,432, 4,893 lincRNAs were categorized as LC-lincRNAs, respectively (Supplemental File 2).

The potential exists for misassembled or fragmented mRNAs or mRNAs with poorly annotated extensions at the 5’ or 3’ end to be misclassified as lincRNAs. To determine the frequency at which we misclassified these transcripts as lincRNAs, we compared independently assembled transcriptomes from Illumina short read and ONT long read derived lincRNA datasets, searching for short read derived “lincRNAs” that mapped to the 3’ or 5’ extension of a coding transcript from our ONT sequencing data. Using this approach, we identified 39 lincRNAs in Arabidopsis that shared at least 1 ONT sequencing read on the same strand as a neighboring mRNA (out of 2,370 lincRNAs for which we obtained ONT coverage ≥ 1). Of the 39 lincRNAs with overlapping sequence reads, only 2 appeared to be mRNA extensions (Supplemental Figure 1A). The other 37 lincRNAs shared sequencing reads due to misassembly or genomic DNA contamination in the sequencing (Supplemental Figures 1B and 1C, asterisks), or are larger variants of Araport IncRNAs. In general, we identified strong agreement between ONT and Illumina derived lincRNA transcript models (Supplemental Figure 1D), suggesting the depth of Illumina sequencing used here was more than sufficient to overcome misassembly common for lowly expressed transcripts. Given the low rate (1.64%) of false positives, we remain confident that the transcripts we have identified are indeed independently transcribed elements within the Arabidopsis genome.
Harmonizing Arabidopsis lincRNA annotations from multiple sources

We next assessed how many of the previously identified Arabidopsis lncRNAs were expressed in our assembled RNA-seq data. Given the comprehensive nature of our dataset, we presumed that a previously annotated lncRNA was misannotated if we did not observe expression above 1 TPM in at least 10 Arabidopsis RNA-seq datasets (out of all Arabidopsis RNA-seq examined). Araport11 lncRNAs comprise 2,455 “lnc_RNAs”, 286 “ncRNAs”, and 726 “novel transcribed regions”. To create a uniform dataset of lincRNAs, we filtered out transcripts that did not fit the most basic definitions of a lincRNA (over 200 nt and not overlapping a protein coding gene), and for which we did not observe expression. Of the 2,455 “lnc_RNAs”, 401 were removed because they overlapped a protein-coding gene, 157 were relabeled as low confidence (LC-Araport) due to lack of sufficient expression levels based on our expression filtering mentioned above (> 1 TPM in 10 experiments). However, we did observe low levels of expression (> 0.1 TPM) for some of these "LC-Araport" lincRNAs in various tissue expression atlases, stress datasets, or our Nanopore sequencing (Table 1). In total, we confirmed 1,897 Araport lnc_RNAs to be HC-lincRNAs. For the 286 annotated “ncRNAs”, 189 (66%) passed the length, intergenic, and expression criteria, and thus were also considered HC-lincRNAs. Finally, we analyzed the novel transcribed regions and first assessed coding capacity. We treated these transcripts to the same set of filters as our lincRNA dataset (ORF < 100 AA, longer than 200 nts, poor coding potential), which resulted in 571 NTRs annotated as lincRNAs and included in further analyses. In total, we reannotated 2,566 of the 3,467 Araport lncRNAs as HC-lincRNAs (Supplemental File 2), while the remaining 901 loci were reannotated as LC-lincRNAs or were discarded as not fitting the definition of a lincRNA (Supplemental File 7).

Annotating lincRNAs based on presence of sORF, structural, and miRNA motifs

The definition used by the community to distinguish between lincRNAs and proteins is arbitrarily set at a length of 100 AA and thus transcripts annotated as lincRNAs may encode small proteins. For example, a few studies have demonstrated previously annotated lincRNAs that are bound to ribosomes and, in some cases, have identified protein products for lincRNAs (Ji et al. 2015; Hsu et al. 2016; Wu et al. 2019). We used Ribo-seq (Ingolia et al. 2009; Wu and Hsu 2021) and protein mass spectrometry (MS; (Domon and Aebersold 2006)) data from Arabidopsis seedlings (PRIDE: PXD026713) to identify translated short ORFs (sORFs) and protein products within our lincRNAs. Of the 1,172 lincRNAs expressed > 0.1 TPM in seedling tissue, 120
appeared in Ribo-seq data and 38 in MS data, ranging in size from 3-136 amino acids (Figure 2A). There is no correlation between transcript and sORF length (Supplemental Figure 2), but we did observe a tendency for Araport11 (n = 81) HC-lincRNAs to contain longer sORFs than Evolinc-derived lincRNAs (n = 77; p-value 0.046; Figure 2A), likely due to the more restrictive criteria used to annotate the Evolinc derived HC-lincRNAs. LincRNAs containing sORFs have been denoted as such in Supplemental File 2, but as they reflect previously unidentified genes that would otherwise have been called lincRNAs, they were retained as a separate set of transcripts for downstream analyses.
**Figure 2:** Identification of functional motifs within Arabidopsis lincRNAs. A) Distribution of the length of the newly identified sORFs within the Araport (previously identified) and Evolinc (this study) lincRNA populations. The largest and smallest sORFs are labeled (red dot) with the length denoted (in amino acids, AA). B) Distribution of identified structured and protein-bound elements within the total Arabidopsis lincRNA dataset based on PIP-seq data from three different tissues (S= Seedling, R-H = Root with hairs, R-NH = Root with no hairs; see Sager et al., 2015 for more details). The percent of unique mRNAs (blue scale) or lincRNAs (red scale) that overlapped with at least one PIP-seq read are shown. C) Frequency of miRNA binding sites within lincRNA populations from each of the four focal species. Total number of lincRNAs in each of the four species containing a putative miRNA binding site is shown in blue, with unique miRNA motifs shown in grey/green.

We next used publicly available transcriptome-wide protein-interaction profile sequencing (PIP-seq, (Foley et al. 2017; Gosai et al. 2015)) data from roots (hair and nonhair) and seedlings (GEOs GSE58974 and GSE86459) to identify lincRNAs in our dataset for which we can infer protein-dependent and independent RNA structural motifs. Across the three datasets, we identified 397 structured and protein-bound lincRNAs. 135 (34%) of these were present in all three datasets, whereas 195 were restricted to a single cell type/tissue (Figure 2B). Of these cell type or tissue-restricted lincRNAs, 119 were found to be structured in root cells, with the vast majority (103; 26% of structured lincRNAs) only present in non-hair root cells (R-NH; Figure 2B). In contrast, most mRNAs (62%) were found to be structured in all three tissues, whereas only 6% were restricted to non-hair root cells (Figure 2B). Thus, we have evidence for structural motifs within a subset of the Arabidopsis lincRNA dataset, a number that will likely only increase as more PIP-seq data are generated. These lincRNAs have been annotated in Supplemental File 2 and the multiple sequence alignment (MSA) files are available in the CyVerse Data Store (DOI).

Some lincRNAs are known to interact with miRNAs, either in a competitive inhibitory fashion (i.e., miRNA sponge; (Zhang et al. 2019)), or to directly regulate the lincRNA itself (e.g., TAS1A; (Howell et al. 2007; Chen, Li, and Wu 2007)). Using the miRNA binding site prediction tool psRNATarget (Dai, Zhuang, and Zhao 2018) we identified 226 Arabidopsis lincRNAs with at least one putative miRNA recognition site (Figure 2C). Importantly, within this set of lincRNAs we recovered previously characterized miRNA-regulated lincRNAs such as TAS1A and TAS1B. We identified a further 668 miRNA-interacting lincRNAs in Camelina, 2,741 in Brassica, and 1,168 in Eutrema (Figure 2C) (Supplemental File 5). In sum, we used a wealth of public information to improve the genome annotations of four agricultural or model Brassicaceae.

**Fundamental features of Brassicaceae lincRNAs:**

We next examined basic characteristics of our lincRNA datasets with the goal of
identifying features that might improve future lncRNA identification efforts. LincRNAs in all four species have significantly lower GC content relative to protein-coding genes (P value for comparison of lincRNA-mRNA for all species < 2.2e-16, Wilcoxon Signed Rank test; Figure 3A). Additionally, transcript length of lincRNAs are significantly shorter than mRNAs (P value for comparison of lincRNA-mRNA for all species < 2.2e-16, Wilcoxon Signed Rank test; Figure 3B).

Figure 3: Basic sequence characteristics of Brassicaceae lincRNAs. A) % GC content and B) transcript length comparisons of mRNAs and lincRNAs in each of our four focal Brassicaceae. C) Exon length distribution for mRNAs and lincRNAs. All comparisons are significant (P < 2e-16) using a Wilcoxon test with Bonferroni multiple testing correction. CpG DNA methylation and H3K27me3 epigenetic profiles for lincRNAs and mRNAs in D) Arabidopsis and E) Brassica. LincRNAs and mRNAs were separated based on expression levels using associated RNA-sequencing data, when available.
Interestingly, when we looked at multi-exonic lincRNAs and mRNAs, we found that the average length of an exon for lincRNAs was significantly longer than the average length of an exon for mRNAs except in Camelina, where lincRNA exons displayed a similar trend to the other species (P value for all species except Camelina < 2.2e-16, Wilcoxon Signed Rank test; Figure 3C). Finally, we analyzed the distribution of exons in lincRNAs in all four species. LincRNAs in Arabidopsis are mostly mono-exonic (~91.1%), while the lincRNAs identified in the other species have a much more balanced distribution of exon counts. Regardless, lncRNAs in Brassica, Camelina, and Eutrema contain fewer exons on average than the mRNAs in these species (Supplemental Figure 3A).

