SEQUENCING AND ANNOTATION OF DIPLOID OAT GENOMES AND THE INVESTIGATION OF *AVENA*-SPECIFIC NUTRIENTS

by

Rachel Nichole Walstead

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2019

Approved by:

_____
Dr. Jessica Schlueter

_____
Dr. Cory Brouwer

_____
Dr. Robert Reid

_____
Dr. Veronica Vallejo

_____
Dr. Ron Sass

ABSTRACT

RACHEL NICHOLE WALSTEAD. Sequencing and annotation of diploid oat genomes and the investigation of *Avena*-specific nutrients. (Under the direction of DR. JESSICA SCHLUETER)

Cultivated hexaploid oat (*Avena sativa*) has held a significant place within the global crop community for centuries; although its cultivation has decreased over the past century, its nutritional benefits have garnered increased interest for human consumption. This dissertation reports the development of fully annotated, chromosome-scale assemblies for the extant progenitor species of the A_s- and C_p-subgenomes, *Avena atlantica* and *Avena eriantha* respectively. The diploid *Avena* species serve as important genetic resources for improving common oat's adaptive and food quality characteristics.

The *A. atlantica* and *A. eriantha* genome assemblies span 3.69 and 3.78 Gb with an N50 of 513 and 535 Mb, respectively. Annotation of the genomes, using sequenced transcriptomes, identified ~50,000 gene models in each species – including 2,965 resistance gene analogs across both species. Analysis of these assemblies classified much of each genome as repetitive sequence (~83%), including species specific repeats, centromeric-specific and telomeric-specific repeats. LTR retrotransposons make up most of the classified elements. Genome-wide syntenic comparisons with other members of the Pooideae revealed orthologous relationships, while comparisons with genetic maps from common oat clarified subgenome origins for each of the 21 hexaploid linkage groups. The utility of the diploid genomes was demonstrated by identifying putative candidate genes for flowering time (HD3A) and crown rust resistance (Pc91). We also investigate the phylogenetic relationships among other A- and C- genome Avena species.

The genomes reported here are the first chromosome scale assemblies reported for the tribe Poeae, subtribe Aveninae. Our analyses provide important insight into the evolution and complexity of common hexaploid oat, including subgenome origin, homoeologous relationships, and major intra- and intergenomic rearrangements. They also provide the annotation framework needed to accelerate gene discovery and plant breeding.

A pipeline was developed to identify species-specific genes and has been applied to *A. atlantica* and *A. eriantha*. Various BLAST algorithms were used to compare gene sets from the Phytozome database, GenBank's nonredundant database, and the two *Avena* species. A custom Python script was written to parse the output of these analyses. This pipeline has identified 2,511 and 3,043 *A. atlantica*- and *A. eriantha*-specific gene models, respectively, from approximately 50,000 each. A domain search was performed on these gene models as a first step in identifying possible functions for these genes. Domains identified in both gene sets include metallothionein family 15, members of which include genes to phytoextract metals from soil and aid in stress and cold response, and eggshell protein signatures, which are found in glycine-rich cell wall structural proteins.

Finally, the relationship between oat and various human diseases was studied using the P2EP Knowledge Base. This study identified several relationships between oat consumption and human pathways that require further investigation, including the HSD11B1, RANKL, PARK7, mTOR, ARID4B, and KMT5C genes. These genes all appear to be affected by oat consumption, but the details of those relationships remain unknown. Further understanding of these relationships could guide the prevention and treatment of heart conditions, diabetes, dermatitises, and cancer.

ACKNOWLEDGEMENTS

I am grateful to all those who I have had the pleasure of working with over the course of this research. Each member of my dissertation committee – Dr. Jessica Schlueter, Dr. Cory Brouwer, Dr. Robert Reid, Dr. Veronica Vallejo, and Dr. Ron Sass – has provided invaluable insight and assistance. I would also like to thank Dr. Adam Whaley, the postdoc who I worked with closely, and the staff of University Research Computing for their frequent assistance.

I have been supported by the Graduate School at the University of North Carolina at Charlotte, the Department of Bioinformatics and Genomics, the National Science Foundation, and the GAANN Fellowship offered by the U.S. Department of Education. This research would not have been possible without the funding and support they have provided.

Of course, this journey would not have been possible without the support of many, many people in my life. My family and friends have been constantly supportive throughout my education and I would not be here without their encouragement. My parents, Chuck and Roberta, have always been my biggest cheerleaders and I am forever grateful for their loving support. Dr. Ron Clouse has been an incredible listener, friend, and mentor during this time and I am grateful for his seemingly infinite insight. Hank the cat has always been there to lay on my computer or papers in the exact moment I need them the most. Countless others have supported me in so many ways over the last five years and I could not have done it without each and every one of them.

Finally, I would like to acknowledge two amazing educators I have had the privilege of learning from. First, Mrs. Lesa Berlinghoff, my high school biology teacher, for igniting a love of science in my heart. And second, Dr. Allison Johnson, my undergraduate mentor, for introducing me to phages and encouraging me to step out of my comfort zone into what has become my passion. To say I wouldn't be here without them would be an understatement.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xvi

CHAPTER 1 : INTRODUCTION 1

1.1 : Overview 1

1.2 : Objectives 3

1.3 : Expected Significance 4

1.4 : Background 4

1.4.1 : Genetic and Genomic Resources in Oat 4

1.4.2 : Nutritive Properties in Oat 7

1.4.3 : DNA Sequencing Methods 9

1.4.4 : Genome Assembly Methods 12

1.4.5 : Gene Prediction and Annotation in Plants 21

1.5 : Dissertation Organization 24

CHAPTER 2 : ASSEMBLY AND ANNOTATION OF TWO DIPLOAD OAT GENOMES: *AVENA ATLANTICA* AND *AVENA ERIANTHA* 25

2.1 : Introduction 25

2.2 : Materials and Methods 27

2.2.1 : Plant Material and Nucleic Acid Extraction 27

2.2.2 : DNA Sequencing and Read Processing 27

2.2.3 : RNA-Seq and Transcriptome Assembly 28

2.2.4 : Genome Size, Assembly, Polishing, and Scaffolding 30

2.2.5 : Repeat Analysis, Genome Completeness and Annotation 31

2.2.6 : Variant Identification and Tree Creation 32
2.2.7 : Genome Comparison 32
2.2.8 : Data availability 33
2.3 : Results and Discussion 33
2.3.1 : Whole Genome Sequencing and Assembly 33
2.3.2 : Chromosome Arm Merging 38
2.3.3 : Analysis of Repetitive Elements 42
2.3.4 : Transcriptome Assembly and Functional Annotation 44
2.3.5 : Genomic Features 45
2.3.6 : Comparative genomics 48
2.3.7 : Utility of the genome assemblies 55
2.3.8 : SNP discovery and Genetic Diversity 59
2.4 : Conclusions 63
CHAPTER 3 : IDENTIFICATION OF *AVENA*-SPECIFIC GENES 66
3.1 : Introduction 66
3.2 : Materials and Methods 68
3.3 : Results and Discussion 71
3.3.1 : Eggshell Protein Signature 79
3.3.2 : Metallothionein Family 15 81
3.3.3 : Pentaxins 83
3.3.4 : Prokaryotic Lipoprotein Lipid Attachment Site Signatures 84
3.3.5 : DUF1110 86
3.3.6 : Reverse Transcriptase Domains 88
3.4 : Conclusions 89

