UNCOVERING STRUCTURAL GENOMIC CONTENTS OF WHEAT

by
HALİSE BÜŞRA ÇAĞIRICI

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

Sabancı University
July 2019

UNCOVERING STRUCTURAL GENOMIC CONTENTS OF WHEAT

Approved by:

Assoc. Prof. Levent Öztürk ……..
(Thesis Supervisor)

Prof. Dr. Hikmet Budak ……..
(Thesis Co-Advisor)

Asst. Prof. Bahar Soğutmaz Özdemir ……..

Prof. Dr. Ali Koşar ……..

Asst. Prof. Hasan Kurt ……..

Assoc. Prof. Meral Yüce ……..

Approval Date: July 12, 2019

ABSTRACT

UNCOVERING STRUCTURAL GENOMIC CONTENTS OF WHEAT

HALİSE BÜŞRA ÇAĞIRICI

MOLECULAR BIOLOGY, GENETICS AND BIOINFORMATICS
PH.D. DISSERTATION, JULY 2019

Dissertation Supervisor: Assoc. Prof. Levent Öztürk
Dissertation Co-Advisor: Prof. Dr. Hikmet Budak

Keywords: wheat, lncRNAs, miRNAs, SNPs, machine learning

Production rate of wheat, an important food source worldwide, is significantly limited by both biotic and abiotic stress factors. Development of stress resistant cultivars are highly dependent on the understanding of the molecular mechanisms and structural elements in wheat and/or wheat interacting species. The huge and complex genome of bread wheat (BBAADD genome) has stood as a vital obstruction for understanding the molecular mechanisms until the recent availability of wheat reference genome. In this study, we provided improved and/or novel methodologies to reveal structural elements in plants. These methodologies include miRNA identification, manual curation of lncRNAs, identification of lncRNAs using wheat specific prediction models and a comparative analysis of WES data analysis tools. Using these techniques, we here focused on the uncovering of structural genomic contents of wheat.

With an improved identification methodologies and manual annotation of lncRNAs, we revealed several miRNAs and lncRNAs in *Triticum turgidum* species and *Wheat stem sawfly* (WSS), a major pest of wheat. We provided a comprehensive transcriptome

analysis of tetraploid wheat varieties and revealed drought responsive transcripts. Additionally, we presented the first clues of miRNA mobility between WSS larva and hexaploid wheat. Thereby, besides enrichment of the genetic information available for wheat species, this study provides important elements driving both abiotic and biotic stress responses in wheat. In this study, we also applied machine learning approaches for the fast and accurate prediction of lncRNAs in wheat species. With annotated genomes of hexaploid and tetraploid wheats, we provided better accuracy scores (99.81%) over the most popular tools available. Finally, we conducted a comparative analysis of the tools used for variant discovery. Among eight aligners and three callers, we chose the best combination for the variant calling in wheat. Later, we performed variant calling in 48 lines of elite wheat cultivars using the best tool sets. Overall, this study focused on the improvements on the identification of miRNAs, lncRNAs and structural variations in wheat.

ÖZET

BUĞDAYIN YAPISAL GENOMİK İÇERİKLERİNİN ORTAYA ÇIKARILMASI

HALİSE BÜŞRA ÇAĞIRICI

MOLEKÜLER BİYOLOJİ, GENETİK VE BİYOİNFORMATİK
DOKTORA TEZİ, JULY 2019

Tez Danışmanı: Assoc. Prof. Levent Öztürk
Tez Eşdanışmanı: Prof. Dr. Hikmet Budak

Anahtar Kelimeler: buğday, lncRNAs, miRNAs, SNPs, yapay zeka ile öğrenme

Dünya genelinde önemli bir gıda olan buğdayın üretim hızı çeşitli stress faktörleri tarafından kısıtlanmaktadır. Strese dayanıklı kültürlerin geliştirilmesi ise buğday ve/veya buğday ile etkileşimde olan türlerin moleküler mekanizmalarının ve yapısal elementlerinin anlaşılmasıyla sağlanabilir. Günümüzdeki referans genomunun yayınlanmasına kadar, buğdayın büyük ve karmaşık genom yapısı bu moleküler mekanizmaların anlaşılmasını zorlaştırıyordu. Bu çalışmada, bitkilerin yapısal parçalarının anlaşılmasını sağlayacak methodlar oluşturmaya, olan methodlarıysa geliştirmeye çalıştık. Bahsi geçen methodlar; miRNA belirleme, tüm özelliklerine bakarak elle lncRNA belirleme, yapay zeka kullanarak buğday genomuna özel lncRNA tanımlama and WES data analizlerinde kullanılan programların karşılaştırılması. Tüm bu methodları kullanarak, buğday genomunun yapısal elementlerini bu çalışmada ortaya çıkartmaya çalıştık.

Geliştirilen tanılama methodları ve lncRNA moleküllerinin manuel belirlenmesi ile durum buğdağı ve ekmeklik buğdayın önemli bir böceği olan ekin sap arısında birçok miRNA ve lncRNA molekülleri ortaya çıkardık. Tetraploid buğday türlerinde kapsamplı bir transkriptom analizi gerçekleştirdik ve kuraklığa duyarlı transcriptleri ortaya çıkardık. Ayrıca, sap arısı larvası ile buğday arasındaki miRNA geçişine yönelik bulguları gösteren ilk çalışmayı sunduk. Böylece, buğday türlerine ait bilinen genetik bilgileri artırmanın dışında, bu çalışma buğdayın biotik ve abyotik stress tepkilerini çalıştıran önemli elementleri de ortaya çıkarmaktadır. Bu çalışmada, aynı zamanda, buğday lncRNA moleküllerinin doğru ve hızlı tanımlayabilmek için yapay zeka kullanılmıştır. Anotasyonları yapılmış hexaploid ve tetraploid buğdağ genomlarını da kullanarak, en sık kullanılan programların üzerinde bir doğruluk payı (%99.81) sağladık. Son olarak, varyant tanımlama için kullanılan programların karşılaştırmalı değerlendirmesini yaptık. Sekiz eşleştirici ve üç tarayıcı arasından buğday için en etkili kombinasyonu seçtik. Sonrasında, bu en iyi kombinasyonu kullanarak, 48 farklı elit buğday kültüründeki varyantları ortaya çıkardık. Genel olarak, bu çalışmada buğday bitkilerindeki yapısal değişkenler, miRNA ve lncRNA molekülleri ortaya çıkarılmıştır.

ACKNOWLEDGEMENTS

I would like to express my sincere gratidudes to both of my supervisors Prof. Dr. Hikmet Budak and Assoc. Prof. Levent Ozturk for their support. Especially, I owe special thanks to Prof. Dr. Hikmet Budak for his extreme patience, continuous guidance and encouragement throughout my doctroral years and my career. I am grateful for the opportunities he has provided me. His great expertise and advises are so valuable and will guide me throughout my life.

I would also like to thank each member of my thesis committee: Prof. Dr. Ali Koşar, Assoc. Prof. Meral Yüce, Assist. Prof. Bahar Soğutmaz Özdemir and Assist. Prof. Hasan Kurt for their valuable comments and support.

I would like to express my gratitude to current and previous members of the Budak Lab; Dr. Babar Hussain, Dr. Bala Anı Akpınar, Burcu Alptekin, Kadriye Kahraman, Dr. Naimat Ullah, Reyyan Bulut, Sezgi Bıyıklıoğlu Kaya, Dr. Stuart James Lucas, Tuğdem Muslu and Dr. Zaeema Khan who kindly shared knowledge and experience with me and made this journey pleasant.

I need to acknowledge the Scientific and Technological Research Council of Turkey (TUBITAK) for the financial support they have provided during my doctoral studies.

Finally, but most important of all, I would like to express my dearest gratitude to each member of my large family. My special thanks belong to my beloved husband, Emre Cagirici, for always being there to support, care and love me whenever I need without questioning. I could not be grateful more to my dearest people in the world; my parents Ethem and Behiye Angın; my sisters Hatice Kübra Baz, Fatma Feyza Angın and Rabia Angın; and my little brother Muhammed Eren Angın. I owe all my achievements to their unconditional love and support.

To my dear family,

TABLE OF CONTENTS

1. GENERAL INTRODUCTION 7
2. GENERAL MATERIALS AND METHODS 13
2.1. SUmir Pipeline 13
3. RNA SEQUENCING and CO-EXPRESSED LONG NON-CODING RNA IN MODERN AND WILD WHEATS 15
3.1. Introduction 15
3.2. Materials and Methods 17
3.2.1. Total transcriptome sequencing, assembly and identification of differentially expressed transcripts 17
3.2.2. Annotation of transcripts and identification of long non-coding RNAs 18
3.2.3. Genome mapping and splicing 21
3.2.4. Quantitative Real Time PCR (QRT-PCR) analysis for miRNA, mRNA and lncRNA transcripts 21
3.2.5. Construction of mRNA-lncRNA-miRNA networks 23
3.3. Results 23
3.3.1. *De novo* assembly of transcriptomes 23
3.3.2. Functional annotation of transcriptomes 24
3.3.3. Putative lncRNAs and their expression pattern under drought stress 29
3.3.4. Characteristics of actively expressed lncRNAs 32
3.3.5. miRNA-related functions of lncRNAs 36
3.3.6. Functional characterization of lncRNAs through lncRNA-miRNA-mRNA networks 38
3.4. Discussion 41
4. ASSEMBLY AND ANNOTATION OF TRANSCRIPTOME PROVIDED EVIDENCE OF MIRNA MOBILITY BETWEEN WHEAT AND WHEAT STEM SAWFLY 45
4.1. Introduction 45

4.2. Materials and Methods 48
4.2.1. *De novo* assembly and differential expression of transcripts 48
4.2.2. Annotation of transcripts and lncRNAs 49
4.2.3. Identification and annotation of miRNAs and tRNAs 50
4.2.4. Prediction of miRNA targets 50
4.3. Results 51
4.3.1. De novo assembly of WSS transcriptome 51
4.3.2. Annotation of WSS transcriptome 52
4.3.3. Identification of lncRNAs 54
4.3.4. Characteristics of lncRNAs and mRNAs 56
4.3.5. tRNA annotation 57
4.3.6. *In silico* miRNA prediction 58
4.3.7. Putative targets of WSS miRNAs 60
4.3.8. lncRNA - miRNA - mRNA network in WSS 61
4.3.9. Bidirectional mobility of miRNA in wheat and WSS 62
4.4. Discussion 64
5. CONSTRUCTION OF LONG NON-CODING RNA IDENTIFICATION MODEL SPECIFIC FOR WHEAT SPECIES USING MACHINE LEARNING APPROACHES 71
5.1. Introduction 71
5.2. Materials and Methods 73
5.2.1. Datasets 73
5.2.2. Feature extraction 74
5.2.3. Model construction 74
5.2.4. Model evaluation 75
5.3. Results 76
5.3.1. Experimental setup and model construction 76
5.3.2. Performance evaluation on tetraploid wheat data 77
5.3.3. Feature importances 79
5.4. Discussion 80
6. COMPARATIVE ANALYSIS OF WHOLE EXOME SEQUENCING (WES) TOOLS IN WHEAT 82
6.1. Introduction 82
6.2. Materials and Methods 83

6.2.1. Preparation of wheat exome capture libraries 83
6.2.2. Alignment parameters 83
6.2.3. Variant calling 84
6.3. Results 87
6.3.1. Datasets and pipelines evaluated 87
6.3.2. Filtering of variants 88
6.3.3. Comparison of aligners and callers on identification efficacy 90
6.3.4. Comparison of filtered results against wheat HAPMAP data 91
6.3.5. Variations identified using bcftools-bwamem pipeline 94
6.4. Discussion 95
7. BIBLIOGRAPHY 97
8. APPENDIX A 118
9. APPENDIX B 127

LIST OF TABLES

Table 3.1. Statistics about quality trimming and assembly construction for the transcriptome assemblies. 24
Table 4.1. Summary statistics of sequencing and combined de novo transcriptome assembly of WSS 51
Table 4.2. Wheat coding targets of WSS larval miRNAs. 63
Table 5.1. Description of datasets used in training and validation of the wheat lncRNA prediction model 76
Table 5.2. Performance of prediction models using training data with 100-fold cross validation 77
Table 5.3. Performance comparison of prediction models on tetraploid wheat data..... 78
Table 6.1. Average run time of aligners. 88
Table 6.2. Variant statistics for all 24 pipelines 90
Table 6.3. Total number of variations identified 90
Table 6.4. Distribution of SNPs across the wheat chromosomes 95

LIST OF FIGURES

Figure 3.1. The pipeline for the identification and annotation of both coding transcripts and lncRNAs. 20
Figure 3.2. Homology pattern between T. turgidum proteins and other plants 26
Figure 3.3. Heatmap for top 20 biological processes represented by stress-responsive coding transcripts in each sample. 28
Figure 3.4. Expression pattern of coding transcript and lncRNAs in three different T. turgidum samples. 30
Figure 3.5. Common and drought specific lncRNAs from different T. turgidum varieties 30
Figure 3.6. Relative normalized expression analysis results for common differentially expressed mRNA and lncRNAs samples 32
Figure 3.7. Repeat content of stress-responsive lncRNAs 36
Figure 3.8. miRNA regulated networks between lncRNAs and coding transcripts 39
Figure 3.9. Distribution of Gene Ontology mapping results of coding targets of putative miRNAs 39
Figure 3.10. Relative normalized expression analysis results for miRNA-mRNA-lncRNA networks involved differentially expressed transcripts 40
Figure 4.1. Comparison of the expressions of transcripts 53
Figure 4.2. Blast2GO term distribution over differentially expressed transcripts 54
Figure 4.3. Structural features of coding and non-coding elements in WSS transcriptome 55
Figure 4.4. tRNA content of mRNA, lncRNA and the remaining transcripts in the WSS transcriptome 58
Figure 4.5. Venn diagram representing sample specific expression of WSS miRNAs. . 59
Figure 4.6. miRNA-mediated lncRNA and mRNA interaction networks 61
Figure 5.1. Accuracy of prediction models on coding and non-coding transcripts

separately on tetraploid wheat data 79
Figure 5.2. Feature ranking of wheat specific prediction model based on logistic regression 79
Figure 6.1. Schematic of the variant calling pipeline used 86
Figure 6.2. Comparison between the number of unique SNPs identified and the number of shared SNPs with wheat HAPMAP data. (a) Distribution of aligners, (b) distribution of callers. The best pipelines selected (by <1200000 unique snps and >32000 shared snps) shown by red circles. 91
Figure 6.3. Distribution of SNPs identified by all 24 pipelines 92
Figure 6.4. Distribution of number of SNPs called using bcftools call function as caller 93

1. GENERAL INTRODUCTION

Wheat (*Triticum ssp.*) is one of the major sources of continuously increasing food demand, ranking second in crop production worldwide (Food and Agriculture Organization of the United Nations 2019). Domestication and cultivation efforts of agricultural practices resulted in an increased yield (Marcussen et al. 2014) with the approximate global production of 700 million tons per year distributed over 200 million hectares (Food and Agriculture Organization of the United Nations 2019). Despite this spread distribution, obtained rate of yield is not sufficient to meet world food demand since production rate is significantly limited by biotic and abiotic stress factors (Budak, Hussain, et al. 2015). Drought is one of the major abiotic stress factors worldwide, causing decrease in grain quality and yield loss through all cereals, including wheat (Bala Ani Akpinar, Lucas, and Budak 2013). Recent studies have suggested a substantial increase in drought caused by climate change and global warming (Fang and Xiong 2015). In order to maintain sufficient amount of yield with an improved nutritional quality, development of new wheat varieties with an increased drought tolerance is urgently needed toward future challenges.

Wheat Stem Sawfly (WSS), *Cephus Cinctus* Norton (Hymenoptera: Cephidae) is, on the other hand, stated as the most damaging pest of wheat in Northern Great Plains, causing crop devastations in Montana region each year (Beres et al., 2011). Female WSS choose the internodes of actively elongating fresh wheat stems to lay their eggs. By tearing the stem with their sharp ovipositors, eggs are placed into the stem where the larvae form after 4-7 days of incubation (Cárcamo et al. 2011). Since the larvae are cannibalistic, only one larva can survive in the stem although there are more eggs deposited. Larva stays and develops in the wheat stem during the growing season, feeding on parenchyma and vascular tissues and, eventually, it moves toward the bottom of the stem to cut a notch, causing plant to lodge in order to overwinter there until the pupation occurs. Stem cutting

cause a dramatic reduction in yield, and even uncut infested plants have low yield due to decreased head weight by 17% (Delaney et al., 2010). However, there are still no effective control method over WSS damage in wheat. Usage of chemicals is limited by the long emergence period of females and the wheat stem protecting the eggs and the larva feeding inside (Knodel et al. 2009). The introduction of solid-stemmed wheat instead of hollow-stemmed wheat maintained a more powerful control on the infestations. Yet, the solid-stemmed cultivars are not preferred by producers because of its low yield and protein content compared to hollow-stemmed cultivars (B. Beres et al. 2011).

With the advances in high-throughput sequencing technologies, vast number of transcripts have been discovered in many different species, including mammals, plants, vertebrates etc. (Mercer et al. 2011; Szymański and Barciszewski 2002; (IWGSC) et al. 2018; Claverie 2005). Transcriptomics and genomics studies revealed majority of these transcripts are not coding for functional proteins although their lengths were greater than 200 nucleotides (Pennisi 2012). Such transcripts were called long non-coding RNAs (lncRNAs). The lack of functional studies and evolutionary conservation raised the concerns about the importance of lncRNAs (Struhl 2007) where these concerns have been answered by the functional characterization of lncRNAs in important biological processes (i.e. COOLAIR/COLDAIR) (Jae Bok Heo and Sung 2011). Studies in the last decade have revealed diverse regulatory functions of lncRNAs as their interactions range from lncRNA:RNA to lncRNA:chromatin interactions (Chekanova 2015). The list of plant lncRNAs with best-studied functions involves several important biological processes, including vernalization (Swiezewski et al. 2009), photo morphogenesis (Y. Wang et al. 2014), reproduction (Ding et al. 2012), nodulation (Campalans 2004) and environmental stress adaptation (J. Liu et al. 2012).

Furthermore, lncRNAs tend to have tissue specific expression and conservation in functionality rather than sequence (Ulitsky et al. 2011; Cabili et al. 2011). Although sequence conservation is almost always accounted for the functionality of the sequence, *vice versa* is not always true (Shannon et al. 2003). Diverse functions of lncRNAs might support the different constraints that might drive conservation of different RNAs, such as mRNAs, miRNA and lncRNAs (Hezroni et al. 2015). Instead of full-length sequence conservation, small binding sites for their interacting partners could be conserved. These lncRNAs could be conserved at structural level to maintain functional interactions with

proteins or other DNA/RNAs (Militti et al. 2014).

Bioinformatics approaches are applied to differentiate lncRNAs from other noncoding RNAs and mRNAs. Due to their considerable lengths (usually >200nt), lncRNAs can easily differentiated from small noncoding RNAs. However, the most challenging aspect of lncRNA identification is that lncRNAs are loosely defined; in fact, lncRNAs are mostly defined with the lack of certain properties. A general definition of lncRNAs is long transcripts without a complete ORF.

Current lncRNA studies are focused on; only ORF size and sequence similarity to known protein sequences; machine learning algorithms, such as support vector machines (SVM); or combination of these and several other features. Several features can be used as selection criteria to distinguish lncRNAs from mRNAs: (1) transcript length; (2) ORF length; (3) homology with known proteins; (4) homology with protein domains; (5) intron-exon structure; (6) genomic location; (7) machine learning techniques (J. Liu et al. 2012; T.-Z. Wang et al. 2015; Boerner and McGinnis 2012; L. Li et al. 2014; Jinhui Chen, Quan, and Zhang 2015). The use of machine learning techniques alone has increased the accuracy of coding potential calculations to over 90% (Kong et al. 2007; Sun et al. 2013; Hoff and Stanke 2013). However, the precise identification of lncRNAs seems impossible due to transcripts that are short and protein coding, and transcripts that are noncoding with long ORFs. Even some lncRNAs are derived from protein coding loci. Using combination of filters can adress some of the challenges in sensitive lncRNA identification; though a volatile solution is to cluster transcripts into two categories as high-confidence lncRNAs and low confidence lncRNAs (L. Li et al. 2014).

These in silico predictions revealed plenty of lncRNAs whose expression need to be corfirmed. qRT-PCR allows detection and quantifiation of the expression in real time; therefore, widely used technique to verify expression of in silico predicted lncRNAs (Shuai et al. 2014). Functional annotation of lncRNAs has been carried on in terms of co-expression patterns and/or interaction networks. An expression based functional prediction can be performed to predict functions of lncRNAs based on co-expressed protein-coding genes (Liao et al. 2011; Guttman et al. 2009). For example, the two lncRNAs, COOLAIR and COLDAIR, are expressed in the FLC loci and control the expression of FLC gene that loci (J. B. Heo and Sung 2011). Moreover, lncRNAs can

serve as sRNA targets, where those lncRNAs prevent interaction between the sRNA and its protein-coding target, thereby enhance the function of a particular protein-coding gene (Shuai et al. 2014; Britton et al. 2014). These interaction network between lncRNA, miRNA and mRNAs could reveal the functions of lncRNAs as endogenous Target Mimics (eTMs) (Jie Chen et al. 2013; Franco-Zorrilla et al. 2007). Moreover, lncRNAs can serve as sRNA precursors, where the downstream patterns of the corresponding sRNA could reveal the functioning of lncRNAs in different molecular pathways (Matzke and Mosher 2014; Ariel et al. 2015).

sRNAs, on the other hand, are double stranded RNAs (dsRNAs) with 20-30nt in length and non-coding regulatory elements of genome. They regulate both genome and transcriptome by targeting both chromatin and the transcripts. sRNAs have a tendency to bind with Argonaute (AGO) family proteins, forming RNA-induced silencing complex (RISC). Upon formation, RISC proteins directs mature sRNAs to their target mRNA. This sRNA-mediated gene silencing is mostly known as RNA interference (RNAi). RNAi is a sequence-specific gene silencing mechanism induced by sRNAs. sRNAs comprises of many subgroups like miRNAs, siRNAs, piwi-interacting RNAs (piRNAs) based on their characteristic hairpin structures, homology to coding sequence or for piRNAs, conserved 5’ motif sequence (Britton et al. 2014).

miRNAs are ~22nt in length and processed from endogenous single-stranded hairpin precursors (Guleria et al. 2011). Primary miRNAs are generated by RNA pol II and processed inside the nucleus to produce mature miRNA. Length of mature miRNAs ranges from 21-24 depending on which DCL family member processes (Budak and Akpinar 2015). Mature miRNA has 2 strands miRNA and miRNA* that has an additional 2nt overhang at 3’end. The miRNA duplex is methylated inside the nucleus to protect miRNA from 3’-exonuclease degradation and 3’-uridylation (Guleria et al. 2011; Budak and Akpinar 2015). The methylated miRNA duplex is exported into the cytosol where a helicase unwinds the duplex and exposes the mature miRNA to RISC. Upon binding RISC, miRNA directs RISC towards the target sequence leading either mRNA degradation in case of full complementarity or translational repression in case of partial complementarity (S. J. Lucas and Budak 2012). A near-perfect complementarity is required for the functioning of miRNAs indicating that miRNAs might have been evolved from the duplicated copies of their targets therefore exhibiting homology to their targets.

Other sources of miRNA formation comprise of transposable elements (TEs), random unstructured sequences, and non-canonical processing. New miRNAs are prone to be lost quickly if complementary sequences do not exist or they exhibit improper processing. Once a miRNA is generated, miRNA families are formed by tandem or segmental duplications.

Genomic variations in coding regions are important factors leading to genome diversity where currently several structural variations are associated with phenotypic traits (Henry et al. 2014). Even certain molecular markers determined for economical physical traits in plants (Zanke et al. 2014). Moreover, these genomic variations are reliable sources for the identification of complex traits as they are not affected by environmental conditions (Hussain et al. 2017).

Until recently, identification of coding and non-coding RNAs was studied via construction of a complementary (cDNA) library and cloning randomly to generate expressed sequence tags (ESTs) (Adams et al. 1991), which are time-consuming and labor-intensive processes (Hennebert et al. 2015; Budak and Akpinar 2015). However, with the recent techniques i.e. chromosome sorting, even individual chromosomes of large complex organisms can be studied (S. J. Lucas and Budak 2012). For the complex genomes like wheat identification of alleles associated with phenotypic traits is not as smooth as crops like Arabidopsis (Cao et al. 2011). However, this complexity can be decreased by exome capture sequencing (Winfield et al. 2012). These techniques fasten the genome studies on complex organisms like wheat and these genomic sequences can be used in many aspects i.e. identification small RNAs, lncRNAs and structural variations.

Therefore, with the advent of NGS technologies, identification of structural elements can be studied in silico in both model and non-model organisms. The process is based on distinctive features of RNAs such as characteristics folding patterns, conservation among known plant RNA molecules and homology to coding sequence. Besides, interrogation of functions or mechnisms of these structural elements is highly dependent on high-quality transcript models where increased throughput and methodological advances are continously improving identification processes.

In this study, we focused on implementation of currently available tools to improve their performances in order to uncover structural genomic contents of wheat. With the recent releases and availability of the wheat reference genome, mining of wheat genome for structural elements with regulatory functions has become more available. Given reference genome annotation and several advance bioinformatics tools, the question has become how to choose the best tools.

In Chapter 2, the general materials and methods section, we introduced an improved pipeline for in silico identification of plant miRNAs. We used this miRNA identification pipeline in the following chapter 3 and 4. Both in chapter 3 and 4, we performed identification of lncRNAs through manual annotation of transcripts. In Chapter 3, we identified both coding and lncRNAs that might drive drought resistance in *Triticum turgidum* species and compare three cultivars with varying drought tolerance levels under drought and control conditions. In Chapter 4, we identified lncRNAs in WSS (wheat stem sawfly) and presented possible interactions of RNAs between larvae and wheat seeds during larval feeding. In Chapter 5, we presented a novel lncRNA prediction model trained on wheat which performs lncRNA prediction in minutes where manual annotation took months during work presented in chapter 3 and 4. Finally, in chapter 6, we performed a comparative analysis of whole exome sequencing data analysis tools. Overall, this study focused on the improvements on the identification of miRNAs, lncRNAs and structural variations in wheat.

2. GENERAL MATERIALS AND METHODS

2.1. SUmir Pipeline

SUmir pipeline were initiated by Lucas et. al. at 2012 (S. J. Lucas and Budak 2012). SUmirFind performed homology screening against a given miRNA query. It uses blastn with parameters optimized for small RNA screenings without any mismatches allowed. Later, SUmirFold evaluates secondary structure of predicted precursor sequences using UNAFold. Afterwards, putative miRNAs were selected manually, which was time consuming lasting ~30 days for large genomes like wheat. We implemented SUmirScreen, a python script which evaluates candidate miRNAs eliminating manual inspection. Additionally, we introduced SUmirLocate, python script, to extract statistics on genomic distribution of predicted miRNAs. Altogether, we made this SUmir pipeline fully automated and error prone from human mistakes.

SUmir pipeline were used in both chapter 3 and chapter 4 in this thesis. Additionally, other studies used this pipeline include the annotation of wheat reference genome (Appels et al. 2018). The scripts were released on GitHub with the following links:

https://github.com/hikmetbudak/miRNA-annotation/blob/master/SUmirScreen_v2.py
https://github.com/hikmetbudak/miRNA-annotation/blob/master/SUmirLocate_v2.py

In general, high confidence mature miRNA sequences of were retrieved from miRBase database (v21, June 2016) (Kozomara and Griffiths-Jones 2011). *In silico* miRNA prediction was performed based on homology and secondary structure predictions. *De novo* assembled transcriptome was subjected to homology screening to predict putative mature miRNA sequences, allowing at most 1 base mismatch using SUmirFind. Predicted

mature miRNA sequences were extended from both ends to predict pre-miRNA sequences after when they can be subjected to UNAFold (Markham and Zuker 2008) to simulate RNA folding. Secondary structure predictions evaluate characteristics of hairpin structure to differentiate miRNAs from other ssRNAs by several parameters including MFEI and GC content using SUmirFold.

Later, final evaluations were performed based on strict criteria of correct folding: (1) max number of mismatches allowed are 4 for miRNA and 6 for miRNA* sequences; (2) no mismatches allowed at Dicer-Like enzyme cut sites; (3) multi-loop structures are not allowed between miRNA and miRNA*; (4) miRNA or miRNA* sequences cannot be involved in the head part of the hairpin, using SUmirScreen.

3. RNA SEQUENCING and CO-EXPRESSED LONG NON-CODING RNA IN MODERN AND WILD WHEATS

3.1. Introduction

Wild plants evolved sophisticated stress tolerance and adaptation mechanisms to drought where the domestication of modern wheat varieties has led to the loss of these valuable genes in the process of domestication (Alptekin and Budak 2016). Introgression of the valuable elements from wild relatives has been an attractive approach for agronomical improvement of modern wheat varieties for decades (Merchuk-Ovnat et al. 2016), due to their rich gene pool for the resistance to many different stress factors. Tetraploid emmer wheat (*T. turdigum ssp. dicoccoides*, 2n=28, AABB) is the wild progenitor of the allohexaploid bread wheat (*T. aestivum*, 6n=42, AABBDD) and the domesticated tetraploid durum wheat (*T. turgidum ssp. durum*, 4n=24, AABB) (Marcussen et al. 2014). Recent studies on tetraploid wheat varieties revealed contrasting drought tolerance in tetraploid wild emmer wheat varieties and domesticated tetraploid durum wheat (Bala Ani Akpinar, Kantar, and Budak 2015; Ergen and Budak 2009). Ergen and colleagues surveyed drought response of several genotypes of wild and domesticated tetraploid wheat varieties; they were able to show that wild emmer wheat, genotype TR39477, exhibits the highest drought tolerance while TTD-22 genotype has the lowest tolerance under drought stress. On the other hand, durum wheat variety Kiziltan showed a moderate tolerance in response to slow drought imposition (Ergen and Budak 2009); however, complete mechanism of these drought responses remains elusive. A better understanding of the genomic background and the molecular mechanisms of drought responses in wild progenitors of wheat might reveal such favorable regulatory elements lost during domestication and cultivation processes.

In recent years, technological advances have made road into the reduction in the cost of sequencing experiments and availability of several genomic and transcriptomic data from bread wheat and its relatives/progenitors (The International Wheat Genome Sequencing Consortium 2014; Budak and Kantar 2015). Particularly, transcriptomic studies shed light into the differential expression and regulation of several transcripts under biotic and abiotic stress conditions, which further provide insights about the molecular mechanism associated with stress tolerance (Bala Ani Akpinar, Lucas, and Budak 2013; Budak, Kantar, et al. 2015; Budak, Khan, and Kantar 2015). Total transcriptome sequencing and annotation possess a potential for detection of protein coding transcripts, differentially regulated under stress conditions, together with their non-coding interacting partners which is associated with a large portion of transcriptomes (Griffiths-Jones 2007). The content and the amount of the non-coding RNAs (ncRNAs) in the genome show an increased correspondence with the genome complexity which further supports their regulatory roles (Guleria et al. 2011; Budak et al. 2016). Over the last decades, extensive studies in both animals and plants have shed light into the functions and mechanisms of ncRNAs such as microRNAs (miRNAs) and small interfering RNAs (siRNAs) in the transcriptional and post-transcriptional regulation of gene expression (Budak and Akpinar 2015; Bala A. Akpinar and Budak 2016). While the miRNAs and siRNAs are referred as small RNAs (sRNAs) based on their small length ranging between 18 to 24 nucleotides, another type of ncRNAs longer than 200 nucleotides has been recently defined as long non-coding RNAs (lncRNAs) (Chekanova 2015; J. Liu et al. 2012). LncRNAs resemble messenger RNAs (mRNAs) in their structure and biogenesis process, i.e., they are mainly transcribed by RNA Pol II and poly-adenylated, as though mRNAs (J. Liu et al. 2012). Additionally, they might possess multiple exons and are subjected to alternative splicing. The major factor distinguishing lncRNAs from mRNAs is the lack of discernable coding potential of lncRNAs (Quinn and Chang 2015). Besides, lncRNAs are composed of ~3 exons on average as opposed to ~11 exons in mRNAs and exhibit a more tissue-specific expression pattern compared to mRNAs where their expression is also relatively less than mRNAs in a given tissue (Quinn and Chang 2015).

Emerging evidence has suggested that lncRNAs have regulatory roles in the major biological processes such as development, vernalization, nodulation and environmental stress adaptation both in direct and indirect manner (J. Liu et al. 2012). As an example, two lncRNAs, the long antisense intragenic RNA (COOLAIR) and the intronic noncoding

RNA (COLDAIR), have been detected as mediating the flowering process in *Arabidopsis* through silencing and epigenetic repression of Flowering Locus C (FLC) (Swiezewski et al. 2009). Additionally, several studies evinced the functions of lncRNAs in the biogenesis and targeting process of small-noncoding RNAs by possessing miRNA-siRNA precursor potential and sRNA target mimicry (Chekanova 2015). RNA-dependent DNA Methylation (RdDM) in plants, for example, utilizes lncRNAs acting as precursors of siRNAs which later target lncRNAs acting as scaffold RNAs recruiting siRNA-AGO4 complex together with RDM1 (RNA-directed DNA Methylation 1) to a target genomic loci for DNA methylation-mediated silencing (Lai and Shiekhattar 2014). In another example, lncRNA IPS1 has been shown to inhibit miR399-mediated cleavage of PHO2 as a competitor for PHO2 mRNA (Shin et al. 2006). LncRNAs have also been identified as differentially expressed under several stress conditions and their regulation on both mRNA and sRNA pool detected as critical for stress tolerance and maintenance of vitality (X. Lu et al. 2016); however, not that much effort has been done in drought responsive lncRNAs and their association with coding and other non-coding RNA species, particularly in cereals. Here, we present a detailed analysis of drought responsive mRNAs and lncRNAs along with their particular interaction with each other in three different tetraploid wheat varieties. Results revealed the presence of more than 200 putative stress responsive lncRNAs per cultivar which provided insights about drought tolerance mechanism in ancestor of modern wheat. Additionally, this study presents a brief method for precise identification and detailed characterization of lncRNAs for plants lacking both an annotated genome and a reference genome.

3.2. Materials and Methods

3.2.1. Total transcriptome sequencing, assembly and identification of differentially expressed transcripts

In a previous study of our group, a number of wild wheat varieties were subjected to slow drought imposition where two wild emmer wheat (*T. turgidum ssp. dicoccoides*) varieties, TR39477 and TTD-22, exhibited contrasting responses as the most tolerant and the most sensitive compared to the cultivated durum wheat (*T. turgidum ssp. durum*) variety

Kiziltan with moderate response (Bala Ani Akpinar, Kantar, and Budak 2015). Total RNA isolation from a pool of three biological replicates of the root samples of control and drought-stressed modern durum wheat, Kiziltan, and wild emmer wheats, TR39477 and TTD-22, conducted with TRI Reagent (Molecular Research Center, Cincinnati, OH, USA) following the manufacturer's recommendations and RNA integrity was controlled using Agilent Bioanalyzer 2100 RNA 6000 Nano Kit (Agilent Technologies, Santa Clara, CA, USA) (Bala Ani Akpinar, Kantar, and Budak 2015). Following, high-throughput sequencing with Illumina HiSeq 2000 were performed with the libraries prepared by using TruSeq RNA Sample Prep Kit v2 (Bala Ani Akpinar, Kantar, and Budak 2015). Illumina HiSeq 2000 paired end reads can be accessible at ENA database with the run ID: ERR1987529.