To better understand the epigenetic mechanisms controlling lincRNA expression, we next examined patterns of epigenetic regulation between lincRNAs and mRNAs. Owing to the wealth of genome-wide epigenetic data in Arabidopsis and Brassica, we identified experiments in both species that allowed us to compare the patterns of CpG DNA methylation and H3K27 trimethylation at lincRNA and protein coding loci (H3K27me3; see methods; Supplemental File 4). We further divided our gene sets based on expression within the tissues examined to better understand the interplay between expression and epigenetic regulation. In Arabidopsis, lincRNA loci show a consistent decrease in CpG methylation across the gene body, a pattern distinct from that of transposable elements and protein-coding loci (Figure 3D; Supplemental Figure 3B). Expressed protein-coding loci show the characteristic dip in CpG methylation at the transcription start site (TSS) and increase near the transcription end site (TES). TEs show elevated CpG methylation across the gene body relative to their surrounding genomic regions. The decrease we detected for Arabidopsis lincRNAs was largely consistent regardless of expression. However, the trend is reversed in Brassica, where protein-coding loci displayed a peak in CpG methylation at the TSS and lncRNAs showed elevated CpG methylation across the gene body (Figure 3E). Expressed protein-coding and lincRNA loci showed decreased levels of H3K27me3 both across the gene body and relative to non-expressed loci in Arabidopsis (Figure 3D) and Brassica (Figure 3E). We also examined H3K9 acetylation, although these data were only available for Arabidopsis (Supplemental Figure 3B). Expressed Arabidopsis protein-coding loci displayed the characteristic peak in acetylation at the TSS and dip at the TES, whereas lincRNA loci displayed an increase in acetylation across the gene body that was positively associated with expression. In general, lincRNAs in Arabidopsis and Brassica are distinguished from protein-coding loci and TEs...
in that they display similar CpG methylation patterns across their gene body to the patterns associated with the TSS of protein-coding loci, a feature that becomes more pronounced with higher expression.

LncRNAs in mammalian systems are often tissue or cell-type specific, and often lowly expressed at the tissue level relative to mRNAs. This has also been observed to a certain extent in plant systems, albeit with far fewer tissue comparisons. Maximum lincRNA expression, in any tissue, was ~10-fold lower compared to mRNAs in all four species (Figure 4A and Supplemental 4A). Tissue specificity (TAU; (Yanai et al. 2005)) was determined based on expressions data from tissue atlases in Arabidopsis ((Klepikova et al. 2016) and Brassica (Tong et al. 2013; Bilichak et al. 2015), as well as from our ONT RNA-seq data. As expected, lincRNAs from all four species were, on average, significantly more tissue-specific than their respective mRNA cohorts (adj-p <2e-16; Figure 4B and Supplemental Figure 4B). We also observed a negative correlation between lincRNA tissue specificity and expression, where more highly expressed lincRNAs tended to be more broadly expressed, a feature that was significantly more pronounced than for mRNAs (p<2.2e-16; Figures 4C and 4D). This negative correlation was observed across multiple tissues (e.g., female reproductive, leaf, and male reproductive; Supplemental Figure 4C), although we did observe tissue-dependent differences, such as high expression associated with high specificity for both lincRNAs and mRNAs in pollen/anther RNA-seq data. The sORF containing lincRNAs displayed expression and tissue specificity values similar to mRNAs (Figures 4A and 4B), further supporting an mRNA assignment. Given this link between lower tissue specificity (broader expression) and coding potential, we more closely examined the Arabidopsis lincRNAs (n = 89) with TAU values lower than the median value for mRNAs (TAU < 0.502; Figure 4B, black box). Based on sequence similarity, these broadly expressed lincRNAs do not appear to be recently pseudogenized protein-coding genes, but, for a subset (n = 61), expression is significantly correlated with a neighboring gene less than 500 bp away (Supplemental Figure 4D). Thus, high tissue specificity and low expression can be considered a defining feature of Brassicaceae lincRNAs and can potentially help to distinguish unannotated sORF containing transcripts.
Figure 4: Expression dynamics of Arabidopsis and Brassica lincRNAs. A) Log2 max TPM for lincRNAs, mRNAs, and sORF containing lincRNAs (Arabidopsis only) using tissue atlas data for the two species. B) Tissue specificity (TAU) for Arabidopsis and Brassica transcripts. The dashed box denotes the 96 Arabidopsis lincRNAs that are below the median TAU value of mRNAs and were inspected further for similarity to protein-coding genes (see text for details). C-D) Correlation between tissue specificity and max expression for Arabidopsis lincRNAs (C) and mRNAs (D) within the Klepikova tissue atlas. E-F) Stacked bar plots describing where lincRNAs or mRNAs are most highly expressed in Arabidopsis (E) and Brassica (F).
In mammalian systems, a large number of lincRNAs are expressed, or show elevated expression, in male reproductive tissues (Hong et al. 2018). This phenomenon is attributed to relaxed epigenetic control within these tissues. We sought to determine if this was also a feature of plant lincRNAs by examining lincRNA expression within the Arabidopsis and Brassica tissue atlases. Approximately 45 and 35% of lincRNAs in Arabidopsis and Brassica, respectively, were most highly expressed in reproductive tissues, with pollen being the predominant source of maximum expression levels (Figure 4E and 4F). A similar percent of mRNAs showed peak expression in reproductive tissues in the two species, suggesting a general transcriptome-wide, instead of lincRNA-specific, phenomenon. Consistent with this transcriptome-wide phenomenon, lincRNAs restricted to pollen were expressed significantly higher than lincRNAs restricted to other tissues (e.g., female reproductive versus leaf tissue; Supplemental Figure 4C, note scales). To aid in exploration of lincRNA and mRNA expression between tissues and experiments, these data have been uploaded to the appropriate BAR eFP Browser (Pro Bart and Zhu 2003), and are explorable through an interactive Clustergrammer (Fernandez et al. 2017) Jupyter notebook binder found at https://github.com/Evolinc/Brassicaceae_lincRNAs (Supplemental Figure 5).

Interestingly, 48% and 60.8% of the complete (HC + LC) Arabidopsis and Brassica lincRNA datasets, respectively, were not expressed above 0.1 TPM in their respective tissue atlas suggesting these lincRNAs are not expressed under “normal” conditions during development. Considering that expression was a requirement for identification, we sought to determine where these “context-specific” lincRNAs (CS-lincRNAs) were expressed. We screened through all of the Arabidopsis and Brassica RNA-seq data looking for experiments of maximal expression. We extracted metadata from those experiments from the NCBI SRA and then grouped lincRNAs into similar categories based on expression (see Materials and Methods). In Arabidopsis, the majority of the CS-lincRNAs showed maximal expression in experiments that performed high-resolution sequencing of root or shoot meristems (n = 5236; Table 1), suggesting these lincRNAs are expressed in very limited cell types. 909 lincRNAs (~4.5%) were found to be expressed under stress (abiotic or biotic) conditions (Table 1). In Brassica, the vast majority of the CS-lincRNAs (n = 40,937; 57.5%) were maximally expressed in sequencing data from recombinant inbred lines (n = 19,097) or hybridization experiments with different Brassica accessions (n = 21,840; Cheng et al., 2016), indicating a high degree of transcriptional variation between genetic backgrounds. We also observed a subset of Arabidopsis lincRNAs (~350) that were only expressed in specific accessions or in crosses between accessions. Finally, 7,407 (10.4%) Brassica CS-lincRNAs were
expressed under stress conditions. To allow researchers to sort lincRNAs based on their own priorities, expression metadata have been assigned to each CS-lincRNA in Supplemental File 2. These data highlight both the extreme tissue specificity possible for lincRNAs, as well as the potential for lincRNAs to be expressed during, and perhaps play a role in, growth and development of recent hybrids.

<table>
<thead>
<tr>
<th>Broad Category</th>
<th>Arabidopsis</th>
<th>Brassica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo-associated</td>
<td>468</td>
<td>143</td>
</tr>
<tr>
<td>Dissected flower tissue</td>
<td>648</td>
<td>77</td>
</tr>
<tr>
<td>Biotic infection</td>
<td>215</td>
<td>NA</td>
</tr>
<tr>
<td>Epigenetic mutants</td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>Root tip or meristem</td>
<td>3,448</td>
<td>NA</td>
</tr>
<tr>
<td>Shoot meristem</td>
<td>1,788</td>
<td>NA</td>
</tr>
<tr>
<td>Mixed accessions/ssp.</td>
<td>363</td>
<td>21,840</td>
</tr>
<tr>
<td>RILs</td>
<td>NA</td>
<td>19,097</td>
</tr>
<tr>
<td>Genetic mutants</td>
<td>2,089</td>
<td>NA</td>
</tr>
<tr>
<td>Circseq</td>
<td>452</td>
<td>NA</td>
</tr>
<tr>
<td>Other</td>
<td>2,773</td>
<td>2,202</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12,291</strong></td>
<td><strong>43,374</strong></td>
</tr>
</tbody>
</table>

**Table 1**: Broad categories of experiment/tissue in which highly context-specific lincRNAs were found to be expressed in Arabidopsis and Brassica.