CHAPTER 4 : NUTRITIONAL PATHWAYS 91

4.1 : Introduction 91

4.2 : Materials and Methods 93

4.3 : Results and Discussion 94

4.3.1 : Oat and Heart Health 94

4.3.2 : Oat and Diabetes 97

4.3.1 : Oat and Dermatitis 99

4.3.2 : Oat and Cancer 102

4.4 : Conclusions 105

CHAPTER 5 : CONCLUSIONS & FUTURE DIRECTIONS 106

REFERENCES 108

APPENDIX A : *AVENA* ACCESSIONS INCLUDED IN RESEQUENCING PANEL 129

APPENDIX B : SUMMARY OF THE REPEAT ELEMENT CONTENT IN THE AMARANTH GENOME ASSEMBLY AS IDENTIFIED BY REPEATMASKER RELATIVE TO THE REPBASE-DERIVED REPEATMASKER LIBRARIES. 131

APPENDIX C : RESISTANCE GENE ANALOGS 135

APPENDIX D : CANDIDATE RESISTANCE GENES 136

APPENDIX E : POLYMORPHISM MARKERS 137

APPENDIX F : LIST OF PHYTOZOME PRERELEASE GENOMES USED IN *AVENA*-SPECIFIC GENES STUDY 138

APPENDIX G : BLAST ANALYSIS PIPELINE 139

APPENDIX H : BLAST PARSER – PYTHON SCRIPT 141

LIST OF TABLES

Table 1.1: Comparison between Canu, FALCON, and Miniasm on a plant genome assembly. Adapted from Koren, et al. 2017 17

Table 2.1: PacBio and Illumina sequencing read statistics for both Avena atlantica and Avena eriantha. 28

Table 2.2: Raw read statistics for RNASeq data for A. atlantica and A. eriantha. All reads were illumina pair-end reads from standard 500 bp insert libraries. 30

Table 2.3: Estimation of genome heterozygosity, repeat contand and genome size using GenomeScope (Vurture et al. 2017) for A. atlantica and A. eriantha at k=21..... 35

Table 2.4: Summary statistics for the Canu (Koren et al. 2017) and Hi-C assemblies for A. atlantica and A. eriantha. 36

Table 2.5: Physical map and Linkage map assignment. Haplotag markers from the consensus map of Latta et al. 2019 where used to assign scaffold assemblies to linkage groups. Two scaffolds mapped to LG 2 and were merged. 40

Table 2.6: Ancestral subgenome groups (A-, C- and D-) designation for each of the 21 consensus linkages groups reported for A. sativa (Bekele et al.). Haplotag markers mapping to (A) A. atlantica and (B) A. eriantha chromosomes, where highest haplotag mapping are colored red and transition to white as the number of haplotags mapping decreases. Subgenome designation for each linkage group are as previously reported by Chaffin et al. and/or Yan et al. 55

Table 2.7: SNPs per chromosome use for maximum likelihood phylogeny produced using SNPhylo (Lee et al. 2014). 60

Table 3.1: Computational Identification of Legume-specific genes 67

Table 3.2: Computational identification of Avena atlantica-specific genes 72

Table 3.3: Computational identification of Avena eriantha-specific genes 73

Table 3.4: Avena sativa genes & accession numbers involved in avenanthramide biosynthesis, which were used as queries for BLASTN analyses. 75

Table 3.5: A summary of domains identified in Avena atlantica-specific genes, showing the number of genes containing at least one of each domain. The number of genes identified as shared with Avena eriantha in parentheses in first column. 77

Table 3.6: A summary of domains identified in Avena eriantha-specific genes, showing the number of genes containing at least one of each domain. The number of genes identified as shared with Avena atlantica in parentheses in first column. 78

Table 4.1: List of queried used in the P2EP KB for identification of diet-disease networks. 94

LIST OF FIGURES

Figure 2.1: Validation of the A. atlantica assembly. The genetic position of mapped markers is plotted as a function of physical distance relative to the A. atlantica genome assembly. The linkage position of six unassigned scaffolds with multiple mapping markers is shown..... 41

Figure 2.2: Genome overview of (A) A. atlantica and (B) A. eriantha. Track 1: Chromosome and sizes; Tracks 2: Annotated gene density; Track 3: Centromeric repeat density; Track 4: Telomeric sub-repeat density; Track 5: C-genome specific repeat (pAm1) density; Track 6: A-genome specific repeat (pAvKB26) density. 44

Figure 2.3: Homoeologous genes were identified between A. atlantica and A. eriantha genomes to detect homoeologous chromosome relationships. Genome synteny was (A) visualized by dotplot analysis, with boxes drawn around syntenic regions, (B) quantified, where the chromosome pairs with the highest amount of syntenic block connections, expressed as a percentage of the total syntenic bases, are colored red and transition to white as the number of connections decreases and (C) correlation between syntenic block sizes between A. atlantica and A. eriantha. 49

Figure 2.4: Rate of synonymous substitutions per synonymous sites (Ks) within duplicated gene pairs from coding sequences predicted from A. atlantica comparisons with A. eriantha, H. vulgare, B. distachyon, O. sativa, and Z. mays. 50

Figure 2.5: Homoeologous genes were identified between A. atlantica and H. vulgare genomes to detect homoeologous chromosome relationships. Genome synteny was (A) visualized by dotplot analysis, with boxes drawn around syntenic regions, (B) quantified, where the chromosome pairs with the highest amount of syntenic block connections, expressed as a percentage of the total syntenic bases, are colored red and transition to white as the number of connections decreases and (C) correlation between syntenic block sizes between A. atlantica and H. vulgare (Hv_IBSC_PGSB_v2; Ensembl Release 36)..... 52

Figure 2.6 Homoeologous genes were identified between A. eriantha and H. vulgare genomes to detect homoeologous chromosome relationships. Genome synteny was (A) visualized by dotplot analysis, with boxes drawn around syntenic regions, (B) quantified, where the chromosome pairs with the highest amount of syntenic block connections, expressed as a percentage of the total syntenic bases, are colored red and transition to white as the number of connections decreases and (C) correlation between syntenic block sizes between A. eriantha and H. vulgare (Hv_IBSC_PGSB_v2; Ensembl Release 36)..... 53

Figure 2.7: Identification of candidate genes putatively underlying heading date in oats. Candidate gene loci were identified using BLAST searches against the A. atlantica genome assembly using makers sequences associated with heading date QTLs located on the homoelogous linkage groups (A) Mrg12 and (B) Mrg02 (Bekele et al.). Markers from both QTLs mapped to the same physical position on chromosome AA1, within an interval containing an FT-like protein (HD3A), suggesting that heading date in modern oat is controlled by two functional homeologs of the flowering time gene. 57

Figure 2.8: Maximum likelihood tree generated from (A) 7,221 SNPs for A-genome diploids rooted to the A. eriantha reference (ER_32) and from (B) 10,894 SNPs for C-genome diploids rooted to the A. atlantica (AT_803) reference. Branch values represent the percentage of 1,000 bootstrap replicates that support the topology. Scale bar represents substitutions per site. Accession names are abbreviated as described in APPENDIX A. 63

Figure 3.1: A flowchart of the methods for identification of Avena-specific genes. 70

Figure 3.2: Venn diagram showing comparison between Avena eriantha and Avena atlantica specific genes. 74

Figure 3.3: Portion of a multiple sequence alignment of select Glycine-Rich eggshell domain proteins and the eggshell domain-containing proteins of the Avena species investigated. From top down: 1) Medicago sativa Cold and drought-regulated protein, 2) Hordeum vulgare (barley) Glycine-rich cell wall structural protein, 3) Phaseolus vulgaris (kidney bean) Glycine-rich cell wall structural protein, 4) Oryza sativa Glycine-rich cell wall structural protein 2, 5) Oryza sativa Putative glycine-rich cell wall structural protein 1, 6) Oryza sativa Putative glycine-rich cell wall structural protein 1, 7) Acanthoscurria gomesiana (tarantula) Acanthoscurrin-1, 8) Arabidopsis thaliana Glycine-rich protein 5, 9) Avena atlantica AT039540, 10) Avena atlantica AT033627, 11) Avena eriantha AE002675, and 12) Avena eriantha AE007819. Eggshell domains enclosed in red boxes. 80