Raw paired-end reads from RNA sequencing of these six samples (three genotypes x two conditions) were quality trimmed using Trimmomatic (v0.32) with default parameters (LEADING:5, TRAILING:5, MINLEN:36) (Bolger, Lohse, and Usadel 2014). De novo assembly for each genotype was generated by Trinity platform (Haas et al. 2013) (release 2014-07-17) from combination of paired-end Illumina reads of control and drought-stressed samples. Assembled transcripts were aligned back to the raw reads using bowtie aligner and the abundance estimation of all transcripts was quantified as FPKM with utilization of RSEM under Trinity pipeline. Individual assembly files for each control and drought-stressed samples were separated based on their corresponding abundance estimates for further analysis. Differential expression analysis was conducted using EdgeR pipeline (Robinson, McCarthy, and Smyth 2010) with the default threshold parameters of p-value=0,001 and log2(fold_change)=2.

3.2.2. Annotation of transcripts and identification of long non-coding RNAs

Following transcriptome assembly, annotation of transcripts and identification of lncRNAs were performed through following rigorous criteria: exclusion of contaminants, open reading frame (ORF) size prediction, *ab initio* predictions and homology screenings. As the first layer of analyses, transcripts were excluded from the assemblies if defined as contaminants after blast screenings against *Triticum turgidum* non-coding RNAs deposited at NCBI and ENA databases (1E-05, -pident 95, -length 30); rRNA, tRNA,

snoRNA, snRNA sequences of *Triticum* families deposited in NCBI database (1E-05, -pident 95, -length 30); *Triticum aestivum* mitochondrion complete genome (NC_007579.1) and *Triticum turgidum* organellar RNAs deposited at NCBI and ENA databases (1E-15, -pident 95, -length 30). Remaining analyses evaluated the coding potential of transcripts and aid to determine either lncRNAs or coding transcripts.

Subsequent to contaminant analysis, the ORF size prediction for each assembly was conducted in order to differentiate between protein coding and non-coding transcripts. Since many transposons have similar ORFs to host genes which may corrupt the coding gene annotation, all assemblies were subjected to repeat-masking prior to ORF content predictions, against the repeat library of MIPS Repeat Element Database v9.3 p for *Poaceae* (ftp://ftp.mips.helmholz-muenchen.de/plants/REdat/) (Nussbaumer et al. 2013) using RepeatMasker v4.0.5 software (Tarailo-Graovac and Chen 2009). Ability of repeat-masked transcripts to construct a full-length protein was evaluated by employing two different software, Transdecoder (*-m* 80) and EMBOSS:getorf (Rice, Longden, and Bleasby 2000)]. Transcripts with a continuous ORF >240 nucleotide in length were accepted as possess a functional ORF. Coding potentials of transcripts were predicted with several *ab initio* methods; CPC online tool (*options:* reverse strand mode was included) (Kong et al. 2007), CNCI (version 2 , *options:* -m pl) (Sun et al. 2013) and AUGUSTUS online tool (Hoff and Stanke 2013) with the pre-established system trained for Triticum/wheat. Transcripts identified as 'coding' by at least one of these tools satisfy the *ab initio* prediction criterion.

In order to identify homolog coding transcripts with other species, assemblies were aligned to a dataset of coding sequences using Blast tool kit (version 31) (Camacho et al. 2009). All transcripts were initially blasted against several datasets; Uniprot/Swissprot database (http://web.expasy.org/docs/swiss-prot_guideline.html) (parameters: *-evalue* 1E-05, *-pident* 80, *-length* 30); *Triticum aestivum* UniGenes (https://www.ncbi.nlm.nih.gov/unigene, build#63) (parameters: *-evalue* 1E-30, *-pident* 98, *-length* 90, *-max-target-seqs* 1); *Triticum turgidum* ESTs and coding sequences deposited at NCBI (https://www.ncbi.nlm.nih.gov/) and ENA (http://www.ebi.ac.uk/ena) databases (parameters: *-evalue* 1E-05, *-pident* 95, *-length* 30) together with fully annotated proteins from *Brachypodium distachyon* (v1.2, http://mips.helmholtz-muenchen.de/plant/brachypodium) (Initiative 2010), *Oryza sativa* (IRGSP-1.0,

http://rapdb.dna.affrc.go.jp(download/irgsp1.html) (Tanaka et al. 2008), *Sorghum bicolor* (v1.4, http://mips.helmholtz-muenchen.de/plant/sorghum) (Paterson et al. 2009), high confidence proteins from *Hordeum vulgare* (http://mips.helmholtz-muenchen.de/plant/barley/) (Mayer et al. 2012), and Triticum UniProt sequences (144,397 entries, http://uniprot.org/) (parameters: *-evalue* 1E-05, *-pident* 95, *-length* 30). Additionally, Transdecoder (*-m* 30) predicted peptide sequences of each transcript were screened using against Swissprot entries (parameters: *-blastp*, *-evalue* 1E-05, *-pident* 80, *-length* 30). Conserved protein domains preserved in these peptides were also controlled with Hmmer (v.3.1b1) against Pfam domains (*-evalue* 1E-05) (Z. Zhang and Wood 2003). Transcripts with homology evidence from any of these screenings were accepted as satisfy homology-based prediction criterion.

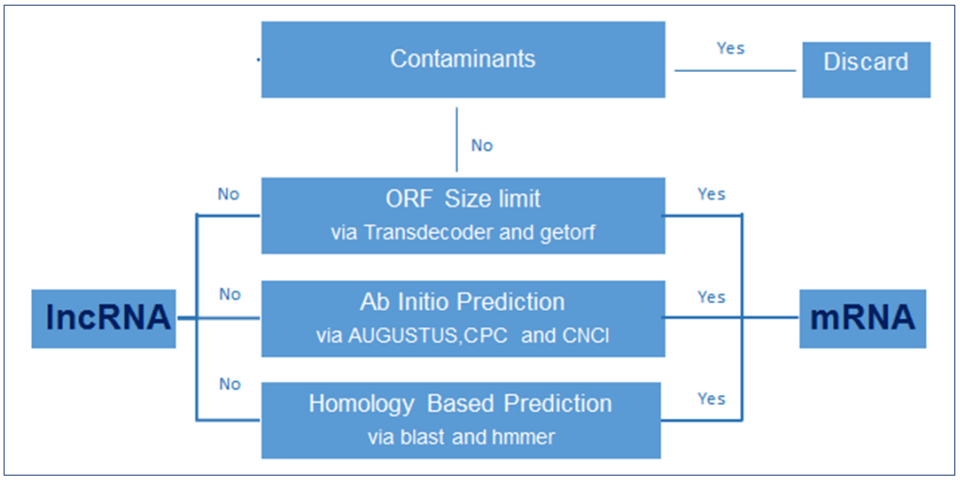


Figure 3.1. The pipeline for the identification and annotation of both coding transcripts and lncRNAs.

As described in Figure 3.1, after exclusion of contaminants, transcripts qualifying remaining criteria were defined as coding transcripts where transcripts with no evidence of coding in ORF size predictions, *ab initio* predictions and homology screenings were defined as lncRNAs. Final functional annotation of coding transcripts were carried out using Blast2GO software (Conesa and Götz 2008) with the initial blast screen run locally against all *Viridiplantae* (taxid: 33090) proteins from the NCBI database (parameters: *blastx*, *-evalue* 1E-5, *-outfmt* 5, *-max_target_seq* 1).

3.2.3. Genome mapping and splicing

Wild emmer wheat genome was obtained from WEWseq consortium (Zavitan_v2_pseudomolecule: http://wewseq.wixsite.com/consortium (Avni et al. 2017)). Transcripts were mapped to Zavitan genome using GMAP (version 2016-07-11, with all parameters set to default except –min-identity=90 –cross-species -f 2). Obtained GFF files were converted into GTF format using gffread. GTF files were submitted to ASTALAVISTA database at default settings to identify alternative splicing events.

3.2.4. Quantitative Real Time PCR (QRT-PCR) analysis for miRNA, mRNA and lncRNA transcripts

In order to show the accordance of differential expression analysis based on RNA sequencing with wet lab, quantitative Real Time PCR (qRT-PCR) experiment was performed. Prior to experiment, a subset of lncRNAs and mRNAs were selected. The differentially expressed transcript across Kiziltan, TR39477 and TTD-22 were compared via blast analysis and transcript sequences showing similarity with 80% identity and query coverage between whole samples was defined as 'common'. Between common transcripts, 2 mRNA (Kiz_mRNA_c17408_g1_i1, Kiz_ mRNA_c55246_g1_i1) and 2 lncRNA sequences (Kiz_lncRNA_c118446_g1_i1 and Kiz_lncRNA_c47700_g1_i1) were chosen randomly and qRT-PCR primers were designed (Appendix A - Supplementary Table 1). Kiziltan seed were surface sterilized in 4% sodium hypochlorite and grown in tall plastic jars for 15 days at 23°C with adequate amount of water. At the end of two weeks, plant seedlings reached the four-leaf stage were dehydration shocked for 4 hours by removing them from plastic jars and leaving on paper towels under the same lighting conditions, while control plants were immediately fast frozen in liquid nitrogen. Both root and whole seedling tissues were collected and stored at -80°C. Total RNA isolation from whole collected tissues from control and drought treated samples was performed with TRI Reagent (Molecular Research Center, Cincinnati, OH, USA) following the manufacturer's recommendations. First strand cDNA synthesis was performed on 1µg of total RNA using RevertAid H Minus Reverse Transcriptase (QuantiTect Reverse Transcription Kit, Qiagen) according to manufacturer's protocols. For quantification of mRNA and lncRNAs transcripts from control/ drought stressed

root/whole seedling samples, the reaction mix containing 3µl of 5X diluted of cDNA, 1µl of forward primers, 1 µl of reverse primer and 5 µl of iTaq™ Universal SYBR® Green Supermix (Bio-Rad) incubated in Bio-Rad CFX 96 Thermal Cycler with following conditions: 95 °C for 10 min followed by 40 cycles of 95 °C for 15 s, 60 °C for 1 min and then 95 °C for 15 s. The constitutive gene of *Triticum aestivum* Actin (TaActin) (Yue et al. 2015) was used as internal standard to normalize the transcripts using a gene-specific primer (Appendix A - Supplementary Table 1). The 2-ΔCt method was used to calculate the difference in expression of chosen genes (Livak and Schmittgen 2001).

For validation of mRNA, miRNA and lncRNAs, miR1436-1 and miR1436-4 were chosen and validated together with their mRNA and lncRNA targets. For qRT-PCR analysis of mRNA and lncRNAs, the same method was utilized described in above. In order to obtain cDNA belonging to mature miRNAs, miRNA-specific stem–loop reverse transcription reactions were performed using RevertAid H Minus Reverse Transcriptase with iScript™ Select cDNA Synthesis Kit (BioRad) following the manufacturer's recommendations with slight modifications. Prior to cDNA synthesis, a mix of 500 ng of DNAase treated RNA and 1 µl of miRNA-specific stem loop PCR was incubated at 65 °C. After incubation, 2 µl of GSP enhancer solution, 4 µl of 5x iScript select reaction mix and 1 µl of iScript reverse transcriptase were added and reaction mix (10 µl) was incubated at 42 °C for 1 hour followed by 5 minutes of 85°C incubation to heat-inactivate the reverse transcriptase. For validation of miRNA expression, the reaction mix containing 3µl of 5X diluted of cDNA, 1µl of forward primers, 1 µl of universal reverse primer and 5 µl of iTaq™ Universal SYBR® Green Supermix (Bio-Rad) incubated in Bio-Rad CFX 96 Thermal Cycler with following conditions: 95°C for 5 min, followed by 35–45 cycles of 95°C for 5 s, 60°C for 10 s, and 72°C for 1 s. For melting curve analysis, samples were denaturated at 95°C, then cool to 65°C at 20°C per second. The fluorescence signals were collected at 530 nm wavelength continuously from 65°C to 95°C at 0.2°C per second. The constitutive gene of *Triticum aestivum* rRNA26 homolog (Tenea et al. 2011)was used as internal standard to normalize the miRNA expression (Appendix A - Supplementary Table 1). For internal control, several control genes including TaU6 were attempted and the rRNA26 was chosen because of its expressional stability under different conditions and tissues. The 2-ΔCt method was used to calculate the difference in expression of chosen miRNAs.

3.2.5. Construction of mRNA-lncRNA-miRNA networks

High confidence and/or experimentally identified mature miRNA sequences from 72 *Viridiplantae* species were collected from miRBase (v21, June 2014) (Kozomara and Griffiths-Jones 2011), suggesting a dataset of 1,404 non-redundant mature miRNA sequences. SUmir pipeline (2. General Materials and Methods) was run using this miRNA dataset as query for *in silico* prediction of miRNAs. Following, a list of lncRNA transcripts and a list of coding transcripts are retrieved as target datasets for each transcriptome. These datasets were screened for relative gene targets of miRNAs, predicted from the assemblies, using psRNAtarget web-tool, with user-defined query and target options at default parameters (http://plantgrn.noble.org/psRNATarget/) (Dai and Zhao 2011). lncRNAs functioning as coding-target mimics were evaluated based on the complementary pairs between miRNA-to-coding transcript targets and miRNAs-to-lncRNA targets. Cytoscape 3.3.0 (Shannon et al. 2003) was used for the visualization of lncRNA-miRNA-mRNA interaction networks.

3.3. Results

3.3.1. *De novo* assembly of transcriptomes

In our previous study, two wild emmer wheat genotypes, *T. turgidum ssp. dicoccoides* TR39477 and TTD-22 showed marked differences in tolerance to drought stress when compared to the modern durum wheat, *T. turgidum ssp. durum var.* Kiziltan. Upon slow drought treatment, Kiziltan exhibited a moderate reaction whereas TR39477 and TTD-22 exhibited the most and the least tolerance, respectively (Ergen and Budak 2009). High-throughput sequencing of root samples from control and drought-stressed Kiziltan, TR39477 and TTD-22 led to more than 27,000,000 raw sequence reads (Bala Ani Akpinar, Kantar, and Budak 2015). In order to remove adaptor sequences and perform the quality trimming, Trimmomatic (Bolger, Lohse, and Usadel 2014) analysis evaluated and more than 95% of raw reads were cleaned after initial processing with Trimmomatic (Table 3.1). The clean reads were assembled using Trinity software (Haas et al. 2013) yielding a total of 243,670, 211,709 and 203,230 transcripts for Kiziltan, TR39477 and

TTD-22, respectively. The average contig lengths detected as altering between 666 and 779 nucleotides long where the total transcriptome size ranges between 99.7 to 146.6 Mb (Table 3.1).

Table 3.1. Statistics about quality trimming and assembly construction for the transcriptome assemblies.

The quality trimming of samples						
Samples	**Kiziltan Control**	**Kiziltan Drought**	**TR39477 Control**	**TR39477 Drought**	**TTD-22 Control**	**TTD-22 Drought**
Before trimming	35463556	36944980	35212424	30670932	27505294	32690630
After trimming	33772655	35171816	33698813	29223580	26200743	31249427
Assembly statistics for the samples						
Samples	**Kiziltan Control**	**Kiziltan Drought**	**TR39477 Control**	**TR39477 Drought**	**TTD-22 Control**	**TTD-22 Drought**
Number of transcripts	204128	18817	169762	159940	168314	155170
Median contig length (b)	482	516	478	483	468	494
Average contig	666.67	779.58	731.65	741.67	726.12	752.61
Total length (Mb)	99.78	146.69	129.32	125.24	122.22	116.78
GC%	49.72	49.45	49.98	49.92	50.63	50.25
N50	1024	1082	1001	1108	1004	1056

3.3.2. Functional annotation of transcriptomes

Gene content of each transcriptome assembly was evaluated through four layers of analyses as described in Figure 3.1. All transcripts were initially screened against known

small non-coding RNA sequences and mitochondria/chloroplast originated sequences of *Triticum* families with Blast tool kit (Camacho et al. 2009). Overall, less than %1 of transcripts with significant hits in these screenings were considered as contaminants and excluded from the Kiziltan, TR39477 and TTD-22 transcriptome assemblies. Subsequent to contaminant analysis, open reading frame (ORF) content of the remaining transcripts was analyzed and transcripts possess ORFs longer than 80 aa were further evaluated for their coding potential through *ab initio* techniques; CPC (Kong et al. 2007), CNCI (Sun et al. 2013) and AUGUSTUS (Hoff and Stanke 2013). Totally, 60% of Kiziltan, 62% of TR39477 and 64% of TTD-22 transcriptome showed coding potential evidence, respectively. These transcripts were further evaluated for detection of their homology to known protein sequences and/or presence of functional protein domains. Ultimately, a total of 84,288, 75,996 and 78,456 putative protein-coding transcripts were identified from Kiziltan, TR39477 and TTD-22, which corresponded to 35%, 37% and 39% of the assemblies, respectively.

Although the assemblies were constructed with the utilization of data from pooled samples of three different biological replicate for each variety, assembled contigs might show none-to-very low expression level; thus, identification of actively-expressed transcripts is necessary for further characterization of transcriptomic data. To determine expression levels of transcripts, transcript abundances were quantified in terms of Fragment Per Kilobase Million mapped reads (FPKM) using RSEM package under Trinity software. Expression activity of protein-coding transcripts was evaluated based on the normalized FPKM. Percent of transcripts that failed to satisfy FPKM cutoffs in both control and drought stressed samples were plotted over a range of FPKM thresholds (Appendix A – Supplementary Figure 1). Overall, 1% change observed between five cut-offs from 0.1 to 0.5 FPKM; however, a sudden 1% change occurred thereafter. The point, 0.5 FPKM, was chosen arbitrarily as this was the point where the slope of the curve changes, indicating the significance of this point, thereby suggesting it as a potential threshold. With this threshold, 95% of each transcriptome were found to be actively-expressed transcripts, indicating the quality of the transcriptome assemblies and good coverage of the sequencing. In total, 81,168 (96%), 73,465 (97%) and 75,861 (97%) actively-expressed protein-coding transcripts (called coding transcript from this point) were identified in Kiziltan, TR39477 and TTD-22, respectively.

The actively-expressed coding transcripts were inspected for their expression patterns in control and drought treated samples. All three *T. turgidum* varieties represented a high portion of common transcripts between controlled and stressed samples where more than 70% of transcripts were detected as common (60,520; 57,012 and 56,164 transcripts for Kiziltan, TR39477 and TTD-22, respectively). Sample specific transcripts were most abundant in control samples, where 14,595, 10,865 and 14,204 transcripts expressed from solely controlled Kiziltan, TR39477 and TTD-22 varieties, respectively, as opposed to that of 7% of transcripts (6,053; 5,588 and 5,493 transcripts for Kiziltan, TR39477 and TTD-22, respectively) expressed only drought-stressed samples. Blast alignments of drought-specific transcripts revealed that 1,034 homologous transcripts (>80% of query identity and coverage) expressed in both tolerant and susceptible varieties. Drought-tolerant TR39477 revealed 36 different transcripts which does not have any similarity to transcripts from drought-susceptible TTD-22 while 4 of the TTD-22 transcripts were detected as TTD-specific (Appendix A - Supplementary Table 2). These transcripts were remarked as effective on the different drought stress tolerance and adaptation mechanism of these wheat varieties.

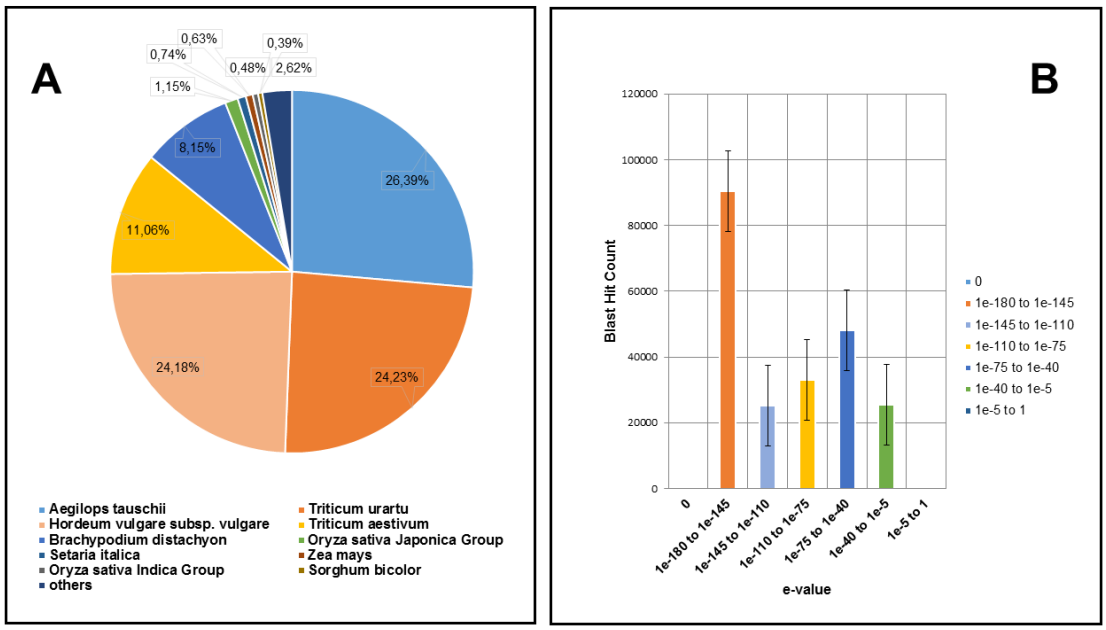


Figure 3.2. Homology pattern between T. turgidum proteins and other plants. (A) The pie chart shows the distribution of top-ten plants which showed the highest homology to T. turgidum proteins. (B) Pie chart shows the distribution of e-values for different blast hits.

Functional annotation of all coding transcripts was conducted by Gene Ontology (GO)

term assignment followed with KEGG pathway and COG analysis via Blast2GO software (Conesa and Götz 2008). In total, 860,446 GO terms were assigned to 157,013 (68% of all coding transcripts) transcripts. Assigned GO annotations were clustered in three main categories; Molecular Function (MF), Cellular Component (CC) and Biological Process (BP), based on the blast hits from the NCBI non-redundant (nr) *Viridiplantae* protein database with an e-value cutoff of 1E-5. Across all GO annotations, 'ATP binding', 'membrane' and 'protein phosphorylation' were predominant in the MF, CC and BP categories, respectively. Since these included coding transcripts from both treated and untreated samples in the three varieties, the categories most represented by transcripts could be considered as representatives of housekeeping genes. These sequences were further inspected in terms of the blast hit distribution through different plants which revealed the homology pattern of coding transcripts of *T. turgidum* varieties (Figure 3.2). Blast hit distributions showed that *T. turgidum* coding sequences possess the highest homology with *Aegilops tauschii* sequences where 26% of transcript revealed as identical to proteins from this species, followed by *Triticum Urartu* (24%) and *Hordeum vulgare ssp. vulgare* (24%) (Figure 3.2A). Additionally, e-value distribution of blast top hits indicated the general quality of the assembled coding transcripts where more than 50% of the hits have e-values smaller than 1e-110 (Figure 3.2B). Following Blast2GO annotations, mapping against KEGG database were performed to retain relative biological pathways of the coding transcripts (Ogata et al. 1999). In total, 50,250 transcripts were assigned to a total of 133 pathways in KEGG database. KEGG pathways the most represented by transcripts were purine metabolism (7,799 transcripts, 15.5%), thiamine metabolism (6,302 transcripts, 12.5%) and biosynthesis of antibiotics (3,147 transcripts, 6.3%) across all transcripts. Additionally, Cluster of Orthologous Groups (COG) screenings were performed using EggNog database, under Blast2GO software (Jensen et al. 2008) and coding transcripts sharing similar functions were classified into 23 functional groups. The largest group represented by transcripts had functions defined as 'unknown' (5,935 transcripts, 23.5%) followed by 'posttranslational modification, protein turnover and chaperones' (2,545 transcripts, 10.1%), 'signal transduction mechanisms' (2,397 transcripts, 9.5%), 'intracellular trafficking, secretion and vesicular transport' (1,600 transcripts, 6.3%) and 'translation, ribosomal structure and biogenesis' (1,506 transcripts, 6%).

The functions of differentially expressed transcripts were further analyzed after the

annotation of all transcripts. In all the three plants, ‘oxidation-reduction’ and ‘protein phosphorylation’ were the most represented BP terms in transcripts exhibited differential expression in response to drought stress (Figure 3.3). Interestingly, drought-susceptible TTD-22 revealed an increased number of upregulated genes which were categorized in ‘response to stress’ group regarding to BP assessment of Blast2GO (from 17 to 37 transcripts) while this category represented less number of associate transcripts in Kiziltan (from 24 to 13 transcripts) and TR39477 (from 37 to 4 transcripts) under drought stress.

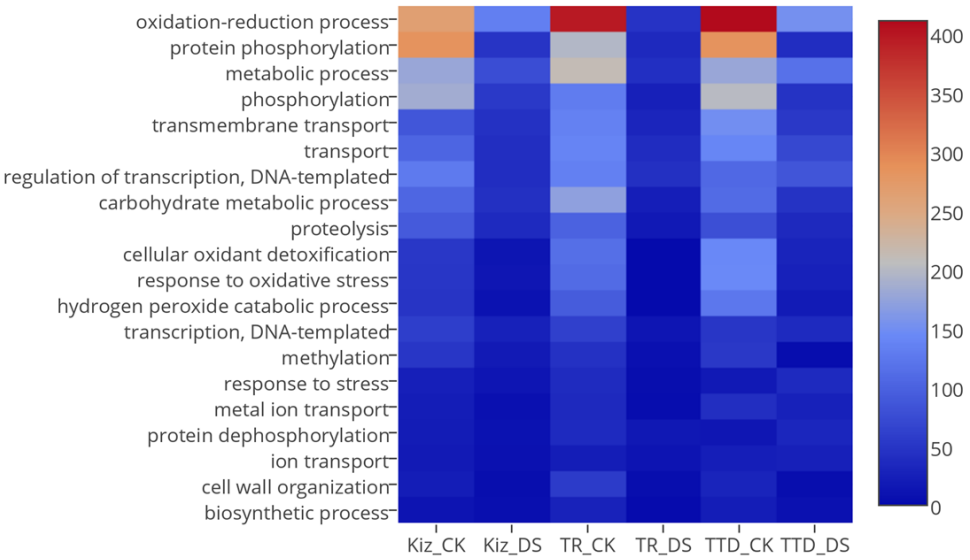


Figure 3.3. Heatmap for top 20 biological processes represented by stress-responsive coding transcripts in each sample. Several biological processed were detected as enhanced under drought stress (Graph legend: Kiz: Kiziltan, TR: TR39477, TTD: TTD-22; CK: control conditions, DS: drought stressed).

The orthologous groups of drought specific transcripts were also analyzed to determine their functional importance. The most representative COG id by all the drought specific transcripts, KOG0987, was associated with a DNA helicase which is functional in cell cycle control and cell division. Another important one, COG0507, connected with ‘exodeoxyribonuclease v alpha’ protein involving in replication and recombination. In order to further understand the drought tolerance of TR39477, the unique transcripts which are specifically expressed under drought stress, 36 transcripts which do not show any resemblance to TTD-22 transcripts, were analyzed in regard to their associated

KEGG pathway. Mostly, pathways associated with secondary metabolite synthesis were detected as enriched by these unique transcripts such as 'Glutathione, Sphingolipid and Thiamine metabolism'. Also, 'Glycosaminoglycan and glycan degradation pathways' were enhanced by unique transcripts of TR39477. Accordingly, identification and annotation of all such transcripts provided insides regarding drought-responsive metabolomic changes in durum wheat together with their transcript partners.

3.3.3. Putative lncRNAs and their expression pattern under drought stress

Identification of long-noncoding RNAs was performed by following the pipeline illustrated in Figure 3.1. Totally; 26% (63,773 transcripts), 29% (61,823 transcripts) and 22% (43,932 transcripts) of the transcriptome assemblies from Kiziltan, TR39477 and TTD-22 varieties, respectively, were associated with lncRNAs. Subsequent to identification, the actively-expressed putative lncRNAs were inspected based on normalized FPKM value obtained from transcripts abundance estimation analysis and total of 59,110 (93%), 57,944 (94%) and 40,858 (93%) putative lncRNAs were identified as actively-expressed putative lncRNAs (called lncRNAs from now on) from Kiziltan, TR39477 and TTD-22 varieties, respectively. The slightly lower ratio of active expression in lncRNAs (93-94%) compared to coding transcripts (96-97%) might arise from the tendency of lower expression of lncRNAs. Additionally, inspection of the expression patterns of transcripts in each control and drought-stressed samples revealed a similar distribution of the expressions of the lncRNA and coding transcripts between the three biological replicates only with little systematic biases (Figure 3.4). Error plots showed lower expression levels of lncRNAs compared to coding transcripts across all three plants which is also supported by literature (Quinn and Chang 2015). Moreover, overall expression pattern of lncRNAs were not altered by drought treatment, except a few transcripts.

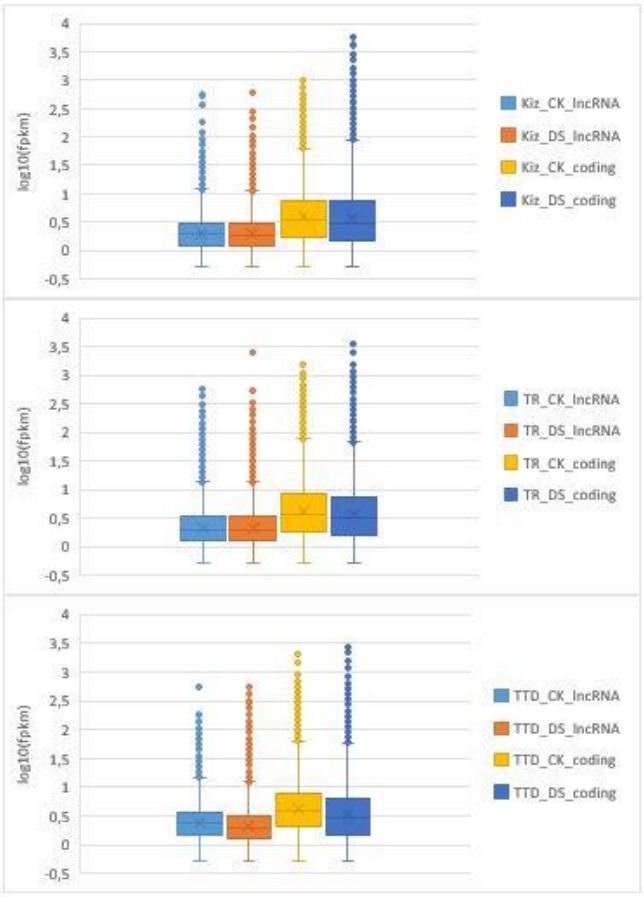


Figure 3.4. Expression pattern of coding transcript and lncRNAs in three different T. turgidum samples. (A) Kiziltan variety which exhibit moderate performance under drought conditions, (B) drought tolerant TR39477 variety, (C) drought susceptible TTD-22 variety.

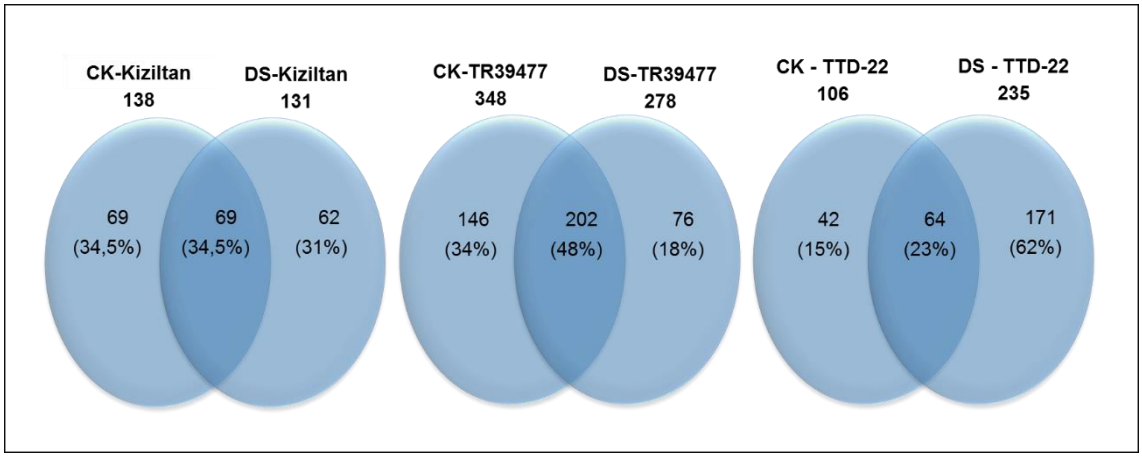


Figure 3.5. Common and drought specific lncRNAs from different T. turgidum varieties. Venn diagrams show the show common and specific differentially expressed lncRNAs between control (CK) and drought-stressed (DS) samples of Kiziltan, TR39477 and TTD-22.

Several lncRNAs from three different plants showed differential expression under drought treatment. These lncRNAs were identified with edgeR software with a p-value smaller than 0,001 and log2 (fold change) greater than 2 (Robinson, McCarthy, and Smyth 2010). Based on these cut-offs, 200 (3% of all lncRNAs), 424 (6% of all lncRNAs) and 277 (4% of all lncRNAs) were detected as 'drought-responsive' from Kiziltan, TR39477 and TTD-22, respectively. Differentially expressed lncRNAs were further evaluated for their sample specific expressions. Most of the differentially expressed lncRNAs showed tendency to exhibit sample specific expressions, indicating distinct molecular functions they might perform. Intriguingly, 66, 52 and 77% of differentially expressed lncRNAs exhibited sample specific expressions in Kiziltan, TR39477 and TTD-22 samples, respectively (Figure 3.5). From 64 to 202 differentially expressed lncRNAs were detected as common between control and drought treated samples of the three *T. turgidum* plants. In Kiziltan, 35% of stress-responsive lncRNAs were common whereas 48% and 23% stress drought-responsive lncRNAs were common between control and drought treated samples of TR39477 and TTD-22 respectively. These results indicate that common transcripts were more abundant in differentially-expressed lncRNAs from TR39477, with the most tolerant profile, than Kiziltan, with moderate reaction, and TTD-22, with the least tolerance. However, further characterizations are required for a complete understanding of the lncRNAs functions under drought stress.

Differential expression of mRNAs and lncRNAs were also confirmed with Quantitative Real Time (qRT-PCR) experiment. Common mRNA and lncRNAs transcripts; defined with %80 identity and query coverage across whole samples; were analyzed and a group of differentially expressed 'common' transcripts was chosen for experimental conformation. From this pool, expression of randomly chosen 2 mRNA and 2 lncRNA transcripts were quantified followed by 4 hours of shock drought treatment with 2-week-old root and whole seedling tissues of Kiziltan genotype. The quantification results with QRT-PCR experiment showed accordance with RNA sequencing differential expression data analysis both for lncRNAs and mRNAs (Figure 3.6). In addition, experimental results showed harmony between root and whole seedling tissues for lncRNAs and mRNAs.