**Evolutionary features of Brassicaceae lincRNAs**

Evolutionary conservation is often considered a proxy for the relative importance of the function a protein-coding gene, and thus we sought to determine the degree to which lincRNAs from each of the four species were evolutionarily conserved in Brassicaceae. Using each of the respective sets of lincRNAs as query, we searched for sequence homologs within the genomes of nine Brassicaceae as well as *Tarenaya hassleriana*, a member of the sister family Cleomaceae (Cheng et al. 2013) using Evolinc-II (Nelson et al. 2017)). This comparative analysis revealed that 32.9% (6,781) of the complete (HC + LC) Arabidopsis lincRNA dataset are species-specific (i.e., no sequence homologs are identified in any other species within the family; Supplemental File 3), while 6,045 (29%) occur only in the genome of at least one other species in the genus (node 1; Figure 5A). Species or genera-specific conservation was largely a feature of LC-lincRNAs, which were significantly less likely to be conserved (nodes 0-1; Supplemental Figure 6A) than HC-lincRNAs. Sequence homologs are present for ~35% (4,336) of the Arabidopsis
lincRNAs back to the coalescence point between Brassicaceae lineages I and II (node 4; Figure 5A; Beilstein et al., 2006; Beilstein et al., 2008), suggesting these lincRNAs originated ≥ 43 MYA (Beilstein et al., 2010). The majority of these sequence homologs corresponded to either Evolinc-identified lincRNAs or unannotated intergenic sequence in each of the other species (Supplemental Figure 6E), suggesting that these lincRNAs have been evolving as lincRNAs, and not as pseudogenized loci, for the last 43 MY.

Figure 5: Sequence and transcriptomic conservation of Arabidopsis lincRNAs and their functional motifs across the Brassicales A) Arabidopsis lincRNA conservation across select Brassicale: Arabidopsis thaliana, Arabidopsis lyrata, 105
Supplemental Figure 7A: Multiple paralogs within the Brassicaceae. A HID1 locus was present in all tested Brassicaceae. ELENA and APOLO were present in the genomes of other species, including Arabidopsis, Eutrema salsugineum, Aethionema arabicum, and Tarenaya hassleriana (representative of Cleomaceae). The inset bar graphs depict the percent of Arabidopsis lincRNAs and mRNAs (yellow bar) restricted to that node (out of 20,416 total lincRNAs and 27,173 mRNAs examined). For lincRNAs, pink bars represent lincRNA sequence homologs found at that node, whereas blue bars represent transcriptional syntelogs, and thus is dependent on lincRNAs having been identified in a species descending from that node. B) Percent of lincRNAs in each of the four focal species for which we could infer gene family expansion, contraction, or for which there was no change (NC) relative to at least 2 of the closest relatives for each species. See Methods for more information. C) Correlation between the nodes at which a lincRNA is conserved (from A) and at which the structured element (from Arabidopsis) is conserved. D) Multiple sequence alignment (MSA) of a structured and protein-bound Arabidopsis lincRNA (Evolinc_tid.00064432) where the functional motif and lincRNA are conserved to the same node. E) Correlation between lincRNA conservation and ORF conservation. F) MSA of a sORF containing Arabidopsis lincRNA (AT1G06113) where the lincRNA and sORF are conserved to the same node. G) Schematic demonstrating the number of putative miRNA binding motifs that were found to be either species-specific (lincRNA and miRNA motif are restricted to Arabidopsis), not conserved (lincRNA is conserved but conservation is not associated with miRNA motif) or conserved. The number in parentheses represents the number of lincRNAs with putative miRNA binding motifs that are also stress-responsive. H) Example MSA of a lincRNA (Arabidopsis TAS1B) with a conserved miRNA binding motif in Cardamine, Capsella, and Arabidopsis.

The percent of species-specific lincRNAs for the other three species ranged from 49% in Camellina to 75% in Eutrema (Supplemental Figures 6B-D). Sequence homologs for 162 (~0.8%) of the Arabidopsis lincRNAs were recovered in T. hassleriana, with similarly low percentages (3, 1.5, and 1) of sequence homologs identified for Camellina, Brassica, and Eutrema lincRNAs, respectively (Supplemental Figures 6B-D). In sharp contrast to lincRNAs, sequence homologs were recovered for ~43% of Arabidopsis protein-coding genes in T. hassleriana (Figure 5A). Thus, while a majority of Brassicaceae lincRNAs are species-specific, for each of the four focal species, a subset of lincRNAs display higher rates of sequence conservation across the family and may depend on sequence for function.

From these data we asked whether we could identify functionally characterized Arabidopsis lincRNAs, suggesting that they were functionally conserved outside of Arabidopsis. We examined five functionally annotated Arabidopsis lincRNAs, including the photo-responsive lincRNA HID1 ((Y. Wang et al. 2014); Supplemental Figure 7A), the salt-responsive lincRNA DRIR ((Qin et al. 2017); Supplemental Figure 7B), the auxin regulated lincRNA APOLO ((Ariel et al. 2020); Supplemental Figure 7C), the pathogen resistance associated lincRNA ELENA (Seo et al. 2017); Supplemental Figure 7D), and the cold-responsive lincRNA SVALKA ((Kindgren et al. 2019); Supplemental Figure 7E). A HID1 locus was present in all tested Brassicaceae. Surprisingly, we recovered an unreported HID1 paralog in Arabidopsis. In fact, HID1 paralogs were present in the genomes of other Brassicaceae (Supplemental Figure 7A). ELENA and SVALKA sequence homologs were present in Camellina (lineage I) but were not recovered in more distantly related species. We identified two previously unreported APOLO paralogs in the genome of Arabidopsis, and multiple paralogs within the A. lyrata genome, but were unable to find...
sequence homologs in other sampled lineage I species. None of the *A. lyrata* APOLO homologs were adjacent to the PID1 locus and thus, if expressed, may not be functionally conserved. Finally, we were unable to identify homologs for DRIR1 in any Brassicaceae genome sampled, including other members of the genus, suggesting that it is specific to Arabidopsis.

Some lincRNAs function as transcriptional regulators in cis, influencing the expression of neighboring genes by recruiting Pol-II or transcription factors (Kopp and Mendell 2018). To identify lincRNAs with conserved expression that may function as transcriptional regulators of adjacent genes, we used SynMap (Haug-Baltzell et al. 2017) to identify collinear blocks between each of our focal species and then searched for lincRNAs transcribed from them (see Materials and Methods). This approach revealed 1,621 Arabidopsis lincRNAs for which a sequence-dissimilar lincRNA was transcribed in Camelina at a syntenic locus, and an additional 3,560 Arabidopsis lincRNAs that shared synteny with lincRNAs in either Eutrema or Brassica (Figure 5A). Interestingly, we identified “transcriptional syntelogs” adjacent to multiple CBF1 loci in Brassica, in a similar orientation and distance to CBF1 from the Arabidopsis lincRNA SVALKA (Supplemental Figure 7E). In total, four putative SVALKA transcriptional syntelogs were identified next to CBF1 paralogs in Brassica. A putative transcriptional syntelog was also identified for DRIR1 in Brassica (Supplemental Figure 7B). Thus, a significant proportion of “species-specific” lincRNAs in Brassicaceae may in fact be transcribed from syntenic loci with diverged sequence and therefore harbor conserved cis-regulatory functions.

Given the apparent expansion of the HID1 and APOLO gene families, we asked how frequently lincRNA gene families expanded or contracted, and whether these dynamics were coincident with known whole genome duplication events (WGD). We examined lincRNAs for which we were able to identify a sequence homolog in at least one other organism, and then asked if the number of lincRNAs in each gene family could have resulted from a WGD, either from a recent WGD or the α-WGD (Bowers et al., 2003) that coincided with the emergence of the Brassicaceae (see Materials and Methods). For Arabidopsis and Eutrema, which have not undergone recent WGDs, lincRNA gene families are predominantly stable (no evidence of expansion for 89% and 81% of lincRNAs respectively; Figure 5B). Camelina and Brassica have both undergone relatively recent whole genome triplication events (Mandakova et al., 2019; Wang et al., 2011). Despite these WGDs, lincRNAs in both species (67% and 98%, respectively) occurred more frequently in one or two copies rather than three, indicating that they were likely removed via fractionation which often occurs following WGD. Interestingly, this extent of copy
loss is greater than the loss observed for protein-coding duplicates (De Smet et al. 2013). Moreover, 71% of Camelina lincRNAs and 85% of Brassica lincRNAs are single copy, suggesting weak selective pressures to retain these genes in multicopy form. In addition, in Brassica, where the least and most dominant subgenomes have been assigned (Cheng et al., 2013; Tang et al., 2012), most single copy lincRNAs, and most lincRNAs in general, fall within the least fractionated subgenome (LF; n = 26,284), vs the medium fractionated (MF1; n = 21,712) and the most fractionated (MF2; n = 15973; Supplemental Figure 6F). For each of these sets of lincRNAs, ~50% are expressed in datasets generated from intra-specific hybrids or RILs (Supplemental Figure 6F). These data suggest that lincRNAs, like mRNAs, are preferentially retained on dominant subgenomes following whole genome duplication events, but that lincRNA hybrid-specific expression is not linked to subgenome of origin.

Camelina has a sufficient number of multi-copy lincRNA gene families to allow us to monitor the impact that whole genome duplication events have on lincRNA expression. Camelina is an allohexaploid (Mandakova et al., 2019), consisting of three subgenomes similar to its two progenitor species (C. hispida and a C. neglecta-like autotetraploid; Brock et al., 2018; Brock et al., 2019), referred to here as the C. hispida, C. neglecta, and C. neglecta (like) subgenomes. In Camelina, C. hispida mRNA paralogs are typically more highly expressed relative to those from the other two subgenomes (Chaudhary et al., 2020). To explore how WGD has impacted lincRNA expression, we performed Illumina short read RNA-sequencing in early embryos of Camelina (n = 5). These data were mapped to the reference genome with an updated gene set (i.e., including lincRNAs). For lincRNA families with paralogs present in the three subgenomes (see Materials and Methods), and with at least one member expressed above 1 TPM in Camelina embryos, we see a significant bias against expression of the paralogs from the C. neglecta (like) subgenome, while expression levels from the paralogs present in the other two subgenomes were similar to each other (Supplemental Figure 8A). This is in contrast to lincRNAs that occur only in single copy (Supplemental Figure 8B). For single copy LincRNAs, those transcribed from the C. neglecta (like) subgenome showed a slight but significant elevated average expression relative to the other two subgenomes (p <0.001; Supplemental Figure 8B). In contrast, paralogous protein-coding genes transcribed from the C. neglecta subgenome showed the lowest average expression relative to the other two subgenomes (Supplemental Figure 8C). These data suggest that lincRNAs with different evolutionary paths (multi-copy vs single copy) are expressed at different levels within the same subgenome.