Figure 3.4: A multiple sequence alignment of select metallothionein family proteins and the metallothionein family genes of the Avena species investigated. From top down: 1) Avena eriantha AE022690, 2) Oryza sativa subsp. indica Metallothionein-like protein 1A, 3) Zea mays Metallothionein-like protein 1, 4) Oryza sativa subsp. japonica Metallothionein-like protein 2C, 5) Avena atlantica AT039122, 6) Avena atlantica AT039123, 7) Theobroma cacao Metallothionein 2A, and 8) Arabidopsis thaliana Metallothionein-like protein 2A. 82

Figure 3.5: Portion of a multiple sequence alignment of select pentaxin domain proteins and the pentaxin domain-containing proteins of the Avena species investigated. From top down: 1) Colletotrichum sublineola (causal agent of sorghum anthracnose) Peroxidase, 2) Avena atlantica AT019368, 3) Cricetulus griseus (Chinese hamster) Neuronal pentraxin-2-like protein, 4) Carassius auratus (goldfish)Pentaxin, 5) Citrus sinensis (sweet orange) F-box domain-containing protein, 6) Triticum aestivum (wheat) PPM-type phosphatase domain-containing protein, 7) Sorghum bicolor Uncharacterized protein, 8) Avena eriantha AE010622. Pentaxin domain shown in red box. 84

Figure 3.6: Portion of a multiple sequence alignment of genes containing Prokaryotic Lipoprotein Lipid Attachment Site Signatures and the Prokaryotic Lipoprotein Lipid Attachment Site Signature-containing proteins of the Avena species investigated. From top down: 1) Ustilago maydis (corn smut) Regulator of itaconic acid biosynthesis (domain in orange box), 2) Plasmopara viticola (grapevine downy mildew) Secreted RxLR effector protein 68 (domain in blue box), 3) Avena atlantica AT012718, 4) Mycobacterium ulcerans Lipoprotein LpqB, 5) Avena atlantica AT022457, 6) Oryza sativa Indole-3-pyruvate monooxygenase YUCCA1 (domain in red box), and 7) Avena eriantha AE013941. 85

Figure 3.7: A multiple sequence alignment between several DUF1110 protein sequences. From top to bottom: 1) Brachypodium distachyon uncharacterized protein, 2) Zea mays uncharacterized protein, 3) Sorghum bicolor uncharacterized protein, 4) Triticum aestivum BTR1-A-like protein, 5) Triticum aestivum Histone domain-containing protein, 6) Triticum monococcum subsp. aegilopoides, 7) Hordeum vulgare subsp. vulgare 8) Avena atlantica AT007255, 9) Avena atlantica AT007252, 10) Avena eriantha AE024564, 11) Leersia perrieri uncharacterized protein, 12) Oryza meridionalis uncharacterized protein, 13) Oryza barthii uncharacterized protein, 14) Oryza glumipatula uncharacterized protein, 15) Oryza punctata uncharacterized protein, 16) Dichanthelium oligosanthes uncharacterized protein, 17) Panicum hallii uncharacterized protein..... 87

Figure 3.8: A multiple sequence alignment of Ribonuclease H domain proteins. From top to bottom: 1) Avena atlantica AT041771, 2) Avena atlantica AT011031, 3) Avena eriantha AE037198, 4) Avena atlantica AT036892, 5) Anaeromyxobacter sp. (strain K) Ribonuclease H, 6) Erwinia tasmaniensis Ribonuclease H, 7) Serratia proteamaculans Ribonuclease H. Domain shown in red box. 89

Figure 4.1: Figure depicting the meta path between Avena sativa and cardiac diseases with multiple hypothetical explanations for the relationships. 96

Figure 4.2: Figure depicting the meta path between Avena sativa and diabetes conditions with multiple hypothetical explanations for the relationships. 97

Figure 4.3: Figure depicting the meta path between Avena sativa and atopic dermatitis with multiple hypothetical explanations for the relationships. 100

Figure 4.4: Figure depicting the meta path between Avena sativa and phototoxic dermatitis and contact dermatitis with multiple hypothetical explanations for the relationships. 101

Figure 4.5: Figure depicting the meta path between Avena sativa various cancers with multiple hypothetical explanations for the relationships. 103

LIST OF ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
DNA	Deoxyribonucleic Acid
GBS	Genotyping by Sequencing
LDL	Low-Density Lipoprotein
mRNA	Messenger RNA
NGS	Next Generation Sequencing
P2EP KB	Plant Pathways Elucidation Project Knowledge Base
RNA	Ribonucleic Acid
SBS	Sequencing by Synthesis

CHAPTER 1: INTRODUCTION

1.1: Overview

Oat (*Avena sativa*) is a nutritionally important crop throughout the world. It is ranked 6th in world cereal production (Ahmad, et al., 2014) and is cultivated as food for animals as well as humans (Oliver, et al., 2013). Among the many nutritional benefits of oats are high levels of fiber and protein. Oat naturally contains no gluten, and is therefore a healthy diet alternative for those who cannot consume gluten for various health reasons. Despite all of its benefits for human health, oat is still primarily used as livestock feed (Ahmad, et al., 2014).

Oat contains high levels of various compounds that have positive health benefits. Among these are avenanthramides, saponins, β-glucan, and calcium (Ahmad, et al., 2014; Fardet, 2010; Peterson, 2001). Avenanthramides are polyphenols that are unique to oat and protect LDL cholesterol from oxidation. Because of this, avenanthramides provide antioxidant, anti-inflammatory, and anti-athrogenic properties (Daou and Zhang, 2012). Saponins are glycosides – sugars bound to another functional group using a glycosidic bond. In oat, two classes of saponins are synthesized: avenacosides (sugars bound to steroids), and avenacins (sugars bound to triterpenoid). Both have been shown to lower cholesterol, stimulate the immune system, and have anti-carcinogenic properties (Fardet, 2010).

β-glucans are a major component of soluble fiber. β(1-3)(1-4)-glucan is found in oat, and the concentration of these soluble fibers vary by the variety of oat. The hull of oat has been shown to contain the most β-glucan. β-glucans can prevent DNA damage by

mutagens and therefore have anti-carcinogenic effects. The absorption of nutrients in the gut is slowed by β-glucan, suggesting that they may help consumers feel full longer, which can assist in weight-loss attempts (Daou and Zhang, 2012). β-glucan has also been shown to reduce the risk of heart disease (Andon and Anderson, 2008; Jenkins, et al., 2002). Even with these indicators, there is still much about the oat fiber development pathway to be discovered.

The FDA has stated that the soluble fiber found in oat, when consumed as part of diet low in saturated fats and cholesterol, may reduce the risk of heart disease (A Food Labeling Guide, 2013). Oat has been suggested as a treatment for many ailments, including tobacco withdrawal, depression, and epilepsy (Yarnell and Abascal, 2001; Yarnell and Abascal, 2001). Oat also has many topical applications having been shown to have a soothing effect on skin and is used to treat dry, itchy skin (Grimalt, et al., 2007). Topical oat products have also been shown to have sun-blocking properties (Potter, et al., 1997) and is often found in products to treat eczema, psoriasis, and other skin conditions (Singh, et al., 2013; Sur, et al., 2008).

Despite the importance of oats and oat-based products, genomic resources for oat have been lagging those of many other crops, especially other cereals. There is a great need for the development of genomic resources in oat, including a functional genome assembly. A genome assembly will unlock the use of advanced genomic breeding methodologies enabling faster breeding cycles and increased genetic gain to more rapidly develop improved oat varieties. A better understanding of the interaction between human nutritional

pathways and oat compounds allows a targeted approach to oat breeding of nutritionally advantaged oat varieties.