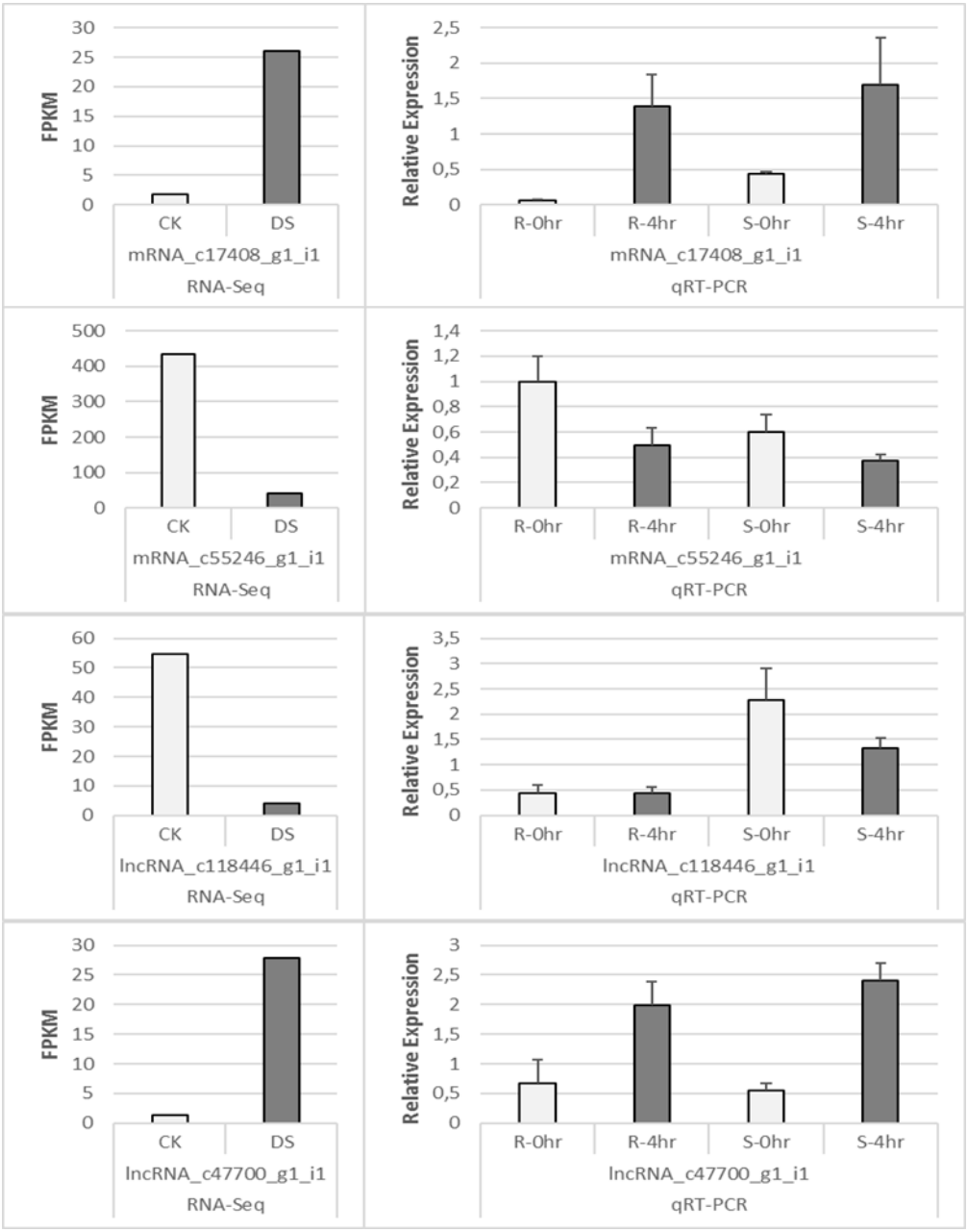


Figure 3.6. Relative normalized expression analysis results for common differentially expressed mRNA and lncRNAs samples. The quantification of transcript expression was performed with both root and whole-seedling tissue. The error bars were constructed based on standard deviation across three replicates of each sample.

3.3.4. Characteristics of actively expressed lncRNAs

All actively expressed lncRNAs were blasted against lncRNAs in *A. thaliana* from NONCODE database (Zhao et al. 2016). We identified 32, 24 and 15 lncRNAs that were

homologous to those lncRNAs in *A. thaliana*, suggesting the weak conservation of lncRNAs between *A. thaliana* and *T. turgidum* species. Actively expressed lncRNAs were further analyzed for their structural features in wild relatives of wheat. To that end, the length distribution and GC content of expressed lncRNAs and coding transcripts were analyzed and compared. The average length of *T. turgidum* lncRNAs was 327 nucleotides long whereas that of coding transcripts was 1,198 nucleotides. Lengths of those lncRNAs ranged from 201 to 2,686 nt, 2,857 nt and 2,540 nt in Kiziltan, TR39477 and TTD-22, respectively. In general, the majority of lncRNAs were relatively short while almost half (47-50 %) of coding transcripts were longer than 1,000 nt in all the *T. turgidum* varieties (Appendix A - Supplementary Figure 2). The average GC content for lncRNAs was detected as ranging 43% to 45% and across all three varieties, the highest ratio of GC content was observed as 82%. Interestingly, highest GC content of lncRNAs was detected in shorter lncRNAs, generally shorter than 1,000 nt which suggest an association between GC content and lncRNAs length (Appendix A - Supplementary Figure 3). On the other hand, the average GC content of coding transcripts was detected as relatively higher than lncRNAs in all three varieties with 52%. Connection between GC content and length distribution was also observed in coding transcripts (Appendix A - Supplementary Figure 3). While the length of the coding transcripts is increasing, the GC content was detected as narrowing around 50%. Overall, these results indicate that lncRNAs as well as coding transcripts share similar structural features in different *T. turgidum* species; yet, lncRNAs slightly differ from coding transcripts in gene structure in terms of structural characteristics.

Tetraploid durum wheat and wild emmer wheat genomes are derived from hybridization of A and B sub-genomes, each contributing to the composition of coding and lncRNA transcripts equally. Since reference genomes for Kiziltan, TR39477 and TTD-22 varieties are not available yet, we analyzed composition of coding and lncRNA transcripts from the recently published assembly of Zavitan (*T. dicoccoides* variety) genome(Avni et al. 2017). Using GMAP, we were able to map 89% of the lncRNA transcripts and 97% of coding transcripts to Zavitan genome. As Zavitan is a different cultivar from our three genotypes, we can expect cultivar dependent lncRNAs, resulting in the slight lower ratio of mapped transcripts of lncRNAs. Among these alignments, 2% of coding and 3% of lncRNA transcripts were mapped to uncharacterized scaffolds. On average, 50 and 48% of coding transcripts and 47 and 50% of lncRNA transcripts were mapped to A and B

sub-genomes, respectively. 48, 53 and 49% of coding transcripts and 46, 50 and 45% of lncRNA transcripts were mapped to A sub-genome whereas 50, 45 and 50% of coding transcripts and 51, 48 and 52% of lncRNA transcripts were mapped to B sub-genome in Kiziltan, TR39477 and TTD-22 varieties, respectively. The results showed enrichment of both coding and lncRNA transcripts in A sub-genome in TR39477 varieties and in B sub-genomes in Kiziltan and TTD-22 varieties. Both coding and lncRNA transcripts were similarly distributed over each chromosome at frequencies varied between 6 to 9%. These transcripts were most abundant at 2A chromosome for TR39477 and at 2B chromosome for Kiziltan and TTD-22 varieties. The results showed each sub-genome and chromosome of tetraploid wheat genome contributed in composition of lncRNAs as in case of coding transcripts.

One transcript can be derived from different loci and from opposite directions on the genome. The results showed similar distribution of coding and lncRNA transcripts on sense and antisense strands. Nearly 24% of all transcripts were shown to be transcribed from both directions whereas remaining alignments were distributed equally on sense and antisense strands. Consistent with previous studies on plants(Tang et al. 2016), most of the lncRNAs (80%) were single-exon transcripts and 6% of lncRNAs could be transcribed as both single-exon and multi-exons transcripts from different loci. On the other hand, coding transcripts tended to have more exons where 76% of coding transcripts transcribed with multi-exons. lncRNA transcripts showed smaller number of exons where maximum exon number can reach up to 16 in a lncRNA transcript as opposed to that of 68 for coding transcripts.

Similar to protein coding transcripts, lncRNAs are also exposed to alternative splicing with a lower rate compared to mRNAs (Xiao et al. 2015). Trinity-constructed isoforms of each gene were accounted for the spliced isoforms and were used to determine alternative splicing ratios of lncRNAs. In Kiziltan, 18% (10,369) of actively expressed lncRNAs were exposed to alternatively splicing where this ratio was detected as 64% (51,634) for that of coding transcripts. Similarly, alternatively spliced lncRNAs were counted as 18% (10,611) and 16% (6,906) of total lncRNAs where that of 60% (45,138) and 59% (44,221) of coding transcripts were identified as alternatively spliced, in TR39477 and TTD-22, respectively. Furthermore, alternative splicing (AS) events were identified from all mapped transcripts to Zavitan genome. AS events occurred in ~14%

of all actively expressed transcripts in each genome. Among the AS events, intron retention with 38% of the events is the predominant over remaining splicing events, followed by 29% to other events, 15% to alternative acceptor, 10% to alternative donor and 8% to exon skipping. Among the transcripts, 25, 22 and 21% of coding transcripts and 5, 5, 4% of lncRNA transcripts were involved in an AS event in Kiziltan, TR39477 and TTD-22 varieties, respectively. Consistent alternative splicing patterns in different *T. turgidum* varieties suggested that alternative splicing is not as prevalent in lncRNAs as it is in coding transcripts.

Among the lncRNAs of trinity-constructed isoforms, the ones with the most abundant splicing events were further inspected. For example, the two lncRNAs, Kiz_both_c65078_g3_i12 and Kiz_both_c65078_g3_i4, were detected as the isoforms of the same gene, Kiz_c65078_g3 which possessed 23 alternatively spliced isoforms in Kiziltan transcriptome. It was also noted that 8 isoforms of this gene showed sample specific expression, where one (Kiz_CK_c65078_g3_i21) of them identified as 'coding transcript' in the control sample. In TR39477 transcriptome, maximum number of gene isoforms was observed as 27 for the gene TR_c63034_g2. Among these isoforms, six of them were detected as actively expressed lncRNAs where only two of these lncRNAs showed differential expression during drought treatment. For TTD-22 transcriptome, the gene TTD_c62818_g1 had the most splicing events with 24 isoforms where six and three of them were identified as lncRNAs and coding transcripts, respectively. None of these isoforms showed differential expression during drought treatment. Yet, all lncRNA isoforms exhibited sample specific expressions where none of them were in common between control and drought treated samples. Since expression levels and significance in p-values were low (FPKM between 0.6 and 4), these sample specific expressions were not defined as differential expression. Regarding to observed alternative splicing patterns, it is tempting to speculate that each alternatively spliced isoform has different expression profiles and might have differential functions during stress response**.**

Repeat-masking of stress-responsive lncRNAs against known Poaceae repeat elements revealed that 37% to 64% of stress-responsive lncRNA sequences contain repetitive elements in three of the replicates. The difference in the repeat content of stress-responsive lncRNAs stem from repeat elements from small RNAs found in *T. dicoccoides* varieties. Once small RNAs excluded from repeat library, percent of stress-responsive

lncRNAs containing repeat elements were decreased to 33-34% in both TR39477 and TTD-22 varieties.

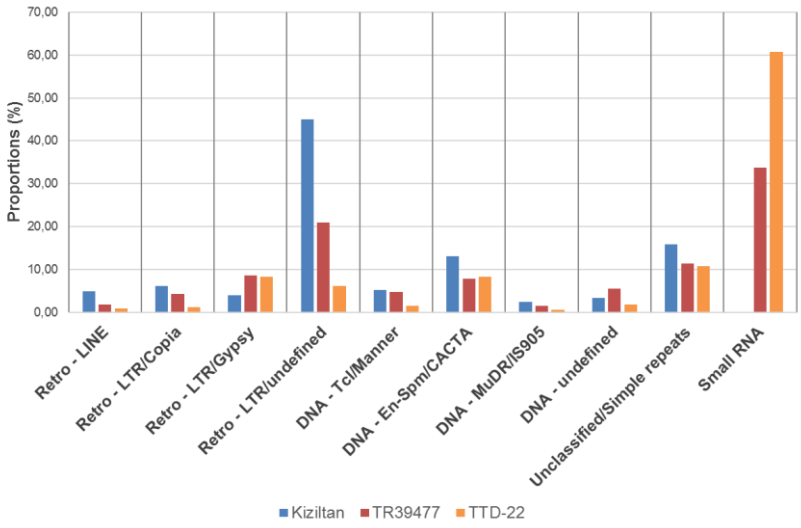


Figure 3.7. Repeat content of stress-responsive lncRNAs. Stress responsive lncRNAs were associated with both DNA transposons and retrotransposons. Each variety was represented with a different color (blue: Kiziltan, pink: TR39477 and orange: TTD-22).

Interestingly, lncRNAs which were from small RNA sequences, made up 34% or 61% of all repeats (Figure 3.7), were detected in *T. dicoccoides* samples, TR39477 and TTD-22 whereas no siRNAs were detected in Kiziltan. These small RNA repeats were from closely related species including *Zea mays*, *Triticum aestivum* and *Oryza sativa*. The most common small RNA sequences found in both TR39477 and TTD-22 samples was ZRSiRGRR00000035, following ZRSiRGRR00000042 and TRSiRGRR00000062. These observations indicate that some lncRNAs involved in stress response may act as siRNA precursors. Besides, their corresponding siRNAs were, therefore, regulated in a stress dependent manner. Excluding small RNAs from repeat content, stress-responsive lncRNAs which were from DNA transposons were marked in all three samples in almost half percent of repeats.

3.3.5. miRNA-related functions of lncRNAs

miRNAs can regulate gene expression at the post-transcriptional level by interacting with the complementary binding sites on target sequences, resulting in cleavage, decoy, or

translation repression (Kurtoglu, Kantar, and Budak 2014; Bala Ani Akpinar, Kantar, and Budak 2015; Budak, Khan, and Kantar 2015). Several studies have suggested that lncRNAs might have functions associated with miRNAs being either their targets or precursors (Chekanova 2015). To explore such functional roles of lncRNAs, *in silico* miRNA prediction was performed from all of three varieties by utilizing a list of 1,404 high confidence and/or experimentally verified plant miRNAs subtracted from miRBase release 21 (Kozomara and Griffiths-Jones 2014). In silico miRNA identification process led to the identification of 54, 58 and 46 lncRNAs in Kiziltan, TR39477 and TTD-22, respectively, as putative precursors of miRNAs belonging to 38 miRNA families. Interestingly, only one of the precursor lncRNAs in each assembly exhibited differential expression during drought treatment. In TR39477 and TTD-22, the stress responsive lncRNAs TR_c65168_g7_i1 and TTD_c34631_g1_i1 were detected as the precursors of miR1127 which do not have any determined target in these transcriptome assemblies. On the other hand, in Kiziltan, the stress-responsive lncRNA, Kiz_c66393_g4_i7, was identified as the putative precursor of the two miRNAs, with 1 or 2 nt changes in mature miRNA sequences of *Triticum aestivum* miRNAs; miR1117 and miR1127a. LncRNAs which have ability to generate miRNA sequences might perform an indirect regulatory function through lncRNAs generated miRNA sequence. In order to determine this indirect regulatory path, target transcripts of lncRNAs-derived miRNAs were analyzed and only targets of miR1117 were identified. miR1117 was associated with one coding; Kiz_c69869_g4_i1: a coding transcript expressed in both control and drought-treated samples without any change in expression; and two noncoding RNA targets; drought-specific Kiz_c106327_g1_i1 and control specific Kiz_c85253_g1_i1.This indicates that lncRNAs-derived miRNAs can perform multiple targeting potential which includes both coding and noncoding transcripts indicating a complex regulatory mechanisms through noncoding RNA performance even though the underlying regulatory network is not completely understood. Moreover, differential expression of precursor transcripts might result in the differential expression of corresponding mature miRNAs, leading to an increased regulation of expression; however, analysis of mature miRNAs at small RNA level is necessary for further validation of differential miRNA expression.

In order to provide more insight into miRNA-lncRNA association, functions of lncRNAs were analyzed in the sense of acting as miRNA targets using psRNATarget webtool at the default settings. It was shown that 1,276 lncRNAs were targeted by 33 miRNAs in

Kiziltan where 1,124 lncRNAs targeted by 24 miRNAs in TR39477 and 560 lncRNAs by 26 miRNAs in TTD-22. In Kiziltan, 9 of the lncRNAs targeted by miRNAs, further suggesting 13 stress-responsive miRNA-lncRNA target pairs, detected as differentially expressed in drought condition (Appendix A - Supplementary Table 3). In TR39477, 15 stress-responsive lncRNAs were detected as putative miRNA targets, building 27 unique miRNA-lncRNA target pairs. Yet, only 4 of the target lncRNA transcripts that established 7 miRNA-lncRNA target pairs in TTD-22 showed differential expression between drought-stressed and control samples.

Intriguingly, miRNAs targetting stress-responsive lncRNAs were mostly dominated by miR1436 and miR1439, where miR1118, miR1122, miR1130, miR1137 and miR1139 possessed putative lncRNA targets in TR39477 samples only, and miR1133 and miR1136 targeted lncRNAs in Kiziltan and TR39477, moderate to high tolerant samples. Additionally, it was shown that, in accordance with drought tolerance profiles of the samples, miR1436 and miR1439 mediated 8, 16 and 5 miRNA-lncRNA target pairs in Kiziltan, TR39477 and TTD-22 samples, respectively. These results suggested that gene regulation of miRNAs on stress-responsive lncRNAs are well corralated with the stress-tolerance profiles of the three genotypes such that stress-responsive miRNA-lncRNA target pairs were prevalent at most in TR39477, the most tolerant genotype and vice versa in TTD-22, the most sensitive genotype to drought. Moreover, the diversity of target lncRNAs in the most tolerant variety, TR39477, might be an indicator of additional regulatory mechanisms mediated by these lncRNAs. Thus, functional characterization of these target lncRNAs may shed light onto the drought tolerance mechanisms in *T. turgidum* species.

3.3.6. Functional characterization of lncRNAs through lncRNA-miRNA-mRNA networks

lncRNAs may interrupt miRNA-based regulation of gene expression by target mimicry where miRNAs bind to lncRNAs instead of their actual mRNA targets (Franco-Zorrilla et al. 2007). Thus, lncRNAs may indirectly enhance functioning of particular coding transcripts by preventing negative regulation of their translation by miRNAs. To explore stress specific association of miRNAs, lncRNAs and mRNAs, stress-responsive lncRNAs

and their miRNA-mRNA network was particularly analyzed. lncRNAs-miRNA-mRNA networks were observed in all three varieties with different levels of complexities (Figure 3.8).

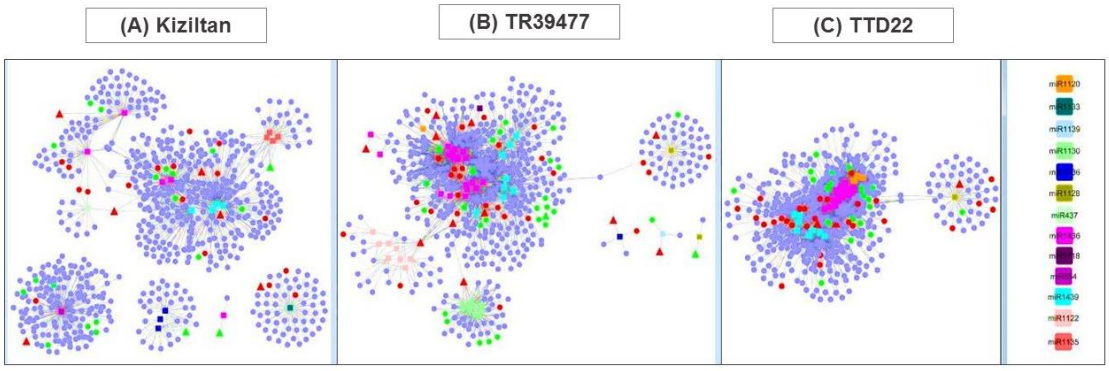


Figure 3.8. miRNA regulated networks between lncRNAs and coding transcripts. miRNA-lncRNA and mRNA networks were represented for each variety, Kiziltan (A), TR39477 (B) and TTD-22 (C). miRNA nodes were presented as rectangle and colored by miRNA family names. lncRNAs and coding transcripts were presented as triangles and circles, respectively. Transcripts that were upregulated by drought were colored as red and downregulated transcripts were colored green.

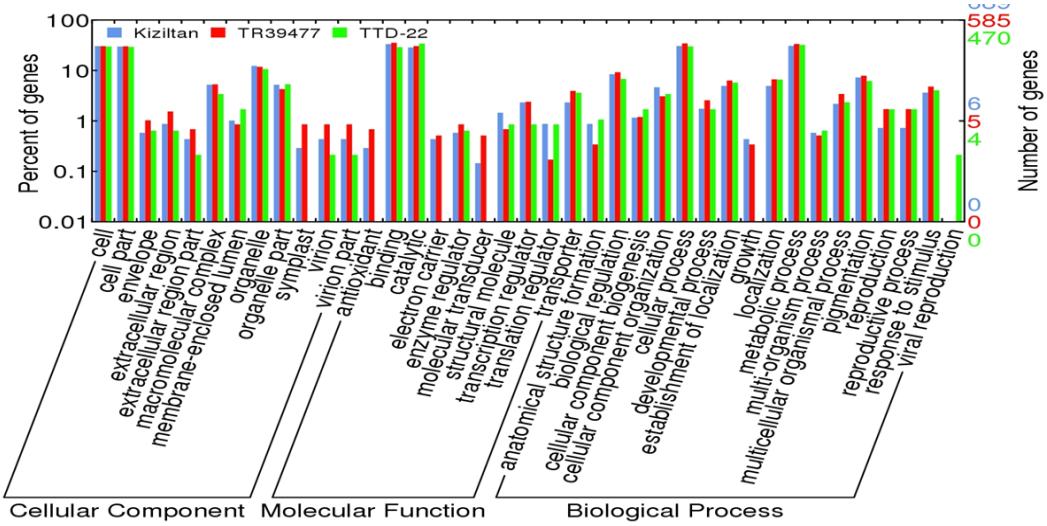


Figure 3.9. Distribution of Gene Ontology mapping results of coding targets of putative miRNAs. Targets from each variety was represented with a different color; blue: Kiziltan, red:TR39477, green: TTD-22. GO terms histogram was prepared through WEGO online tool.

The most complex interaction networks were established by miR1436 and miR1439. Intriguingly, these two miRNAs were detected as sharing similar target transcripts in all samples, indicating a dual-regulation of gene expression. Besides it was shown that miR437 and miR1135; miR1120, miR1122, miR1128 and miR1130; and miR1120 and miR1128 were also contributing to the interaction circuitry of miR1436 and miR1439 in Kiziltan, TR39477 and TTD-22 samples, respectively, suggesting additional players in these complex networks. Among these, interactions through miR1120 and miR1128 were conserved at the interspecies level.

The putative functions of coding transcripts were elucidated through GO mapping annotations. Intriguingly, antioxidant, electron carrier and molecular transducer molecular functions and growth biological process were highly enriched in Kiziltan and TR39477 varieties, but no evidence was found in TTD-22 varieties (Figure 3.9). These results suggested that increased regulation in these functions might be involved in drought stress response in *T. turgidum* varieties.

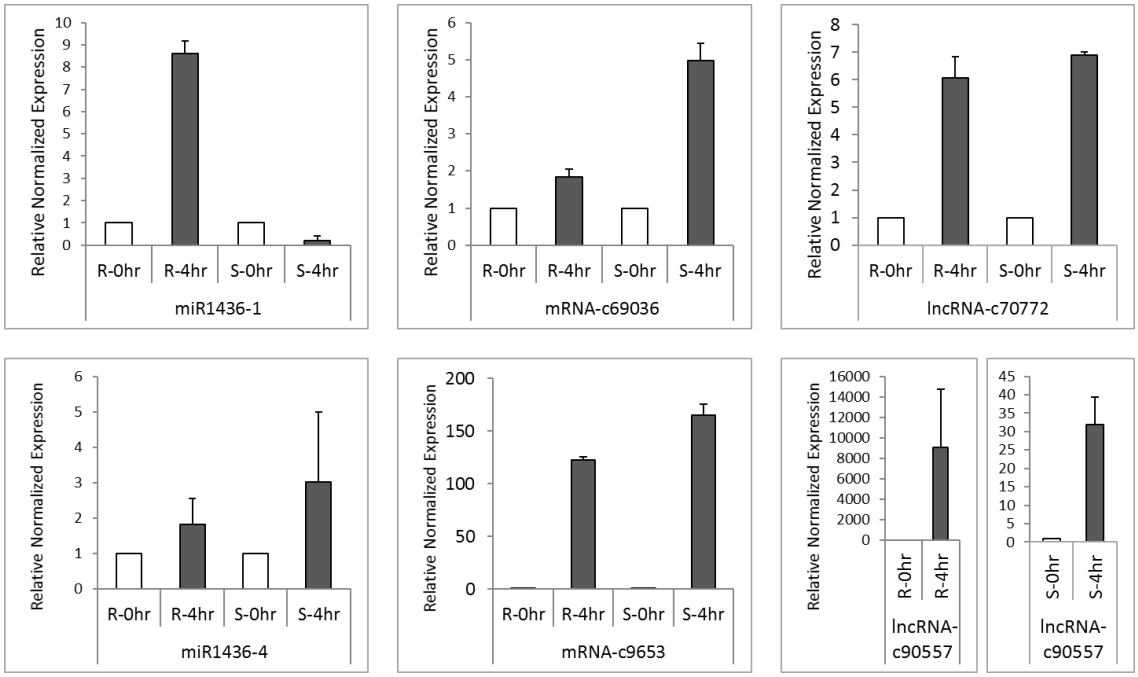


Figure 3.10. Relative normalized expression analysis results for miRNA-mRNA-lncRNA networks involved differentially expressed transcripts. The quantification of transcript expression was performed with both root and whole-seedling tissue. The error bars were constructed based on standard deviation across three replicates of each sample.

Among the interaction networks, expression profiles of miRNAs (mir1436-1 and miR1436-4) and their corresponding lncRNA (c70772_g2_i1 and c90557_g1_i1) and mRNA (c69036_g1_i1 and c9653_g1_i2) targets were shown in Figure 3.10. The expression of mRNA and lncRNAs molecules were concordant with the RNA-Seq data where both mRNAs and lncRNAs are upregulated under drought stress. Additionally, Q-RT-PCR results proved the drought specific expression of lncRNA c90557_g1_i1 where no expression for this lncRNAs was detected under control condition in variety Kiziltan.

3.4. Discussion

The increased effects of drought stress, caused by climate change, compel the improvement of major crop species such as wheat. However, complex genome of hexaploid wheat, combined from three different sub-genomes, A, B and D, becloud the understanding of gene regulations and molecular pathways underlying stress adaptation mechanisms, which is essential for establishing better crop performance. Alternatively, tetraploid wheat species, possessing a less complex genomic organization, stand as good candidates to pave the way for a deeper understanding of such mechanisms in wheat. Several varieties of wild tetraploid wheat have already been shown to exhibit differential drought tolerance, which might enhance our understanding of the drought-tolerance mechanisms in bread wheat (Ergen and Budak 2009). With the aim of providing further insights to drought response mechanisms and associated stress tolerance profiles of tetraploid wheat, transcriptomic changes in the roots of three different T. turgidum samples under slow drought imposition were analyzed at both coding and non-coding levels. Overall, this study showed the differential regulation of both coding and non-coding transcripts in response to drought stress, which might further be used for a better crop performance under drought conditions.

Sequenced reads from both control and drought-treated samples were assembled together and analyzed in the sense of differentially expressed coding transcripts and lncRNAs. A stringent filtering of transcripts (Figure 3.1) enabled the identification of 35, 36 and 39% of actively-expressed coding transcripts over all actively-expressed transcripts besides 26, 29 and 21% of actively-expressed lncRNAs over all actively-expressed transcripts in

Kiziltan, TR39477 and TTD-22 varieties, respectively (Appendix A - Supplementary Table 4). Overall, 2 mRNA and 2 lncRNA transcripts were validated with qRT-PCR experiment and the expression trend of transcript showed a similar sense with RNA-Seq data analysis even though the fold changes are different. Since the shock drought stress treatment was used for validation of presence of these transcripts, this is an expected result; particularly for lncRNAs, which the expression is highly dependent on condition. In the RNA-Seq analysis, interestingly, the number of transcripts was detected as decreased in all three varieties, regardless of their stress tolerance (Table 3.1). The stress-induced protein breakdown is a known phenomenon in plants where the accumulated amino acids support the osmotic balances in cells (Krasensky and Jonak 2012). Thus, it might be possible that the number of transcripts which leads to translation of several proteins may decrease to further support this breakage and osmotic balance. Additionally, it was noted that although having the highest percent of coding transcripts, the transcriptome of TTD-22, the most sensitive genotype, contained the lowest percent of lncRNAs. As several studies have provided evidence of the functional importance of lncRNAs for drought stress response (Muthusamy et al. 2015; Qi et al. 2013), the low abundance of lncRNAs in TTD-22 might be associated with its low drought tolerance; however, further characterization of stress responsive lncRNAs, particularly from drought tolerant variety TR39477, is essential to fully understand the role of lncRNAs in drought response.

To understand the function of differentially expressed genes under drought stress, the functional annotation was conducted via BlastX and Blast2GO. Analysis of homology patterns in *T. turgidum* proteins revealed a scattered homology of proteins across different *Poaceae* members. Although a high homology of *T. urartu* proteins is expected because of heritage of tetraploid wheat, high homology with *A. tauchii* stands as unexpected and further examination of these homolog proteins might provide insight into evolution of *T. turgidum.* The conservation of different coding transcripts was also observed between different accessions of *T. turgidum*. Additionally, more than 70% of transcripts were detected as common under control and drought-treated samples. Since plant cells tries to keep basal reaction rate for cellular maintenance, it is normal for high conservation of proteins under drought treatment (Kantar, Lucas, and Budak 2011). Moreover, even though a small portion of transcripts was differentially expressed under drought stress, they might have serious effect on other proteins.

Drought specific transcripts from TR39477 and TTD-22 were further analyzed in regards to their function to further understand the differences in the drought tolerance. Comparison of these transcripts from two varieties revealed that only approximately 20% of these transcripts are conserved. These common transcripts, expressed in response to drought regardless of the degree of drought tolerance, might be associated with general proteins which are expressed in the stress conditions such as ABA-responsive transcription factors of ROS scavengers (Krasensky and Jonak 2012). On the other hand, TR39477 revealed 36 transcripts which are specific to this cultivar and do not possess any degree of homology to TTD-22 transcripts. These transcripts related to several osmolytes and secondary metabolite such as 'Glutathione' and 'Thiamine' metabolisms regarding to KEGG maps. For instance, glutathione metabolism was associated with proline production, which is an important osmolytes accumulated in drought stress (Liang et al. 2013). Thiamine is an important molecule which involves in to phenylpropanoid pathway and this pathway cause the generation of several secondary metabolites which enhance the performance of plants under drought stress (Boubakri et al. 2013; Krasensky and Jonak 2012). Thus, it is tempting to speculate that TR39477 utilize these transcripts to regulate osmolytes production together with secondary metabolites to survive under drought stress. Further characterization of these transcripts may provide more insights into the molecular mechanisms of these events.

Besides the stress responsive transcripts, lncRNAs exclusive to drought stress and their relation with miRNAs and mRNAs provided further insight to the molecular mechanism of drought tolerance. The most complex networks were detected for miR1436 and miR1439, which indicates their important function in drought stress response. Among these, miR1439 was detected as targeting an aquaporins proteins which is conserved in wheat, rice and Brachypodium (Su et al. 2014). Under drought stress, lncRNAs might embed the inhibition of aquaporin translation, via target mimicry to miR1439 family members, and enhance the function of this protein for further transport of water from roots. Interestingly, miR1120 and miR1128 detected as conserved at interspecies level suggesting its important function. In another study, Yao and colleagues also detected ubiquitous expression pattern of miR1128 (misnamed as miR504 in the publication) even though no information about the targets of these miRNAs again not suggested (Yao et al. 2007). Computational inspection and experimental validation of targets of these miRNAs

might shed light onto their presence in these networks. Here, with Q-RT-PCR, we validated expression of miRNA1436-1 and miR1436-4 and their corresponding targets, supporting the existence of lncRNA-miRNA-mRNA networks (Figure 3.10).

Gene regulation is not limited to protein-coding genes where most of the genes transcribed in complex organisms are in fact non-protein-coding genes with important regulatory functions. Increasing number of studies has showed that both sRNAs and lncRNAs are important players of gene regulation in various vital biological processes, including stress responses in plants. Drought is a major stress factor to crops, causing serious yield losses to wheat (*Triticum ssp.*), and an important food source worldwide. On top of being an important limiting factor to the yield already, the effect of drought has been expected to increase by climate changes. Improved crop varieties that are tolerant to drought could sustain increased yield and quality of crops. In order to obtain improved varieties with enhanced productivity and stress tolerance, introgression of favorable elements into domesticated crop varieties has been suggested as a viable approach for decades. However, understanding of the molecular mechanisms behind drought response is crucial in determining these elements. The current study provides a comprehensive transcriptome analysis of tetraploid wild wheat varieties with diverse stress tolerance profiles, revealing drought-responsive genes and lncRNAs, thereby enriching the genetic information available for *T. turgidum* varieties. Further in silico predictions of miRNAs and their target interactions exploited the putative functional roles of lncRNAs. Besides, identification and characterization of lncRNAs in the present study expands the current knowledge of lncRNAs and their regulatory roles in drought response in plants in general.

4. ASSEMBLY AND ANNOTATION OF TRANSCRIPTOME PROVIDED EVIDENCE OF MIRNA MOBILITY BETWEEN WHEAT AND WHEAT STEM SAWFLY

4.1. Introduction

Wheat Stem Sawfly (WSS), *Cephus Cinctus* Norton (Hymenoptera: Cephidae) is stated as the most damaging pest of wheat in Northern Great Plains, causing crop devastations in Montana region each year (Beres et al., 2011). Female WSS choose the internodes of actively elongating fresh wheat stems to lay their eggs. By tearing the stem with their sharp ovipositors, eggs are placed into the stem where the larvae form after 4-7 days of incubation (Cárcamo et al. 2011). Since the larvae are cannibalistic, only one larva can survive in the stem although there are more eggs deposited. Larva stays and develops in the wheat stem during the growing season, feeding on parenchyma and vascular tissues and, eventually, it moves toward the bottom of the stem to cut a notch, causing plant to lodge in order to overwinter there until the pupation occurs. Stem cutting cause a dramatic reduction in yield, and even uncut infested plants have low yield due to decreased head weight by 17% (Delaney et al., 2010). However, there are still no effective control method over WSS damage in wheat. Usage of chemicals is limited by the long emergence period of females and the wheat stem protecting the eggs and the larva feeding inside (Knodel et al. 2009). The introduction of solid-stemmed wheat instead of hollow-stemmed wheat maintained a more powerful control on the infestations. Yet, the solid-stemmed cultivars are not preferred by producers because of its low yield and protein content compared to hollow-stemmed cultivars (B. Beres et al. 2011).

Advances in next-generation sequencing technologies have revealed that most of the genomes of higher eukaryotes is transcribed, of which only a small percent corresponds

to protein-coding genes. Until recent years, non-coding RNAs (ncRNAs) had been overshadowed by the interest on protein-coding RNAs and their pathways. As bioinformatics tools and experimental technologies brought new aspects in our understanding of RNA world, the structures and regulatory functions of ncRNAs came to light and most of the recent studies extended their focuses on microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) (Charon et al. 2010; S. J. Lucas and Budak 2012; Alptekin, Akpinar, and Budak 2016). Discovery and functional characterization of the remaining ncRNAs is yet in its infancy.

Both plant and animal miRNAs are ~22 nucleotide-long molecules and are derived from transcripts that fold on themselves to form stem-loop structures. In animals, the primary sequences transcribed by RNA polymerase II are processed by Drosha and Dicer-1 enzymes to produce pre-miRNAs and finally, mature miRNA/miRNA* duplexes (Bartel 2009). In plants, both processes are performed by Dicer-like protein (DCL) since plants lack Drosha enzyme (Budak and Akpinar 2015). Upon unwinding of the duplex, mature miRNA is exposed to RNA-Induced Silencing Complex (RISC) to recruit them towards its target (Budak and Akpinar 2015). miRNAs can bind their target mRNAs from either 3’ or 5’ UTR regions, with an imperfect complementarity (Bartel 2009; Budak and Akpinar 2015), resulting in transitional repression or degradation of the target (Alptekin et al., 2016b). The interactions between mature miRNAs and their target mRNAs provide an additional control on gene expression regulation. The first miRNA reported, lin-4, was shown to regulate timing of development through targeting lin-14 mRNA in *Caenorhabditis elegans* (Alvarez-Garcia and Miska 2005; He and Hannon 2004). Since then, distinct roles have been characterized for a vast number of miRNAs from animals and plants. Functional characterization of miRNAs in insect species have revealed the importance of miRNAs in several regulatory processes, including metabolism (K. Lucas and Raikhel 2013), growth and development (Bilak, Uyetake, and Su 2014), survival (Jones et al. 2013). miRNAs from one specie may function at interspecies level, targeting genes or genomes of organisms which they have physical contact. Very recently, independent studies have been reported several examples of trans-kingdom delivery of sRNAs from; plant to virion (Iqbal et al. 2017), oomycetes to plant (Jia et al. 2017), plant to nematodes (Tian et al. 2016). Similar to what these studies suggested, miRNAs might also be effective in regulating insect-host interactions at WSS larval stages once larva gets into the stem of the host plant.