We observed a number of distinct features within the Arabidopsis lincRNA dataset,
including structured regions, sORFs, and miRNA interaction motifs that may act as functional motifs (Lucero et al., 2020). If these elements are important for lincRNA function, then we would expect them to be conserved. Structural elements, inferred from PIP-seq data, strongly and positively correlated with conservation (p< 0.01; Figure 5C). Of the 415 lincRNAs for which structured elements were identified, 324 were conserved in another Brassicaceae genome. For 70% of these sequence conserved lincRNAs, the structured region was conserved to the same node as the lincRNA itself, suggesting the structural element is driving conservation of the lincRNA. An example of this is shown in Figure 5D, where the two structural elements of lincRNA “Evolinc_tID.00064432” overlap with deeply conserved (i.e., Node 5) portions of the lincRNA.

We next addressed the degree to which the 158 sORF-containing Arabidopsis lincRNAs are conserved, as conservation of the sORF would lend support to the idea that they are actually protein-coding transcripts. Of the 127 sORF lincRNAs conserved outside of Arabidopsis, there was no significant variation in the overall rate of conservation relative to non-sORF lincRNAs, indicating that sORF-lincRNAs are not preferentially retained. Of the 158 sORF lincRNAs tested, 53 sORFs were conserved in at least one other species, and 39 were conserved to the same node as the lincRNA from which they were derived (see Materials and Methods; Figure 5E; Supplemental File 4). There was no clear bias towards the length of sORF or the encoding transcript (Supplemental Figure 2). Some of the conserved sORFs were quite short, such as the sORF within AT1G06113, which encodes for a nine amino acid peptide and lies within a region of the sORF-lincRNA that shares almost 100% identity across the 11 species present in the MSA (Figure 5F). Although most sORF-lincRNAs (26/36) were previously annotated lincRNAs (i.e., Araport11 lincRNAs), a subset were identified in this study, suggesting that current filtering schemes are not entirely sufficient for removing short protein-coding transcripts from our dataset.

Finally, we determined the degree to which the predicted miRNA interaction sites within our Arabidopsis lincRNA dataset were conserved. Of the 226 lincRNAs with predicted miRNA interaction sites, 68 were species-specific (Figure 5G). A further 83 were sequence-conserved in at least one other Brassicaceae, but the conserved region did not overlap with the putative miRNA interaction site. The remaining 75 lincRNAs contained sequence conserved miRNA interaction motifs, with an example for this shown for AT1G50055 (TAS1B) in Figure 5H. Multiple sequence alignments supporting our conservation assignments for structure, sORFs, and miRNA interaction sites can be found in the CyVerse Data Store (DOI). LincRNAs with conserved domains are annotated in Supplemental File 4. In sum, our evolutionary approach has uncovered conserved
lincRNA functional elements and shed additional light on how plant lincRNAs evolve in the face of WGD.

**Assigning putative function to Brassicaceae IncRNAs**

Basic characterization of lincRNA expression, along with a conservation analysis, can provide clues as to which lincRNAs in our datasets are potentially functional, but these data alone do not permit the formation of robust functional hypotheses. To better clarify when and where the lincRNAs in our catalogs are functioning, we took three approaches. The first was to determine which lincRNAs are stress responsive based on pairwise comparisons of publicly available RNA-seq data (stress vs. control). The second was to use weighted gene co-expression networks (WGCNA) of larger, more complex experiments, to identify modules of similarly expressed protein-coding and IncRNA genes (i.e., guilt-by-association) to infer in which molecular pathway a lincRNA might be acting. Third, as many lincRNAs regulate the expression of neighboring genes (Khyzha et al. 2019; Gil and Ulitsky 2020), we examined correlation of expression of lincRNA-adjacent mRNA gene pairs across tissue and stress expression atlases to identify candidate gene pairs in which the lincRNA has the potential to regulate its neighbor.

We first searched for lincRNAs in each species that were differentially expressed in response to stress. For Arabidopsis and Brassica, we chose publicly available datasets with multiple independently generated stress experiments (Supplemental File 4). In both species, most of the stress-responsive lincRNAs were specific to a particular stress (Figures 6A, 6B, and Supplemental Figure 9A), with the highest proportion associated with temperature stress (cold, heat, or cold + heat). We observed a similar pattern for protein-coding genes in both species (Supplemental Figures 9B and 9C), therefore we sought to determine which of the Arabidopsis stress-responsive lincRNAs were also ABA responsive. Abscisic acid mediated phytohormone signaling is a key regulatory pathway known to promote abiotic stress tolerance (Vishwakarma et al., 2017) and thus we would expect at least a subset of abiotic stress-responsive lincRNAs to also be ABA-responsive. We screened through Arabidopsis RNA-seq data associated with seedlings and roots treated with exogenous ABA (5-100uM) and identified 672 ABA-responsive lincRNAs, 105 of which overlap with our stress-responsive lincRNAs, suggesting these lincRNAs may be stress-responsive in an ABA-dependent manner (Figure 6A, inset). As lincRNAs were predominantly responsive to temperature stress (heat and cold), we next asked how many lincRNAs, out of the total heat/cold responsive lincRNAs, showed an anti-correlated response to temperature stress (i.e., up in heat and down in cold). In both species, heat/cold responsive
lincRNAs were predominantly upregulated by heat and repressed by cold (Figures 6C and 6D). This pattern was specific for lincRNAs, as most mRNAs were either up in both or down in both conditions (Supplemental Figure 9D). Taken together, we observed that a substantial fraction of lincRNAs were differentially regulated during temperature stress in both Arabidopsis and Brassica. Stress and ABA-responsive lincRNAs for each species have been denoted in Supplemental File 2. Additionally, all differential expression results from the 4 focal Brassicaceae can be found in Supplemental File 8.
**Figure 6:** Inferring lincRNA function from transcriptomic data. A-B) Upset plots depicting the number of stress-responsive lincRNAs in Arabidopsis (A) and Brassica (B). The vertical tan bars depict the number of lincRNAs found in each stress, or combination of stresses, shown below. The horizontal-colored bars depict the total number of lincRNAs associated with that stress across all combinations. For Arabidopsis, an inset Venn diagram depicts the number of lincRNAs found to be both stress (heat, cold, salt, or drought stresses) and ABA responsive. C-D) Scatterplots of temperature responsive lincRNAs in Arabidopsis (C) and Brassica (D). E-F) Modules of similarly expressed mRNAs and lincRNAs from the Arabidopsis Klepikova tissue atlas (E) or stress atlas (F). “M” = meristem tissue, “Male Rep.” = male reproductive tissues (stamens and anthers). G-J) Density plots showing the distribution of expression correlation between different gene pairs in Arabidopsis and Brassica tissue atlases (G-H) as well as Arabidopsis and Brassica heat stress experiments (I-J). K) Summary diagram of lincRNAs that are both hypoxia-stress responsive and bound by the HRE2 transcription factor. On the left are Z-transformed expression data for each lincRNA and their closest protein-coding gene. On the right is a depiction of the arrangement and orientation of each lincRNA and protein-coding gene set. L) Bar plot depicting the number of lincRNAs for which we have varying degrees of functional evidence.

WGCNAs help to identify clusters of genes that are coordinated in their expression and thus potentially regulated by, or regulate, similar pathways. This allows us to assign putative functions to lincRNAs based on significant co-expression with functionally characterized mRNAs or lincRNAs, a process referred to as guilt-by-association (Tian et al. 2008). To remove noise from normalizing across many disparate experiments, we grouped experiments by tissue or, where available, by project (as in the case of the tissue atlases; see Materials and Methods). In total, we identified 987 lincRNAs in Arabidopsis and 3,473 lincRNAs in Brassica whose expression profiles were sufficient to classify them into at least one co-expression module. For example, when we examine the Arabidopsis Klepikova tissue atlas (Klepikova et al., 2016), we identified a module of 233 mRNAs and 11 lincRNAs (8 of which were annotated as part of this study) whose expression peaked in flowers and male reproductive tissues (i.e., anther and pollen; Figure 6E). Within this module, gene ontology (GO) terms associated with fertilization were enriched, suggesting that lincRNAs within this module may function during fertilization (Module associated GO terms found in Supplemental Figure 10A). We also observed lincRNAs that were members of co-expression modules determined from the Klepikova tissue atlas stress experiments. One of these modules contains 182 transcripts (176 mRNAs, 6 lincRNAs) whose expression peaks rapidly after wounding (Figure 6F). As expected, the GO terms we see with these member mRNAs are highly enriched for response to wounding and jasmonic acid regulation (Supplemental Figure 10B), a hormone that is released in response to herbivory and biotic stress (J. Wang et al. 2020). We identified a separate module of 140 genes, six of which are lincRNAs, specifically induced under cold stress (Supplemental Figure 10C). Importantly, one of these six lincRNAs was SVALKA, a previously reported cold-induced transcript critical for the freezing response in
Arabidopsis (Supplemental Figure 10D). By expression patterns and association, we hypothesize that the other five lincRNAs are also regulating the Arabidopsis cold response. Thus, through WGCNA and guilt-by-association, we have generated putative annotations for ~1,000 Arabidopsis and ~3,000 Brassica lincRNAs that may guide future in vivo functional analyses. These lincRNAs have been annotated with expression modules in Supplemental File 2. Detailed WGCNA results can be found in Supplemental Dataset 2.
Expression for a subset of LincRNAs is significantly correlated with adjacent mRNAs:

LincRNAs are known to regulate the expression of other genes, either in cis or in trans, through a variety of mechanisms (Kindgren et al. 2019; Gil and Ulitsky 2020). One signature of cis-regulatory lincRNAs is correlation in expression relative to neighboring genes across a diverse transcriptomic dataset. To identify putative cis-regulatory lincRNAs, we searched for correlation between all Arabidopsis and Brassica lincRNAs and their immediate neighboring mRNAs that were expressed above 0.1 TPM (i.e., both lincRNA and mRNA > 0.1 TPM) in either their respective tissue atlases or heat stress experiments. In the Arabidopsis tissue atlas, we identified 252 lincRNA-mRNA pairs in which both genes in the pair were expressed and for which we could calculate expression correlation. This correlation was significantly more positive than mRNA-mRNA pairs or random pairs of genes (Figure 6G; P value=1.62e-14; Wilcoxon rank sum test with Bonferroni multiple testing correction). When examining all genes that fall within 10 Kb of an expressed lincRNA, we observe even stronger positive correlation, in contrast to mRNA-mRNA pairs within the same region, which show very little correlation across all distances measured (up to 10 Kb; Supplemental Figure 11A). We observed even more lincRNA-mRNA pairs with correlated expression during heat stress in Arabidopsis (n = 2,544), again with a significant positive correlation relative to mRNA-mRNA pairs (p<2e-16; Figure 6I). We also observed positive correlation for Brassica lincRNA-mRNA pairs in the Brassica tissue atlas (3,757 out of 23,756 expressed lincRNAs; Figure 6H and Supplemental Figure 11B) and heat experiments (n = 6,514), although this correlation was less pronounced than in Arabidopsis. In sum, we identify a subset of lincRNAs whose expression appears to be positively correlated with neighboring genes up to at least 10 Kb away, suggesting that these lincRNAs might be cis-regulatory RNAs. LincRNA-mRNA pairs with a strong correlation (r > 0.5 or r < -0.5), as well as all correlated neighboring pairs are listed in Supplemental File 2.

Synthesizing our functional assignment approach:

We searched for experiments that took a holistic approach towards the analysis of stress responsiveness where we could assess the active regulation and response of lincRNAs (i.e., not just RNA-seq, but ChIP-seq or other methods to study the regulation of gene expression). Lee and Serres (2019) performed such an integrative approach to understand hypoxia responses in Arabidopsis seedlings. We set out to re-analyze these data in the context of both mRNAs and
lincRNAs, and recovered 153 Arabidopsis lincRNAs that are differentially expressed in response to hypoxic stress. We also observe 62 lincRNAs that fall into a co-expression module with significant enrichment of GO terms associated with hypoxia (Supplemental Figure 13A-H). In Arabidopsis, changes in gene expression in response to hypoxia is regulated predominantly by the transcription factor HRE2 (Hypoxia Responsive Ethylene Responsive Factor 2; AT2G47520). We searched for the presence of our hypoxia stress-associated lincRNAs in Arabidopsis HRE2 ChIP-seq data in and found evidence that HRE2 was bound to the promoter regions of 20 differentially expressed lincRNAs. We then examined correlation in expression between the HRE2-bound lincRNAs and their adjacent mRNAs and identified four lincRNAs with a positive expression correlation, one with a negative expression correlation, and two where the correlation differed between the two adjacent genes (Figure 6K). Thus, these seven lincRNAs appear to be specifically regulated by HRE2 in response to hypoxic stress and may act as cis-regulatory elements (annotated in Supplemental File 2).

In sum, we have used a wealth of public data, supplemented with short and long-read RNA-seq, to identify and provide putative functional annotations for lincRNAs across four Brassicaceae. We combined our transcriptomics data with comparative genomic and evolutionary analyses to determine conservation of not just the full-length lincRNAs, but also putative functional elements within them, such as sORFs, structured regions, and miRNA interaction motifs. Using these approaches, we have identified >100,000 Brassicaceae lincRNAs with multiple lines of functional and/or contextual clues that will facilitate downstream functional analyses.

Discussion:
A comprehensive and unified lincRNA annotation effort for the mustard lineage:
Here we generated an expansive catalog of high confidence lincRNAs for four agricultural and model Brassicaceae species by processing > 20,000 publicly available RNA-seq datasets for those species. We supplemented these publicly available data with our own ONT long read sequencing data, and further annotated the identified lincRNAs with epigenetic, genomic, structural, translational, and evolutionary information. These efforts build on previous efforts to catalog novel transcribed elements within plant genomes (Liu et al., 2012; Moghe et al., 2013), and serve as the most exhaustive lincRNA identification and annotation effort to date in any plant species.

Due to the scale of our efforts and the wealth of data available for these four species, we were able to uncover defining features for Brassicaceae lincRNAs, features that may guide future
discovery and annotation efforts in other plant lineages. LincRNAs tend to be mono-exonic, but when multi-exonic, harbor longer exons relative to those seen in spliced transcripts. LincRNAs appear to be epigenetically regulated in a distinct manner from both protein-coding genes and transposable elements. And, as expected based on prior observations in plants and mammals, lincRNAs in all four species were, on average, expressed at low levels and displayed significantly higher tissue specificity relative to protein-coding genes in tissue atlases and our ONT data. The exception to this observation were the sORF containing lincRNAs, which behave more similar to protein-coding genes in terms of both higher expression levels and tissue specificity. Interestingly, many of the lincRNAs we identified displayed high expression in, or were restricted to, very specific cell types (e.g., meristematic tissue) or experimental conditions (e.g., environmental stress) suggesting that 1) lincRNA expression is highly context and cell-type specific, and 2) sampling bulk tissues may not accurately reflect a lincRNA’s contribution to the transcriptome. The lincRNAs restricted to inter-accession crosses as in *B. rapa* may be the result of improper transcriptional control given their relatively even distribution across the genome or, albeit less likely, may reflect transcripts that help mediate compatibility of two subtly different genomes.

**Using comparative genomics to provide functional insights:**

Given that we identified thousands of lincRNAs in each of our four focal species, functional analyses will need to be prioritized. In order to facilitate that prioritization, we used a comparative genomic approach to assess the degree to which each identified lincRNA is conserved, and if there are any particular motifs of interest within those conserved lincRNAs. As expected based on prior observations in plants and mammals (Nelson et al., 2016; , we observed low levels of sequence conservation for lincRNAs identified in each of the four species relative to protein-coding genes. However, when sequence homologs were detected between two species (e.g., Arabidopsis to Brassica), those sequence homologs were predominantly annotated as lincRNAs and not protein-coding genes. Inspired by a smaller comparison between Arabidopsis and Aethionema (Mohammadin et al., 2015), we also searched for and observed a cohort of lincRNAs that are transcriptional syntelogs in that they are transcribed from similar genomic positions in multiple species but share little sequence conservation. LincRNAs that regulate gene expression in cis are an interesting class of transcripts from an evolutionary perspective in that positional and transcriptional conservation may be more critical than sequence conservation. Although additional
study is needed, we posit that these lincRNAs may be functionally conserved in regulating expression of the orthologous genes to which they are adjacent in each species. An exciting set of candidates for further study are the putative SVALKA transcriptional syntelogs we identified in Brassica. In Arabidopsis, SVALKA regulates an adjacent, non-overlapping, protein-coding gene through transcriptional interference. This mode of function in particular might depend more on conservation of transcription, and from where transcription arises, than it does on sequence similarity.

Identifying lincRNAs in species with recent WGD events (e.g., Camelina and Brassica) allowed us to more closely examine how lincRNAs evolve following these genomic events. LincRNAs are not typically retained as multicopy loci following WGD events, supporting prior results from a smaller set of Arabidopsis-specific lincRNAs (Nelson et al., 2016). In Brassica, lincRNAs are predominantly retained as single copy from the least fractionated genome. The fractionation - and retention, of a certain set of lincRNAs may suggest functional interactions (e.g., genetic or molecular) are preferentially retained following WGD events - similar to that observed for protein-coding genes (Emery et al., 2018; Schnable et al., 2012). When lincRNAs are retained as multicopy, their expression appears to be more sensitive to the influence of subgenome dominance than protein-coding genes - perhaps explaining why they are fractionated from the genome. However, the retention, and expression, of paralogous lncRNAs such as HID1 may suggest that lncRNAs can be functionally retained post-duplication in a similar manner as protein-coding genes. Further studies are needed to determine if these paralogous lncRNAs (e.g. HID1) have sub or neo-functionalized as is often the case for retained proteins.

Using multiple sequence alignments for our sets of conserved lincRNAs, we also determined if the identified structural, putative miRNA binding, or sORFs were within those conserved regions. Although we did identify examples of conserved sORFs, to our surprise we did not observe strong correlation between sORF and lincRNA conservation. One particularly interesting conserved sORF is found within the lincRNA AT1G06113. The Ribo-seq identified sORF within this lincRNA is only nine amino acids long, but is almost perfectly conserved across the Brassicaceae and even in T. hassleriana. The functional significance of this peptide, as well as the other lincRNA-derived small proteins remains to be determined. In contrast to the sORF-containing lincRNAs, the regions we identified to be protein-bound and structured were typically conserved to the same degree as the lincRNA itself. This conservation suggests these structured...
regions are important for function and may bind similar proteins in multiple species. Thus, identifying the protein binding partner in Arabidopsis might help provide functional insights for these lincRNAs across the family as well as develop a protein-RNA interaction database for improving functional predictions.