1.2: Objectives

This project seeks to utilize high-throughput sequence technologies to fully sequence and annotate two diploid oat genomes. Further, we know that there are compounds that are unique to oat that have nutritional benefits to consumers of these grains. This project will work to identify and characterize the potential function of more of these oat-specific compounds. In addition, we seek to identify how oat consumption can affect human health.

To achieve these goals, this dissertation pursues the following objectives:

Objective I – Assembly, Annotation, and Analysis of *Avena atlantica* and *Avena eriantha* diploid genomes:

This objective is to annotate and analyze two diploid *Avena* genomes. Genomic resources for these two diploids will provide insight into the structure of the hexaploid oat genome as well as a framework for understanding relationships among other Avena diploid species.

Objective II – *Avena*-specific Genes:

This objective is to identify the genes from *A. atlantica* and *A. eriantha* that are specific to the genus *Avena*. The identification of these genes will inform research into what makes oat unique from other grasses and identify potential breeding targets.

Objective III – Nutritional Pathways in Oat:

Human disease pathways affected by oat will be investigated using the Plant Pathways Elucidation Project Knowledge Base (P2EP KB).

1.3: Expected Significance

The overall goal of this research is to make genomic information available to scientists across the oat and grass research communities, as well as to better understand the oat genome structure and unique pathways found in oat. Objective I will provide a much needed resource to researchers within the oat community and beyond. Assembled diploid oat genomes aid in the further efforts to fully sequence and assemble the hexaploid oat genome. Objective II will provide better understanding of the unique pathways and resulting compounds found in oat. This research could provide more insight into the benefits of consuming oat. Objective III will identify potentially unknown relationships between oat compounds and human health.

1.4: Background

1.4.1: Genetic and Genomic Resources in Oat

Oat is a member of the Poaceae (grass) family. The Poaceae family has more than 12,000 species in 771 genera, subdivided in to 12 subfamilies, 51 tribes, and 80 subtribes (Soreng, et al., 2015). One of these subfamilies is the pooideae family, which contains oat, wheat, and barley and is the subfamily from which rice diverged. The Triticeae tribe, containing wheat and barley, then diverged from oat, followed by the divergence between wheat and barley.

Avena sativa L., cultivated oat, is an allohexaploid. The *Avena* genus, however, contains 30 recognized species which vary in ploidy levels (diploid, tetraploid, and hexaploid). The diploid oat species fall into two distinct genome groups: A and C. Tetraploid oat have been categorized into AA, AB, AC, CC, and DC groups, though diploid B and D genomes are not known to exist at this time. All known hexaploid oat plants are categorized as having ACD genomes (Yan, et al., 2016).

The C genome of the hexaploid oat is more diverged from the A and D genomes. Evidence supports that DC tetraploid oat likely shares a recent DC (previously designated as AC) ancestor with ACD hexaploid oat. An A genome diploid oat is likely the other ancestor. It is possible that multiple polyploidization events occurred to lead to the extant hexaploid oat currently cultivated. *Avena sativa* has an estimated haploid genome size of 12,567 Mbp (Yan, et al., 2016).

Liu, et al. (2017) used sequence data from oat plastid and three nuclear genes to study the polyploidization of 109 different *Avena* species. The results of this study indicate that it is possible that AB and DC tetraploids had different A-genome diploid ancestors, and that multiple A-genome diploids were involved in the origin of the DC-genome tetraploid. The ancient DC tetraploidization event occurred between the ancient (or diverged) A-genome and C-genome diploids and then a more recent hexaploidization between that DC-genome tetraploid and a more recent A-genome diploid. This study identified that the C-genome *A. ventricosa* is likely the progenitor for the DC tetraploid and ACD hexaploid genomes. Additionally, the study found that the levels of variation

between hexaploid oat species are consistent with recurrent hexaploidization, with multiple, diverse diploid progenitors introducing variation into the hexaploids.

Genomic resources for oat have lagged those of other grain crops. A consensus map resulting from genotype-by-sequencing (GBS) markers across 12 oat mapping populations has been developed providing further insight in chromosome structure; specifically showing that large structural rearrangements exist between different varieties (Chaffin, et al., 2016). As of August 2019, there are 146,953 Nucleotide, 2,112 SRA, 94 Bioproject, and 1 Genome sequencing project *Avena* entries in NCBI. The genome sequencing project is *Avena sativa* cultivar Victoria whole genome shotgun sequencing project and is currently unpublished. An EST study has been performed on a wild oat (*Avena barbata*) at the Joint Genome Institute (Grigoriev, et al., 2012).

Whole genome sequencing efforts of four *Avena* genomes (*Avena sativa*, *A. brevis*, *A. hirtula*, and *A. strigosa*) have identified that >70% of *Avena* genome sequences cluster into ~200 highly related repetitive sequence clusters. 92% of sequence repeats were found in all four of the genomes examined, while other repeats were specific to subgenomes (Liu, et al., 2019).

Esvelt Klos, et al. (2017) performed a genome-wide association study mapping crown rust resistance genes to the *A. sativa* consensus map created by Chaffin, et al. (2016), and those results may be useful in marker assisted breeding projects. Another study identified crown rust resistance markers in *A. strigosa* using SNP genotyping and introgressed them into hexaploid *A. sativa* (Rines, et al., 2018). Bekele, et al. (2018) used

haplotag-enabled GBS to create an updated consensus map and the first haplotype map of oat.

Two hexaploid genome sequencing efforts are ongoing. The first is an NSF funded project (NSF Award #1444575) to sequence and assemble hexaploid oat variety 94197A1-9-2-2-2-5 (GS-7) strictly using Pacific Bioscience long read technologies. This project has generated over 68x coverage and employed Dovetail sequencing to aid in anchoring the assembly. To date, the assembly is nearly 1/3 smaller than the expected size for the genome. Investigations into how collapse occurred during assembly have shown that the error correction stage during PacBio assembly collapsed the raw reads into chimeric corrected reads causing loss of parts of the genome. A second hexaploid oat sequencing effort using the variety Belinda has been sequenced and assembled in collaboration with NRGene, but those data have not been released and no release date has been announced. This genome does not show the same collapse issues seen in the long-read assembly, however, it is very fragmented due to the use of short-read Illumina based technology for assembly.

1.4.2: Nutritive Properties in Oat

Among the nutritive properties of oat are dietary fiber, which is only found in plant foods. β-glucan is the primary soluble fiber found in oat (Butt, et al., 2008). β-glucan has multiple properties making it beneficial for human consumption. It is known to be an immune modulator, meaning it binds to surface receptors and cases activation of macrophages, white blood corpuscles, and lymphocytes (leading to its antitumor and antimicrobial properties). Oat samples higher in β-glucan have been shown to have a lower glycemic index, or a slower increase in blood sugar levels, as well as the ability to lower

the glycemic index of other glucose consumed at the same time (Butt, et al., 2008; Tapola, et al., 2005). Consumption of whole grain oat products was shown in a large-scale, 10-year study to be inversely related to type 2 diabetes risk. Further, a higher intake of refined grains or lower intake of whole grains was also associated with a significantly higher risk of type 2 diabetes (Liu, et al., 2000).

Plasma and LDL cholesterol levels have been shown to be lowered with consumption of oat bran. Since blood cholesterol is highly associated with coronary artery disease, this suggests that regular consumption of oat could decrease the risk of developing heart disease (Ahmad, et al., 2014; Butt, et al., 2008). It has also been observed that elevated blood pressure can be lowered significantly, especially in patients with higher body mass indices, with the regular consumption of oat bran and oatmeal (Maki, et al., 2006). Regulatory agencies around the world have approved health claims associated with the consumption of oat. The U.S. Food and Drug Administration has approved the labeling of oat products containing β-glucan soluble fiber may reduce the risk of heart disease (A Food Labeling Guide, 2013). The European Commission has approved similar claims, as well as claims regarding the consumption of β-glucan and the maintenance of healthy cholesterol levels and the reduction of blood-glucose spikes following meals (EU Register on Nutrition Health Claims, 2018).