As being another important class of ncRNAs, lncRNAs draw attention with their mRNA-like structural features and biogenesis processes. Like mRNAs, they are expected to be longer than 200 nucleotide, subjected to alternative splicing and 5' capping, and mainly transcribed by RNA polymerase II (Legeai and Derrien 2015). None-to-very low coding-potential of lncRNAs is the major factor to differentiate lncRNAs from mRNAs. Several remarkable features of lncRNAs include the tendency to exhibit tissue and sample specific expressions (reviewed in Quinn and Chang, 2015), which can be speculated to the importance of lncRNAs in regulatory mechanisms. It has been shown that lncRNAs are indeed involved in key regulatory mechanisms across diverse biological processes, such as dosage compensation (Militti et al. 2014), developmental- and epigenetic-regulation (Schmitz, Grote, and Herrmann 2016) in various species. For example, a yellow-achaete intergenic RNA (yar) was found to be an effective component of the sleep behavior in *D. melanogaster* (Soshnev et al. 2011). *Drosophila melanogaster*, as a model organism, has been extensively investigated for its lncRNA genes (M. Li et al. 2012; Soshnev et al. 2011), although functions of the majority of lncRNAs in flies remain unknown (Xiao et al. 2015).

The interactions between miRNAs and lncRNAs are also critical for the regulation of gene expression since lncRNAs might act as miRNA precursors or miRNA targets. By binding on the complementary sites on the target lncRNAs, miRNAs decrease the stability of the target, controlling their abundance and regulatory function in the cell (J.-H. Yoon, Abdelmohsen, and Gorospe 2014). miRNAs and lncRNAs are both known to form decoys, titrating the transcription factors from the environment (Banks et al. 2012; K. Wang and Chang 2011). Moreover, lncRNAs can function as endogenous Target mimics (eTMs) of miRNAs (Franco-Zorrilla et al. 2007) or competing endogenous RNAs (ceRNAs) (Salmena et al. 2011) of mRNAs where the target lncRNA titrates the miRNA to inhibit its pairing with the target mRNA.

In this study, transcriptome data from eight WSS samples were utilized to generate the assembly and, later, to identify miRNA, lncRNA and mRNA molecules from larvae, female and male WSS. In total, we obtained 11 miRNA families, 40,185 coding transcripts and 59,676 lncRNA transcripts from the WSS transcriptome. Additionally, we constructed differential expression library of WSS transcripts to compare expression

profiles of larva and adult WSS samples. Annotations and the expression profiles of transcripts will be useful resources in the understanding of the molecular mechanisms of WSS. Considering the effect of WSS larvae on wheat, we have focused on the action mechanisms of RNAs in larvae and their targets in wheat and compared them with female and male adult data. Understanding the role of RNAs in infestation of wheat crop fields by WSS will give insight for future strategies in fighting with the pests and increasing the wheat yield.

4.2. Materials and Methods

4.2.1. *De novo* assembly and differential expression of transcripts

RNA-Sequencing (RNA-Seq) of eight WSS samples (larvae, antennae, female and male) from infected wheats using Illumina HiSeq 2000 Sequencer was obtained from NCBI SR database (Appendix B – Supplementary Table 1; Sequence Read Archive (SRA) accession number SRP067708). Trimmomatic (v0.32) with default parameters (LEADING:5, TRAILING:5, MINLEN:36) was used for adaptor trimming and quality trimming of reads (Bolger, Lohse, and Usadel 2014). A single assembly containing reads from all eight WSS samples was generated *de novo* using Trinity software (release 2014-07-17) (Grabherr, Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., and Friedman 2013). All transcripts were restricted to be >200 bp in length.
Trimmed raw reads were aligned back to the assembled transcripts using Bowtie assembler and abundance estimates of transcripts were quantified as Fragment Per Kilobase Million mapped reads (FPKM) using RSEM (version 3.2) (B. Li and Dewey 2011) under Trinity pipeline. Differential expression analysis was performed using EdgeR (Robinson, McCarthy, and Smyth 2010) pipeline with the default threshold parameters of p-value=0,001 and log2 (fold_change)=2. Assembly files of larvae, female and male pooled whole samples were separated based on their corresponding abundance estimates for further analyses.

4.2.2. Annotation of transcripts and lncRNAs

Annotation of transcripts were performed by analyzing the reads in a four-step process; eliminating contaminants, separation by ORF size criteria, coding potential calculations and homology-based predictions. All assembled transcripts were aligned with known small non-coding RNA sequences of all hexapoda species deposited in NCBI (1711 sequences) using blastn (-evalue 1E-05). Since the focus was on the coding and lncRNA sequences, transcripts with homology to small non-coding RNAs were defined as contaminants and eliminated. The abilities of transcripts to code for a full-length protein was evaluated using Transdecoder under Trinity software. Transcripts with predicted open reading frames (ORFs) longer than 100 amino acids passed the ORF size criteria for annotation process. Coding potentials of the transcripts were calculated using two prediction techniques; CPC (online version, reverse strand included, 2016) (Kong et al. 2007) and CNCI (-s ve) (Sun et al. 2013). Transcripts predicted as 'coding' by at least one of these tools were accepted to be satisfied the coding potential prediction criteria. Homology-based predictions were performed through homology screenings against functional coding sequences using Blast (version 2.2.26) and against known protein domains with Pfam identification using Hmmer (v.3.1b1) (Z. Zhang and Wood 2003). All assembled transcripts were screened for homology to known mRNA sequences of WSS, protein sequences of *Cephus*, *Apis*, *Hymenoptera* families and Swissprot entries (all deposited at NCBI) using blast (-e-value 1E-05, -length 90, -identity 80). Peptide sequences of assembled transcripts with an ORF size longer than 30 amino acids were predicted using Transdecoder. These peptide sequences were further screened using blastp against Swissprot entries (1E-05, -length 30, -identity 80) and using Hmmer (v.3.1b1) against Pfam domains (1E-05). Transcripts with a homology to functional sequences or a predicted Pfam domain passed the homology-based prediction criteria.

Following this multi-layered analysis, putative coding transcripts were identified by excluding contaminant transcripts and selecting transcripts that passed ORF size, coding potential prediction and homology-based prediction analyses. On the other hand, knowing that lncRNAs do not possess open reading frames or protein-coding potentials, transcripts which failed in all homology-based, coding potential and ORF size prediction analyses were identified as putative lncRNAs. Actively-expressed transcripts were extracted

according to the fpkm threshold of 0.5. Differential-expression analysis was performed through pair-wise comparison of sample-specific expressions of each transcript using edgeR software with p-value of 0.001 and fold-change of 4 thresholds. Actively-expressed mRNA and lncRNA transcripts were provided in Supplementary Data 1 and 2 (Cagirici, Biyiklioglu, and Budak 2017).

4.2.3. Identification and annotation of miRNAs and tRNAs

High confidence mature miRNA sequences of hexapoda species were retrieved from miRBase database (v21, June 2016) (Kozomara and Griffiths-Jones 2011). *In silico* miRNA prediction was performed using SUmir pipeline (2. General Materials and Methods) with this set of 562 mature miRNA sequences as query. The genes encoding tRNA species were extracted using the local version of tRNAscan-SE software (Lowe and Eddy 1996) with the default parameters for eukaryotic genomes.

4.2.4. Prediction of miRNA targets

Target transcripts of newly identified miRNAs were predicted using two algorithms, RNAhybrid (Krüger and Rehmsmeier 2006) and miRanda (Enright et al. 2003). Filtering criteria were applied to each prediction as follows: RNAhybrid: p-value adjusted to 3utr_fly, mfe<=-25 kcal/mol; miRanda: total score >=140, total energy<=-25 kcal/mol. Putative target transcripts were accepted from those predicted by the two software. The resulting putative mRNA targets were aligned to NCBI non-redundant (nr) protein database (blastx, -evalue 10-5, -outfmt 5) where blast top hits were functionally annotated using Blast2GO software. A list of target transcripts from lncRNAs and mRNAs targeted by the same mature miRNA sequences was gathered together to construct an interaction network between lncRNAs, miRNAs and mRNAs, which was visualized using Cytoscape 3.3.0 (Shannon et al. 2003).

Identified larval mature miRNA sequences of WSS were further evaluated for their putative mRNA targets within wheat coding sequences by using psRNATarget webtool (Dai and Zhao 2011). Functional annotation of the target sequences was performed using

Blast2GO software following homology screening against protein sequences of 72 *Viridiplantae* species (blastx, -evalue 10-5, -outfmt 5, -max_target_seq 1).

4.3. Results

4.3.1. De novo assembly of WSS transcriptome

RNA-sequencing data from eight WSS samples, including larvae, antennae, females and males, from infected plants were retrieved from NCBI database (Sequence Read Archive (SRA) accession number SRP067708). Initially, all reads were subjected to adaptor and quality trimming using Trimmomatic, revealing a total of 28.799 Gbp clean reads. Despite reducing the number of reads, this step improved the quality and the process time of the assembly.

Table 4.1. Summary statistics of sequencing and combined de novo transcriptome assembly of WSS

Read processing	
Reads before trimming	50.248 Gb
Reads after trimming	28.799 Gb
Assembly statistics	
Number of 'genes'	116560
Number of transcripts	165284
Percent GC	40.65
N50 (bp)	3304
Median contig length	523
Average contig	1380.63
Total assembled bases	228196136

All trimmed reads were then assembled into one assembly using Trinity *de novo* assembler, resulting in 165,284 transcripts with a N50 length of 3,304 bases (Table 4.1), indicating the high-quality of the transcripts that could construct full-length protein

sequences. GC content of the assembly was 40.65 %, which is similar to the GC content of the raw reads (39-43 %). A detailed summary of the assembly statistics can be found in Table 4.1. Clean raw reads were aligned back to the assembly to determine the expression levels of each transcript, which were scaled to fragment per kilobase million (fpkm). Based on the normalized fpkm values greater than 0.5 in at least one of the eight WSS samples, 143,483 (86.8%) transcripts were defined as actively-expressed WSS transcripts.

4.3.2. Annotation of WSS transcriptome

To elucidate interactions of noncoding RNAs (lncRNAs and miRNAs) with protein-coding sequence content of WSS, all actively expressed transcripts were subjected to a selection process, following the transcriptome assembly. Transcripts satisfying the criteria of having homology to known coding sequences, a predicted coding potential and an ORF region that is at least 100 amino acid-long were defined as candidate mRNA transcripts (called mRNA transcripts from now on). Thus, 40,185 mRNA transcripts were identified, of which 38,934 (96.86%) of them showed significant resemblance to known WSS mRNAs with 80% or more identity, indicating 1,251 novel mRNAs were identified. These novel mRNAs were screened through NCBI non-redundant (nr) protein database for similarity to a known protein from other organisms, thereby revealing potential functions of transcripts. Functions of proteins with significant hits included tRNA ligases, histone proteins, kinases and more (Cagirici, Biyiklioglu, and Budak 2017).

Although all novel mRNAs showed significant homology to at least one known protein, only 868 of them were mapped to 15,947 Gene Ontology (GO) terms. These GO terms represented molecular functions (MF) of newly identified mRNAs as binding, catalytic activity and structural molecule activity where their biological processes (BP) were predicted as metabolic, cellular or single-organism processes at level 2. At a multi-level classification, ion binding and biosynthetic process were the most predominant annotations in the MF and BP categories, respectively.

Varying sets of expressed mRNA transcripts showed differential expression between larva and adult WSS samples, reflecting the effect of developmental stage on the WSS

transcriptome. Differential-expression analysis performed through pair-wise comparison of sample-specific expressions of each transcript revealed 16,291 and 16,928 mRNAs that were differentially-expressed between larva-adult male and larva-adult female samples, respectively, where 12,453 of them were common in both comparison pairs, totaling 20,766 mRNAs differentially expressed between larva and male or female samples.

A list of differentially expressed transcripts has been compiled combining ten transcripts with the highest levels of expression from each of the larva, male and female samples. Three of the top 10 highly expressed transcripts of female and male samples coincided, totaling 27 differentially expressed transcripts with the top 10 highest levels of expression in one of the three samples (Figure 4.1).

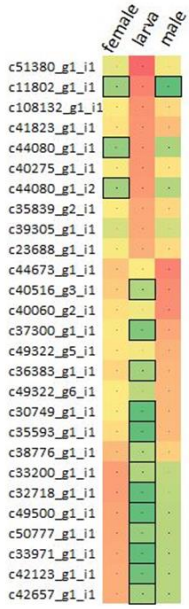


Figure 4.1. Comparison of the expressions of transcripts. Top 10 differentially expressed transcripts with the highest expressions were collected from pooled larva, male and female samples, totaling 27 non-redundant list of transcripts. Expressions were presented in terms of log10(fpkm) from red to green, representing high to low expression. Transcripts having low-to-none expressions (<2fpkm) were highlighted with the boxes.

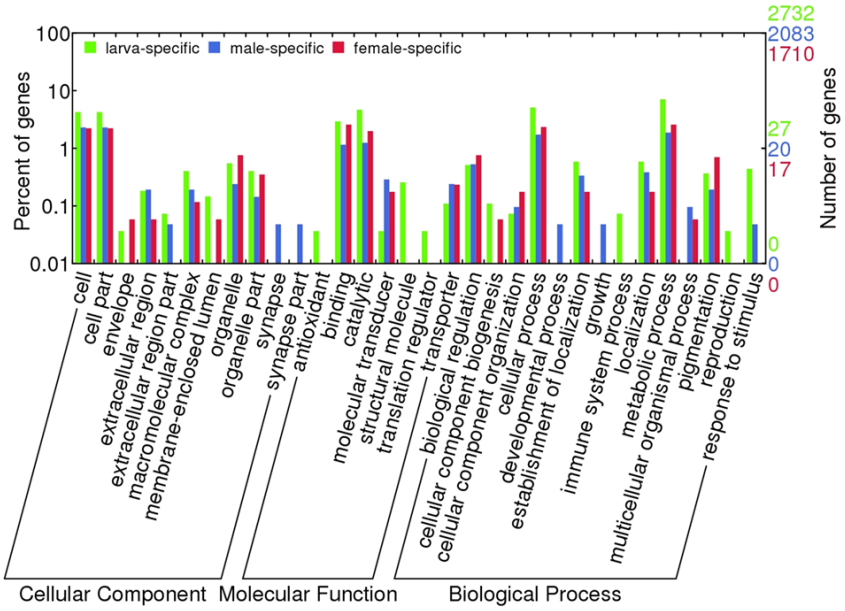


Figure 4.2. Blast2GO term distribution over differentially expressed transcripts. Transcripts with sample specific expressions were shown.

Comparative functional annotation of mRNA transcripts revealed that 2,732, 2,083 and 1,710 transcripts were exclusively expressed in larva, male and female samples, respectively. These mRNA transcripts were composed of proteins known to be involved in various biological processes (Figure 4.2), which exclusively were in immune system process and reproduction in larva, and developmental process and growth in males. Besides, antioxidant and translation regulation molecular functions were identified only in larva samples. Unfortunately, hypothetical, predicted and unknown proteins made up to 25% of these transcripts, which points out to that there might be many additional pathways that these differentially expressed transcripts play roles in.

4.3.3. Identification of lncRNAs

The analyses for lncRNA identification yielded a total of 71,220 putative lncRNAs, which corresponded to 4.09% of all transcripts of the Trinity-assembled transcriptome of WSS. Based on normalized fpkm which was greater than 0.5 in at least one of the eight WSS samples, actively-expressed lncRNA transcripts (named as lncRNAs from this point) were identified for further analyses. The results showed that 83.79% (59,676) of lncRNAs

passed the threshold of active expression as opposed to 92.21% (40,185 out of 43,581) of annotated transcripts, illustrating the tendency of lncRNAs to exhibit lower expressions. All lncRNAs were further examined in terms of expression patterns in larva and adult WSS samples to discover larva-specific and adult-specific lncRNAs in WSS. Among a total of 59,676 actively-expressed lncRNA transcripts, 55,946 (56.88%) of them possessed a normalized fpkm greater than 0.5 in at least one of the larva, male or female WSS samples. It appeared that lncRNAs were the most abundant in larva followed by male and female WSS transcriptomes. 16,965 (34%) of 49,943 actively-expressed larva transcripts were defined as lncRNAs as opposed to 17,554 (27%) of 63,837 male transcripts and 9,110 (19%) of 47,042 female transcripts (Figure 4.3A). Moreover, most of the larva and male lncRNAs were sample-specific whereas most of the female lncRNAs were common in either one of the samples. This comparison of lncRNA content of the three samples indicated that larva showed the highest and female the lowest, transcriptional diversity and specificity. These results suggested the abundance of lncRNAs in larvae compared to adult WS, indicating the functional importance of lncRNAs in different levels of WSS life cycle, especially in the larval stages.

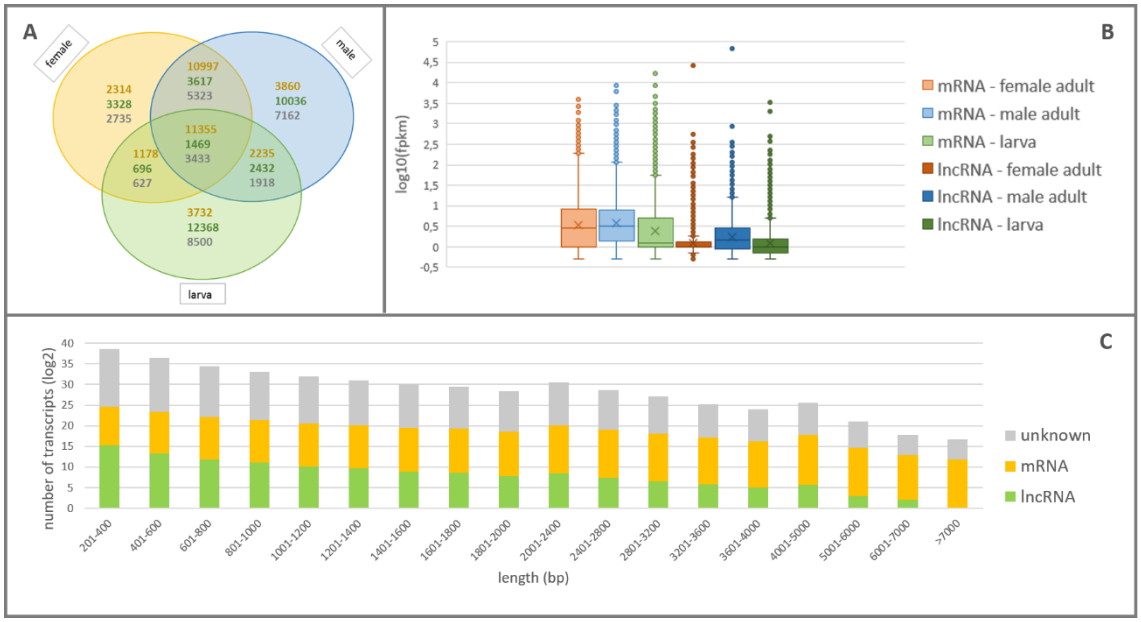


Figure 4.3. Structural features of coding and non-coding elements in WSS transcriptome. (A) Venn diagram shows the numbers of common and specific elements in larva, male and female WSS samples. The numbers of mRNAs, lncRNAs and unknown transcripts were written in orange, green and gray colors, respectively. (B) The expression patterns of mRNAs and lncRNAs in larva, male and female samples. (C) Length distribution of

the transcripts expressed in any WSS samples.

mRNA transcripts, on the other hand, showed less sample-specific expressions than lncRNA transcripts. In fact, 88.77% (35,671) of actively-expressed mRNA transcripts exhibited expression evidence in at least two of the larva, male and female samples as opposed to 56.88% of actively-expressed lncRNA transcripts (Figure 4.3A). Besides, 31.83% of these mRNAs were common in all three WSS samples and 66.62% of them were shared by more than one samples whereas that of 2.63% of common lncRNAs and 14.68% of shared lncRNAs. Further examination of expression levels of lncRNA and mRNA transcripts showed lower levels of lncRNA expression in all three WSS samples (Figure 4.3B). These results indicated sample-specific expression patterns as well as lower expression levels of lncRNAs than of mRNAs.

To determine lncRNAs that were either upregulated, downregulated or showed no differential expression between different WSS samples, a pairwise differential expression analysis was performed using edgeR package under Trinity software. It was found that 1,893 of the lncRNAs were differentially expressed between larva and adult WSS samples. 728 of those differentially expressed lncRNAs were upregulated in larva samples whereas 686 and 1,059 of them showed upregulation in female and male adult samples when compared to larva. Although there were more sample specific lncRNAs identified, these differentially expressed lncRNAs were the ones that passed the strict criteria.

4.3.4. Characteristics of lncRNAs and mRNAs

We analyzed structural features of all actively-expressed lncRNA transcripts and compared with the ones for mRNA transcripts in WSS. The lengths of the lncRNAs ranged from 201 to 6,465 bp. Most of the lncRNAs, however, had shorter transcripts such that 93.6% of the lncRNAs were shorter than 1,000 bp (Figure 4.3C). On the other hand, mRNA transcripts were remarked by longer sequences such that longest mRNA transcripts contained 27,058 nt and half of them were longer than 2,990 nt. Average transcript length of lncRNAs was 444 bp as opposed to that of 3614 bp for mRNA transcripts. In addition, GC contents were ranging between 8 and 70% for lncRNAs and

26 to 72% for mRNA transcript, the majority of which (83% and 91% for lncRNA and mRNA transcripts, respectively) were around 30 to 50% (Appendix B - Supplementary Figure 1). Average GC content for lncRNA and mRNA transcripts were 39% and 42%, respectively. The longest transcripts, of both lncRNAs and mRNAs, were the ones with average GC content. We could not detect any significant correlation between length and GC content of both mRNA and lncRNA transcripts.

Alternative splicing is one of the common features between lncRNAs and mRNAs although lncRNAs have lower splicing ratio than protein-coding genes in mammals. Consistent with their counterparts in the mammals, WSS lncRNAs showed less splicing than annotated transcripts. Alternatively-spliced isoforms were identified for only 11% (6,376) of the lncRNA transcripts in this assembly, which is significantly lower than 83% (33,537) of the ratio observed in annotated transcripts. Among the lncRNAs having alternatively spliced isoforms, 20% (1,286) of them shared at least 4 isoforms, which is less than one third the ratio of 69% (23,079) for mRNA transcripts having alternatively spliced isoforms. Such low levels of splicing events in lncRNA transcripts indicated that it is not as common as in mRNA transcripts of WSS. As an exception, 76 of the putative lncRNAs showed high splicing events with at least twelve isoforms. The maximum number of alternative splicing in lncRNAs was 23, observed in the gene, c49416_g1. 5 isoforms of this gene were identified as putative lncRNAs. Two of these lncRNAs failed to pass expression threshold in larva, male, female WSS samples**.** Remaining lncRNAs exhibited sample specific expressions where c49416_g1_i22 expressed only in male, and c49416_g1_i23 and c49416_g1_i6 expressed only in larva samples. These estimated abundances of transcripts over different samples revealed the unique expression profiles of the alternatively spliced isoforms in the different stages of WSS life cycle.

4.3.5. tRNA annotation

The analysis of tRNA gene content of WSS transcriptome revealed that the majority of tRNA gene families were represented by more than a single copy in the WSS transcriptome. A total of 159 putative tRNA genes were identified, 41 and 50 of which were encoded by actively-expressed mRNA and lncRNA transcripts, respectively (Figure 4.4). These tRNA genes correspond to 21 putative tRNA gene families with a specificity

for 45 anticodons. With a total of 18 loci, tRNA:Met-CAT was marked as the most abundant tRNA species among all WSS transcriptome as well as among mRNA (8) and lncRNA (7) transcripts. The codon it decodes, AUG, is the most common canonical start codon. Moreover, several tRNA species were encoded by only mRNAs or lncRNAs but not by any other transcripts. 8 and 14 tRNA species were found to be either mRNA or lncRNA specific, respectively. For the remaining tRNA species, we could not detect any correlation between mRNA and lncRNA transcripts.

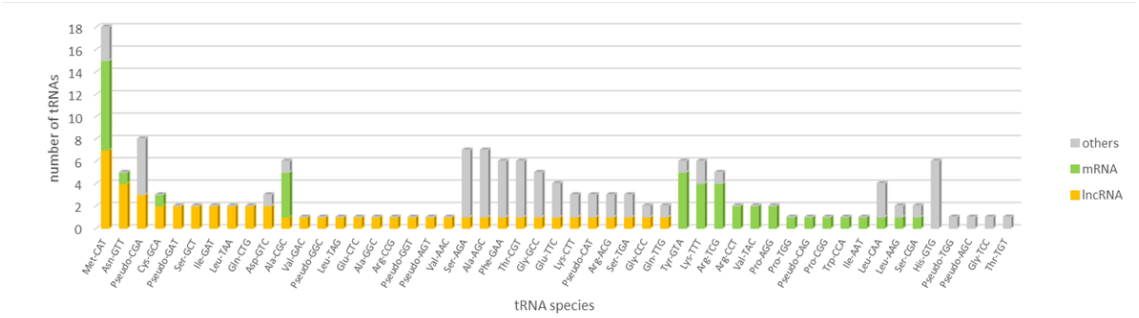


Figure 4.4. tRNA content of mRNA, lncRNA and the remaining transcripts in the WSS transcriptome. tRNA species sorted by their abundance in lncRNAs, mRNAs and others in order.

4.3.6. *In silico* miRNA prediction

Using 562 high confidence mature miRNA sequences from hexapoda species deposited at miRbase, a total of 18 mature miRNA sequences referring to 11 miRNA families were identified from the assembly of WSS transcriptome. Among these miRNA families, four miRNA families, miR-281 (4), miR-8 (3), miR-10 (2) and miR-14 (2), were represented with more than one stem-loops (Supp. Table 4). Predicted mature miRNA and pre-miRNA sequences were ranging between 21-23 nt and 94-125 nt, respectively. Average length of all putative mature miRNA sequences was 22 nt where that of 99 nt for their respective pre-miRNA sequences. These values are consistent with the 80-100 nt mean sequence length of animal miRNAs (Greenberg et al. 2012).

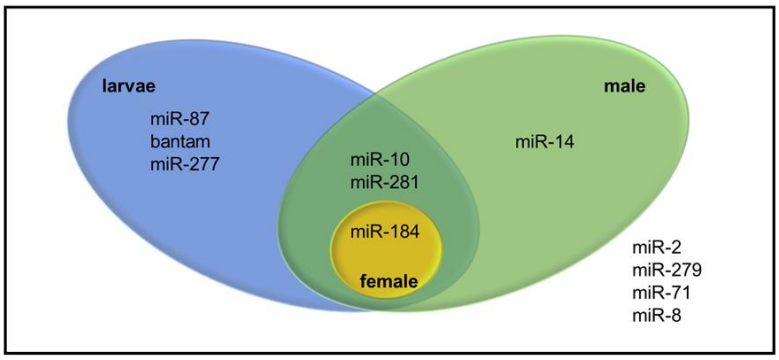


Figure 4.5. Venn diagram representing sample specific expression of WSS miRNAs. The four miRNAs listed outside the venn were identified from WSS samples other than pooled larva, male and female samples.

Pre-miRNA sequences were also examined in terms of the direction and the location on the transcriptome where one stem-loop might arise from different locations on the transcriptome. 38 transcripts were identified indeed as putative precursors of 18 mature miRNAs (Cagirici, Biyiklioglu, and Budak 2017). While 22 of them stemmed from sense strand, 16 of them were found in antisense strand. Among putative miRNAs, only miRNAs from miR-184 and miR-281 families were identified from both sense and antisense strands. Since expression of the precursor transcripts in different WSS samples might reveal sample-specific miRNAs, all precursor transcripts were discriminated by the evidence of expression in larva and pooled adult WSS samples. 12 mature miRNAs belonging to 7 miRNA families were identified in either larva, male or female samples. Among them, only one mature miRNA was found in female as opposed to that of 9 mature miRNA sequences (4 miRNA families) in male and 10 mature miRNA sequences (6 miRNA families) in larva (Figure 4.5). The results showed that miR-184 was expressed in all three samples, whereas miR-14 was male-specific; and miR-87, bantam and miR-277 were larva-specific miRNAs. miR-10 and miR-281, on the other hand, were identified in both larva and male samples.

Further examination on sources of putative miRNAs suggested six lncRNA transcripts as putative precursors of miRNAs belonging to six miRNA families; miR-10, miR-14, miR-2, miR-279, miR-71 and miR-8. These lncRNAs were the only precursors identified for the respective miRNAs in WSS transcriptome. Among them, the lncRNA transcript, c46526_g1_i1, was identified as the precursor of miR-10 in both male and larva samples where c106582_g1_i1 was identified as the precursor of miR-14 in male sample only.

Nevertheless, none of the lncRNA transcripts in female samples were identified as miRNA precursors. Expressions of remaining precursor lncRNA transcripts were detected in at least one of the remaining five WSS samples, supporting the expression of respective miRNAs at a sample specific level in WSS. These results also point out the functional importance of lncRNAs as being miRNA precursors.

4.3.7. Putative targets of WSS miRNAs

miRNAs regulate gene expression at the post-transcriptional level by interrupting expression through binding to the complementary sites on the target sequences. For 18 mature miRNAs, 32,149 and 6,458 miRNA-mRNA pairs were predicted using RNAhybrid and miRanda, respectively. A total of 5,070 unique mature miRNA-mRNA pairs, predicted by both algorithms were selected as reliable interaction pairs. From the larva miRNAs, miR-281 involved in the highest number of interactions with mRNAs (1,654), where bantam miRNA contributed in 70 interactions which was the lowest number between larval miRNAs (Cagirici, Biyiklioglu, and Budak 2017). ~282 mRNA targets were assigned per mature miRNA sequence on average. These large set of putative mRNA targets indicated the extend of the functional roles of miRNAs in WSS. Homology screenings against NCBI non-redundant (nr) protein database revealed sequence similarity of target mRNAs to the genes involved in several important molecular functions including binding, catalytic, molecular transducer, transporter and structural molecule. The most abundant term in biological process category was cellular and metabolic processes followed by biological regulation.

Other targets of putative mature miRNAs involved lncRNA transcripts. RNAhybrid predicted 20,788 mature miRNA-lncRNA pairs and miRanda predicted 1,075 miRNA-lncRNA pairs. 774 lncRNA transcripts suggesting 965 unique mature miRNA-lncRNA pairs predicted by the two algorithms. While the highest number of interactions was made by miR-184 within the larval miRNAs, miR-87 was involved in the least number of interactions with lncRNAs. ~54 lncRNA targets were estimated per mature miRNA, indicating potential functions of lncRNAs as being miRNA targets although target mRNAs were shown to be more prevalent in WSS.

4.3.8. lncRNA - miRNA - mRNA network in WSS

lncRNAs might involve in miRNA-mediated gene regulation through an indirect protection of target mRNAs, which called as target mimicry. By mimicking the binding site on the target mRNA sequence, lncRNAs might recruit miRNAs to enhance the expression of respective mRNAs. To have a broader aspect about these regulatory mechanisms, interaction networks between miRNA, lncRNA and mRNAs were established combining miRNAs and their lncRNA and mRNA targets predicted here.

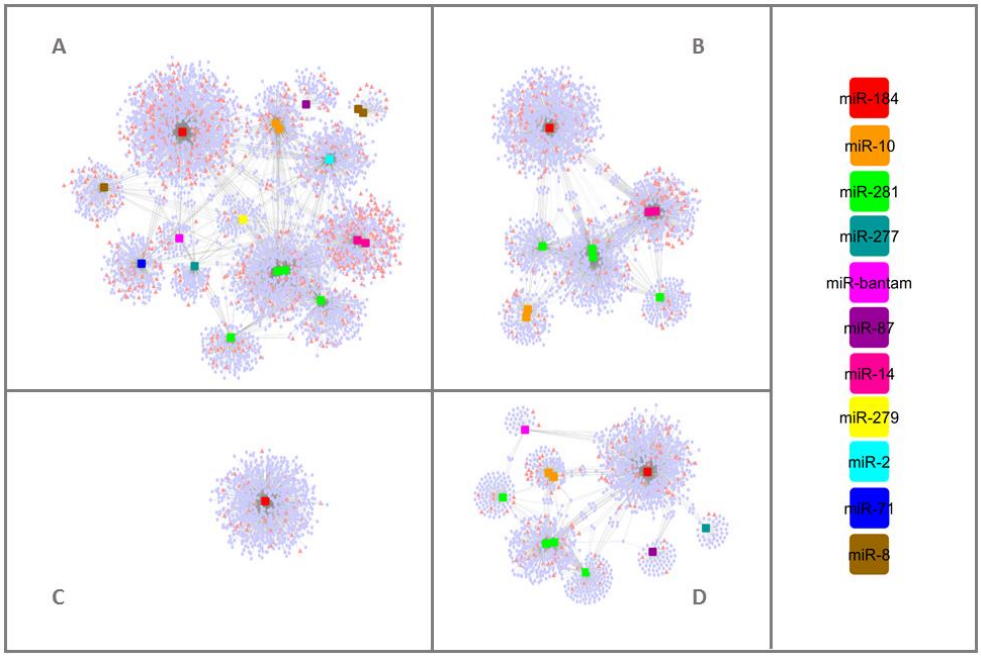


Figure 4.6. miRNA-mediated lncRNA and mRNA interaction networks. Networks constructed using; all WSS miRNAs (A), male miRNAs (B), female miRNAs (C) and larval miRNAs (D). lncRNA transcripts were represented as pink triangles whereas pale blue circles were denoted to mRNA transcripts. miRNAs were shaped as squares and colored based on the color scale shown at right.

Remarkable, all miRNA families had both mRNAs and lncRNAs as interacting partners. Figure 4.6 illustrated that lncRNAs differentially expressed between larva and adult WSS samples were involved in one complex interaction network with miRNAs and mRNAs. All miRNAs identified from each growth stage of WSS contributed to the interaction network constructed in its respective stage. Functional annotation of mRNAs involved in

any part of these networks was performed using Blast2GO to elucidate potential functions of lncRNAs as competing endogenous RNAs (ceRNAs). All mRNA and lncRNA targets of miRNAs were included in the combined network which build up one large and complex network. The results indicated that response to stimulus biological process was highly enriched in larva whereas structural molecule and transporter molecular functions in male. No enrichment was detected in female samples as all female miRNAs shared by larva and male. Overall, the interaction networks between miRNA, lncRNA and mRNAs suggest putative roles of lncRNAs to increase regulation in variety of molecular processes through target mimicry for miRNAs.

4.3.9. Bidirectional mobility of miRNA in wheat and WSS

WSS larva accommodates in wheat stem and feeds from there until pupae stage of its life cycle (Delaney, Weaver, and Peterson 2010). Given the evidence of cross-kingdom regulation by miRNAs (L. Zhang et al. 2012; Jia et al. 2017; Tian et al. 2016), the interaction between intracellular molecules of WSS larvae and wheat cannot be underestimated due to these two organisms being in contact and trying to defeat each other. To assess possible effects of larval miRNAs on wheat gene expression and its response to WSS pathogen, target analysis for larval miRNAs was performed against *T. aestivum* coding sequences deposited at ensemble plant database using psRNAtarget tool. We identified 10 putative wheat targets for 3 miRNAs expressed at larvae.