Using omics-approaches to assign putative function to Brassicaceae lincRNAs:
Our ultimate goal, beyond identifying lincRNAs in each of these species, was to annotate these lincRNAs so as to aid in future functional studies. We used expression data to assign lincRNAs into broad regulatory categories, such as stress-responsive, cis-regulatory, or others associated with GO-terms extracted from network analyses. As most functionally described lincRNAs to date are associated with changes in the environment (i.e., biotic/abiotic stress; reviewed in Ariel et al., 2015; Chekanova, J., 2015), our initial expectations were that most lincRNAs would be stress responsive. Interestingly, this was not the case. Roughly 10% of the lincRNAs identified in Arabidopsis and Brassica are stress-responsive, with most responding to temperature stress. While this could be linked to changes in genome-wide epigenetic control that is not specific to lincRNAs, there does appear to be a degree of response specificity. A majority of the temperature (cold or heat) responsive lincRNAs were either specific to one stress or the other, or showed opposite responses to the two stresses. Furthermore, we also identified a set of lincRNAs whose response appears to be ABA-dependent. The preponderance of lincRNAs associated with temperature stress in our dataset may simply reflect sampling bias as our analyses were dependent on publicly available data. However, given the lincRNAs and NAT-lncRNAs that have already been functionally described as temperature responsive in Arabidopsis (Kindgren et al., 2018; Castaings et al., 2014; Zhao et al., 2018), the potential for widespread adaptation to environmental conditions by lincRNAs remains an exciting avenue for future research.

Guiding future lincRNA annotation efforts across the plant lineage:
Most transcriptomic analyses ignore lincRNAs, not because they are not present in the data, but because the genome annotations used are lacking information about these transcript classes. Thus, many of the most impactful plant lincRNA functional studies to date have relied on de novo lincRNA classification because their locus of interest was unannotated. Our approach is species-independent, and because we are repurposing available RNA-seq data, naturally focuses on the
experimental questions in which the plant community is interested. We aim, and encourage others, to expand these annotation efforts to all plant species with significant RNA-sequencing data in order to fully understand how lincRNAs contribute to the biology of plants.

**Methods:**

**Plant materials and growth**

*Arabidopsis thaliana* (Col-0; (Lamesch et al. 2012), *Brassica rapa* (R-0-18; (Howe et al. 2021), *Camelina sativa* (cultivar Ames), and *Eutrema salsugineum* (Shandong; (Yang et al. 2013) seeds were surface sterilized by washing with 70% ethanol followed by soaking in 30% bleach and 1% Tween 20 for 10 minutes before being rinsed and plated on ½ MS media supplemented with 0.5% sucrose. Plates were placed in the dark at 4°C for 5 days before being moved to a long day (16 hour light 22°C/8 hour dark 20°C) growth chamber. Ten days after germination, seedlings were either collected in liquid nitrogen or transplanted to soil and placed into the same growth chamber. For leaf samples, leaves were either collected 4 weeks after germination, or at the mature most vegetative stage, whichever came first. Finally, for flower samples, opened flowers with no sign of developing fruit were collected. All plant samples were immediately frozen in liquid nitrogen and stored in a -80°C freezer until ready for processing.

**RNA extraction and ONT library preparation**

Frozen plant samples were pulverized in liquid nitrogen using a chilled mortar and pestle until a fine powder was obtained. RNA was extracted using the RNeasy Plant Mini kit (Qiagen) following the manufacturer’s instructions. Purified RNA was used as input for the Dynabeads mRNA Purification kit (Invitrogen). Purified poly-A RNA was used as input for the Nanopore direct cDNA sequencing kit (SQK-DCS109) following the manufacturer’s instructions. Nanopore libraries were sequenced on a MINion sequencer (R9.4.1 flowcell). Raw reads were basecalled using a GPU-enabled version of Guppy in the command line.

**Illumina RNA-sequencing of Camelina sativa seeds**

Developing seeds of four *Camelina sativa* accessions were collected in biological triplicate at ~15 days post anthesis and immediately placed in liquid nitrogen. Total RNA was isolated from developing seeds using the PureLink® Plant RNA Reagent (Thermo Fisher Scientific, Waltham,
MA, USA) and its associated protocol. Extracted RNA was then purified further using an RNeasy RNA clean-up kit (Qiagen, Valencia, CA, USA) and quantified on a Qubit fluorometer (Life Technologies, Carlsbad, CA, USA). Sequencing libraries were prepared with the SENSE mRNA-seq library prep kit and protocol, using up to 1,000 ng total RNA per sample (Lexogen GmbH, Vienna, Austria). Individual transcriptome libraries were quantified using a Qubit fluorometer and fragment size, distribution, and overall library quality was determined with an Agilent Bioanalyzer (Agilent, Santa Clara, CA, USA) system. Samples were pooled into three final libraries and sequenced by Novogene (Sacramento, CA, USA) on an Illumina HiSeq platform (Illumina, San Diego, CA, USA) producing 150 bp paired-end reads.

**LincRNA identification and basic characterization**

The RMTA (Peri et al. 2020) pipeline was used to process all available short read RNA-seq experiments as of December 2018 within the CyVerse Discovery Environment (Merchant et al. 2016) using the HiSat2 and Stringtie (Pertea et al., 2016) mapping and assembly options. Assembled transcripts were then processed through the Evolinc (Nelson et al. 2017) pipeline to identify lincRNAs. For *Arabidopsis thaliana*, the TAIR-10 assembly was used as a reference for the initial RMTA workflow (including mapping, quantification, and transcript assembly), for *Brassica rapa* the Ensembl v1.0, for *Camelina sativa* v2.0 from Ensembl (Plant Release 51), and for *Eutrema salsugineum* Phytozone v1.0 (Yang et al. 2013). An updated annotation including newly identified lincRNAs for each species can be downloaded from the CyVerse Data Store: (DOI from CyVerse).

Basecalled Nanopore reads were demultiplexed and processed following (A Eccles 2019). To identify lincRNAs with Evolinc, processed reads were aligned to each species’ genome using Minimap2 (H. Li 2018). Mapped reads were assembled into transcripts using Stringtie2 (Kovaka et al. 2019) using the -L parameter. Transcript assemblies were then used as input for Evolinc for lincRNA identification.

The BEDTools suite (Quinlan and Hall 2010), “nuc” function) was used to characterize the GC content and gene lengths of mRNAs and lincRNAs. Exon counts were determined using the R (Team and Others 2013) R Core Team, 2013, version 4.1.0) package GenomicFeatures (Lawrence et al. 2013), “exonsBy” function, v 1.44.1).
**Analysis of DNA methylation patterns and histone modification dynamics**

LincRNA and mRNA epigenetic profiles were monitored by reprocessing publicly available whole genome bisulfite sequencing (WGBS) datasets as well as chromatin immunoprecipitation with sequencing (ChIP-seq) experiments (see Supplemental File 6). WGBS data was processed using the Bismark tool (Krueger and Andrews 2011) with default parameters to generate BedGraph files which were then converted to bigWig format. These bigWig files were used in deepTools (Ramírez et al. 2014) with the “computeMatrix" function and “scale-regions" option to visualize CpG methylation over lincRNA and mRNA genes. ChIP-seq datasets were processed by first aligning the raw reads to the respective genome using BWA-MEM (H. Li 2013) which is deployed as a CyVerse app in the Discovery Environment (BWA_mem_0.7.15 with default settings). SAM files from BWA-MEM were converted to sorted BAM files. The Picard Toolkit (Picard Toolkit 2019) was used to remove PCR duplicates using the “lenient” setting for the “VALIDATION_STRINGENCY” option. These processed BAM files were then used as input for the deepTools “bamCompare" function with a ChIP input sample (if available) as a comparison experiment. For Arabidopsis, paired RNA-seq experiments were used to determine which genes were expressed for plotting for WGBS and ChIP-seq experiments. For Brassica, the defined tissue atlas was used instead.

**Characterization of lincRNA expression patterns**

To characterize expression from ONT-seq data, Minimap2 was used to map ONT-reads to transcriptomes for each of the respective species' updated gene sets (prior annotated genes + Evolinc lincRNAs) using similar parameters as above. Minimap2 produced BAM files were used as input for Salmon in alignment-based mode, specifying the --noErrorModel option (Patro et al. 2017) Soneson et al. 2019, (Patro et al. 2017). TPM values were aggregated from each experiment using the tximport R package (Soneson, Love, and Robinson 2015) to obtain gene level expression estimates.

Specific Illumina short read datasets from Arabidopsis and Brassica were used to gain additional resolution of tissue specific expression. For Arabidopsis, the Klepikova et al., 2016 (NCBI PRJNA314076) tissue expression atlas was reprocessed, and for Brassica two datasets were combined to create a tissue atlas similar to Arabidopsis (PRJNA253868 & PRJNA185152). RNA-sequencing reads (FASTQ) associated with each dataset were re-aligned to transcripts with Salmon
using XXX parameters to generate transcript-level expression values (TPM). Gene level expression values were obtained as above using tximport. To calculate the tissue specificity metric $\tau$ TAU, TPM values were first averaged across replicates. TAU was then calculated as described by (Yanai et al. 2005) using quantile normalized TPM values generated from the preprocessCore R package (Bolstad n.d.). To assess tissue of maximum expression, variance stabilized transformed expression values generated from DESeq2 were utilized (Love, Huber, and Anders 2014).