Avenanthramides are low molecular weight phenolic compounds found in oat. Oat is the only cereal to contain these compounds. There are around 40 different avenanthramides found in oat grain and leaves. They are reported to have anti-inflammatory, anti-itch, and anti-irritant properties. Though oat has been used as a topical

treatment for many skin conditions for many years, the role avenanthramides play in these anti-itch effects were not identified until recently. Avenanthramides have also been shown to inhibit the growth of cancer cells and may activate apoptosis. Given avenanthramides' antioxidant abilities, they are able to prevent oxidative stress to consumers and therefore may play a role in preventing many human diseases, including atherosclerosis, diabetes, and more. Avenanthramides have been shown to protect DNA in skin cells from UV radiation (Perrelli, et al., 2018)

Saponins are found in many plant species, including oat. They have been shown to have a hypoglycemic effect in animals, as well as the ability to lower serum cholesterol levels. Various saponins have been shown to boost immune systems and stimulate antibody production. In plants, saponins have an anti-fungal effect and therefore may be a part of the plant defense system (Francis, et al., 2007).

1.4.3: DNA Sequencing Methods

Next-Generation Sequencing (NGS) is known by many names. It is sometimes called Massively Parallel Sequencing (Reis-Filho, 2009), as well as High-Throughput Sequencing (Henson, et al., 2012). While older sequencing methods could only sequence 96 sequences at a time, NGS methods can sequence millions of DNA fragments in parallel. Fragments of DNA formed into libraries, using specific adapters on either end of the DNA or RNA fragment. NGS methods require far less DNA than older methods (Mardis, 2008).

Next Generation Sequencing methods vary in their chemistries and amplification approaches. Early NGS methods produced read lengths from around 35bp to 250bp (Mardis, 2008). Newer long-read NGS methods can produce read lengths up to 20,000bp.

The sequencing chemistries and amplification approaches vary between next-generation sequencing technologies, as well as costs (Goodwin, et al., 2016). Newer technologies (sometimes called "third-generation sequencing") sequence individual DNA molecules using sequencing-by-synthesis methods, instead of amplifying prior to sequencing (Henson, et al., 2012; Reis-Filho, 2009).

For the purposes of this project, we have used both Illumina short-read sequencing as well as Pacific Biosciences single-molecule sequencing. Illumina's NGS methods employ sequencing-by-synthesis (SBS) methods. SBS entails synthesizing new DNA from the template strands and using fluorescent nucleotides to determine sequence. HiSeq 2000 is one of Illumina's SBS platforms. One run of the machine can produce 600 Gb of sequence in just 11 days, and the cost is quite low, around $7 per Gb. The quality is good, and the raw error rate is <1%. Maximum read lengths are around 150 bases and the amount of DNA required is low, around 50-1000 ng (Goodwin, et al., 2016; Quail, et al., 2012).

Low cost and DNA requirements make these Illumina SBS methods attractive for sequencing projects. The short read-lengths, however, make it difficult to resolve complex, repetitive sequences. Because these reads do not always span repetitive regions, they often do not bridge to unique regions spanning repetitive ones. This makes it difficult to place the regions properly in the genome, even in cases where genome coverage is high (Henson, et al., 2012).

Pacific Biosciences, or PacBio, has developed a single-molecule real-time (SMRT) sequencing method. SMRT sequencing offers longer reads than second generation sequencing methods, and at faster run speeds. RSII machines can produce average read

lengths of 20Kb with runtimes around 4 hours. The cost per gigabase is around $1,000. The newer Sequel machines can produce average read lengths up to 30Kb with even shorter runtimes. (Goodwin, et al., 2016; PacBio Sequel System) Sequence information is captured during replication of target DNA. A SMRTbell is created by ligating hairpin adapters to either end of a double-stranded DNA segment. This creates a closed loop of DNA for sequencing. One polymerase molecule begins replication at one hairpin adapter of the SMRTbell. As nucleotides are added to the molecule, fluorescent light is emitted, identifying the nucleotide added. This is repeated all the way around the SMRTbell. This process is done in zero-mode waveguide sequencing unit, which is able to detect the light emitted (Rhoads and Au, 2015).

Read lengths are a very important quality of PacBio SMRT sequencing. Depending on the chemistry release used, SMRT sequencing can provide mean read lengths on average of 15 kbp, and maximum read lengths up to 60 kb. Conversely, the Illumina MiSeq boasts a maximum read length of 2 x 300 bp, the longest read length available with Illumina technology (Rhoads and Au, 2015; Sequencing power for every scale). This increased read length is critical for sequencing large, repetitive genomes. The lower accuracy of PacBio sequencing can be compensated for by increasing PacBio coverage or performing hybrid assemblies with higher-accuracy reads. Many experiments have been performed to assess the quality of the PacBio assemblies. Many small genomes can be assembled in single chloroplasts, while gaps in larger genomes are closed or shortened. It is also more cost effective to use PacBio sequencing for closing gaps (especially of lengths longer than

2.5kb) in one sequencing round than several rounds of Sanger sequencing (Rhoads and Au, 2015).

Approaches to assembling genomes once they are sequenced vary depending on the target genome size and structure. Repetitive regions of DNA complicate this process. If repetitive sequences are longer than the reads in that region, it can be difficult to determine where they belong in the assembly, or how many times they may occur across the genome. Different assemblers tackle this and other problems in different ways. Due to these different methods, some assemblers are more fit for certain types of genomes than others (Henson, et al., 2012).

1.4.4: Genome Assembly Methods

While next-generation sequencing has increased the depth and availability of sequence at a greatly reduced cost, short-read methods do not lend themselves to an easy means to resolve repetitive regions of genomes (Henson, et al., 2012). Plants are known to have complicated, highly repetitive genomes, making a complete assembly with short reads nearly impossible. These challenges can be overcome by long-read technologies such as PacBio. Because the read lengths are so long, repetitive regions can sometimes be completely resolved by one read (Rhoads and Au, 2015). While the reads are long, the throughput of long-read technologies, such as the PacBio RSII system, is much lower than other methods, such as Illumina HiSeq 2500. Additionally, the read error rate of PacBio systems is also much higher, requiring at least 30x coverage per haploid genome to achieve 99% accuracy during error-correction (Rhoads and Au, 2015). Many assemblers have been developed for PacBio-only and hybrid sequencing projects.

SOAPdenovo was developed as an assembly algorithm for short-read Illumina data. It was first used in the assembly of human genomes (Li, et al., 2010). As with many other assemblers, it begins with an error-correction step. While rates of errors are low in Illumina data, the most common error is substitution errors, and sequencing reads are often preprocessed to account for this as more accurate reads allow for better assemblies. This preprocessing is often done by comparing all reads of a given position, and when a majority of reads call a certain base, any reads calling any other base can be corrected to match the majority (Yang, et al., 2012). Error-corrected reads are used to create a de Bruijn graph. This graph contains 25-mers as nodes, and read path information as edge connections. This information is used to merge sequences or paths with very minor differences into contigs. These contigs are then realigned to the initial reads and this information was used to scaffold multiple contigs together. Paired-end reads are used to close gaps for the final assembly (Li, et al., 2010).

ABySS is a two-stage assembly algorithm for short-read data. The first stage includes generating all possible *k*-mers from the sequence reads and error removal, as well as the building of contigs. Contigs are built by extending them as far as possible without using paired-end sequence information. The second stage involves attempting to extend contigs utilizing paired-end information. (Simpson, et al., 2009).

For long-read assembly, there are several algorithms available. The MaSuRCA assembler was developed by the University of Maryland Genome Assembly Group. The program aims to assemble genome sequences using super-reads. These super-reads are developed by extending original reads in either direction as long as the extension is unique.