As shown in Table 4.2, a larva specific miRNA, miR-277, specifically targets several transcripts on the three sub-genomes of chromosome 3. Among the chromosome 3 targets, three transcripts were from chromosome 3B which was characterized with the wheat stem solidness (Nilsen et al. 2016). Blast screening of these chromosome 3 targets revealed similarity to methyltransferase PMT11 and ankyrin-like proteins (Table 4.2). Another larva specific miRNA, miR-87, has shown to have putative targets on chromosome 5BL and the only target for the male and larva shared miRNA, miR-281, was a transcript from chromosome 2AL of wheat. These 2A and 5B chromosomes were associated before with larval mortality. Although the predicted targets of miR-281 does not share homology with a protein with known function, targets of miR-87 was defined as vacuolar protein sorting-associated protein 22 homolog 1 (Table 4.2). Overall, these

findings suggested that putative wheat targets of larval miRNAs were likely to be involved in defence mechanisms of wheat against insects.

Table 4.2. Wheat coding targets of WSS larval miRNAs.

Source	Mirna Acc.	Wheat_target_Acc.	Target Annotation
Larva	miR-277	TRIAE_CS42_3AL_TGACv1_193846_AA0620850.1	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-277	TRIAE_CS42_3AL_TGACv1_193846_AA0620850.2	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-277	TRIAE_CS42_3B_TGACv1_224524_AA0797630.1	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-277	TRIAE_CS42_3B_TGACv1_224524_AA0797630.2	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-277	TRIAE_CS42_3B_TGACv1_224524_AA0797630.3	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-277	TRIAE_CS42_3DL_TGACv1_251571_AA0882620.1	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-277	TRIAE_CS42_3DL_TGACv1_251571_AA0882620.2	methyltransferase PMT11 & ankyrin-like protein
Larva	miR-87	TRIAE_CS42_5BL_TGACv1_408620_AA1363930.1	vacuolar protein sorting-associated protein 22
Larva	miR-87	TRIAE_CS42_5BL_TGACv1_408620_AA1363930.2	vacuolar protein sorting-associated protein 22
Larva, Male	miR-281	TRIAE_CS42_2AL_TGACv1_094608_AA0300450.1	hypothetical protein F775_10692 [Aegilops tauschii]

miRNAs might pass from wheat to larva during their close contact. To assess putative larva targets, wheat mature miRNA sequences (119 entries) were retrieved from miRbase database. Using miRanda and RNAhybrid tools in combination, we identified 12,535 larval coding transcripts as putative targets of wheat miRNAs. The number of predicted targets varied widely between miRNAs, ranging from 2 to 6,174. Homology screening of the putative targets were performed based on blast hits from the NCBI non-redundant (nr) protein database with an e-value cutoff of 1E-5. Blast hits suggested that the genes targeted by wheat-derived miRNAs were likely to be involved in several functions such as kinases, helicases and transcription initiation factors. Among them, the two proteins

with known functions targeted by more than 10 miRNAs were “Endothelin-converting enzyme 1-like isoform X1” and “N-acetylglucosamine-6-phosphate deacetylase”. Besides, digestive enzymes, i.e., lipases and glycogen synthases, were among the putative targets of wheat miRNAs.

4.4. Discussion

Wheat production is severely limited by the a/biotic stress factors and biotic stress can account for up to 20% yield loss in wheat. Wheat Stem Sawfly (WSS; *Cephus cinctus* Norton) is the most harmful pest of wheat in North America (B. Beres et al. 2011), due to larval mining inside the plant stem. Although understanding their mechanisms of action is critical to fight effectively with WSS infestations and help farmers to reduce the devastation, very little is known on the genetic information and molecular mechanisms of WSS. To expand our knowledge, a detailed noncoding RNAs and their interactions with transcriptome has been conducted for WSS larvae and adults. Here we utilized a different method which is combining all reads from all tissues/samples. As many non-coding elements tend to show tissue specific expressions (Quinn and Chang 2015), combining raw reads from different samples is important for the richness of the genetic elements available and the completeness of the transcriptome. Here, transcriptome-guided mRNA, lncRNA and miRNA identification was performed with a focus on larvae transcriptomics and differential expression of transcripts between larvae and, female and male samples since most of the damage is caused from the larvae growing and feeding inside the wheat stem. Furthermore, the network between these RNA molecules besides the potential passage of WSS miRNA molecules towards wheat cells to target wheat coding sequences and to regulate the gene expression there as a part of its damaging effect has been disclosed.

With a stringent filtering of 165,284 transcripts in the *de novo* assembled WSS transcriptome, we identified 40,185 (24%) actively expressed protein-coding sequences. Of these transcripts, 1,251 transcripts were selected as novel mRNA candidates with lack of homology to known WSS mRNAs. To provide a broader aspect of their functions with

non-coding RNA, these novel mRNA transcripts were classified in three GO categories, molecular function, biological process and cellular function. The functional annotations revealed proteins from many different molecular pathways, reflecting the complexity of eukaryotic cells. A significant number of these annotated proteins were ribosomal subunits, transcription and translation initiation factors, kinases, histone proteins which have important roles in the basic cellular mechanisms for the survival of the cell. In addition, six transcripts were identified as chemo-response-related proteins which might function in olfactory pathways that is important in sexual and social interactions of insects as discovered in honeybees (Pelosi et al. 2017; Benton 2006). Another protein affecting insect behavior was longitudinals lacking (lola) protein which had three isoforms in WSS transcriptome assembly. This protein was found to be important in neuronal system development by maintaining proper axon guidance (Kuzin et al. 2005) and mutation studies in *Drosophila melanogaster* resulted in aggressive behaviors on the insects (Edwards et al. 2006). These novel findings shed light on the undiscovered mechanisms in the cells of WSS and the organism being a social insect and reflected a potential to manipulate the developmental pathways of WSS in order to find more effective ways to cope with the infestations.

Dynamic changes in gene expression reflect the response of an organism to intrinsic and environmental signals. Thus, expression of genes varies over the course of a species' life cycle; between stages of growth and development and between different sexual categories. Here, a total of 20,766 differentially expressed mRNAs were identified through pair-wise comparison of female and male samples to larva (Supp. Table 2). Intriguingly, one fourth (6,525) of these transcripts showed sample specific expressions, indicating the distinct patterns of regulation between larva and adult developmental stages of WSS. While 6,019 of these differentially expressed transcripts were upregulated in larva when compared to adults, 14,824 of them were upregulated in adults, which could be a sign of a more complex cellular system in the adult stage of WSS life cycle. The cellular activity in larval stages of insect species was found to be less complex than it is in adults (Python and Stocker 2002), which might have caused from the lack of complex behaviors in the larval stage while adult individuals are more motile and they involve in social interactions more often. The transcripts that showed a great differential expression between larva and both adult samples also emphasized the distinct cellular activities between larva and adults. Comparison of the expression levels of these transcripts in each

sample revealed similar patterns of expression between male and female transcripts when compared to larva. Figure 4.1 showed that transcripts upregulated in male compared to larva were also likely to be upregulated in female, although the level of regulation may differ. Intriguingly, most of these transcripts (16 out of 27) that showed the top 10 highest expression in one of the samples exhibited low-to-none expression (<2 fpkm) in any other samples, indicating the abundance of distinct regulatory mechanisms in different WSS life stages, thereby pointing out the functional importance of sample specific expressions of transcripts. We also included functional annotations of differentially expressed transcripts between larva and adult samples. Among them, allatostatin-A-receptor was one of the proteins that were encoded from the transcripts upregulated in larva. Allatostatin-A proteins were discovered to inhibit juvenile hormones in cockroach and criket (Hergarden, Tayler, and Anderson 2012) which preserve the larval characteristics (Riddiford 2012). Therefore, the upregulation of allatostatin-A-receptor might be a part of the passage through adult stage by contributing the inhibition of juvenile hormones. A number of transcripts upregulated in larva were encoding proteins related to circulatory system and central nervous system (CNS) development. Neurofibromin was one of the proteins annotated from two upregulated transcripts form larva, which was identified with its role in body size determination during larval development of *Drosophila melanogaster* (Lee et al. 2013). In addition, chitinase was also encoded by 13 transcripts that were upregulated in larva. As an insect larva grows to form an adult individual, chitin molecules within the cuticle surrounding its body should be degraded by chitinases and synthesized again (Khajuria et al. 2010). Therefore, together with this information, it can be concluded that the cellular metabolism of the larva is focused on growth and formation of critical body systems leading to a complete adult development.

mRNAs are not the only players of molecular mechanisms where non-coding elements such as long non-coding RNAs (lncRNAs) were involved in various biological processes, including cell fate decision, developmental processes, sex-specific functions and growth (Keniry et al. 2012; Militti et al. 2014; M. Li et al. 2012). With the advent of high-throughput sequencing technologies, RNA-seq has boosted the identification and characterization of lncRNAs in several species. Despite the extensive studies on the functions of lncRNAs in *Drosophila* (Ecker et al. 2012; Soshnev et al. 2011; M. Li et al. 2012), little is known about characteristics and functions of lncRNAs in other flies (Xiao et al. 2015), including WSS. Major challenge in the identification of lncRNAs were that

lncRNAs are not conserved between species. In fact, these are the non-conserved long transcripts that are not able to construct a full-length protein (Yan and Wang 2012). A total of 59,676 novel lncRNAs were identified in this study that will likely be useful for further genomics research. Analysis for sample-specific expression profiles of lncRNAs showed the transcriptional diversity of lncRNAs between larva, female and male WSS, supporting the evidence of the transcriptional diversity and specificity of lncRNAs in several species provided by recent studies (Cesana et al. 2011; Quinn and Chang 2015). Interestingly, the results revealed that lncRNAs were much more abundant in larva than the adults (Figure 4.3A). This high abundancy of lncRNAs coincides with the high activity of developmental processes of larval stage of WSS life cycle. Thus, the results of this study supported the previous findings that the transcriptional diversity of lncRNAs could be related to developmental processes and sex-specific functions, even though further experiments are required to validate this conclusion. Notably, only 774 (1.3%) lncRNAs were common in all eight samples whereas 31,556 (52.9%) lncRNAs exhibited sample specific expressions. Thus, it is likely that a number of lncRNAs with tissue- or condition- specific expression exist and will be discovered through additional RNA-seq analyses at larger scales. In addition, the expression levels of lncRNAs are significantly lower when compared to the expression levels of protein-coding transcripts (Z. Wu et al. 2014). The comparison of expression levels of WSS mRNAs and lncRNAs revealed that mRNAs from larva and adult stages were expressed relatively higher than the lncRNA molecules (Figure 4.3B), supporting the previous observations.

The major factor discriminating lncRNAs from mRNAs is lack of a discernable coding potential. Our tRNA analysis revealed that tRNA gene with anticodon CAT (tRNA-Met-CAT) decoding AUG start codon was found for both lncRNAs and mRNAs. Therefore, we identified that lncRNAs might do encode translation start codon, indicating the initiation of translation into proteins, as mRNAs do. 14% of tRNAs in lncRNAs corresponded to anticodon CAT (tRNA:Met-CAT) as opposed to that of 20% for mRNAs (Figure 4.4). With the highest abundance in each group, we could not correlate the initiation of translation with the potential of protein coding; however, distribution of remaining tRNA-anticodons differs broadly between mRNA and lncRNA transcripts. Content of the remaining tRNA species might regulate construction of the full-length and functional proteins.

Functions of lncRNAs can be inferred from their association with other non-coding elements. Several lncRNAs have shown to generate miRNAs, such as H19 lincRNA functioning as the precursor of miR-675 which in turn suppresses the growth promoting Insulin-like growth factor 1 receptor (*Igf1r*) (Keniry et al. 2012). Here, six lncRNA transcripts were identified as the only precursors of the six miRNAs; miR-10, miR-14, miR-2, miR279, miR-71 and miR-8. Besides being miRNA precursors, some lncRNAs act as miRNA targets. Through direct targeting, miRNAs might regulate the abundance of lncRNAs which are involved in different cell functions (J.-H. Yoon, Abdelmohsen, and Gorospe 2014). Several lncRNAs targeted by miRNAs have been uncovered recently, such as lincRNA-p21 (J. H. Yoon et al. 2012) and H19 (Kallen et al. 2013). The assessment of the possible miRNA-lncRNA target interactions identified 54 putative lncRNA targets per miRNA agents. Having the miRNA binding site, lncRNAs might enhance the functioning of miRNA target genes by titrating shared miRNAs from environment. As lncRNAs targeted by miRNAs could be involved in a regulatory circuitry between lncRNAs, miRNAs and mRNAs, we investigated putative target mimicry functions of these lncRNAs. The first evidences of target mimicry were discovered in plants (Franco-Zorrilla et al. 2007). Later, several examples were identified in mammals in the name of competing endogenous RNAs (ceRNAs) of miRNA targets (Karreth et al. 2011; Kallen et al. 2013; Cesana et al. 2011). Here, we also constructed a putative interaction network between lncRNAs, miRNAs and mRNAs in WSS to identify putative lncRNAs acting as ceRNAs (Figure 4.6). Experimental validation of target lncRNAs might shed light of the regulatory functions of these networks. We believe that importance of lncRNAs and such regulatory networks will emerge further.

The journey through understanding the functions of miRNAs has started with the discovery of lin-4 and its role in larval development in *C. elegans*. Lin-4 miRNA was upregulated in *C. elegans* larvae in one of the four larval stages, targeting lin-14 mRNA, suggesting that it has a regulatory role in larval development (He and Hannon 2004; Alvarez-Garcia and Miska 2005). The importance of miRNAs in developmental-timing of larvae was also shown in vertebrates in several studies. Here, we identified three miRNAs specifically expressed at larval stages of WSS; miR-87, bantam and miR-277. miR-87 was suggested as a regulator of the immune responses of mosquitoes against viral infections (Y. Liu et al. 2015). Later, its expression was identified in the nematode, *Meloidogyne incognita* (Y. Zhang et al. 2016); however, its function in insects remains

elusive. Functions of putative targets of miR-87 includes transferase activity, topoisomerase activity, binding and extracellular matrix structural constituent, suggesting its structural and functional importance. Both of miR-277 and bantam miRNA were associated with anti-apoptotic activities in insects (Jones et al. 2013; Bilak, Uyetake, and Su 2014). Although direct targets and function of miR-277 requires further evidence, miRNA bantam was linked directly to protective functions ensuring cell proliferation (Bilak, Uyetake, and Su 2014). As the larval stages of WSS are the most stressed periods in WSS stages, increased regulation through bantam miRNA and miR-277 in the larva samples supported its anti-apoptotic activities. On the other hand, miR-14, the only adult male-specific miRNA identified is expressed greatly in testicular tissues of immature and fully-mature adult *B. dorsalis* flies, and its target was putatively identified as β2-tubulin (Tariq et al. 2015). The function of β2-tubulin was first revealed in *D. melanogaster* as maintaining the mobility of sperms (Zimowska, Nirmala, and Handler 2009). These findings support the idea that miR-14 is a male-specific miRNA functioning in WSS adult male testes.

Plants have evolved mechanisms to protect themselves from herbivorous feeding. In the case of an insect attack, defense mechanisms in plants are triggered by signals such as touch, oviposition, tissue damage and molecules coming from the insect (Chung et al. 2013). On the other hand, insects use effector molecules to suppress or manipulate defense response in host plant (Erb, Meldau, and Howe 2012; Hogenhout and Bos 2011). For example, a recent study showed that small RNA molecules of a fungi species, *B. cinerea*, inhibiting the RNAi machinery and silencing the genes for plant immunity through binding to AGO1 protein of its host plant Arabidopsis (Weiberg et al. 2013). Another study showed that host target sequences of *P. parasitica* sRNAs were transcribed at the low or undetectable levels (Jia et al. 2017). In the light of these findings, we considered larval miRNAs affecting host wheat plants to regulate gene expression in favor of larval survival. Target prediction analysis of larvae miRNAs brought out the possible interactions with wheat protein-coding sequences, which may result in the blockage of resistance to larval feeding. Intriguingly, miR-277 was shown to target several loci on chromosome 3B, which has been associated with the stem solidness feature of wheat. Predicted wheat targets of these transcripts showed significant similarity to methyltransferases and Ankyrin-like proteins. Ankyrin, a repeat domain, is important for several protein-protein interactions (Becerra et al. 2004). One of the best-studied

functions of Ankyrin-like proteins is pathogen resistance through regulation of salicylic acid-induced gene expression (H. Lu et al. 2003; Despres et al. 2000). Thus, it is tempting to speculate on the interactions between larval miR-277 and plant RNAs, potentially affecting stem solidness and plant defense, thus decreasing resistance to larval feeding inside the stem.

Since WSS larvae eat plant tissues for survival, it is very likely that plant miRNAs are taken inside of the insect body within their dietary consumptions. Several studies have provided evidence of trans-kingdom transfer of sRNAs from plant to other species which are in close contact; plant to virion (Iqbal et al. 2017), plant to nematodes (Tian et al. 2016), and plant to animal during feeding (L. Zhang et al. 2012). Wheat miRNAs might also act as the regulators of insect metabolism. Here, we showed potential larval targets of wheat miRNAs. However, these initial findings are needed to be validated to conclude on cross-kingdom miRNA regulation between WSS and wheat species.

5. CONSTRUCTION OF LONG NON-CODING RNA IDENTIFICATION MODEL SPECIFIC FOR WHEAT SPECIES USING MACHINE LEARNING APPROACHES

5.1. Introduction

Long non-coding RNAs (lncRNAs) can be easily distinguished from small noncoding RNAs, like miRNAs, snoRNAs and sRNAs, by the size of transcripts. However, although there are certain structural and functional differences between lncRNAs and mRNAs, they both are long transcripts and share similar splicing and poly-A tailed structure (Ulitsky and Bartel 2013). Therefore, it is hard to distinguish them through sequencing as they are sequenced together with the current sequencing techniques. Besides, lncRNA transcripts could not be identified by homology as lncRNA sequences appear less conserved than protein-coding genes (Pang, Frith, and Mattick 2006).

Another difficulty is the presence of open reading frames in lncRNAs. There are growing evidence showing lncRNAs coding for short functional peptides. The best-known example is the RNA called early nodulin 40 (ENOD40) (Campalans 2004), whose conserved nucleotide sequence at the 5' end encodes two short peptides with 12 and 24 amino acids in length (Rohrig et al. 2002). Recently, proteogenomic and mass spectrometry have been carried on to identify peptides identified from small ORFs (Zhu et al. 2018; Andrews and Rothnagel 2014). Nevertheless, molecular functions and biological significance of most lncRNAs is far from clear comparing with coding RNAs as correct identification of them remains a challenge at the first place.

In recent years, several predictive tools have been developed to distinguish between lncRNAs and coding RNAs using different features and different algorithms. The popular

tools, Coding Potential Calculator (CPC) (Kong et al. 2007), Coding Non-Coding Index (CNCI) (Sun et al. 2013) and Coding Potential Assessment Tool (CPAT) (L. Wang et al. 2013), are among the ones most accurate and informative.

CPC uses support vector machine (SVM) with a standard radial basis function kernel to differentiate coding RNAs from ncRNAs based on three ORF related and three sequence alignment related features (Kong et al. 2007). CPC has been updated to an alignment-free CPC2 in 2017 (Kang et al. 2017) which became much more faster and accurate in the identification of ncRNAs. The selected features were evolved in CPC2 to ORF length, ORF integrity, isoelectric point and Fickett score adapted from CPAT. Fickett score refers the asymmetrical distribution of each base favored in a sequence (L. Wang et al. 2013).

CPAT evaluates coding potential using an alignment-free logistic regression model (L. Wang et al. 2013). Its features include ORF length, Fickett score and Hexamer score. Hexamer score captures the score for the codon usage bias of adjacent amino acids in a sequence (L. Wang et al. 2013). CPAT has an advantage over CPC2 as it allows users to create a model with their own data.

CNCI is another alignment-free tool using SVM with radial basis function kernel. It differentiates coding RNAs and ncRNAs based on the intrinsic composition of the sequence (Sun et al. 2013). Similar to hexamer score in CPAT, CNCI estimates the codon bias using unequal distribution of adjoining nucleotide triplets (ANTs) with a sliding window approach. The most likely coding domain sequence (MLCDS) is selected after scanning a sequence for six times for each potential reading frames. Although having similarities with hexamer score, ANT approach conducts more comprehensive downstream analysis to handle the classification of partial transcripts (Han et al. 2016).

There are also several other tools exist each using different prediction models and different feature sets; PLEK, lncRNA-ID, DeepLNC etc. In short, PLEK facilitates support vector machine using k-mer based features to distinguish lncRNAs from coding RNAs (A. Li, Zhang, and Zhou 2014). BASINET uses decision tree algorithms trained with alignment-free features (Ito et al. 2018). DeepLNC facilitates deep-learning (Tripathi et al. 2016) whereas LncRNA-ID uses random forest (Achawanantakun et al. 2015). Even some tools construct an ensemble of models such as gradient boosting and

random forest are utilized by Simopoulos et al. for the prediction of plant lncRNAs (Simopoulos, Weretilnyk, and Golding 2018).

Although current computational methods have yielded encouraging results, they are facing certain limitations. The predictions are highly dependent on the training data. These tools aim to achieve high overall accuracy in several species from human to plants. Although some of them allow specification to plants, recent studies have showed that species-specific predictions perform best on its own data or on closely related species (Singh et al. 2017). Singh et al. showed that the model built for monocots achieved higher accuracy in predictions of lncRNAs in monocots rather than dicots and *vice versa*.

Here we developed a lncRNA prediction model specific to wheat species. Wheat is one of the major crops ranking second in human consumption worldwide (Food and Agriculture Organization of the United Nations 2019). To accurately identify both lncRNA transcripts and coding transcripts, we developed an alignment free prediction model specific for wheat. This model takes several features proposed by popular tools along with basic statistics like length, GC content and k-mer (1-3) distribution of nucleotides as feature set. Using these feature set and a comprehensive training data, we first evaluated prediction accuracies of ten different algorithms including popular ones like support vector machine, logistic regression and random forests. Of the ten algorithms, eight provided prediction accuracy over 99% with a 100-fold cross validation. After selecting best performing algorithms, we also compared our models with the popular tools.

5.2. Materials and Methods

5.2.1. Datasets

Training and validation datasets were collected from GrainGenes database which provides direct links to latest genome annotations of small grains, serving as their central data repository (Blake et al. 2016). A comprehensive annotation of hexaploid common wheat genome has become available almost three years after the sequencing of IWGSC

wheat genome reference sequence ((IWGSC) et al. 2018). We compiled a list of 87,511 lncRNA sequences and 137,056 high confidence coding domain sequences available through IWGSC refseq annotation v1.0 for the training of the prediction model specific for wheat species. For performance evaluation, we utilized lncRNAs and high confidence CDS available by the recent release of tetraploid durum wheat cultivar Svevo reference genome (Maccaferri et al. 2019). A total of 115,437 lncRNA and 196,153 high confidence cds sequences were retrieved from Svevo annotation.

5.2.2. Feature extraction

We extracted 92 features based on sequence intrinsic properties to use in prediction model construction. We executed known software for 6 of the features. Transdecoder (m -30) was executed for the prediction of longest putative ORF. The features used in the prediction models, CPAT and CPC2, were also included in the feature set. Final feature set was prepared based on sequence nucleotide composition using custom python scripts.

1	*ORF length*
2	*ORF coverage*
3	*sequence length*
4	*GC%*
5-8	*k-mer (k=1) frequencies; monomer frequencies of the four nucleotides*
9-24	*k-mer (k=2) frequencies; dimer frequencies of the four nucleotides*
25-88	*k-mer (k=3) frequencies; trimer frequencies of the four nucleotides*
89	*Fickett score*
90	*Hexamer score*
91	*ORF integrity*
92	*isoelectric point*

The features 1 and 2 were derived from the longest ORF predicted by Transdecoder. Features 89 and 90 were generated using CPAT and features 91 and 92 using CPC2.

5.2.3. Model construction

All data preprocessing, machine learning, and prediction evaluation of models were performed using python scikit-learn library. Data was preprocessed before starting training or validation by scaling. Scaling was performed to standardize features to avoid breaking the sparsity since many algorithms assume that all features are centered around 0 with a variance in the same order.

A total of ten machine learning algorithms were initiated and compared for accuracy in these predictive models. These algorithms include: (1) LogisticRegression, (2) RandomForest, (3) neural networks (NeuralNet), (4) NearestNeighbors, (5) support vector machine with linear kernel (linearSVM), (6) support vector machine with radial basis kernel (rbfSVM), (7) DecisionTree, (8) gaussion naive bayes (NaiveBayes), (9) AdaBoost, and (10) quadric discriminant analysis (QDA).

5.2.4. Model evaluation

For evaluation of the prediction accuracy of all ten machine learning algorithms, a 100-fold cross validation was conducted using cross_val_score function in the model_selection package. Cross validation works by selection of different test and train set in each run. Prediction performance was assessed by the mean and the standard deviations of the accuracy scores in these runs.

We proceeded to validation with the models with top three performing algorithms. The prediction models created using hexaploid wheat data as training were validated using tetraploid wheat data for the top three performing algorithms.

The classification performance was evaluated based on the statistic metrics; accuracy (ACC), precision (PRE), sensitivity (SN), specificity (SP) and Fscore which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Fscore = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

TP: true positive, TN: true negative, FP: false positive, FN: false negative.

For comparison of the prediction performances with other tools, we utilized CPC2, CPAT and CNCI. CPC2 were run at its default settings using pre-built training model. CNCI were run at its plant mode and using 20 threads. CPAT were trained with hexaploid wheat data. The cutoff for the identification of coding and noncoding transcripts were identified as described in its manual (L. Wang et al. 2013).

5.3. Results

5.3.1. Experimental setup and model construction

We constructed list of features including basic characteristics like length and GC content, k-mer patterns, and the features proposed by CPC2 and CPAT. As a preprocessing step, these features were scaled prior to construction of prediction models.

Table 5.1. Description of datasets used in training and validation of the wheat lncRNA prediction model

Dataset	Source	Reference	Transcript	Size	Max length	Min length	Mean length
Training	Hexaploid wheat	(IWGSC) et al., 2018	mRNA	137,056	16,080	96	1,122
			lncRNA	87,511	5,508	200	355
Validation	Tetraploid wheat	Maccaferri et al., 2019	mRNA	196,153	16,083	192	1,161
			lncRNA	115,437	4,407	72	245

To test the performance of prediction models, we selected the data for two wheat genomes recently released; hexaploid common wheat, Chinese Spring and tetraploid durum wheat, Svevo (Table 5.1). As training data, we used the data for hexaploid wheat which comprised of 137,056 coding transcripts and 87,511 lncRNA transcripts. For the

validation data, tetraploid wheat annotation containing 196,153 coding transcripts and 4,407 lncRNA transcripts was used. The training set was used to train the prediction models and the validation set was used to test the actual prediction capability of the models.

Table 5.2. Performance of prediction models using training data with 100-fold cross validation

Algorithm	**Training Accuracy (%)**	**Std (%)**
LinearSVM	99.94	+/- 0.11
LogisticRegression	99.89	+/- 0.14
NeuralNet	99.84	+/- 0.23
rbfSVM	99.57	+/- 0.41
RandomForest	99.36	+/- 0.44
AdaBoost	99.32	+/- 0.48
DecisionTree	99.23	+/- 0.44
QDA	98.81	+/- 1.00
NaiveBayes	96.80	+/- 2.16
NearestNeighbors	93.91	+/- 3.52

As different algorithms were suggested as the best fit for the classification of lncRNAs by different studies, we compared classification accuracies of ten machine learning algorithms. Most of the algorithms resulted in over 99% accuracy (Table 5.2) indicating the good fit of the selected features in the prediction models. We selected top three algorithms for validation of prediction models, which are SVM with linear kernel, logistic regression and neural networks.

5.3.2. Performance evaluation on tetraploid wheat data

We compared the performance of top three prediction models with popular coding potential prediction tools: CPC2, CPAT and CNCI. We retrained the classification model in CPAT for hexaploid wheat data. CNCI was run with its plant mode. As CPC2 don't provide training option, we used its pre-built model.

Table 5.3. Performance comparison of prediction models on tetraploid wheat data

Model	**ACC**	**PRE**	**SN**	**SP**	**F-score**
LinearSVM	99.77	99.46	99.90	99.68	99.68
LogisticRegression	**99.81**	99.62	99.86	99.78	**99.74**
NeuralNet	99.72	99.41	99.85	99.65	99.63
CPC2	97.35	93.36	99.97	95.81	96.55
CPAT	99.70	**99.69**	99.50	**99.82**	99.60
CNCI	96.54	91.47	**100.00**	94.51	95.54

Abbreviations are as follows; ACC: accuracy, PRE: precision, SN: sensitivity, SP: specificity. Highest values of the metrics were shown in bold.

Table 5.3 shows the performance comparison of the prediction models on tetraploid wheat data. The model created by logistic regression resulted in the best accuracy and F-score. CNCI had the highest sensitivity where among 115,437 lncRNAs, only three were classified falsely. Although providing the best sensitivity to lncRNA identification, CNCI provides the lowest values of the remaining metrics. Its accuracy in prediction of coding transcripts (specificity) were only 94.51% while the highest value for this metrics was 99.82% in CPAT. CPAT also provides the best precision although the values for precision were pretty close between logistic regression model (99.62) and CPAT (99.69) as well as the values for specificity; 99.78 and 99.82 for logistic regression model and CPAT, respectively.

Although all three proposed models were performed well, we selected the model created by logistic regression for further use as it provides the best results in general. Figure 5.1 shows detailed results for the classification predictions of the selected logistic regression model and the three popular tools, CPC2, CPAT and CNCI.

All models performed well in the classification of non-coding transcripts. Only a small percent of non-coding transcripts noticeable were misclassified as coding transcripts by CPAT. For the classification of coding transcripts, our model and CPAT outperformed CNCI and CPC2. Overall, our logistic regression model performed well in both coding and noncoding transcript predictions, whereas other tools favor one.

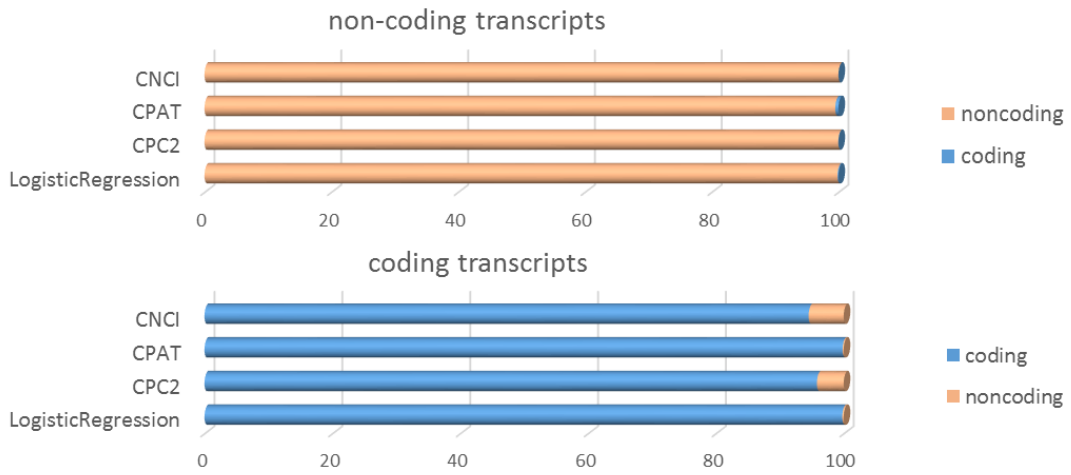


Figure 5.1. Accuracy of prediction models on coding and non-coding transcripts separately on tetraploid wheat data. Above figure showed the percentage of non-coding transcripts classified as coding (blue) and non-coding (orange). Below figure showed the percentage of coding transcripts classified as coding (blue) and non-coding (orange).

5.3.3. Feature importances

We further investigated feature importance in our prediction model. Scaling of the data prior to analyses makes the features comparable; therefore, relative feature ranking indicates their contribution to prediction accuracies. We extracted feature ranking using coef_ attribute in logistic regression classifier.

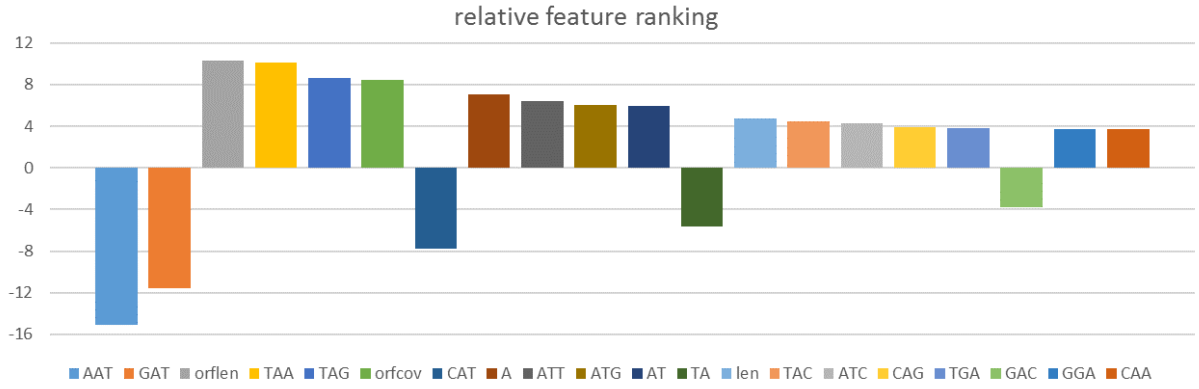


Figure 5.2. Feature ranking of wheat specific prediction model based on logistic regression. Relative feature ranking indicates the coefficient of the features in the decision function.

Among the top 20 features (Figure 5.2), 14 are trimer percentages, 2 are dimer percentages, 1 is monomer distribution (A nucleotide composition), and the remaining features are ORF length, ORF coverage and length of the transcripts. Interestingly, neither of features from CPAT and CPC2 was among the top 20 features.

5.4. Discussion

The performance of machine learning models highly depends on the training data and the features used. Several features proposed by different studies have been shown to be informative in the classification of coding and noncoding transcripts. These features include k-mers, basic structural features like length and GC content, Fickett score, hexamer score, ORF integrity and isoelectric point.

In this study, we proposed an accurate model for lncRNA and mRNA identification in wheat species. As training data, we used the comprehensive annotation of wheat reference genome. As for the feature set, we incorporated all these features listed to achieve better prediction accuracies. With a comprehensive training data and substantial list of features, we compared 10 different algorithms for their prediction performances using the same training data and the same feature sets. Interestingly, training accuracies were over 99% in eight of the algorithms (Table 5.2), indicating importance of training data and feature sets over prediction algorithm. Among these algorithms, logistic regression, support vector machine with linear kernel and neural network performed best in the prediction of wheat lncRNAs.

Comparison of the top three algorithms and popular tools like CPAT, CPC2 and CNCI revealed that our model created based on logistic regression classifier had the best performance overall. Among all the tools, CNCI and CPC2 showed the worst performance with the lowest accuracy and the lowest precisions. We found only CPAT as competitor whereas our model provides better results for all metrices except for specificity and precision, which had very close results in our model (Table 5.3).

We found that other tools might provide better sensitivity but with a cost of lowest

precision (Table 5.3). For example, similar to CPC2, CNCI provide 100% sensitivity in the prediction of lncRNAs; however, its precision was only 91.47%. On the contrary, our model provided 99.86% of sensitivity together with a 99.62% of precision. Therefore, our model is consistent with both lncRNA and coding transcripts.

Although we performed these analyses on wheat data, the same model with the defined features can be used in other plant and mammalian data. Creating a feature set based on sequence intrinsic composition should allow accurate prediction in other datasets too, although that is not in the scope of this work.