The DESeq2 package (in R), with the DESeq and results functions, were used to identify differentially expressed genes in pair-wise comparisons. For time-course studies, only the first and last treatments were examined, treating each of them as separate analyses (e.g. early stress response vs. late stress response). Genes were considered to be differentially expressed if they had a $\log_2$ fold change ($\log_2$FC) greater or less than 1 or -1, respectively, as well as an adjusted p-value (q-value, FDR) of 0.05 or lower.

**Analysis of co-expression modules**
To assess co-expression modules, raw RNA-seq data from select stress experiments (Supplemental File 6) was re-processed using Salmon as performed above. The tximport and DESeq2 R packages were used to import Salmon quantification files and convert expression estimates to a variety of normalized values (TPM, VST, normalized counts, etc.) The R package CEMiTool (Russo et al. 2018), v 1.10.2) was used to perform gene co-expression network analyses. In most cases, $\log_2 + 1$ converted TPM values obtained from tximport were used as input to CEMiTool. Gene ontology (GO) information was provided to CEMiTool from the biomaRt R package (Durinck et al. 2005), and protein-protein interaction data was obtained from the Arabidopsis Protein Interaction Network (Brandão, Dantas, and Silva-Filho 2009).

**Measuring expression correlation of adjacent genes**
Arabidopsis and Brassica expression data from the above described tissue atlases or heat experiments (Arabidopsis: PRJNA324514, Brassica: PRJNA298459) were normalized using the DESeq2 “vst” function with the “blind” parameter set to false. Additionally, for the tissue atlases, replicates for each tissue were averaged, if applicable. Genes that did not vary substantially across the input experiments were removed. This was performed by calculating the interquartile range for expression of all genes and only those in the top 50% of IQR values were retained (50% most
variable genes). Pearson correlation coefficients of expression were then calculated between all remaining genes post filtering using the corrr R package (Kuhn and Wickham 2020), v 0.4.3. Relevant correlations were then filtered for between lincRNAs and their nearest upstream and/or downstream mRNA neighbors. LincRNA-mRNA pairs separated by fewer than 100 base pairs were removed before subsequent analyses. Random gene pairs were generated from all pairwise correlations using the slice_sample function from the dplyr R package.

To analyze all gene pairs within defined distances, the bedmap function from the BEDOPS suite (Neph et al. 2012) was used with the range, echo, and echo-map-id options. This generated all lincRNA-mRNA or mRNA-mRNA pairs within 200, 500, 1000, 2000, 5000, and 10000 base pairs of each other. These gene pairs were used to filter out the pairwise correlations generated above.

**Assessing lincRNA function from multi-omics hypoxia datasets**
The multi-omics datasets generated by Lee and Bailey-Serres (2019) were re-processed using the above methods. After processing ChIP-seq data as above, the output BAM files were used as input for HOMER motif analysis (Heinz et al. 2010) following along closely with the provided “Next-Generation Sequencing Analysis” tutorial provided by HOMER. Differentially expressed genes were generated using a basic design strategy in DESeq2. Briefly, the early hypoxia stress was compared to the early control samples (2 hour hypoxia vs 2 hour control using the contrasts option with the results command from DESeq2), and the later hypoxia stress was compared to the later control experiments. Differentially expressed genes for the reoxygenation experiments were not analyzed, but the expression data was used for constructing the DESeq data set and running the differential expression analysis. For identifying relevant co-expression modules including lincRNAs that may be involved in the hypoxia response, CEMiTool was used as above using log transformed normalized counts from DESeq2.

**Identifying translated sORFs from Ribo-seq**
Translated sORFs within the lincRNAs were identified using our recent Ribo-seq and RNA-seq data in Arabidopsis seedling (GEO accession no. GSE183264; (Wu and Hsu 2021). Briefly, BAM files of the Ribo-seq and RNA-seq and a GTF containing the lincRNAs and Araport11 annotated genes were imported into RiboTaper (Calviello et al., 2015). The Ribo-seq read lengths and offsets
for RiboTaper were 24, 25, 26, 27, 28 and 8, 9, 10, 11, 12, respectively, as previously described (Wu and Hsu 2021). RiboTaper then computed 3-nucleotide periodicity, which corresponds to translating ribosomes move 3-nucleotide per codon, in each possible ORF within the transcripts. The sORFs were considered translated if they displayed significant 3-nucleotide periodicity and the translated ones were extracted from the RiboTaper output ORF_max_filt file.

To identify lincRNAs harboring putative sORFs based on mass spectrometry data, proteomic experiments, PXD026713 and PXD009714, were retrieved from the PRIDE repository. Raw chromatograms were analyzed using MaxQuant software (Version 1.6.0.16) with Andromeda- an integrated peptide search engine (Cox at al., 2011). Following search settings were applied: a maximum of two missed cleavages was allowed, and the threshold for peptide validation was set to 0.01 using a decoy database. In addition, methionine oxidation and N-terminal acetylation were considered variable modifications, while cysteine carbamidomethylation was a fixed modification. The minimum length of a peptide was set to at least seven amino acids. Moreover, label-free protein quantification (LFQ) was applied. Peptides were identified using the Arabidopsis Information Resource, [www.Arabidopsis.org](http://www.Arabidopsis.org) and a library of all Arabidopsis lincRNA ORFs (positive strand) obtained using Transdecoder.

**Evolutionary analyses**
LincRNA sequence homologs were identified using the Evolinc-II module (v2.0, [https://github.com/Evolinc/Evolinc-II](https://github.com/Evolinc/Evolinc-II); e-value of -10), with the following genomes: *Arabidopsis thaliana* (TAIR10), *Arabidopsis lyrata* (Ensembl v1.0, Hu et al., 2011), *Capsella grandiflora* (Phytozome v1.1, Slotte et al., 2013), *Capsella rubella* (Phytozome v1.1, Slotte et al., 2013), *Camelina sativa* (Ensembl v2.0), *Cardamine hirsuta* (v1.0, Gan et al., 2016), *Brassica rapa* (Ensembl v1, Wang et al., 2011), *Schrenkiella parvula* (Phytozome v2.0, Dassanayake et al., 2011 and Oh et al., 2014), *Eutrema salsugineum* (Phytozome v1.0, Yang et al., 2013), *Aethionema arabicum* (CoGe vVEGI 2.5. gID 20243, Haudry et al., 2013 and Nguyen et al., 2019), *Tarenaya hassleriana* (CoGe v4, gID 20317, Cheng et al., 2013). For each of the four species, the entire lincRNA list (LC + HC) were included as query in the analyses. LincRNAs were determined to be restricted to a particular node if no sequence homolog was identified in a more distantly related species. LincRNAs were determined to be conserved as lincRNAs or mRNAs in other species if they overlapped by 50% or more with an annotated gene on the same strand. If not, they were
considered to be unannotated. Multiple sequence alignments produced by Evolinc-II (using MAFFT) were imported into Geneious (Genious Prime 2021.1.1, https://www.geneious.com) for downstream structure, sORF, and miRNA motif analysis.

Transcriptional syntelogs were identified by downloading the DAGChainer output, with genomic coordinates, from pairwise CoGe SynMap (Nelson et al., 2018; Haug-Baltzell et al., 2017) analyses between Arabidopsis and each of the three other species (links to regenerate analyses: Camelina https://genomevolution.org/r/1fjg7, Eutrema https://genomevolution.org/r/1f7si, and Brassica https://genomevolution.org/r/1f79g). LincRNAs that were found within syntenic blocks (10 colinear protein-coding genes), between orthologous genes in either of the pairwise SynMap analyses, and in the same orientation to at least one of the neighboring orthologous genes were considered transcriptional syntelogs. To infer lincRNA gene family contraction or expansion, a rudimentary ancestral state reconstruction was performed. For Arabidopsis, ancestral gene copy number for each Arabidopsis lincRNA was inferred by averaging the number of recovered sequence homologs in (at minimum) A. lyrata, C. rubella, and C. grandiflora. Species-specific lincRNAs were not examined. For Camelina, C. rubella, C. grandiflora, A. thaliana, and A. lyrata were used to determine the copy number in the last common ancestor. This value was then multiplied by three (to account for the Camelina-specific whole genome triplication event). Values above or below this value were considered to be expansions or contractions, respectively. A similar approach was performed for Brassica and Eutrema.

MSAs were manually scanned to infer depth of conservation of sORFs, putative miRNA binding motifs, and structural/protein-binding elements. On top of lincRNA sequence homology and synteny requirements, for a sORF to be considered conserved, the start and stop sites within the annotated Arabidopsis lincRNA must be positionally conserved (within +/- three AA). In addition, the translated amino acid sequence must be 75% identical in pairwise alignments between Arabidopsis and each putative homologous sORF. To identify putative miRNA binding sites, all lincRNAs were scanned for motifs using psRNATarget (Dai et al., 2018) using an expectation score of 2.5 as cutoff. LincRNAs with putative miRNA binding motifs were then compared against the list of lincRNAs that were conserved outside of Arabidopsis. MSAs were then scanned for the presence of miRNA motifs. Motifs with complete coverage and no more than two (pairwise) mismatches in at least one other species were considered for evolutionary comparisons. For conservation of structural/protein-binding motifs, structured regions inferred by PIP-seq (GEOS
GSE58974 and GSE86459; Gosai et al., 2015 and Foley et al., 2017) were intersected with lincRNAs using Bedtools intersect (Quinlan et al., 2010). Arabidopsis lincRNAs, their sequence homologs (from Evolinc-II) and structured regions were combined into a MSA using MAFFT (Nakamura et al., 2018) for manual inspection. PIP-seq motifs were considered conserved if the entire motif was contained within an alignable region of a sequence homolog from another species. For a motif (sORF, miRNA, or structural) to be considered conserved to a particular node, at least one species that shares that node with Arabidopsis was required to contain those motifs under the parameters described above.