To this end, k-mer count lookup tables are utilized to quickly identify the number of times a give k-mer occurs in the reads. The last k-1 bases of the read are identified, and A, C, G, or T is appended to the end to make a k-mer. If only one of these four k-mers exists in the k-mer lookup table, that base is appended to the super-read. There can be many reads that yield the same super-read. Reads that yield the same super-read will be replaced by the super-read. If two reads differ by any number of bases internally, they will be incorporated into different super-reads. This helps to separate divergent haplotypes or non-identical repeats (Zimin, et al., 2013).

Utilizing the MaSuRCA super-read algorithm, no information is lost from original reads, and the super-reads lead to a reduced dataset. This is because original reads are replaced by the super-reads that they form. Following the creation of super-reads, an overlap-based assembly approach is used. Maximal super-reads, or super-reads that are not identical substrings of another super-read, are assembled using a modified CABOG (Miller, et al., 2008) assembler (Zimin, et al., 2013).

MaSuRCA is often used for hybrid assemblies – or those using both long and short reads. When being used for hybrid assemblies, short reads are transformed into super-reads as described above. Alignments between these super-reads and PacBio reads (or other long reads such as Oxford Nanopore) are created using k-mers. This starts by identifying which k-mers in the PacBio read are found in any super-read. Then, any ordered k-mers found in both the super-read and the PacBio reads are identified. The super-read with the longest common subsequent k-mers is labeled plausible if the score is above a pre-set threshold. Paths are constructed using paths along each PacBio read, where super-reads are nodes and

the k-overlaps are edges. The longest common subsequence score is calculated for each connected component of paths, and the highest score is chosen as a pre-mega-read (Zimin, et al., 2017).

Before the pre-mega-reads can be assembled, they must be tiled. Starting with the longest pre-mega-read, they are tiled, with overlaps being shorter than k bases. Pre-mega-reads are chosen by maximizing the longest common subsequence scores. PacBio reads can be used to join neighboring pre-mega-reads if three conditions are met. First, three PacBio read tilings must have identical gaps. Secondly, the pre-mega-reads that surround the gap must have identical sequence. Finally, the gap lengths, or the length of PacBio sequence between aligned pre-mega-reads, must be nearly identical. Sometimes, gaps are not filled because filling them may cause a low-quality region within the mega-reads. When a gap remains within the mega-reads, a linked pair of "reads" that spans the gap is created. This is done by linking two 500 bp sequences from either side of the gap, creating artificial mates, or linked pairs (Zimin, et al., 2017). Finally, the CABOG (Celera Assembler with the Best Overlap Graph) assembler (Miller, et al., 2008) is used for final assembly (Zimin, et al., 2017).

FALCON is an assembler developed by Pacific Biosciences to assemble genomes sequences using PacBio methods. The first of these steps is error-correction of raw sequences using sequence alignments between long reads obtained from DALIGNER (Myers) and construction of a graph of overlapping reads. This graph will contain "bubbles," which are highly divergent regions between homologous sequences. The FALCON-sense algorithm (or the consensus module) parallelizes the pairwise alignment

step and removes maximum read length limits. This includes aligning supporting reads to seed sequences, generating tags from these alignments, and grouping these tags as an alignment graph. The optimum path is found and considered the consensus path. The FALCON-phasing module of FALCON-Unzip will then review "bubbles" formed during the error-correction step and identify phased SNPs and assign haplotypes when there are enough of these phased SNPs in a given block of sequence. These will be combined and re-assembled into "haplotigs" and integrated with contigs in the assembly (Chin, et al., 2016).

The final step of FALCON-unzip is running the Quiver algorithm, which was developed by PacBio as a consensus caller. It uses a reference alignment to identify reads that correspond to a reference window. A candidate template sequence is made. This template is modified by a series of single base mutations in efforts to increase the likelihood that the reads could come from the given template. This is repeated until convergence (Chin, et al., 2013).

Another assembler, Canu, has recently been developed. Canu includes the same basic three stages of assembly (correct, trim, and assemble), but using a slightly different algorithm. During the correction phase, there are two filtering steps to determine which overlaps should correct each read. A global filter chooses where reads will provide correction evidence, and a local filter accepts or rejects evidence supplied by other reads for correction. The trim phase recalculates overlaps for corrected read. Sequences that are not supported by other reads are removed. The final assembly phase is a has multiple steps. First, poor overlaps and suspicious reads are removed. A best overlap graph is created and

previously removed reads are reincorporated, if possible. A template sequence is constructed for each contig based on the best overlap graph. There is not yet a strong process for resolving assembly bubbles, created by haplotype differences (Koren, et al., 2017).

Canu has been shown to be faster at the assembly process in the testing of both assemblers, as shown in Table 1.1 (Koren, et al., 2017). This is likely due to the local alignment step in Canu, which differs from FALCON, and makes it faster than the one-to-one methods used in the FALCON assembly process. This time savings may be crucial for large genome assembly projects. It is possible to use Canu-corrected reads as input for FALCON, which could be useful if the FALCON-Unzip module is desired.

Table 1.1: Comparison between Canu, FALCON, and Miniasm on a plant genome assembly. Adapted from Koren, et al. 2017

Genome	Assembly & Polish Tools	Max Contig Size (Mbp)	Genome N50 (Mbp)	% of Reference Genome Covered by Assembly	Number of breakpoints	Time to assemble (CPU hours)	% Identity to Reference
Arabidopsis thaliana	Canu + Quiver	15.95	8.31	82.94%	220	925.31	99.0710%
	FALCON + Quiver	15.94	8.17	82.72%	222	1,132.25	99.0710%
	Miniasm + Quiver	11.61	5.07	82.88%	205	976.43	99.0710%

SMARTdenovo is another *de novo* assembler developed for PacBio and Oxford Nanopore data. While other assemblers include an error-correction step, SMARTdenovo does not. Instead, if the user prefers to use error-corrected reads, those reads must be corrected using another tool. SMARTdenovo's assembly is produced utilizing an all-vs-all read alignments and generating a consensus sequences. Quiver can still be utilized for PacBio data to obtain a polished, more accurate, assembly

(https://github.com/ruanjue/smartdenovo). It runs very quickly and uses very little memory (compared to other assemblers). For the yeast genome, it often provides the lowest number of contigs in the final assembly, and has high reference coverage when the read coverage is sufficient (Giordano, et al., 2017).

Another assembler, designed for PacBio data, is wtdbg (https://github.com/ruanjue/wtdbg). This assembler utilizes a fuzzy Bruijn graph approach to assembling long, noisy reads. The wtdbg developers claim that a 10Gbp plant genome with 20X coverage can be assembled in less than one day. Error correction is not performed by wtdbg, and if desired, must be done using another tool (Ruan, 2017). An updated version of this assembler, wtdbg2, was released in early 2019 (Ruan and Li, 2019).

The diploid *Vitis vinifera* cv. Cabernet Sauvignon was sequenced using PacBio to a coverage of 140x and assembled using Canu, FALCON, and FALCON-Unzip. The estimated genome size of *V. vinifera* is approximately 500 Mbp, but all the assembled genomes were larger than this. FALCON-Unzip had the most contiguous genome, which was 591 Mbp in 718 contigs. The FALCON assembly was 633 Mbp in 1,314 contigs, while the Canu assembly was 1,066 Mbp over 14,489 contigs (Chin, et al., 2016).

A new maize genome was assembled using purely PacBio data. The genome was sequenced at 65x and assembled using a comparative approach with FALCON and PBcR-MHAP (Berlin, et al., 2015) to determine the best assembly method. These assemblies were aligned to BioNano maps to determine the best match. For the maize genome, this was PBcR-MHAP, as it had the fewest conflicts with the genome map (Jiao, et al., 2017).