6. COMPARATIVE ANALYSIS OF WHOLE EXOME SEQUENCING (WES) TOOLS IN WHEAT

6.1. Introduction

It is currently a very exciting time for wheat genomics study. With the recent releases and availability of the wheat reference genomes ((IWGSC) et al. 2018), the mining of the wheat exome for the variants responsible for important traits of interest is becoming ever more readily available. However, as with all analysis, there is a need to ensure that results are both reliable and reproducible. Although calling for variants is relatively straightforward with two major steps; read alignment and variant calling, choice of the best tools at each stage of the analysis is not. There have been concerns raised in the literature regarding that impact of method choice on these ever-important metrics of result quality (Cornish and Guda 2015).

In response, research has been conducted to better characterize such impacts in organisms including human (Abecasis et al. 2010), where exome sequencing plays a large role in the clinic settings, and *Arabidopsis* (Cao et al. 2011). However, such attention has yet to be turned to wheat; an organism that relies still on the mining of exome data to characterize underlying variation. As such, there is therefore a need to better characterize and understand how different methods affect the analysis of whole exome capture and sequencing in common bread wheat (*Triticum* aestivum).

This study aims to meet these requirements by assessing the outcome of various methods at all stages in a whole exome sequencing (WES) analysis pipeline. We sequenced exome of 48 elite wheat cultivars, analysed and compared different tools. This WES data from

hexaploid wheats to compare the bioinformatics pipelines with the help of IWGSC RefSeq V1.

6.2. Materials and Methods

6.2.1. Preparation of wheat exome capture libraries

The exome regions were captured with the SeqCap EZ Developer Reagents (Roche). The libraries for 48 Montana elite wheat cultivars were quantitated by qPCR and sequenced on one lane for 101 cycles from each end of the fragments on a HiSeq 4000 using a HiSeq 4000 sequencing kit version 1. Generated fastq files were demultiplexed with the bcl2fastq v2.17.1.14 which removes adaptors from the 3'-end of the reads.

FastQC returned supporting evidence on its quality. No need for additional trimming (from either adaptors and low-quality regions).

6.2.2. Alignment parameters

Figure 6.1 shows the overall pipeline starting from raw fastq files and ending with filtered vcf files. After quality controls, WES reads for 48 wheat cultivars were aligned to wheat reference genome separately using 8 different aligners. Aligners included were bowtie2 (Langmead 2010), bowtie2 --local (bowtie2local), bwa –sampe (bwa) (H. Li and Durbin 2010), bwa –mem (bwamem), gsnap (T. D. Wu and Nacu 2010), hisat2 (Pertea et al. 2016), STAR (Dobin et al. 2013) and novoalign . These aligners converted fastq files into raw SAM files. All parameters were set to defaults. Multiple threads were used where available. Versions of aligners used:

bwa-0.7.17-gcc-8.2.0-o6zcgoi
bowtie2-2.3.4.1-gcc-7.2.0-hp2vf2y
gmap-gsnap-2018-03-25-gcc-7.2.0-tj6cfa7
hisat2-2.1.0-gcc-7.2.0-bradwj6

star-2.6.1a-gcc-8.2.0-stezqss
novoalign-v3.09.00

The choice of aligner and command line arguments to run alignments were as follows:

1. bowtie2 -x ../wheat_bowtie2_db -p 20 -1 $read1 -2 $read2 -S $filename.bowtie2.sam
2. bowtie2 --local -x ../wheat_bowtie2_db -p 20 -1 $read1 -2 $read2 -S $filename.bowtie2local.sam
3. bwa aln -t 20 -q 20 $genomefasta $read1 > $filename.bwa1.sam
   bwa aln -t 20 -q 20 $genomefasta $read2 > $filename.bwa2.sam
   samtools index $filename.bwa1.sam > $filename.bwa1.sai
   samtools index $filename.bwa2.sam > $filename.bwa2.sai
   bwa sampe -a 200 $genomefasta $filename.bwa1.sam $filename.bwa2.sam $read1 $read2 > $filename.bwa.sam
4. bwa mem -M -t 20 $genomefasta $read1 $read2 > $filename.bwamem.sam
5. gsnapl -A sam -d wheat_gsnap_db -D ../gsnapindex -t 20 $read1 $read2 > $filename.gsnap.sam
6. hisat2 -x ../wheat_hisat2_db -p 20 -1 $read1 -2 $read2 -S $filename.hisat2.sam
7. STAR --genomeDir ../index --readFilesIn $read1 $read2 --runThreadN 20 > $filename.star.sai
8. novoalign -d wheat_novoindex_db.ndx -f $read1 $read2 -o SAM > $filename_novoalign.sam

Alignments were processed further before variant calling to prepare a sorted and clean bam file. To do so, duplicates were removed first from each SAM file using SAMblaster v0.1.24 (Faust and Hall 2014). Clean SAM file converted into BAM file and sorted using samtools v1.9 (H. Li et al. 2009). Finally, group IDs were inserted according to file names.

6.2.3. Variant calling

Three variant calling methods were executed for each alignment (48 samples * 8 aligners): freebayes v1.2.0 (Garrison and Marth 2012), bcftools call (bcftools) v1.8 and varscan v2.4.2 (Koboldt et al. 2009). We performed variant calling separately, as freebayes and varscan returned memory error for a merged file, indicating 240 GB of

memory is not sufficient to perform variant calling for a merged file for 48 samples. All parameters were set to defaults except for freebayes -p 6 -0; bcftools mpileup --redo-BAQ; bcftools call -m. Multiple threads were used where available.
The choice of aligner and command line arguments to run alignments were as follows:

1. Freebayes (parameters -p 6 (ploidy=6) -0):

   freebayes -f $genomefasta -p 6 -0 $filename.$aligner_RG.bam > $filename.$aligner.free.vcf

2. Varscan:

   samtools mpileup -f $genomefasta $filename.$aligner_RG.bam | varscan mpileup2snp --output-vcf 1 > $filename.$aligner.varscan.vcf

3. Bcftools (parameters: --redo-BAQ, call -m)

   bcftools mpileup --redo-BAQ -f $genomefasta $filename.$aligner_RG.bam | bcftools call -m -o $filename.$aligner.bcftools_rm.vcf -O v -v

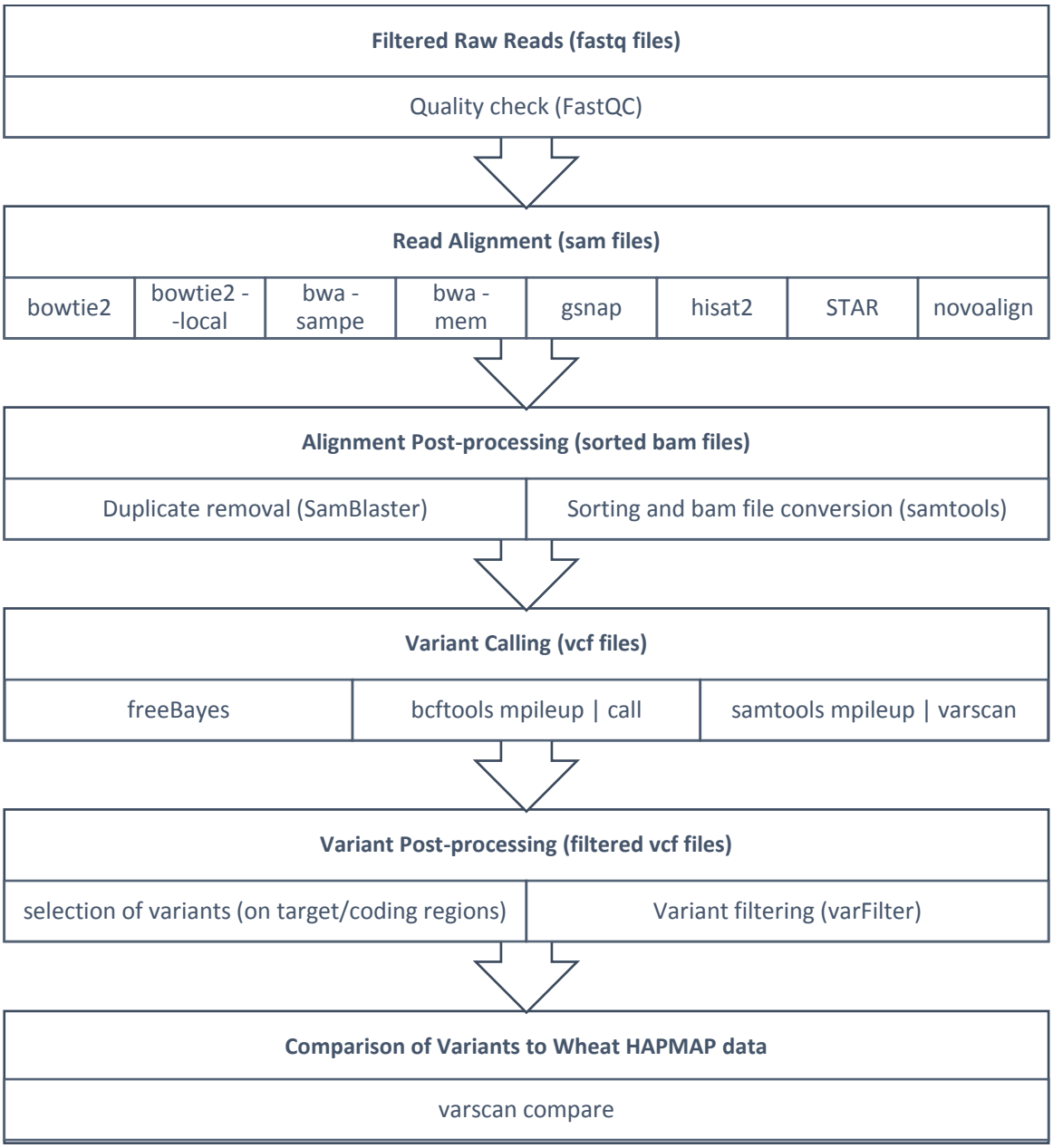


Figure 6.1. Schematic of the variant calling pipeline used. Only one of the read aligners and one of the variant callers used for one pipeline, totaling twenty-four different pipelines (8 aligners x 3 callers).

We performed merging of variant for 48 samples for 24 combinations of tools (8 aligners * 3 callers). Using bcftools, we zipped, indexed and merged variant files for all samples. Later, we performed variant postprocessing; we filtered variations at the target exon sites using bed files provided by reference genome annotation (refseq v1.1) of wheat; We extracted only SNPs from these variants using bcftools view (-v snps). Finally, variants were filtered using varFilter at default settings.

Although there is not any complete haplotype map of wheat, we used wheat HAPMAP data as the gold standard genotype data. We compared newly identified variants to wheat HAPMAP data to find out which the pipeline identifies most of the known structural variants. However, this data is not complete and could not provide us even near complete map. Therefore, we used the terms ‘shared’ and ‘unique’ variations when compared to HAPMAP data, instead of ‘true positives’ and ‘false positives’. snpEff used to annotate variants based on wheat refseq annotation.

6.3. Results

6.3.1. Datasets and pipelines evaluated

To reduce bias introduced by samples and conduct a more complete haplotype map, we included whole exome capture data for 48 wheat cultivars. Using eight aligners and three variant callers (Figure 6.1), 24 pipelines were assessed in terms of detecting variations from WES data. The aligners include both older and newest version of bwa; bwa sampe and bwa mem, which from this point will be names as bwa and bwamem, respectively. Both options of bowtie2; bowtie2 and bowtie2 --local were also included. Remaining aligners were hisat2, gsnap, STAR and novoalign. For variant calling, we included three tools (freebayes, varscan and bcftools).

Although freebayes and varscan contains some default filters, bcftools call function does not apply a default filtering. As a result, total number of variations at the target site varies as much as 23 folds. Running varFilter at default settings compensate this inequity. varFilter default options include min mapping quality of 10, min depth of 2, min number of alternate bases of 2 etc. These options only indicate a soft filtering where several studies prefer filtering options like 30 mapping quality and 10 read depth.

Run time is highly divergent among aligners while being highly similar among callers. As run time for callers did not exceed half a day and similar to each other, we did not perform a run time comparison of callers. Table 6.1 illustrates the average time spent running each aligner.

Table 6.1. Average run time of aligners.

Aligner	Average Run Time	# of threads
star	00-00:22:17	20
hisat2	00-00:24:36	20
bwamem	00-01:10:26	20
bowtie2	00-01:29:16	20
bowtie2local	00-01:39:30	20
bwa	00-04:50:49	20
gsnap	05-05:36:37	20
novoalign	14-16:30:00	1

Run time: days-hours:minutes:seconds

Novoalign supports multiple threads too but only in payed versions. Without payment, we were able to run novoalign on single threads which increased run time a lot more. If multiple threads available, similar to bowtie2local, we expect an average run time of ~2 hours. Star and hisat2 were the fastest aligners, completed in less than half an hour for all samples. Assuming running novoalign with 20 threads, bwamem, bowtie2, bowtie2local and novoalign provided similar computational complexity. However, novoalign is not suitable without multiple threads due to its run time with single thread. Gsnap, on the other hand, lasted ~5 days on average with 20 threads which is not even comparable with the remaining aligners.

6.3.2. Filtering of variants

We included results both for raw vcf files and targeted vcf files in Table 6.2. While those off-target reads (raw vcf files) are also highly valuable, we filtered them to evaluate performance of variant calling pipelines rather than the targeting efficiency of sequencing platforms. The number of SNPs identified decreased at least 5-fold by filtering for coding regions, resulting in ~870,000 SNPs per pipeline on average (Table 6.2).

Noticable, the number of unique variations in either raw or targeted data is a lot higher than the number of shared variations (Table 6.2), which is mostly due to the

incompleteness of the HAPMAP data. As the complete haplotype map for wheat is not available yet, we used the terms ‘shared’ and ‘unique’ variations compared to HAPMAP data available, instead of ‘true positives’ and ‘false positives’.

Table 6.2. Variant statistics for all 24 pipelines, including raw and target specific variants. The best pipelines were colored by red.

		Raw			**Targeted**		
Aligner	**Caller**	**Unique**	**Shared**	**Total**	**Unique**	**Shared**	**Total**
hisat2	varscan	870 890	113 536	984 426	129 413	18 914	148 327
bwamem	varscan	2 045 196	156 588	2 201 784	250 604	25 244	275 848
novoalign	varscan	1 453 292	142 508	1 595 800	250 716	24 279	274 995
star	varscan	1 888 143	158 815	2 046 958	313 425	27 440	340 865
gsnap	varscan	1 720 189	147 939	1 868 128	301 326	26 028	327 354
bowtie2-local	varscan	2 185 316	157 326	2 342 642	314 006	26 611	340 617
bowtie2-local	bcftools	9 681 921	282 928	9 964 849	505 035	41 752	546 787
bwa	varscan	1 527 162	147 957	1 675 119	341 912	27 126	369 038
bowtie2	varscan	1 930 915	154 350	2 085 265	358 231	27 234	385 465
bowtie2	bcftools	18 976 316	267 761	19 244 077	517 771	38 854	556 625
hisat2	bcftools	18 976 316	267 761	19 244 077	517 771	38 854	556 625
star	bcftools	10 148 130	301 127	10 449 257	692 755	47 284	740 039
bwa	bcftools	17 323 257	304 225	17 627 482	723 058	47 465	770 523
bowtie2-local	freebayes	3 891 690	114 415	4 006 105	196 136	11 371	207 507
bowtie2	freebayes	4 686 479	137 713	4 824 192	296 725	16 667	313 392
novoalign	bcftools	27 621 517	312 976	27 934 493	1 103 083	49 260	1 152 343
bwamem	bcftools	27 713 219	314 608	28 027 827	1 122 459	49 570	1 172 029
gsnap	bcftools	43 107 654	345 910	43 453 564	1 811 226	55 477	1 866 703
novoalign	freebayes	17 921 696	273 072	18 194 768	1 503 327	42 067	1 545 394
bwa	freebayes	14 591 569	251 208	14 842 777	1 312 765	36 671	1 349 436
hisat2	freebayes	21 932 016	259 341	22 191 357	1 552 361	38 657	1 591 018
bwamem	freebayes	21 164 838	279 046	21 443 884	1 763 897	42 937	1 806 834
star	freebayes	24 872 748	287 906	25 160 654	1 866 353	44 786	1 911 139
gsnap	freebayes	33 162 705	303 252	33 465 957	2 372 777	47 780	2 420 557
	Average	13 724 716	228 428	13 953 143	838 214	35 514	873 728
	Std	12 029 744	76 883	12 097 637	662 944	11 964	671 806

6.3.3. Comparison of aligners and callers on identification efficacy

Total number of SNPs identified, from targeted sites, using different combination of aligners and callers were shown in Table 6.3. As can be seen from standard deviations, aligners tend to deviate a lot more compared to callers. For example, hisat2 provided 765,323 SNPs on average but its standard deviation was also 743,643. The scores were similar between aligners indicating that aligners were not the limiting factor in variant calling.

On the other hand, standard deviations were almost half of the average numbers in bcftools and freebayes and were almost one quarter in varscan (Table 6.3). The results showed that varscan provided similar number of SNPs with all eight aligners, possible due to its high default filtering parameters. Average number SNPs identified were ~1,300,000 by freebayes whereas it was around only ~300,000.

Table 6.3. Total number of variations identified on the targeted site using different aligners and callers. The lowest and the highest total number of variations indicated by orange and green, respectively.

	bcftools	freebayes	varscan	AVERAGE	STDEV
bowtie2	556625	313392	340617	403545	133268
bowtie2local	546787	207507	385465	379920	169708
bwamem	770523	1349436	275848	798602	537345
bwa	1172029	1806834	369038	1115967	720536
gsnap	1866703	2420557	327354	1538205	1084577
hisat2	556625	1591018	148327	765323	743643
novoalign	1152343	1545394	274995	990911	650403
Star	740039	1911139	340865	997348	816147
AVERAGE	920209	1393160	307814	873728	
STDEV	458546	768196	75464		671806

Among aligners, gsnap and among callers, freebayes provided highest number of variations. As expected, highest number of SNPs were also identified using gsnap-freebayes pipeline. However, interestingly, although the lowest number of SNPs were identified using bowtie2local or bowtie2 as aligner, hisat2-varscan pipeline resulted in

the lowest number of SNPs. These results indicated that combination of tools also affect the results (although callers have more influence on the results).

6.3.4. Comparison of filtered results against wheat HAPMAP data

We compared targeted variations against wheat HAPMAP data to determine the best pipeline. The motivation behind this comparison was to select the pipelines which identify the highest number of known structural variations with lowest number of total variations since we could not make sure of the validity of unique variations at this point.

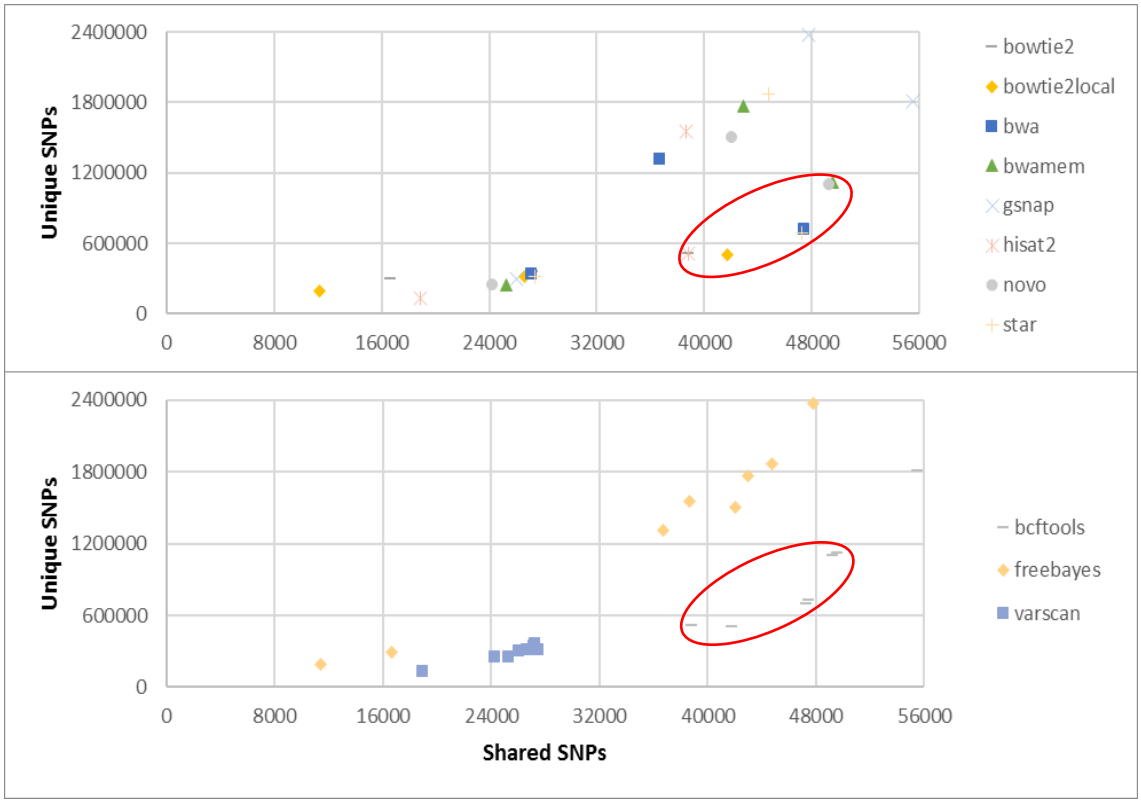


Figure 6.2. Comparison between the number of unique SNPs identified and the number of shared SNPs with wheat HAPMAP data. (a) Distribution of aligners, (b) distribution of callers. The best pipelines selected (by <1200000 unique snps and >32000 shared snps) shown by red circles.

Figure 6.2 shows distribution of unique and shared variations with wheat HAPMAP data. Unlike callers' graph, unique variations over shared variations graphs for aligners does not indicate any clustering but rather a random distribution of aligners (Figure 6.2). These

results also suggest that variant discovery is highly dependent on the caller rather than the aligner. Interestingly, two freebayes pipelines, bowtie2 and bowtie2local, were outside of the freebayes cluster, with the lowest numbers of shared SNPs.

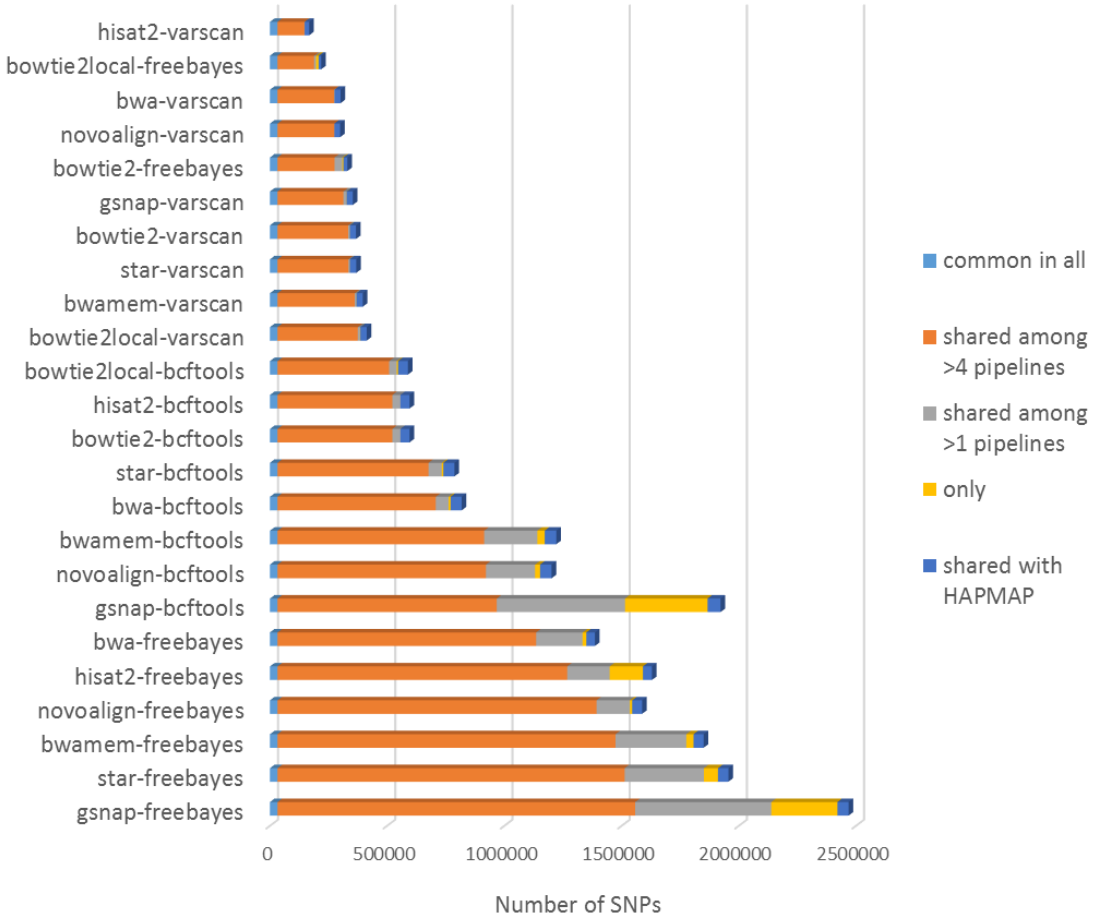


Figure 6.3. Distribution of SNPs identified by all 24 pipelines. SNPs identified by all 24 pipelines were colored as blue, remaining SNPs identified by more than 4 pipelines were colored in orange, remaining SNPs identified by at least 2 pipelines were colored in gray and the novel SNPs identified were colored in yellow. In addition to the total number of SNPs, the SNPs shared with HAPMAP data were included for representative purposes (colored by dark blue). The pipelines were sorted by the number of SNPs shared among more than four pipelines.

It should also be noted that although providing the lowest number of variations, bowtie2-freebayes and bowtie2local-freebayes pipelines suggested 4785 and 10901 novel SNPs, respectively. Median number of novel SNPs identified by each pipeline was 6512. All varscan predictions together with bowtie2-bcftools (0) and hisat2-bcftools (0) pipelines

were resulted in lower number of novel predictions. These results suggest that freebayes tend to call novel variations irrespective of total number of variations identified, which might be an indicator of unreliability of freebayes in WES data.

Variant calling using varscan resulted in almost the same SNPs in all pipelines. In fact, varscan provided highest percent of variations shared with wheat HAPMAP data. If we define sensitivity as the percent of shared variations over all predicted variations, we achieved 12.8% of sensitivity at most among 24 pipelines (8 aligners * 3 callers). The pipelines using varscan as caller together with bowtie2local-bcftools pipelines were the ones with the highest sensitivity, which were over 7%. However, these pipelines, except bowtie2local-bcftools pipeline, provided the lowest number of shared SNPs.

It can be noticed that varscan filters low confidence variations based on depth, quality etc. to achieve highest sensitivity but at the cost of elimination of known SNPs. We can suggest using varscan as caller only if you are interested in limited number of high confidence variations. The choice of aligner does not interfere with the results as varscan filtering results in shared variations only (Figure 6.3).

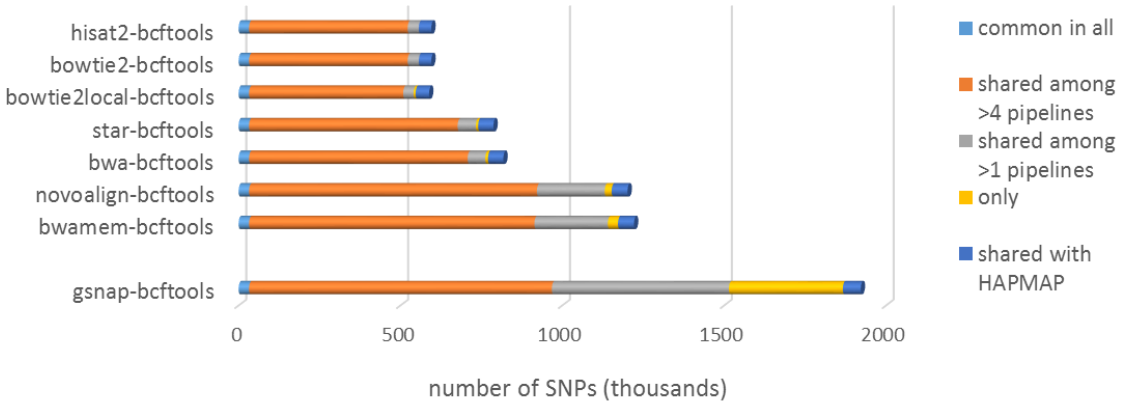


Figure 6.4. Distribution of number of SNPs called using bcftools call function as caller. The selected seven pipeline were separated from remaining gsnap-bcftools pipeline. 'Shared with HAPMAP' is additional to the total number of SNPs.

On the other hand, both bcftools call function and freebayes returned diverse set of variations (Figure 6.2). However, the use of bcftools call function instead of freebayes significantly enhanced prediction accuracies by decreasing the number of unique snps

while increasing the number of shared snps (bwamem, novoalign, bwa, gsnap, star, hisat2). For the remaining two aligners (bowtie2 and bowtie2local), the use of bcftools call function over freebayes greatly increased total number of shared SNPs with a slight increase in the number of unique SNPs.

As we determine quality of our predictions based on the similarities to the wheat HAPMAP data, the best pipelines were limited to five pipelines using bcftools call function (Figure 6.4). The cut-offs were applied as unique variations <1.200.000 and shared variations >32.000 in targeted regions to determine the best pipeline. The best pipelines can also be noticeable by red color in Table 6.2.

6.3.5. Variations identified using bcftools-bwamem pipeline

Total number of SNPs identified by bwamem-bcftools pipeline was 1209440, where 47465 of them were common with wheat HAPMAP data. Total number of SNPs and SNP density over wheat chromosomes were shown in Table 6.4. Average SNP density was highest in 2B, 2D and 2A chromosomes and lowest in 3B, 4D and 4B chromosomes (Table 6.4).

Average SNP density was 4,5% where 10 chromosomes had SNP densities between 4 and 5. Total variant rate was 12,028 meaning that there were 1 variant for every 12,028 bases. SNPs identified by bwamem-bcftools resulted in a missense-to-silent mutation rate of 1,05 and in a transitions (Ts) over transversions (Tv) ratio (ts/tv) of 1.89.

Table 6.4. Distribution of SNPs across the wheat chromosomes following bwamem-bcftools pipeline

Chromosome	Length	Variants	Variants rate	Variants percentage
1A	594 102 056	58 789	10,105	5,170
1B	689 851 870	73 997	9,322	6,507
1D	495 453 186	56 389	8,786	4,959
2A	780 798 557	80 097	9,748	7,043
2B	801 256 715	117 275	6,832	10,313
2D	651 852 609	80 760	8,071	7,102
3A	750 843 639	51 379	14,613	4,518
3B	830 829 764	72, 327	11,487	0,006
3D	615 552 423	58 069	10,600	5,106
4A	744 588 157	54 501	13,661	4,793
4B	673 617 499	20 351	33,099	1,790
4D	509 857 067	8 771	58,129	0,771
5A	709 773 743	28 988	24,485	2,549
5B	713 149 757	46 293	15,405	4,071
5D	566 080 677	30 525	18,544	2,684
6A	618 079 260	50 986	12,122	4,484
6B	720 988 478	74 833	9,634	6,581
6D	473 592 718	44 380	10,671	3,903
7A	736 706 236	64 419	11,436	5,665
7B	750 620 385	51 357	14,615	4,516
7D	638 686 055	49 440	12,918	4,348
Un	480 980 714	35 514	13,543	3,123
Total	**14 547 261 565**	**1 209 440**	**12,028**	100,000

6.4. Discussion

WES tends to produce large portion of off-target reads, including intronic and intergenic reads, although being a sequencing method targeting exon regions. Wheat genome studies revealed that ~85% of wheat genome is composed of repeat elements. Besides, wheat

chromosomes share high sequence similarity. Both repeat content and homologous chromosomes might drive off-target sequencing of exone regions. Additionally, filtering of variations as in varscan (at least 2 supporting reads at variant site, min 15 base quality, at least 8 read depth) decreased off-target effect (5 to 8 folds), indicating low quality of the variations at the off-target sites.

Another posibility is that the sequencing performed at the late 2017s. At that time reference genome sequence was not published yet. They might used probes for the older version of wheat genome during sequencing. Therefore, some regions might be involved unintentionally.

Our initial screening suggest that the choice of variant caller has more influence on the results than the aligner (Figure 6.2). Therefore, this study can be extended by including a few more callers to widen our comparative analysis of WES analysis pipelines. Besides, our results pointed out the importance of the sample size, as the best tool combinations differ by each single sample. Sample size is important for concrete conclusions. Given the less influence, the aligners having long run times and/or not supporting multiple threads can be eliminated for future analysis.

We believe this study will provide benefits to all plant scientist, especially to the wheat and barley community. So far, our recommendation is the use of bcftools call function as variant caller with bwamem or novoalign as aligner. The pipelines hisat2-bcftools and bowtie2-bcftools can be used to limit results to high confidence variations as those did not returned any novel variations but still kept number of variations shared with HAPMAP data at highest. For the identification of novel variants, one can prefer using freebayes.

7. BIBLIOGRAPHY

(IWGSC), The International Wheat Genome Sequencing Consortium, IWGSC RefSeq principal investigators:, Rudi Appels, Kellye Eversole, Catherine Feuillet, Beat Keller, Jane Rogers, et al. 2018. "Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome." *Science* 361 (6403): eaar7191. https://doi.org/10.1126/SCIENCE.AAR7191.

Abecasis, Gonçalo R, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, R A Gibbs, Matt E Hurles, and Gil a McVean. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature*. https://doi.org/10.1038/nature09534.

Achawanantakun, Rujira, Jiao Chen, Yanni Sun, and Yuan Zhang. 2015. "LncRNA-ID: Long Non-Coding RNA IDentification Using Balanced Random Forests." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv480.

Adams, M D, J M Kelley, J D Gocayne, M Dubnick, M H Polymeropoulos, H Xiao, C R Merril, a Wu, B Olde, and R F Moreno. 1991. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project." *Science (New York, N.Y.)* 252 (5013): 1651–56. https://doi.org/10.1126/science.2047873.

Akpinar, Bala A., and Hikmet Budak. 2016. "Dissecting MiRNAs in Wheat D Genome Progenitor, Aegilops Tauschii." *Frontiers in Plant Science* 7 (May): 1–17. https://doi.org/10.3389/fpls.2016.00606.

Akpinar, Bala Ani, Melda Kantar, and Hikmet Budak. 2015. "Root Precursors of MicroRNAs in Wild Emmer and Modern Wheats Show Major Differences in Response to Drought Stress." *Functional & Integrative Genomics* 15 (5): 587–98. https://doi.org/10.1007/s10142-015-0453-0.

Akpinar, Bala Ani, Stuart J. Lucas, and Hikmet Budak. 2013. "Genomics Approaches for Crop Improvement against Abiotic Stress." *The Scientific World Journal* 2013. https://doi.org/10.1155/2013/361921.

Alptekin, Burcu, Ani Akpinar, and Hikmet Budak. 2016. "A Comprehensive Prescription for Plant MiRNAs Annotation." *Frontiers in Plant Science* 7 (January): 2058. https://doi.org/10.3389/fpls.2016.02058.

Alptekin, Burcu, and Hikmet Budak. 2016. "Wheat MiRNA Ancestors: Evident by

Transcriptome Analysis of A, B, and D Genome Donors." *Functional & Integrative Genomics*. https://doi.org/10.1007/s10142-016-0487-y.

Alptekin, Burcu, Peter Langridge, and Hikmet Budak. 2016. "Abiotic Stress MiRNomes in the Triticeae." *Functional & Integrative Genomics*, 1–26. https://doi.org/10.1007/s10142-016-0525-9.

Alvarez-Garcia, Ines, and Eric A. Miska. 2005. "MicroRNA Functions in Animal Development and Human Disease." *Development* 132 (21): 4653–62. https://doi.org/10.1242/dev.02073.

Andrews, Shea J., and Joseph A. Rothnagel. 2014. "Emerging Evidence for Functional Peptides Encoded by Short Open Reading Frames." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3520.