Author Contributions and Acknowledgements:
KRP, EL, MAB, and ADLN developed the project. KRP, LY, LW, ES, and ADLN performed the analyses. ACND performed RNA extractions and ONT-sequencing. JRB performed the Camelina RNA-seq. HLW and PYH performed the sORF identification from Ribo-seq data. ES and AS examined the MS data. KRP, EL, MAB, and ADLN wrote the manuscript. The authors would like to acknowledge the NSF Graduate Research Fellowship Grant DGE-1746060 (awarded to K.R.P.), NSF-MCB 2051885 (awarded to PYH), NSF-IOS 1758532 (awarded to ADLN), NSF-IOS 1444490 (awarded to EL and MAB), NSF-DBI-1743442 to EL, and NSF-IOS 2023310 (awarded to ADLN and EL). The authors would like to thank the plant scientists who have contributed genomic and transcriptomic data to the NCBI SRA, making this work possible. The authors, particularly JRB, would like to thank Dr. Ken Olsen, Washington University in St. Louis, for thoughtful discussion regarding the Camelina experiments. Finally, the authors would like to thank members of the Stress Architecture and RNA Biology Cluster and Skirycz lab at BTI and the PaBeBaMo group at the University of Arizona School of Plant Sciences for helpful discussion.

Supplemental File 1: List of SRAs examined for all four species
Supplemental File 2: Functional annotations for each lincRNA from all four species
Supplemental File 3: Evolinc II results for all four species
Supplemental File 4: sORF and structural motif conservation and characteristics
Supplemental File 5: Predicted miRNA binding motifs for all four species
Supplemental File 6: SRAs, with associated metadata, used in targeted transcriptomic studies
Supplemental File 7: Araport11 lincRNAs that were removed from analysis
Supplemental File 8: Expression information for mRNAs and lincRNAs in paired analyses
References:


Travis A. Lee, Julia Bailey-Serres, Integrative Analysis from the Epigenome to Translatome Uncovers Patterns of Dominant Nuclear Regulation during Transient Stress, The Plant Cell, Volume 31, Issue 11, November 2019, Pages 2573–2595


Seki, Masahide, Eri Katsumata, Ayako Suzuki, Sarun Sereewattanawoot, Yoshitaka Sakamoto,


Zhao, Xinyue, Jingrui Li, Bi Lian, Hanqing Gu, Yan Li, and Yijun Qi. 2018. “Global Identification of Arabidopsis lncRNAs Reveals the Regulation of MAF4 by a Natural Antisense RNA.” Nature Communications 9 (1): 5056.
**Supplemental Figure 1:** Assessing the assembly quality of Arabidopsis lincRNAs. 

**A)** Illumina short read RNA-seq lincRNA which was reassessed as an UTR extension of a neighboring mRNA based on Nanopore long read sequencing. 

**B-C)** Two different lincRNAs initially believed to be mRNA associated, but upon closer inspection were miscalled due to apparent genomic DNA contamination in the ONT-sequencing data. 

**D)** Comparing the annotated gene structure of lincRNAs assembled in both long and short sequencing reads (n = 357). 334 of the lincRNAs assembled in both technologies were in complete agreement regarding 5’ and 3’ positions, as well as exon structure. 13 lincRNAs were annotated as being, on average, 190 nt longer by ONT-sequencing, whereas 11 were annotated as being 217 nt longer in the 3’ direction.
Supplemental Figure 2: Scatterplots describing the lack of correlation between sORF length and lincRNA transcript length for Araport (top) lincRNAs and Evolinc (bottom) lincRNAs.

For Araport lincRNAs (n = 81):

\[ \hat{y} = 1040 - 2.15x \quad R^2 < 0.01 \]

For Evolinc lincRNAs (n = 77):

\[ \hat{y} = 1040 - 3.9x \quad R^2 = 0.01 \]
Supplemental Figure 3: Further basic characterization of lincRNAs. A) Exon per transcript distribution of lincRNAs and mRNAs in each of the 4 focal species (mRNAs in blue and lincRNAs in red.) B) Metagene plots of CpG and H3K27me3 for transposable elements, as well as H3K9 acetylation for lincRNAs, mRNAs, and TEs. Transcripts are separated based on expression from paired RNA-seq data.
Supplemental Figure 4: Additional expression characteristics. A) ONT-sequencing derived maximum TPM values for lincRNAs and mRNAs for each focal species. Each species’ mRNA-lincRNA comparison is significantly different at $P < 1.0e^{-9}$ using a pairwise Wilcoxon rank sum test with Bonferroni multiple testing correction. B) Tissue specificity comparisons from Nanopore TPM values for mRNAs and lincRNAs from all four species. Each species’ mRNA-lincRNA comparison is significantly different at $P < 2e^{-16}$ using a pairwise Wilcoxon rank sum test with Bonferroni multiple testing correction. C) Relationship between maximum expression (TPM) and tissue specificity (TAU) between all expressed Arabidopsis lincRNAs (left) and mRNAs (right) within female reproductive tissues (top) leaf tissue, male reproductive tissue, and all tissues combined (bottom). Black lines represent the best fit line of the data. D) Relationship between expression levels of low TAU (broadly expressed) lincRNAs and their neighboring mRNAs divided into two groupings based on distance to closest mRNA.
Supplemental Figure 5: Example screenshot of Clustergrammer Jupyter notebook in which users can examine normalized expression data based on Z-score.
Supplemental Figure 6: Evolutionary features of Brassicaceae lincRNAs. A) Percent of low-confidence and high-confidence Arabidopsis lincRNAs that are sequence conserved at each evolutionary node. Asterisks denote significant difference between observed homolog recovery for the two classes of lincRNAs (p-value <<< 0.01; Student’s t-test). B-D) Conservation of lincRNAs from Camelina (B), Brassica (C), and Eutrema (D) across representative Brassicales. The purple wedge in the pie chart in each panel represents the percent of lincRNAs for which sequence homologs were recovered at each node, thus indicating that each lincRNA was conserved to at least that node. E) Number of Arabidopsis lincRNAs for which the sequence homolog corresponded to another lincRNA (green bar), mRNA (orange bar) or unannotated sequence (pink bar) at each particular node. F) Number of lincRNAs originating from each of the Brassica rapa subgenomes (LF = least fractionated, MF1 = medium fractionated, and MF2 = most fractionated). Coordinates for
Supplemental Figure 7: Evolution of functionally characterized lincRNAs. A) Schematic representing conservation of the HID1 locus across representative Brassicaceae. In Arabidopsis HID1A represents the published HID1 locus. All other green boxes represent loci inferred based on sequence homology and synteny. B) Conservation of DRIR1. Although no sequence homologs were identified for DRIR1, a putative transcriptional syntelog (Blue box) was identified in Brassica at a syntenic locus. C) Conservation of APOLO. Although APOLO sequence homologs (i.e., paralogs) were identified in Arabidopsis lyrata, none were adjacent to the PID1 ortholog, the protein-coding gene known to be regulated by APOLO in Arabidopsis thaliana. D) Conservation of ELENA. ELENA sequence homologs were identified in species as distantly related as Camelina sativa, where they were situated in syntenic positions adjacent to CBL6 orthologs. E) Conservation of SVALKA. Sequence homologs of SVALKA were identified in Camelina adjacent to CBF1. In Brassica, no sequence homologs were identified, but several putative transcriptional syntelogs were recovered adjacent to CBF1 orthologs.
Supplemental Figure 8: Subgenome expression dominance of *Camelina* lincRNAs. A) Expression values of lincRNAs found in all three subgenomes of *C. sativa*. B) Expression values of lincRNAs specific to one subgenome. C) Expression values of protein coding genes found in all three subgenomes of *C. sativa*. D) Expression values of protein coding genes specific to one subgenome. Numbers represent Student’s t-test $P$ values between groups of expression values.
Supplemental Figure 9: Differential expression during stress. A) Upset plot of Arabidopsis lincRNAs differentially expressed in a variety of broad stress categories (an expanded set of stresses compared to Figure 6A). B) Upset plot of Arabidopsis mRNAs found to be differentially expressed in various abiotic stresses. C) Upset plot of Brassica mRNAs found to be differentially expressed in various broad stress categories. D) Scatterplot comparing log2FC of Arabidopsis mRNAs in cold and heat stress when mRNAs are DE in both, or just a single stress. Note that a majority of differentially expressed mRNAs are up or down in both conditions.
Supplemental Figure 10: Enriched GO terms of mRNAs found in WGCNA modules from Figure 6E (A) and Figure 6F (B). C) Co-expression module of cold-induced genes from the Klepikova stress atlas data. This module includes six lincRNAs, one of which is SVALKA. D) Enriched GO terms of mRNAs found in (C).
Supplemental Figure 11: Gene expression correlation (Pearson) between lincRNA-mRNA and mRNA-mRNA pairs within defined distances in Arabidopsis and Brassica tissue atlases. A) Arabidopsis gene expression correlation of all expressed lincRNA/mRNAs with nearby expressed mRNAs within defined distances (x-axis). B) Brassica gene expression correlation of all expressed lincRNA/mRNAs with nearby expressed mRNAs within defined distances (x-axis). Note, all pairs within smaller distances are contained within larger distances.
Supplemental Figure 12: WGCNA modules of biological interest from the Lee and Serres (2019) hypoxia datasets containing lincRNAs. Expression profiles and enriched GO terms from module 1 (A and B), module 7 (C and D), module 9 (E and F), and module 10 (G and H).