The diploid tomato (*Solanum pennellii*) genome was assembled recently using Oxford Nanopore data. These reads are similar in length to PacBio reads. The genome is estimated to be 1-1.1 Gb. The assembly was attempted using Canu, SMARTdenovo, and Miniasm. BUSCO was used to assess the completeness of these genomes (Simão, et al., 2015). The BUSCO completeness estimations of the genome after these first attempts were all low: the assemblies were estimated to be 26.46, 26.74, and 0.21% complete using Canu, SMARTdenovo, and Miniasm, respectively. Canu-corrected reads were then used as input for SMARTdenovo, and the resulting assembly was 889.92 Mb and had a completeness of 29.1%. Illumina reads were used to polish these assemblies, a total of 5 times, utilizing the Pilon software (Walker, et al., 2014). This improved all the assemblies tremendously, and the Canu-SMARTdenovo assembly was improved to 915.6 Mb and an estimated 96.46 completeness. The assembly from Canu, after polishing, was improved to 961.83 Mb (from 922.94 before polishing) and estimated BUSCO completeness of 96.46, the same as the Canu-SMARTdenovo assembly (Schmidt, et al., 2017).

A 2017 study evaluated several non-hybrid PacBio assemblers across various organisms. While the wtdbg assembler is very fast, with nearly the shortest CPU time across the board, the assemblies produced were consistently ranked low compared to other assemblers, including CANU, SMARTdenovo, and Falcon (Jayakumar and Sakakibara, 2017). Other studies have found N50 values to be subpar and therefore explored other methods of assembly (Schmidt, et al., 2017). A wild peanut assembly using wtdbg was improved by merging with CANU results (Yin, et al., 2018).

A 2019 study by Fu, et al. (2019) evaluated ten hybrid error correction methods for long reads. Hybrid correction methods utilize short-read sequences to "rescue" more error-prone long-read sequences. There are three main methods used for hybrid error correction. The first is alignment-based, which computes a consensus sequence by aligning the long reads to either short reads, or sequences assembled from short reads. ECTools (Lee, et al., 2014), LSC (Au, et al., 2012), Nanocorr (Goodwin, et al., 2015), pacBiotoCA (Koren, et al., 2012), and proovread (Hackl, et al., 2014) are alignment-based methods. The second strategy is graph-based methods, which use short reads to construct a de Bruijn graph, and then find paths to long reads for correction. FMLRC (Wang, et al., 2018), Jabba (Miclotte, et al., 2015), and LoRDEC (Salmela and Rivals, 2014) use graph-based approaches. Finally, there are dual approaches, which use both alignment-based and graph-based methods to correct the long reads. HALC (Bao and Lan, 2017) and CoLoRMap (Haghshenas, et al., 2016) use dual approaches (Fu, et al., 2019).

Several of the correction methods crashed when applied to the larger datasets, particularly when higher coverage Illumina short-read coverage was added. All five alignment-based methods did not have reported results for the larger datasets (*Drosophila melanogaster* and *Arabidopsis thaliana*) as they took longer than 20 days to run on servers with 20 machines of 16 cores and 256GB memory. Ultimately, it was determined that graph-based and alignment-based methods are comparable, but alignment-based methods are more robust when short-read coverage is lower. The coverage levels of long-reads were not found to be a major factor in the performance of the hybrid correction methods. Perhaps unsurprisingly, performance was better on smaller datasets than on the large ones, and

genome complexity and repetitive regions likely play a role in lowered performance on these large datasets. (Fu, et al., 2019).

1.4.5: Gene Prediction and Annotation in Plants

Gene prediction and annotation is a non-trivial essential step to providing a research community with a genomic resource. There are two main methods for predicting genes: transcript evidence based and *ab initio*. Transcript evidence, or sequence similarity method, employ local or global sequence alignments of the target genome to transcripts (such as ESTs, full-length cDNA's or RNA-Seq data), protein sequences, or previously annotated genomes. This method is limited to sequence-based evidence available, and has the potential to miss predicting genes that are simply not expressed in those datasets or available in those databases. *Ab initio* prediction methods are designed to look for sequence signatures within the DNA sequence itself that indicate gene structure to detect possible genes. Sensors such as sequence motifs, start and stop codons, and patterns of codon usage are used to predict gene start and stop locations, as well as the location of exons within that sequence. Combining both these methods produces far superior annotations, and pipelines have been created that use multiple tools to obtain these results. Once predicted genes are identified, functional information, or biological function, can be assigned to them. This is done by performing sequence similarity searches and domain analyses to known genes (Bolger, et al., 2019; Wang, et al., 2004).

SNAP was developed as a new *ab initio* gene predictor. The program, like many other gene prediction algorithms, utilizes weight matrices, weight array matrices, and Markov models. Weight matrices are developed for splice acceptor, splice donor, and

translation start and stop sites. Markov models are developed for coding, intron, and intergenic sequence regions. These matrices and models are tuned for the organism being studied. This makes it more effective than previous algorithms for identifying genes within a genome (Korf, 2004).

MAKER is a genome annotation pipeline developed by the Yandell Lab. The MAKER pipeline includes external programs which, by default, are RepeatMasker (Smit, et al., 1996-2010), BLAST (Altschul, et al., 1997), Exonerate (Slater and Birney, 2005), and SNAP. The pipeline uses RepeatMasker to identify repeats and low complexity DNA sequences. Then BLAST is used to further identify repeats as well as ESTs, mRNAs, and proteins with similarity to the input genome sequence. Exonerate is used due to its ability to align nucleotide and protein sequences, while taking into account the splice sites of mRNAs. Once the collected information is polished, SNAP is trained using the known information. Then SNAP can perform *ab initio* prediction based upon species specific traning (Cantarel, et al., 2008).

MAKER2 builds upon MAKER as an annotation tool for second-generation sequencing projects. The developers added a metric known as Annotation Edit Distance to help improve quality control. The MAKER2 pipeline also has support for mRNA-Seq and allows for the pass-through of gene models. This give re-annotation capabilities to the pipeline, while it is still able to perform *de novo* annotations. Other features of the MAKER2 update are the ability to evaluate quality of entire genome annotations, as well as to identify annotations that should be reviewed by the user. It also has the ability to be parallelized for larger projects. During experimentation, it was determined that SNAP

performs better and gives more reliable results when under the supervision of MAKER2, especially in cases of limited or low quality data (Holt and Yandell, 2011).

BRAKER1, another improved suite of tools, uses GeneMark-ET (Lomsadze, et al., 2014) and AUGUSTUS (Stanke, et al., 2008). RNA-Seq data is aligned to the assembled genome for the organism in question. The resulting BAM file is used as input for BRAKER1. GeneMark-ET builds upon GeneMark-ES by including RNA-Seq alignments into the training procedures. After this, *ab initio* gene predictions are made. This can then be used by AUGUSTUS, which requires a strong training set. BRAKER1 was compared to MAKER2 on three organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. BRAKER1 gained 15% on average in accuracy on the gene level when using only RNA-Seq as input evidence (Hoff, et al., 2016).

As test sets, the MAKER2 team did several *de novo* and re-annotations. The Argentine ant (*Linepithema humile*) genome was annotated using MAKER2 utilizing ESTs from both the species itself as well as other ESTs from the ant family and wasps and bees. MAKER2 produced 13,785 gene annotations. A 22 megabase region of the maize genome was re-annotated utilizing MAKER2. Annotations were input into the program, along with ESTs and cDNAs. *Arabidopsis thaliana* was used as the protein homology dataset, along with any non-maize proteins from the UniProt/Swiss-Prot database. The re-annotation identified 304 (of 493) gene models that needed updating suggested by alignments and 88 new gene models that previously did not exist. 89 reference gene models were found to not have sufficient support and needed review (Holt and Yandell, 2011).

The barley genome was annotated utilizing sequenced transcriptomes from various stages of barley growth. These were anchored to contigs utilizing GenomeThreader (Gremme, et al., 2005) and CuffLinks (The International Barley Genome Sequencing Consortium, 2012; Trapnell, et al., 2010).