Appels, Rudi, Kellye Eversole, Catherine Feuillet, Beat Keller, Jane Rogers, Nils Stein, Curtis J. Pozniak, et al. 2018. "Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome." *Science* 361 (6403): eaar7191. https://doi.org/10.1126/science.aar7191.

Ariel, Federico, Natali Romero-Barrios, Teddy Jégu, Moussa Benhamed, and Martin Crespi. 2015. "Battles and Hijacks: Noncoding Transcription in Plants." *Trends in Plant Science*. https://doi.org/10.1016/j.tplants.2015.03.003.

Avni, Raz, Moran Nave, Omer Barad, Kobi Baruch, Sven O Twardziok, Heidrun Gundlach, Iago Hale, et al. 2017. "Wild Emmer Genome Architecture and Diversity Elucidate Wheat Evolution and Domestication." *Science (New York, N.Y.)* 357 (July): 93–97.

Banks, Isaac R., Yuanji Zhang, B. Elizabeth Wiggins, Greg R. Heck, and Sergey Ivashuta. 2012. "RNA Decoys." *Plant Signaling & Behavior* 7 (9): 1188–93. https://doi.org/10.4161/psb.21299.

Bartel, David P. 2009. "MicroRNA Target Recognition and Regulatory Functions." *Cell* 136 (2): 215–33. https://doi.org/10.1016/j.cell.2009.01.002.MicroRNA.

Becerra, Cristian, Torben Jahrmann, Pere Puigdomènech, and Carlos M. Vicient. 2004. "Ankyrin Repeat-Containing Proteins in Arabidopsis: Characterization of a Novel and Abundant Group of Genes Coding Ankyrin-Transmembrane Proteins." *Gene* 340 (1): 111–21. https://doi.org/10.1016/j.gene.2004.06.006.

Behura, Susanta K. 2007. "Insect MicroRNAs: Structure, Function and Evolution." *Insect Biochemistry and Molecular Biology* 37 (1): 3–9. https://doi.org/10.1016/j.ibmb.2006.10.006.

Benton, R. 2006. “On the ORigin of Smell: Odorant Receptors in Insects.” *Cellular and Molecular Life Sciences* 63 (14): 1579–85. https://doi.org/10.1007/s00018-006-6130-7.

Beres, B L, H a Cárcamo, and J R Byers. 2007. “Effect of Wheat Stem Sawfly Damage on Yield and Quality of Selected Canadian Spring Wheat.” *Journal of Economic Entomology* 100 (1): 79–87. https://doi.org/10.1603/0022-0493(2007)100[79:EOWSSD]2.0.CO;2.

Beres, BL, LM Dosdall, DK Weaver, HA Cárcamo, and DM Spaner. 2011. “Biology and Integrated Management of Wheat Stem Sawfly and the Need for Continuing Research.” *The Canadian Entomologist* 143 (2): 105–25. https://doi.org/10.4039/n10-056.

Bilak, Amber, Lyle Uyetake, and Tin Tin Su. 2014. “Dying Cells Protect Survivors from Radiation-Induced Cell Death in Drosophila.” *PLoS Genetics* 10 (3). https://doi.org/10.1371/journal.pgen.1004220.

Blake, Victoria C., Clay Birkett, David E. Matthews, David L. Hane, Peter Bradbury, and Jean-Luc Jannink. 2016. “The Triticeae Toolbox: Combining Phenotype and Genotype Data to Advance Small-Grains Breeding.” *The Plant Genome*. https://doi.org/10.3835/plantgenome2014.12.0099.

Boerner, Susan, and Karen M. McGinnis. 2012. “Computational Identification and Functional Predictions of Long Noncoding RNA in Zea Mays.” *PLoS ONE* 7 (8). https://doi.org/10.1371/journal.pone.0043047.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Boubakri, Hatem, Anne Poutaraud, Mohamed Ali Wahab, Celine Clayeux, Raymonde Baltenweck-Guyot, Damien Steyer, Christophe Marcic, Ahmed Mliki, and Isabelle Soustre-Gacougnolle. 2013. “Thiamine Modulates Metabolism of the Phenylpropanoid Pathway Leading to Enhanced Resistance to Plasmopara Viticola in Grapevine.” *BMC Plant Biology* 13 (1): 31. https://doi.org/10.1186/1471-2229-13-31.

Britton, Collette, Alan D. Winter, Victoria Gillan, and Eileen Devaney. 2014. “MicroRNAs of Parasitic Helminths - Identification, Characterization and Potential as Drug Targets.” *International Journal for Parasitology: Drugs and Drug Resistance* 4 (2): 85–94. https://doi.org/10.1016/j.ijpddr.2014.03.001.

Budak, Hikmet, and B Ani Akpinar. 2015. “Plant MiRNAs: Biogenesis, Organization and Origins.” *Functional & Integrative Genomics* 15 (5): 523–31. https://doi.org/10.1007/s10142-015-0451-2.

Budak, Hikmet, Reyyan Bulut, Melda Kantar, and Burcu Alptekin. 2016. “MicroRNA Nomenclature and the Need for a Revised Naming Prescription.” *Briefings in Functional Genomics*. https://doi.org/10.1093/bfgp/elv026.

Budak, Hikmet, Babar Hussain, Zaeema Khan, Neslihan Z Ozturk, and Naimat Ullah. 2015. “From Genetics to Functional Genomics: Improvement in Drought Signaling and Tolerance in Wheat.” *Frontiers in Plant Science* 6 (November): 1012. https://doi.org/10.3389/fpls.2015.01012.

Budak, Hikmet, and Melda Kantar. 2015. “Harnessing NGS and Big Data Optimally: Comparison of MiRNA Prediction from Assembled versus Non-Assembled Sequencing Data--The Case of the Grass Aegilops Tauschii Complex Genome.” *Omics : A Journal of Integrative Biology* 19 (7): 407–15. https://doi.org/10.1089/omi.2015.0038.

Budak, Hikmet, Melda Kantar, Reyyan Bulut, and Bala Ani Akpinar. 2015. “Stress Responsive MiRNAs and IsomiRs in Cereals.” *Plant Science* 235 (FEBRUARY): 1–13. https://doi.org/10.1016/j.plantsci.2015.02.008.

Budak, Hikmet, Zaeema Khan, and Melda Kantar. 2015. “History and Current Status of Wheat MiRNAs Using Next-Generation Sequencing and Their Roles in Development and Stress.” *Briefings in Functional Genomics* 14 (3): 189–98. https://doi.org/10.1093/bfgp/elu021.

Cabili, Moran, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L. Rinn. 2011. “Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses.” *Genes and Development*. https://doi.org/10.1101/gad.17446611.

Cagirici, Halise B., Sezgi Biyiklioglu, and Hikmet Budak. 2017. “Assembly and Annotation of Transcriptome Provided Evidence of MiRNA Mobility between Wheat and Wheat Stem Sawfly.” *Frontiers in Plant Science* 8 (September): 1653. https://doi.org/10.3389/fpls.2017.01653.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Campalans, A. 2004. “Enod40, a Short Open Reading Frame-Containing MRNA, Induces Cytoplasmic Localization of a Nuclear RNA Binding Protein in Medicago Truncatula.” *THE PLANT CELL ONLINE* 16 (4): 1047–59. https://doi.org/10.1105/tpc.019406.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, et al. 2011. “Whole-Genome Sequencing of Multiple Arabidopsis Thaliana Populations.” *Nature Genetics* 43 (10): 956–63. https://doi.org/10.1038/ng.911.

Cárcamo, Héctor A, Brian L Beres, Carolyn E Herle, Hugh McLean, and Sean McGinne. 2011. “Solid-Stemmed Wheat Does Not Affect Overwintering Mortality of the Wheat Stem Sawfly, Cephus Cinctus.” *Journal of Insect Science (Online)* 11: 129. https://doi.org/10.1673/031.011.12901.

Cesana, Marcella, Davide Cacchiarelli, Ivano Legnini, Tiziana Santini, Olga Sthandier, Mauro Chinappi, Anna Tramontano, and Irene Bozzoni. 2011. “A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA.” *Cell* 147 (2): 358–69. https://doi.org/10.1016/j.cell.2011.09.028.

Charon, Celine, Ana Beatriz, Florian Bardou, and Martin Crespi. 2010. “Non-Protein-Coding RNAs and Their Interacting RNA-Binding Proteins in the Plant Cell Nucleus.” *Molecular Plant* 3 (4): 729–39. https://doi.org/10.1093/mp/ssq037.

Chekanova, Julia A. 2015. “Long Non-Coding RNAs and Their Functions in Plants.” *Current Opinion in Plant Biology* 27: 207–16. https://doi.org/10.1016/j.pbi.2015.08.003.

Chen, Jie, Zhikun Liang, Yongkang Liang, Rui Pang, and Wenqing Zhang. 2013. “Conserved MicroRNAs MiR-8-5p and MiR-2a-3p Modulate Chitin Biosynthesis in Response to 20-Hydroxyecdysone Signaling in the Brown Planthopper, Nilaparvata Lugens.” *Insect Biochemistry and Molecular Biology* 43 (9): 839–48. https://doi.org/10.1016/j.ibmb.2013.06.002.

Chen, Jinhui, Mingyang Quan, and Deqiang Zhang. 2015. “Genome-Wide Identification of Novel Long Non-Coding RNAs in Populus Tomentosa Tension Wood, Opposite Wood and Normal Wood Xylem by RNA-Seq.” *Planta* 241 (1): 125–43. https://doi.org/10.1007/s00425-014-2168-1.

Chung, S. H., C. Rosa, E. D. Scully, M. Peiffer, J. F. Tooker, K. Hoover, D. S. Luthe, and G. W. Felton. 2013. “Herbivore Exploits Orally Secreted Bacteria to Suppress

Plant Defenses." *Proceedings of the National Academy of Sciences* 110 (39): 15728–33. https://doi.org/10.1073/pnas.1308867110.

Claverie, Jean-Michel. 2005. "Fewer Genes, More Noncoding RNA." *Science (New York, N.Y.)* 309 (5740): 1529–30. https://doi.org/10.1126/science.1116800.

Conesa, Ana, and Stefan Götz. 2008. "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics." *International Journal of Plant Genomics* 2008: 619832. https://doi.org/10.1155/2008/619832.

Cornish, Adam, and Chittibabu Guda. 2015. "A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference." *BioMed Research International* 2015. https://doi.org/10.1155/2015/456479.

Dai, Xinbin, and Patrick Xuechun Zhao. 2011. "PsRNATarget: A Plant Small RNA Target Analysis Server." *Nucleic Acids Research* 39 (SUPPL. 2). https://doi.org/10.1093/nar/gkr319.

Delaney, K J, D K Weaver, and R K D Peterson. 2010. "Photosynthesis and Yield Reductions From Wheat Stem Sawfly (Hymenoptera: Cephidae): Interactions With Wheat Solidness, Water Stress, and Phosphorus Deficiency." *Journal of Economic Entomology* 103 (2): 516–24. https://doi.org/Doi 10.1603/Ec09229.

Despres, C, C DeLong, S Glaze, E Liu, and P R Fobert. 2000. "The Arabidopsis NPR1/NIM1 Protein Enhances the DNA Binding Activity of a Subgroup of the TGA Family of BZIP Transcription Factors." *THE PLANT CELL* 12 (2): 279–90. https://doi.org/10.1016/S1369-5266(00)80026-6.

Ding, Jihua, Jianqiang Shen, Hailiang Mao, Weibo Xie, Xianghua Li, and Qifa Zhang. 2012. "RNA-Directed DNA Methylation Is Involved in Regulating Photoperiod-Sensitive Male Sterility in Rice." *Molecular Plant* 5 (6): 1210–16. https://doi.org/10.1093/mp/sss095.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts635.

Ecker, Joseph R., Wendy a. Bickmore, Inês Barroso, Jonathan K. Pritchard, Yoav Gilad, and Eran Segal. 2012. "Genomics: ENCODE Explained." *Nature* 489 (7414): 52–55. https://doi.org/10.1038/489052a.

Edwards, Alexis C., Stephanie M. Rollmann, Theodore J. Morgan, and Trudy F C Mackay. 2006. "Quantitative Genomics of Aggressive Behavior in Drosophila

Melanogaster.” *PLoS Genetics* 2 (9): 1386–95. https://doi.org/10.1371/journal.pgen.0020154.

Enright, Anton J, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. 2003. “MicroRNA Targets in Drosophila.” *Genome Biology* 5 (1): R1. https://doi.org/10.1186/gb-2003-5-1-r1.

Erb, Matthias, Stefan Meldau, and Gregg A. Howe. 2012. “Role of Phytohormones in Insect-Specific Plant Reactions.” *Trends in Plant Science* 17 (5): 250–59. https://doi.org/10.1016/j.tplants.2012.01.003.

Ergen, Neslihan Z, and Hikmet Budak. 2009. “Sequencing over 13 000 Expressed Sequence Tags from Six Subtractive CDNA Libraries of Wild and Modern Wheats Following Slow Drought Stress.” *Plant, Cell & Environment* 32 (3): 220–36. https://doi.org/10.1111/j.1365-3040.2008.01915.x.

Fang, Y, and L Xiong. 2015. “General Mechanisms of Drought Response and Their Application in Drought Resistance Improvement in Plants.” *Cellular and Molecular Life Sciences* 72: 673–89. https://doi.org/10.1007/s00018-014-1767-0.

Faust, Gregory G., and Ira M. Hall. 2014. “SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction.” In *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu314.

Food and Agriculture Organization of the United Nations. 2019. “FAO Statistics.” FAOSTAT Statistics Database. 2019. http://www.fao.org/faostat/en/#data/QC.

Franco-Zorrilla, José Manuel, Adrián Valli, Marco Todesco, Isabel Mateos, María Isabel Puga, Ignacio Rubio-Somoza, Antonio Leyva, et al. 2007. “Target Mimicry Provides a New Mechanism for Regulation of MicroRNA Activity.” *Nature Genetics* 39 (8): 1033–37. https://doi.org/10.1038/ng2079.

Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing -- Free Bayes -- Variant Calling -- Longranger.” *ArXiv Preprint ArXiv:1207.3907*. https://doi.org/arXiv:1207.3907 [q-bio.GN].

Grabherr, Manfred G.;, Nir Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., and and Aviv Regev Friedman. 2013. “Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data.” *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.Trinity.

Greenberg, J. K., J. Xia, X. Zhou, S. R. Thatcher, X. Gu, S. A. Ament, T. C. Newman, et al. 2012. “Behavioral Plasticity in Honey Bees Is Associated with Differences in Brain MicroRNA Transcriptome.” *Genes, Brain and Behavior* 11 (6): 660–70. https://doi.org/10.1111/j.1601-183X.2012.00782.x.

Griffiths-Jones, Sam. 2007. “Annotating Noncoding RNA Genes.” *Annual Review of Genomics and Human Genetics* 8: 279–98. https://doi.org/10.1146/annurev.genom.8.080706.092419.

Guleria, Praveen, Monika Mahajan, Jyoti Bhardwaj, and Sudesh Kumar Yadav. 2011. “Plant Small RNAs: Biogenesis, Mode of Action and Their Roles in Abiotic Stresses.” *Genomics, Proteomics and Bioinformatics* 9 (6): 183–99. https://doi.org/10.1016/S1672-0229(11)60022-3.

Guttman, Mitchell, Ido Amit, Manuel Garber, Courtney French, Michael F. Lin, David Feldser, Maite Huarte, et al. 2009. “Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals.” *Nature* 458 (7235): 223–27. https://doi.org/10.1038/nature07672.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis” 8 (X). https://doi.org/10.1038/nprot.2013.084.

Han, Siyu, Yanchun Liang, Ying Li, and Wei Du. 2016. “Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination.” *BioMed Research International*. https://doi.org/10.1155/2016/8496165.

He, Lin, and Gregory J Hannon. 2004. “MicroRNAs: Small RNAs with a Big Role in Gene Regulation.” *Nature Reviews. Genetics* 5 (7): 522–31. https://doi.org/10.1038/nrg1415.

Hennebert, Elise, Barbara Maldonado, Peter Ladurner, Patrick Flammang, and Patrick Flammang. 2015. “Experimental Strategies for the Identification and Characterization of Adhesive Proteins in Animals : A Review.” *Interface Focus* 5: 20140064. https://doi.org/10.1098/rsfs.2014.0064.

Henry, I. M., U. Nagalakshmi, M. C. Lieberman, K. J. Ngo, K. V. Krasileva, H. Vasquez-Gross, A. Akhunova, et al. 2014. “Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing.” *The Plant Cell*. https://doi.org/10.1105/tpc.113.121590.

Heo, J. B., and S. Sung. 2011. “Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA.” *Science* 331 (6013): 76–79. https://doi.org/10.1126/science.1197349.

Heo, Jae Bok, and Sibum Sung. 2011. “Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA.” *Science* 331 (6013): 76–79. https://doi.org/10.1126/science.1197349.

Hergarden, A. C., T. D. Tayler, and D. J. Anderson. 2012. “Allatostatin-A Neurons Inhibit Feeding Behavior in Adult Drosophila.” *Proceedings of the National Academy of Sciences* 109 (10): 3967–72. https://doi.org/10.1073/pnas.1200778109.

Hezroni, Hadas, David Koppstein, Matthew G. Schwartz, Alexandra Avrutin, David P. Bartel, and Igor Ulitsky. 2015. “Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species.” *Cell Reports* 11 (7): 1110–22. https://doi.org/10.1016/j.celrep.2015.04.023.

Hoff, Katharina J., and Mario Stanke. 2013. “WebAUGUSTUS--a Web Service for Training AUGUSTUS and Predicting Genes in Eukaryotes.” *Nucleic Acids Research* 41 (Web Server issue). https://doi.org/10.1093/nar/gkt418.

Hogenhout, Saskia A., and Jorunn I B Bos. 2011. “Effector Proteins That Modulate Plant-Insect Interactions.” *Current Opinion in Plant Biology* 14 (4): 422–28. https://doi.org/10.1016/j.pbi.2011.05.003.

Hussain, Babar, Stuart James Lucas, Levent Ozturk, and Hikmet Budak. 2017. “Mapping QTLs Conferring Salt Tolerance and Micronutrient Concentrations at Seedling Stagein Wheat.” *Scientific Reports*. https://doi.org/10.1038/s41598-017-15726-6.

Initiative, International Brachypodium. 2010. “Genome Sequencing and Analysis of the Model Grass Brachypodium Distachyon.” *Nature* 463 (7282): 763–68. https://doi.org/10.1038/nature08747.

Iqbal, Muhammad Shahzad, Basit Jabbar, Muhammad Nauman Sharif, Qurban Ali, Tayyab Husnain, and Idrees A. Nasir. 2017. “In Silico MCMV Silencing Concludes Potential Host-Derived MiRNAs in Maize.” *Frontiers in Plant Science* 8 (March): 1–9. https://doi.org/10.3389/fpls.2017.00372.

Ito, Eric Augusto, Isaque Katahira, Fábio Fernandes da Rocha Vicente, Luiz Filipe Protasio Pereira, and Fabrício Martins Lopes. 2018. “BASiNET—BiologicAl Sequences NETwork: A Case Study on Coding and Non-Coding RNAs Identification.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky462.

Jensen, Lars Juhl, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. 2008. “EggNOG: Automated Construction and Annotation of Orthologous Groups of Genes.” *Nucleic Acids Research* 36 (SUPPL. 1). https://doi.org/10.1093/nar/gkm796.

Jia, Jinbu, Wenqin Lu, Chengcheng Zhong, Ran Zhou, Junjie Xu, Wei Liu, Xiuhong Gou, et al. 2017. “The 25-26 Nt Small RNAs in Phytophthora Parasitica Are Associated with Efficient Silencing of Homologous Endogenous Genes.” *Frontiers in Microbiology* 8 (MAY): 1–15. https://doi.org/10.3389/fmicb.2017.00773.

Jones, Christopher I, Dominic P Grima, Joseph A Waldron, Sue Jones, Hannah N Parker, and Sarah F Newbury. 2013. “The 5’-3’ Exoribonuclease Pacman (Xrn1) Regulates Expression of the Heat Shock Protein Hsp67Bc and the MicroRNA MiR-277-3p in Drosophila Wing Imaginal Discs.” *RNA Biology* 10 (8): 1345–55. https://doi.org/10.4161/rna.25354.

Kallen, Amanda N., Xiao Bo Zhou, Jie Xu, Chong Qiao, Jing Ma, Lei Yan, Lingeng Lu, et al. 2013. “The Imprinted H19 LncRNA Antagonizes Let-7 MicroRNAs.” *Molecular Cell* 52 (1): 101–12. https://doi.org/10.1016/j.molcel.2013.08.027.

Kang, Yu Jian, De Chang Yang, Lei Kong, Mei Hou, Yu Qi Meng, Liping Wei, and Ge Gao. 2017. “CPC2: A Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx428.

Kantar, Melda, Stuart J. Lucas, and Hikmet Budak. 2011. *Drought Stress. Molecular Genetics and Genomics Approaches*. *Advances in Botanical Research*. 1st ed. Vol. 57. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-387692-8.00013-8.

Karreth, Florian A., Yvonne Tay, Daniele Perna, Ugo Ala, Shen Mynn Tan, Alistair G. Rust, Gina Denicola, et al. 2011. “In Vivo Identification of Tumor- Suppressive PTEN CeRNAs in an Oncogenic BRAF-Induced Mouse Model of Melanoma.” *Cell* 147 (2): 382–95. https://doi.org/10.1016/j.cell.2011.09.032.

Keniry, Andrew, David Oxley, Paul Monnier, Michael Kyba, Luisa Dandolo, Guillaume Smits, and Wolf Reik. 2012. “The H19 LincRNA Is a Developmental Reservoir of MiR-675 That Suppresses Growth and Igf1r.” *Nature Cell Biology* 14 (7): 659–65. https://doi.org/10.1038/ncb2521.

Khajuria, Chitvan, Lawrent L. Buschman, Ming Shun Chen, Subbaratnam Muthukrishnan, and Kun Yan Zhu. 2010. “A Gut-Specific Chitinase Gene Essential for Regulation of Chitin Content of Peritrophic Matrix and Growth of

Ostrinia Nubilalis Larvae." *Insect Biochemistry and Molecular Biology* 40 (8): 621–29. https://doi.org/10.1016/j.ibmb.2010.06.003.

Knodel, J J, P B Beauzay, E D Eriksmoen, and J D Pederson. 2009. "Pest Management of Wheat Stem Maggot (Diptera: Chloropidae) and Wheat Stem Sawfly (Hymenoptera: Cephidae) Using Insecticides in Spring Wheat." *Journal of Agricultural and Urban Entomology* 26 (4): 183–97. https://doi.org/10.3954/1523-5475-26.4.183.

Koboldt, Daniel C., Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. 2009. "VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp373.

Kong, Lei, Yong Zhang, Zhi Qiang Ye, Xiao Qiao Liu, Shu Qi Zhao, Liping Wei, and Ge Gao. 2007. "CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine." *Nucleic Acids Research* 35 (SUPPL.2). https://doi.org/10.1093/nar/gkm391.

Kozomara, Ana, and Sam Griffiths-Jones. 2011. "MiRBase: Integrating MicroRNA Annotation and Deep-Sequencing Data." *Nucleic Acids Research* 39. https://doi.org/10.1093/nar/gkq1027.

———. 2014. "MiRBase: Annotating High Confidence MicroRNAs Using Deep Sequencing Data." *Nucleic Acids Research* 42 (D1). https://doi.org/10.1093/nar/gkt1181.

Krasensky, Julia, and Claudia Jonak. 2012. "Drought, Salt, and Temperature Stress-Induced Metabolic Rearrangements and Regulatory Networks." *Journal of Experimental Botany* 63 (4): 1593–1608. https://doi.org/10.1093/jxb/err460.

Krüger, Jan, and Marc Rehmsmeier. 2006. "RNAhybrid: MicroRNA Target Prediction Easy, Fast and Flexible." *Nucleic Acids Research* 34 (WEB. SERV. ISS.). https://doi.org/10.1093/nar/gkl243.

Kurtoglu, Kuaybe Yucebilgili, Melda Kantar, and Hikmet Budak. 2014. "New Wheat MicroRNA Using Whole-Genome Sequence." *Functional and Integrative Genomics* 14 (2): 363–79. https://doi.org/10.1007/s10142-013-0357-9.

Kuzin, Alexander, Thomas Brody, Adrian W. Moore, and Ward F. Odenwald. 2005. "Nerfin-1 Is Required for Early Axon Guidance Decisions in the Developing Drosophila CNS." *Developmental Biology* 277 (2): 347–65. https://doi.org/10.1016/j.ydbio.2004.09.027.

Lai, Fan, and Ramin Shiekhattar. 2014. “Where Long Noncoding RNAs Meet DNA Methylation.” *Cell Research* 24 (3): 263–64. https://doi.org/10.1038/cr.2014.13.

Langmead, Ben. 2010. “Aligning Short Sequencing Reads with Bowtie.” *Current Protocols in Bioinformatics*, no. SUPP.32. https://doi.org/10.1002/0471250953.bi1107s32.

Lee, Siu F., Ying Chen Eyre-Walker, Rahul V. Rane, Caroline Reuter, Giovanna Vinti, Lea Rako, Linda Partridge, and Ary A. Hoffmann. 2013. “Polymorphism in the Neurofibromin Gene, Nf1, Is Associated with Antagonistic Selection on Wing Size and Development Time in Drosophila Melanogaster.” *Molecular Ecology* 22 (10): 2716–25. https://doi.org/10.1111/mec.12301.

Legeai, F, and T Derrien. 2015. “Identification of Long Non-Coding RNAs in Insects Genomes.” *Current Opinion in Insect Science* 7: 37–44. https://doi.org/10.1016/j.cois.2015.01.003.

Li, Aimin, Junying Zhang, and Zhongyin Zhou. 2014. “PLEK: A Tool for Predicting Long Non-Coding RNAs and Messenger RNAs Based on an Improved k-Mer Scheme.” *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-15-311.

Li, Bo, and Colin N Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12 (1): 323. https://doi.org/10.1186/1471-2105-12-323.

Li, Heng, and Richard Durbin. 2010. “Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp698.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352.

Li, Lin, Steven R Eichten, Rena Shimizu, Katherine Petsch, Cheng-Ting Yeh, Wei Wu, Antony M Chettoor, et al. 2014. “Genome-Wide Discovery and Characterization of Maize Long Non-Coding RNAs.” *Genome Biology* 15 (2): R40. https://doi.org/10.1186/gb-2014-15-2-r40.

Li, Meixia, Shengyun Wen, Xiangqian Guo, Baoyan Bai, Zhefeng Gong, Xiaojun Liu, Yijin Wang, et al. 2012. “The Novel Long Non-Coding RNA CRG Regulates Drosophila Locomotor Behavior.” *Nucleic Acids Research* 40 (22): 11714–27. https://doi.org/10.1093/nar/gks943.

Liang, Xinwen, Lu Zhang, Sathish Kumar Natarajan, and Donald F Becker. 2013. "Proline Mechanisms of Stress Survival." *Antioxidants & Redox Signaling* 19 (9): 998–1011. https://doi.org/10.1089/ars.2012.5074.

Liao, Qi, Changning Liu, Xiongying Yuan, Shuli Kang, Ruoyu Miao, Hui Xiao, Guoguang Zhao, et al. 2011. "Large-Scale Prediction of Long Non-Coding RNA Functions in a Coding-Non-Coding Gene Co-Expression Network." *Nucleic Acids Research* 39 (9): 3864–78. https://doi.org/10.1093/nar/gkq1348.

Liu, Jun, Choonkyun Jung, Jun Xu, Huan Wang, Shulin Deng, Lucia Bernad, Catalina Arenas-Huertero, and Nam-Hai Chua. 2012. "Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis." *The Plant Cell* 24 (11): 4333–45. https://doi.org/10.1105/tpc.112.102855.

Liu, Yanxia, Yanhe Zhou, Jinya Wu, Peiming Zheng, Yiji Li, Xiaoying Zheng, Santhosh Puthiyakunnon, Zhijian Tu, and Xiao-Guang Chen. 2015. "The Expression Profile of Aedes Albopictus MiRNAs Is Altered by Dengue Virus Serotype-2 Infection." *Cell & Bioscience* 5 (1): 16. https://doi.org/10.1186/s13578-015-0009-y.

Livak, Kenneth J., and Thomas D. Schmittgen. 2001. "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method." *Methods* 25 (4): 402–8. https://doi.org/10.1006/meth.2001.1262.

Lowe, Todd M., and Sean R. Eddy. 1996. "TRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence." *Nucleic Acids Research* 25 (5): 955–64. https://doi.org/10.1093/nar/25.5.0955.

Lu, Hua, Debra N Rate, Jong Tae Song, and Jean T Greenberg. 2003. "ACD6, a Novel Ankyrin Protein, Is a Regulator and an Effector of Salicylic Acid Signaling in the Arabidopsis Defense Response." *The Plant Cell Online* 15 (10): 2408–20. https://doi.org/10.1105/tpc.015412.

Lu, Xuke, Xiugui Chen, Min Mu, Junjuan Wang, Xiaoge Wang, Delong Wang, Zujun Yin, et al. 2016. "Genome-Wide Analysis of Long Noncoding Rnas and Their Responses to Drought Stress in Cotton (Gossypium Hirsutum L.)." *PLoS ONE* 11 (6). https://doi.org/10.1371/journal.pone.0156723.

Lucas, Keira, and Alexander S. Raikhel. 2013. "Insect MicroRNAs: Biogenesis, Expression Profiling and Biological Functions." *Insect Biochemistry and Molecular Biology*. https://doi.org/10.1016/j.ibmb.2012.10.009.

Lucas, Stuart J, and Hikmet Budak. 2012. "Sorting the Wheat from the Chaff:

Identifying MiRNAs in Genomic Survey Sequences of Triticum Aestivum Chromosome 1AL." *PloS One* 7 (7): e40859. https://doi.org/10.1371/journal.pone.0040859.

Maccaferri, Marco, Neil S. Harris, Sven O. Twardziok, Raj K. Pasam, Heidrun Gundlach, Manuel Spannagl, Danara Ormanbekova, et al. 2019. "Durum Wheat Genome Highlights Past Domestication Signatures and Future Improvement Targets." *Nature Genetics*. https://doi.org/10.1038/s41588-019-0381-3.

Macedo, Tulio B, Robert K D Peterson, David K Weaver, and Wendell L Morrill. 2005. "Wheat Stem Sawfly , Cephus Cinctus Norton , Impact on Wheat Primary Metabolism : An Ecophysiological Approach Wheat Stem Sawfly , Cephus Cinctus Norton , Impact on Wheat Primary Metabolism : An Ecophysiological Approach." *Environ. Entomol.* 34 (3): 719–26.

Marcussen, Thomas, Simen R Sandve, Lise Heier, Manuel Spannagl, Matthias Pfeifer, International Wheat Genome Sequencing Consortium, Kjetill S Jakobsen, et al. 2014. "Ancient Hybridizations among the Ancestral Genomes of Bread Wheat." *Science (New York, N.Y.)* 345 (6194): 1250092. https://doi.org/10.1126/science.1250092.

Markham, Nicholas R., and Michael Zuker. 2008. "UNAFold." *Bioinformatics*. https://doi.org/10.1007/978-1-60327-429-6_1.

Matzke, Marjori A., and Rebecca A. Mosher. 2014. "RNA-Directed DNA Methylation: An Epigenetic Pathway of Increasing Complexity." *Nature Reviews Genetics* 15 (8): 570–570. https://doi.org/10.1038/nrg3794.

Mayer, Klaus F X, Robbie Waugh, Peter Langridge, Timothy J Close, Roger P Wise, Andreas Graner, Takashi Matsumoto, et al. 2012. "A Physical, Genetic and Functional Sequence Assembly of the Barley Genome." *Nature*, 1–83. https://doi.org/10.1038/nature11543.

Mercer, Tim R, Daniel J Gerhardt, Marcel E Dinger, Joanna Crawford, Cole Trapnell, Jeffrey A Jeddeloh, John S Mattick, and John L Rinn. 2011. "Targeted RNA Sequencing Reveals the Deep Complexity of the Human Transcriptome." *Nature Biotechnology* 30 (1): 99–104. https://doi.org/10.1038/nbt.2024.

Merchuk-Ovnat, Lianne, Vered Barak, Tzion Fahima, Frank Ordon, Gabriel A Lidzbarsky, Tamar Krugman, and Yehoshua Saranga. 2016. "Ancestral QTL Alleles from Wild Emmer Wheat Improve Drought Resistance and Productivity in Modern Wheat Cultivars." *Frontiers in Plant Science* 7: 452.

https://doi.org/10.3389/fpls.2016.00452.

Militti, Cristina, Sylvain Maenner, Peter B Becker, and Fátima Gebauer. 2014. "UNR Facilitates the Interaction of MLE with the LncRNA RoX2 during Drosophila Dosage Compensation." *Nature Communications* 5: 4762. https://doi.org/10.1038/ncomms5762.

Muthusamy, M., S. Uma, S. Backiyarani, and M. S. Saraswathi. 2015. "Genome-Wide Screening for Novel, Drought Stress-Responsive Long Non-Coding RNAs in Drought-Stressed Leaf Transcriptome of Drought-Tolerant and -Susceptible Banana (Musa Spp) Cultivars Using Illumina High-Throughput Sequencing." *Plant Biotechnology Reports* 9 (5): 279–86. https://doi.org/10.1007/s11816-015-0363-6.

Nilsen, Kirby T., John M. Clarke, Brian L. Beres, and Curtis J. Pozniak. 2016. "Sowing Density and Cultivar Effects on Pith Expression in Solid-Stemmed Durum Wheat." *Agronomy Journal* 108 (1): 219–28. https://doi.org/10.2134/agronj2015.0298.

Nussbaumer, Thomas, Mihaela M. Martis, Stephan K. Roessner, Matthias Pfeifer, Kai C. Bader, Sapna Sharma, Heidrun Gundlach, and Manuel Spannagl. 2013. "MIPS PlantsDB: A Database Framework for Comparative Plant Genome Research." *Nucleic Acids Research* 41 (D1). https://doi.org/10.1093/nar/gks1153.

Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. 1999. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*. https://doi.org/10.1093/nar/27.1.29.

Pang, Ken C., Martin C. Frith, and John S. Mattick. 2006. "Rapid Evolution of Noncoding RNAs: Lack of Conservation Does Not Mean Lack of Function." *Trends in Genetics*. https://doi.org/10.1016/j.tig.2005.10.003.

Paterson, Andrew H, John E Bowers, Rémy Bruggmann, Inna Dubchak, Jane Grimwood, Heidrun Gundlach, Georg Haberer, et al. 2009. "The Sorghum Bicolor Genome and the Diversification of Grasses." *Nature* 457 (7229): 551–56. https://doi.org/10.1038/nature07723.

Pelosi, Paolo, Immacolata Iovinella, Jiao Zhu, Guirong Wang, and Francesca R. Dani. 2017. "Beyond Chemoreception: Diverse Tasks of Soluble Olfactory Proteins in Insects." *Biological Reviews*. https://doi.org/10.1111/brv.12339.

Pennisi, E. 2012. "ENCODE Project Writes Eulogy for Junk DNA." *Science*. https://doi.org/10.1126/science.337.6099.1159.

Pertea, Mihaela, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, and Steven L Salzberg.

2016. “Transcript-Level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown.” *Nature Protocols* 11 (9): 1650–67. https://doi.org/10.1038/nprot.2016-095.

Python, François, and Reinhard F. Stocker. 2002. “Adult-like Complexity of the Larval Antennal Lobe of D. Melanogaster despite Markedly Low Numbers of Odorant Receptor Neurons.” *Journal of Comparative Neurology* 445 (4): 374–87. https://doi.org/10.1002/cne.10188.

Qi, Xin, Shaojun Xie, Yuwei Liu, Fei Yi, and Jingjuan Yu. 2013. “Genome-Wide Annotation of Genes and Noncoding RNAs of Foxtail Millet in Response to Simulated Drought Stress by Deep Sequencing.” *Plant Molecular Biology* 83 (4–5): 459–73. https://doi.org/10.1007/s11103-013-0104-6.

Quinn, Jeffrey J., and Howard Y. Chang. 2015. “Unique Features of Long Non-Coding RNA Biogenesis and Function.” *Nature Reviews. Genetics* 17 (1): 47–62. https://doi.org/10.1038/nrg.2015.10.

Rice, Peter, Ian Longden, and Alan Bleasby. 2000. “EMBOSS: The European Molecular Biology Open Software Suite.” *Trends in Genetics* 16 (1): 276–77. https://doi.org/10.1016/j.cocis.2008.07.002.

Riddiford, Lynn M. 2012. “How Does Juvenile Hormone Control Insect Metamorphosis and Reproduction?” *General and Comparative Endocrinology* 179 (3): 477–84. https://doi.org/10.1016/j.ygcen.2012.06.001.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics (Oxford, England)* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Rohrig, H., J. Schmidt, E. Miklashevichs, J. Schell, and M. John. 2002. “Soybean ENOD40 Encodes Two Peptides That Bind to Sucrose Synthase.” *Proceedings of the National Academy of Sciences* 99 (4): 1915–20. https://doi.org/10.1073/pnas.022664799.

Salmena, Leonardo, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. 2011. “A CeRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language?” *Cell* 146 (3): 353–58. https://doi.org/10.1016/j.cell.2011.07.014.

Schmitz, Sandra U., Phillip Grote, and Bernhard G. Herrmann. 2016. “Mechanisms of Long Noncoding RNA Function in Development and Disease.” *Cellular and Molecular Life Sciences*, 2016. https://doi.org/10.1007/s00018-016-2174-5.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Shin, Heungsop, Hwa Soo Shin, Rujin Chen, and Maria J. Harrison. 2006. "Loss of At4 Function Impacts Phosphate Distribution between the Roots and the Shoots during Phosphate Starvation." *Plant Journal* 45 (5): 712–26. https://doi.org/10.1111/j.1365-313X.2005.02629.x.

Shuai, Peng, Dan Liang, Sha Tang, Zhoujia Zhang, Chu Yu Ye, Yanyan Su, Xinli Xia, and Weilun Yin. 2014. "Genome-Wide Identification and Functional Prediction of Novel and Drought-Responsive LincRNAs in Populus Trichocarpa." *Journal of Experimental Botany* 65 (17): 4975–83. https://doi.org/10.1093/jxb/eru256.

Simopoulos, Caitlin M.A., Elizabeth A. Weretilnyk, and G. Brian Golding. 2018. "Prediction of Plant LncRNA by Ensemble Machine Learning Classifiers." *BMC Genomics*. https://doi.org/10.1186/s12864-018-4665-2.

Singh, Urminder, Niraj Khemka, Mohan Singh Rajkumar, Rohini Garg, and Mukesh Jain. 2017. "PLncPRO for Prediction of Long Non-Coding RNAs (LncRNAs) in Plants and Its Application for Discovery of Abiotic Stress-Responsive LncRNAs in Rice and Chickpea." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx866.

Soshnev, A A, H Ishimoto, B F McAllister, X G Li, M D Wehling, T Kitamoto, and P K Geyer. 2011. "A Conserved Long Noncoding RNA Affects Sleep Behavior in Drosophila." *Genetics* 189 (2): 455-U497. https://doi.org/10.1534/genetics.111.131706.

Struhl, Kevin. 2007. "Transcriptional Noise and the Fidelity of Initiation by RNA Polymerase II." *Nature Structural & Molecular Biology* 14 (2): 103–5. https://doi.org/10.1038/nsmb0207-103.

Su, Chun, Xiaozeng Yang, Shiqing Gao, Yimiao Tang, Changping Zhao, and Lei Li. 2014. "Identification and Characterization of a Subset of MicroRNAs in Wheat (Triticum Aestivum L.)." *Genomics* 103 (4): 298–307. https://doi.org/10.1016/j.ygeno.2014.03.002.

Sun, Liang, Haitao Luo, Dechao Bu, Guoguang Zhao, Kuntao Yu, Changhai Zhang, Yuanning Liu, Runsheng Chen, and Yi Zhao. 2013. "Utilizing Sequence Intrinsic Composition to Classify Protein-Coding and Long Non-Coding Transcripts."

Nucleic Acids Research 41 (17). https://doi.org/10.1093/nar/gkt646.

Swiezewski, Szymon, Fuquan Liu, Andreas Magusin, and Caroline Dean. 2009. "Cold-Induced Silencing by Long Antisense Transcripts of an Arabidopsis Polycomb Target." *Nature* 462 (7274): 799–802. https://doi.org/10.1038/nature08618.

Szymański, Maciej, and Jan Barciszewski. 2002. "Beyond the Proteome: Non-Coding Regulatory RNAs." *Genome Biology* 3 (5): reviews0005. https://doi.org/10.1186/gb-2002-3-5-reviews0005.

Tanaka, Tsuyoshi, Baltazar A Antonio, Shoshi Kikuchi, Takashi Matsumoto, Yoshiaki Nagamura, Hisataka Numa, Hiroaki Sakai, et al. 2008. "The Rice Annotation Project Database (RAP-DB): 2008 Update." *Nucleic Acids Research* 36 (Database issue): D1028-33. https://doi.org/10.1093/nar/gkm978.

Tang, Wei, Yi Zheng, Jing Dong, Jia Yu, Junyang Yue, Fangfang Liu, Xiuhong Guo, et al. 2016. "Comprehensive Transcriptome Profiling Reveals Long Noncoding RNA Expression and Alternative Splicing Regulation during Fruit Development and Ripening in Kiwifruit (Actinidia Chinensis)." *Frontiers in Plant Science* 7 (March): 1–15. https://doi.org/10.3389/fpls.2016.00335.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics* 4 (Supplement 25): 1–14. https://doi.org/10.1002/0471250953.bi0410s25.

Tariq, K., W. Peng, G. Saccone, and H. Zhang. 2015. "Identification, Characterization and Target Gene Analysis of Testicular MicroRNAs in the Oriental Fruit Fly *Bactrocera Dorsalis*." *Insect Molecular Biology* 00 (1): n/a-n/a. https://doi.org/10.1111/imb.12196.

Tenea, Gabriela N, Adrian Peres Bota, Fernando Cordeiro Raposo, and Alain Maquet. 2011. "Reference Genes for Gene Expression Studies in Wheat Flag Leaves Grown under Different Farming Conditions." *BMC Research Notes* 4 (September): 373. https://doi.org/10.1186/1756-0500-4-373.

The International Wheat Genome Sequencing Consortium, (IWGSC). 2014. "A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (Triticum Aestivum) Genome." *Science (New York, N.Y.)* 345 (6194): 1251788. https://doi.org/10.1126/science.1251788.

Tian, Bin, Jiarui Li, Thomas R. Oakley, Timothy C. Todd, and Harold N. Trick. 2016. "Host-Derived Artificial MicroRNA as an Alternative Method to Improve Soybean Resistance to Soybean Cyst Nematode." *Genes* 7 (12).

https://doi.org/10.3390/genes7120122.

Tripathi, Rashmi, Sunil Patel, Vandana Kumari, Pavan Chakraborty, and Pritish Kumar Varadwaj. 2016. “DeepLNC, a Long Non-Coding RNA Prediction Tool Using Deep Neural Network.” *Network Modeling Analysis in Health Informatics and Bioinformatics*. https://doi.org/10.1007/s13721-016-0129-2.

Ulitsky, Igor, and David P. Bartel. 2013. “XLincRNAs: Genomics, Evolution, and Mechanisms.” *Cell*. https://doi.org/10.1016/j.cell.2013.06.020.

Ulitsky, Igor, Alena Shkumatava, Calvin H. Jan, Hazel Sive, and David P. Bartel. 2011. “Conserved Function of LincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution.” *Cell*. https://doi.org/10.1016/j.cell.2011.11.055.

Wang, KC, and HY Chang. 2011. “Molecular Mechanisms of Long Noncoding RNAs.” *Molecular Cell* 43.

Wang, Liguo, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean Pierre Kocher, and Wei Li. 2013. “CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt006.

Wang, Tian-Zuo, Min Liu, Min-Gui Zhao, Rujin Chen, and Wen-Hao Zhang. 2015. “Identification and Characterization of Long Non-Coding RNAs Involved in Osmotic and Salt Stress in Medicago Truncatula Using Genome-Wide High-Throughput Sequencing.” *BMC Plant Biology* 15 (1): 131. https://doi.org/10.1186/s12870-015-0530-5.

Wang, Y., X. Fan, F. Lin, G. He, W. Terzaghi, D. Zhu, and X. W. Deng. 2014. “Arabidopsis Noncoding RNA Mediates Control of Photomorphogenesis by Red Light.” *Proceedings of the National Academy of Sciences* 111 (28): 10359–64. https://doi.org/10.1073/pnas.1409457111.

Weiberg, Arne, Ming Wang, Feng-Mao Lin, Hongwei Zhao, Zhihong Zhang, Isgouhi Kaloshian, Hsien-Da Huang, et al. 2013. “Fungal Small RNAs Suppress Plant Immunity by Hijacking Host RNA Interference Pathways.” *Science (New York, N.Y.)* 342 (6154): 118–23. https://doi.org/10.1126/science.1239705.

Winfield, Mark O., Paul A. Wilkinson, Alexandra M. Allen, Gary L.A. Barker, Jane A. Coghill, Amanda Burridge, Anthony Hall, et al. 2012. “Targeted Re-Sequencing of the Allohexaploid Wheat Exome.” *Plant Biotechnology Journal*. https://doi.org/10.1111/j.1467-7652.2012.00713.x.

Wu, Thomas D., and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of

Complex Variants and Splicing in Short Reads." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq057.

Wu, Zhuomin, Xiaoxia Liu, Li Liu, Houliang Deng, Jingjing Zhang, Qian Xu, Bohong Cen, and Aimin Ji. 2014. "Regulation of LncRNA Expression." *Cellular and Molecular Biology Letters* 19 (4). https://doi.org/10.2478/s11658-014-0212-6.

Xiao, Huamei, Zhuting Yuan, Dianhao Guo, Bofeng Hou, Chuanlin Yin, Wenqing Zhang, and Fei Li. 2015. "Genome-Wide Identification of Long Noncoding RNA Genes and Their Potential Association with Fecundity and Virulence in Rice Brown Planthopper, Nilaparvata Lugens." *BMC Genomics* 16 (1): 749. https://doi.org/10.1186/s12864-015-1953-y.

Yan, Biao, and Zhenhua Wang. 2012. "Long Noncoding RNA: Its Physiological and Pathological Roles." *DNA and Cell Biology* 31 (S1): S-34-S-41. https://doi.org/10.1089/dna.2011.1544.

Yao, Yingyin, Ganggang Guo, Zhongfu Ni, Ramanjulu Sunkar, Jinkun Du, Jian-Kang Zhu, and Qixin Sun. 2007. "Cloning and Characterization of MicroRNAs from Wheat (Triticum Aestivum L.)." *Genome Biology* 8 (6): R96. https://doi.org/10.1186/gb-2007-8-6-r96.

Yoon, Je-Hyun, Kotb Abdelmohsen, and Myriam Gorospe. 2014. "Functional Interactions among MicroRNAs and Long Noncoding RNAs." *Seminars in Cell & Developmental Biology* 0: 9–14. https://doi.org/10.1016/j.semcdb.2014.05.015.

Yoon, Je Hyun, Kotb Abdelmohsen, Subramanya Srikantan, Xiaoling Yang, Jennifer L. Martindale, Supriyo De, Maite Huarte, Ming Zhan, Kevin G. Becker, and Myriam Gorospe. 2012. "LincRNA-P21 Suppresses Target MRNA Translation." *Molecular Cell* 47 (4): 648–55. https://doi.org/10.1016/j.molcel.2012.06.027.

Yue, Jieyu, Hong Sun, Wei Zhang, Dan Pei, Yang He, and Huazhong Wang. 2015. "Wheat Homologs of Yeast ATG6 Function in Autophagy and Are Implicated in Powdery Mildew Immunity." *BMC Plant Biology* 15 (1): 1–15. https://doi.org/10.1186/s12870-015-0472-y.

Zanke, Christine, Jie Ling, JÃ¶rg Plieske, Sonja Kollers, Erhard Ebmeyer, Viktor Korzun, Odile Argillier, et al. 2014. "Genetic Architecture of Main Effect QTL for Heading Date in European Winter Wheat." *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2014.00217.

Zhang, Lin, Dongxia Hou, Xi Chen, Donghai Li, Lingyun Zhu, Yujing Zhang, Jing Li, et al. 2012. "Exogenous Plant MIR168a Specifically Targets Mammalian

LDLRAP1: Evidence of Cross-Kingdom Regulation by MicroRNA." *Cell Research* 22 (1): 107–26. https://doi.org/10.1038/cr.2011.158.

Zhang, Yanqiong, Yunsheng Wang, Fuliang Xie, Chao Li, Baohong Zhang, Robert L Nichols, and Xiaoping Pan. 2016. "Identification and Characterization of MicroRNAs in the Plant Parasitic Root-Knot Nematode Meloidogyne Incognita Using Deep Sequencing." *Functional & Integrative Genomics*, no. April: 127–42. https://doi.org/10.1007/s10142-015-0472-x.

Zhang, Zemin, and William I. Wood. 2003. "A Profile Hidden Markov Model for Signal Peptides Generated by HMMER." *Bioinformatics* 19 (2): 307–8. https://doi.org/10.1093/bioinformatics/19.2.307.

Zhao, Yi, Hui Li, Shuangsang Fang, Yue Kang, Wei Wu, Yajing Hao, Ziyang Li, et al. 2016. "NONCODE 2016: An Informative and Valuable Data Source of Long Non-Coding RNAs." *Nucleic Acids Research* 44 (D1): D203–8. https://doi.org/10.1093/nar/gkv1252.

Zhu, Yafeng, Lukas M. Orre, Henrik J. Johansson, Mikael Huss, Jorrit Boekel, Mattias Vesterlund, Alejandro Fernandez-Woodbridge, Rui M.M. Branca, and Janne Lehtiö. 2018. "Discovery of Coding Regions in the Human Genome by Integrated Proteogenomics Analysis Workflow." *Nature Communications*. https://doi.org/10.1038/s41467-018-03311-y.

Zimowska, Grazyna J., Xavier Nirmala, and Alfred M. Handler. 2009. "The B2-Tubulin Gene from Three Tephritid Fruit Fly Species and Use of Its Promoter for Sperm Marking." *Insect Biochemistry and Molecular Biology* 39 (8): 508–15. https://doi.org/10.1016/j.ibmb.2009.05.004.

8. APPENDIX A

Supplementary Information

Supplementary Figure 1. The distribution of percent change of the number of transcripts over a set of FPKM cutoffs. Plot of percent change over 0 to 2 FPKM cutoffs **(A)**. Closer look at the graph between the cutoffs of 0 to 1 **(B)**.

Supplementary Figure 2. Length distribution of lncRNAs and coding transcripts from each T. turgidum variety. (Graph legend: Kiz: Kiziltan, TR: TR39477, TTD: TTD-22; CK: control conditions, DS: drought stressed)

Supplementary Figure 3. Association between GC% content and length of lncRNAs from each T. turgidum variety.

Supplementary Table 1. QRT-PCR primers for common DE transcripts from Kiziltan.

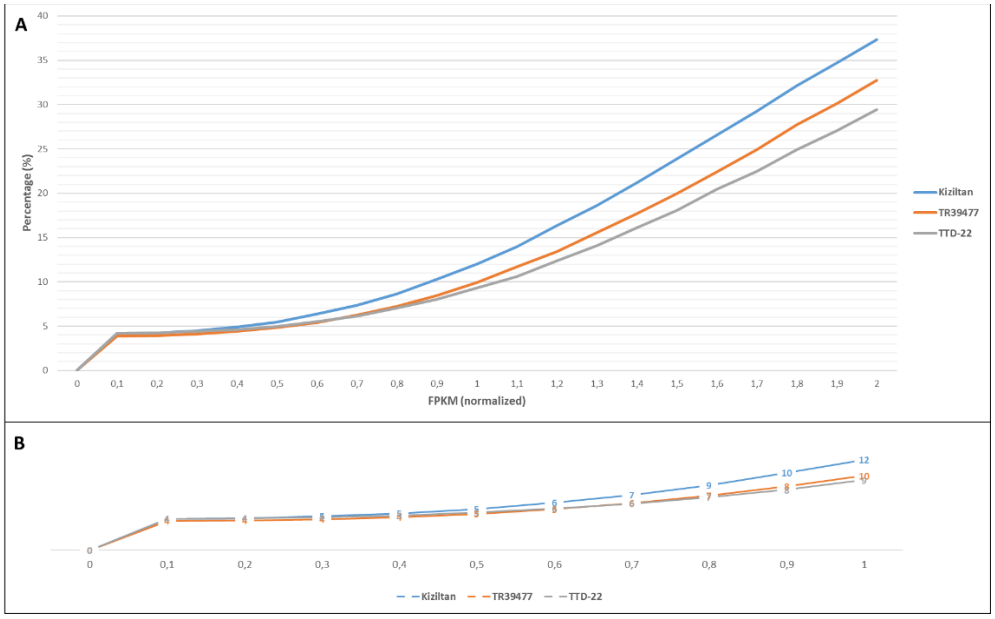
Supplementary Table 2. TR39477 and TTD-22 specific transcripts expressed under drought stress. **(A)** and **(B)** Drought specific transcripts of TR39477 and TTD-22 were blasted and the transcripts which do not exhibit any similarity to each other were listed in below. **(C)** KEGG maps for TR39477 specific transcripts which are expressed in response to drought stress are listed.

Supplementary Table 3. miRNAs which targets the lncRNAs from variety Kiziltan **(A)** TR39477**(B)** and TTD-22**(C)**.

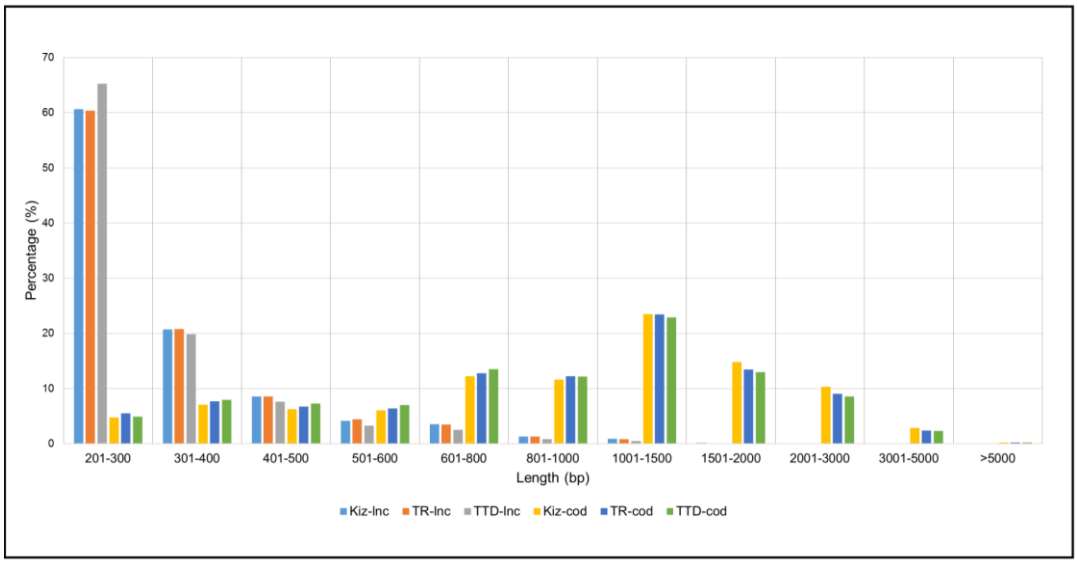
Supplementary Table 4. Number of transcripts across the three T. turgidum varieties. Number of transcripts listed either without filtering or with a filtering of >0.5 FPKM.

Supplementary Figures

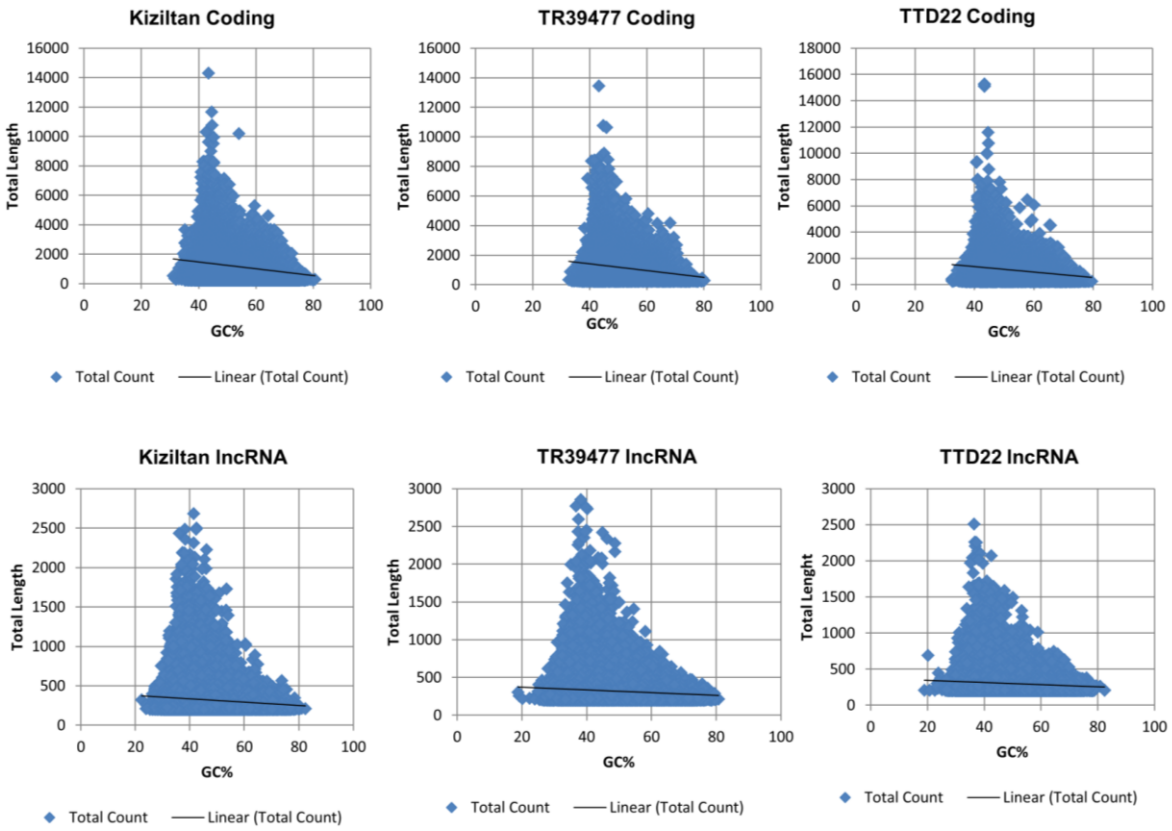
Supplementary Figure 1. The distribution of percent change of the number of transcripts over a set of FPKM cutoffs. Plot of percent change over 0 to 2 FPKM cutoffs (A). Closer look at the graph between the cutoffs of 0 to 1 (B).



Supplementary Figure 2. Length distribution of lncRNAs and coding transcripts from each T. turgidum variety. (Graph legend: Kiz: Kiziltan, TR: TR39477, TTD: TTD-22; CK: control conditions, DS: drought stressed)



Supplementary Figure 3. Association between GC% content and length of lncRNAs from each T. turgidum variety.



Supplementary Tables

Supplementary Table 1. QRT-PCR primers for common DE transcripts from Kiziltan.

lncRNA ID	lncRNA Length	Sequence	Primer Start Site in Target	Primer End Site in Target	Len	GC %	Amplicon Length
lncRNA_c118446_g1_i1-Forward	341	TCTCTGGCCTAAGCAACTTTAC	100	122	22	45.5	127
lncRNA_c118446_g1_i1-Reverse	341	GCTTTCCCAAAGCCCTGATA	207	227	20	50	127
lncRNA_c47700_g1_i1-Forward	472	GGAACAGCGACAGTACAGTAAG	188	210	22	50	139
lncRNA_c47700_g1_i1-Reverse	472	TGTGTGACTGTGAGAGAGAGATA	304	327	23	43.5	139
mRNA_c17408_g1_i1-Forward	879	CTCAGACCTTCGATCAAAGACG	110	132	22	50	97
mRNA_c17408_g1_i1-Reverse	879	TCCATGTACGTCCACCTAGAG	186	207	21	52.4	97
mRNA_c55246_g1_i1-Forward	1734	CGACGTGTAAGCATCAGAGAA	154	175	21	47.6	105
mRNA_c55246_g1_i1-Reverse	1734	AGCCTATGCACTTCCCTAAATC	237	259	22	45.5	105

Supplementary Table 2. TR39477 and TTD-22 specific transcripts expressed under drought stress (A and B). KEGG map for TR39477 specific transcripts expressed under drought (C).

(A) Transcript IDs for drought specific coding transcripts of TR39477 which do not exhibit any similarity to TTD-22 transcripts
TR_DS_c100335_g1_i1,TR_DS_c100721_g1_i1,TR_DS_c10663_g1_i1,TR_DS_c111097_g1_i1,TR_DS_c112175_g1_i1,TR_DS_c115342_g1_i1,TR_DS_c118586_g1_i1,TR_DS_c119351_g1_i1,TR_DS_c121962_g1_i1,TR_DS_c122655_g1_i1,TR_DS_c124109_g1_i1,TR_DS_c133829_g1_i1,TR_DS_c139001_g1_i1,TR_DS_c13972_g1_i1,TR_DS_c18086_g1_i1,TR_DS_c23531_g1_i1,TR_DS_c32200_g1_i2,TR_DS_c5817_g1_i1,TR_DS_c59003_g1_i2,TR_DS_c68118_g1_i1,TR_DS_c7239_g1_i1,TR_DS_c76507_g1_i1,TR_DS_c77362_g1_i1,TR_DS_c77721_g1_i1,TR_DS_c79749_g1_i1,TR_DS_c81581_g1_i1,TR_DS_c83024_g1_i1,TR_DS_c83819_g1_i1,TR_DS_c84249_g1_i1,TR_DS_c86337_g1_i1,TR_DS_c87167_g1_i1,TR_DS_c87458_g1_i1,TR_DS_c87491_g1_i1,TR_DS_c93476_g1_i1,TR_DS_c94439_g1_i1,TR_DS_c9969_g1_i1
(B) Transcript IDs for drought specific coding transcripts of TTD-22 which do not exhibit any similarity to TR39477 transcripts
TTD_DS_c115797_g1_i1,TTD_DS_c53857_g1_i1,TTD_DS_c54409_g1_i4,TTD_DS_c86980_g1_i1

(C) KEGG maps for TR39744 specific transcripts (expressed only under drought) and their description	
map00190	oxidative phosphorylation
map00480	Glutathione metabolism
map00230	Purine metabolism
map00600	Sphingolipid metabolism
map00240	Pyrimidine metabolism
map00730	Thiamine metabolism
map00604	Glycosphingolipid biosynthesis - ganglio series
map00531	Glycosaminoglycan degradation
map00511	Other glycan degradation
map00052	Galactose metabolism

Supplementary Table 3. miRNAs which targets the lncRNAs from variety Kiziltan **(A)** TR39477**(B)** and TTD-22**(C)**.

(A) miRNAs and their lncRNAs targets from Kiziltan variety					
miRNA ID	**lncRNA ID**	**Target Start**	**Target End**	**Aligned Target Fragment**	**Target Inhibition Mode**
miR1436	Kiz_both_c123855_g1_i1	150	169	CUUCUUCCAUCCCAUGAUUG	Cleavage
miR1136	Kiz_both_c38021_g1_i1	3	26	UAGAUACAUCCAUUUCUGCGAUGA	Cleavage or Translation repression
miR854	Kiz_both_c50149_g1_i1	577	597	CUUCUUCUUCCUCUUCUUCUU	Cleavage
miR1135	Kiz_CK_c50944_g2_i1	496	519	UCCUUUCCAAAUUACUUGUCGUGG	Cleavage
miR1439	Kiz_CK_c50944_g2_i1	488	507	CUACUCCCUCCUUUCCAAAU	Translation repression
miR1133	Kiz_DS_c681200_g9_i2	120	141	UUAGGAACGGAGGGAGUAGGUC	Cleavage
miR1436	Kiz_both_c70772_g2_i1	132	151	ACUCCCUCCGUUCC-UAAAUA	Cleavage
miR1439	Kiz_both_c70772_g2_i1	130	149	CUACUCCCUCCGUUCCUAAA	Cleavage
miR1439	Kiz_DS_c79679_g1_i1	201	221	[A]CUACUCCCUCCGUUCCGAAU	Cleavage
miR1436	Kiz_DS_c90557_g1_i1	361	381	ACUCCCUCCGUUCCUUUAUGU	Cleavage
miR1436	Kiz_both_c96195_g1_i1	262	282	ACUCCCUUUGUUCCAGAAUAA	Cleavage
miR1439	Kiz_both_c96195_g1_i1	259	279	UUUACUCCCUUUGUUCCAGAA	Cleavage
miR437	Kiz_both_c96195_g1_i1	304	324	AACUCAAUCUUGUUUAAGUUU	Translation repression
(B) miRNAs and their lncRNAs targets from TR39477 variety					
miRNA ID	**lncRNA ID**	**Target Start**	**Target End**	**Aligned Target Fragment**	**Target Inhibition Mode**
miR1137	TR_both_c118135_g1_i1	189	209	AGUGUCUCAAAUUUUGUACUA	Translation repression

miR1436	TR_both_c11835_g1_i1	169	189	ACUACCUCCGUCCUAAAAUAA	Cleavage
miR1439	TR_both_c11835_g1_i1	140	160	GAUACUCCCUCCGUCUUAAAA	Cleavage
miR1439	TR_both_c11835_g1_i1	167	186	UUACUACCUCCGUCCUAAAA	Cleavage
miR1436	TR_CK_c12385_g1_i1	275	295	ACUCCCUUUGUCUUAAAAUGA	Cleavage
miR1120	TR_DS_c34320_g1_i1	121	143	UCCGUUUCAUAAUAUAACAGCGU	Cleavage
miR1436	TR_DS_c34320_g1_i1	115	135	[A]CUUCAUCCGUUUCAUAAUAU	Cleavage
miR1439	TR_DS_c34320_g1_i1	113	132	GUACUUCAUCCGUUUCAUAA	Cleavage
miR1128	TR_DS_c44260_g1_i1	328	348	UUAGGGACGGAGGGAGUAGUU	Cleavage
miR1128	TR_both_c56781_g1_i5	1266	1285	UUUAUAUGGAGGGAGUAUUU	Cleavage
miR1133	TR_both_c56781_g1_i1	1266	1286	UUUAUAUGGAGGGAGUAUUUA	Cleavage
miR1436	TR_DS_c58890_g4_i2	448	467	CUCCCUCCCUCUUUUAAUAU	Cleavage
miR1139	TR_both_c60784_g1_i5	461	480	GUUACUAG-CUAAGUUACUCC	Cleavage
miR1128	TR_CK_c61658_g1_i1	4	24	UUCGGAACGGAGGGAGUAGUA	Cleavage
miR1436	TR_both_c63034_g2_i23	24	44	ACUUCCUCGGUUCCAAAAUUC	Cleavage
miR1439	TR_both_c63034_g2_i23	22	41	CUACUUCCUCGGUUCCAAAA	Translation repression
miR1122	TR_DS_c63034_g2_i3	744	763	UCUAAAUGCGGAUGUAUCUA	Cleavage
miR1436	TR_DS_c63034_g2_i3	709	729	ACUCCCUCCGUCUCAAAAUUC	Cleavage or Translation repression
miR1439	TR_DS_c63034_g2_i3	706	726	[A]GUACUCCCUCCGUCUCAAAA	Cleavage
miR1118	TR_DS_c63271_g2_i1	831	853	UCCCUCCAUUCCAAAAUAUAGCG	Cleavage

miR1436	TR_DS_c63271_g2_i1	829	849	[A]CUCCCUCCAUUCCAAAAUAU	Cleavage
miR1439	TR_DS_c63271_g2_i1	826	846	[C]UUACUCCCUCCAUUCCAAAA	Cleavage or Translation repression
miR1136	TR_DS_c63631_g1_i7	662	681	AACAUUCAUAUGUGUGACAU	Cleavage
miR1436	TR_DS_c65082_g1_i1	244	264	ACUCCCUCCGUUCCUUUAUAU	Cleavage
miR1439	TR_CK_c67474_g1_i1	401	421	ACUGCUCCCUCCGUUUCUAAA	Cleavage
miR1130	TR_both_c94590_g1_i1	210	232	[AU]UCUUAUAUUAUGGGACGGAGG	Cleavage or Translation repression
miR1436	TR_both_c94590_g1_i1	188	207	ACUCCUUCUGUCCC-UAAUGC	Cleavage
(C) miRNAs and their lncRNAs targets from TTD-22 variety					
miRNA ID	**lncRNA ID**	**Target Start**	**Target End**	**Aligned Target Fragment**	**Target Inhibition Mode**
miR1436	TTD_DS_c58970_g2_i3	225	245	[A]CUCCCUCCGUUCCAAAAUAG	Cleavage
miR1439	TTD_DS_c58970_g2_i3	223	242	[G]UACUCCCUCCGUUCCAAAA	Cleavage
miR1128	TTD_DS_c60722_g1_i1	4	24	UUUGGGACGGAGGGAGUACUA	Cleavage
miR1120	TTD_CK_c62154_g1_i2	682	705	CUCCGUCCCAUAAUAUAACAGCGU	Cleavage
miR1436	TTD_CK_c62154_g1_i2	677	697	[A]UUCCCUCCGUCCCAUAAUAU	Cleavage or Translation repression
miR1436	TTD_DS_c64640_g2_i1	2490	2510	ACUCCCUCCGUCCCAAAAUUC	Cleavage
miR1439	TTD_DS_c64640_g2_i1	2487	2507	[A]CUACUCCCUCCGUCCCAAAA	Cleavage

Supplementary Table 4. Number of transcripts across the three T. turgidum varieties. Number of transcripts listed either without filtering or with a filtering of >0.5 FPKM.

	The number of	Kiziltan	TR39477	TTD-22
Without filtering	All transcripts	243670	211709	203230
	Coding transcripts	84288	75996	78456
	lncRNA transcripts	63773	61823	43932
Actively-expressed	All transcripts	230359	201499	193087
	Coding transcripts	81168	73465	75861
	lncRNA transcripts	59110	57944	40858

9. APPENDIX B

Supplementary Information

Supplementary Table 1. Sequence Read Archieve (SRA) run table. Instrument for all data was Illumina HiSeq 2000.

Samples	# of spots	# of bases	Layout	Run	Accession	SRA
pooled whole larvae	35M	7G	paired	SRR3051777	SRX1497658	SRS1219310
pooled adult males	44.9M	9G	paired	SRR3052012	SRX1497656	SRS1219308
pooled adult male antennae	43.1M	8.6G	paired	SRR3052016	SRX1497648	SRS1219305
pooled adult females	38.5M	7.7G	paired	SRR3052013	SRX1497638	SRS1219289
pooled adult female antennae	45.3M	9.1G	paired	SRR3052011	SRX1497613	SRS1219281
single whole larva	21.2M	4.2G	paired	SRR3051636	SRX1497599	SRS1219265
pooled adult whole males	17.8M	3.6G	paired	SRR3048775	SRX1497595	SRS1219261
pooled adult whole females	17.7M	3.5G	paired	SRR3048750	SRX1497594	SRS1219258

Supplementary Figures

Supplementary Figure 1. Correlation between length and GC content in lncRNA and mRNA transcripts. **(A)** GC content distribution of mRNA and lncRNA transcripts. **(B, C)** Association between length and GC content in lncRNA and mRNA transcripts.