The diploid strawberry genome was re-annotated using the MAKER2 pipeline in 2015. For the annotation, they input data from multiple sources into MAKER. These included: assembled transcripts, SNAP *ab initio* predictions, plant gene models from Augustus (Stanke and Morgenstern, 2005), old strawberry annotations, reference-based assemblies (from the alignments of RNA-Seq data to an existing strawberry genome assembly), and plant reference proteins from UNIPROT's database. Utilizing and combining all of these resources identified 2,286 new gene models (Darwish, et al., 2015).

The polyploid sugarcane genome (cultivar SP80-3280) was annotated using BRAKER1 which resulted in 153,078 predicted protein-coding genes (Riaño-Pachón and Mattiello, 2017). BRAKER1 was also utilized in the annotation of the allotetraploid quinoa (*Chenopodium quinoa* Willd.) genome (Yasui, et al., 2016).

1.5: Dissertation Organization

The remainder of this dissertation will be organized as follows. Chapter 2 will address the assembly and annotation of the *Avena atlantica* and *Avena eriantha* genomes and further analyses. Chapter 3 will address the identification of *Avena*-specific genes. This work was completed using a BLAST pipeline. Chapter 4 will address the unique nutritive properties of oat and human pathways affected. Here we utilize the P2EP Knowledge Base. Chapter 5 will include conclusions and recommendations for future work.

CHAPTER 2: ASSEMBLY AND ANNOTATION OF TWO DIPLOAD OAT GENOMES: *AVENA ATLANTICA* AND *AVENA ERIANTHA*

2.1: Introduction

Common oat (*A. sativa*) and red oat (*A. byzantina* C. Koch) are allohexaploids (2*n*=6*x*=42, AACCDD subgenomes) belonging to the Avenae Tribe of the Poaceae and were domesticated from wild-weedy *A. sterilis* L. (Zhou, et al., 1999), a species that arose from hybridization between a CCDD allotetraploid closely related to modern *A. insularis* Ladiz. and an A_sA_s diploid (Yan, et al., 2016). Several variants of the A-subgenome diploids exist (A_c, A_d, A_l, A_p, and A_s; (Loskutov and Rines, 2011)) and are known to harbor several genetic features of significance, including major crown rust resistance genes such as Pc23 and Pc94 that have been sexually transferred into hexaploid oat cultivars (Aung T, 1996; Dyck and Zillinsky, 1963). The A-genome diploids have also been identified as potential gene sources for improving soluble fiber and protein (Welch, et al., 2000). The A-subgenome is also part of a major intergenomic translocation (7C-17A) in *A. sativa-A. sterilis* that has been associated with adaptation to daylength insensitivity and winter hardiness - key elements in oat production that likely contributed to the plant's ability to shift from Mediterranean winter ecology to Eurasian spring-summer cultivation (Jellen and Beard, 2000).

The C-subgenome chromosomes have a high amount of diffuse heterochromatin along their entirety (Fominaya, et al., 1988); this is a genetic feature not seen in the A and D chromosomes, where heterochromatin is localized and seemingly concentrated around the centromeres, at the telomeres and flanking secondary constrictions where rRNA genes

are located. Among the important genetic features in the C-subgenome is a terminal translocation segment on the long arm of 21D which carries a putative *CSlF6c* locus that likely has a negative effect on seed soluble fiber content (Coon, 2012; Jellen, et al., 1994). Linkage has also been demonstrated between the chromosome 5C telomeric knob in allotetraploid *A. magna* Murphy et Terrell (CCDD subgenomes) and co-segregating genes controlling awn production and basal abscission layer formation which have been implicated in the domestication of common oat (Oliver, et al., 2011).

Despite the historical importance of oat and the renewed interest in its nutritional value, a complete genome sequence of oat has yet to be reported. The *A. sativa* genome is large (> 12 Gb (Bennett and Smith, 1976)), complex and highly repetitive, and characterized by several major intra- and inter-genomic rearrangements - making full genome assembly of the hexaploid difficult (Sanz, et al., 2010). Here we report the development of fully annotated, chromosome-scale assemblies for the extant progenitor species of the A_s- and C_p-subgenomes, *A. atlantica* B.R.Baum & Fedak and *Avena eriantha* Durieu., respectively. Using the assemblies we i) identified and quantified repetitive element content in the genome, including centromeric and telomeric repeats; ii) analyzed syntenic relationships with other cereal grains and homoelogous relationship within oats using existing consensus linkage maps (Chaffin, et al., 2016); iii) identified putative candidate genes for flowering time (Bekele, et al., 2018) and crown rust resistance (Klos, et al., 2017) using recently published Genome-Wide Association Studies (GWAS); iv) estimated the age of the evolutionary split between the A- and C-subgenomes using synonymous substitution rates (K_s) analysis and v) examine genetic diversity and

phylogenetic relationship from a resequencing panel of 76 A- and C-genome *Avena* species.

2.2: Materials and Methods

2.2.1: Plant Material and Nucleic Acid Extraction

For whole-genome assembly, young leaf tissue (~14-21 days post emergence), dark treated for 72 h, from *A. atlantica* (CC7277; T. Langdon, Aberystwyth University, Wales, UK) and *A. eriantha* (BYU132; EN Jellen, Brigham Young University, Provo, UT) was flash-frozen and sent to the Arizona Genomics Institute (AGI; Tucson, AZ, USA) for high molecular weight DNA extraction, in preparation for PacBio sequencing. For the diversity panel, DNA from 76 accessions of diploid A- and C-subgenome species (APPENDIX A) was extracted from 30 mg of freeze-dried leaf tissue using a protocol devised by Sambrook et al. (Sambrook, 1989.) with modifications described by Todd and Vodkin (Todd and Vodkin, 1996). All plants were grown in the greenhouses at Brigham Young University (BYU) using Sunshine Mix II (Sun Gro, Bellevue, WA, USA) supplemented with Osmocote fertilizers (Scotts, Marysville, OH, USA) and maintained at 25°C under broad-spectrum halogen lamps, with 12-hour photoperiods.

2.2.2: DNA Sequencing and Read Processing

For whole genome sequencing, large-insert SMRTBell libraries (>20kb), selected using a BluePippin System (Sage Science, Inc., Beverly, MA, USA) were prepared according to standard manufacture protocols. Libraries were sequenced using P6-C4 chemistry on either the RS II or Sequel sequencing instruments (Pacific BioSciences, Menlo Park, CA, USA; Table 2.1). Sequencing was performed for *A. atlantica* at the DNA

Sequencing Center (DNASC) at BYU (Provo, UT, USA) and at RTL Genomics (Lubbock, TX), while the sequencing for *A. eriantha* was performed at the Los Alamos National Laboratory (Los Alamos, NM) and the BYU DNASC. For the diversity panel and whole genome polishing, extracted DNA was sent to the Beijing Genomic Institute (BGI; Hong Kong, China) for 2 X 150 bp paired end (PE) sequencing from standard 500 bp insert libraries (Table 2.1). Trimmomatic v0.35 (Bolger, et al., 2014) was used to remove adapter sequences and leading and trailing bases with a quality score below 20 or average per-base quality of 20 over a four-nucleotide sliding window. After trimming, any reads shorter than 75 nucleotides in length were removed.

Table 2.1: PacBio and Illumina sequencing read statistics for both Avena atlantica and Avena eriantha.

PacBio	Species	Technology	Number of Cells	Number of Reads	Total Size (bp)	Longest Read (bp)	Mean Read Size (bp)	Median Read Size (bp)	N50 Read Length (bp)	Genome Coverage*
DNASC	Atlantica (AT 808)	Sequel	40	23,475,393	246,554,592,835	194,884	10,706	8,162	18,242	63.2X
RTL Genomics	Atlantica (AT 808)	RS II	72	7,737,947	75,297,173,119	76,481	9,432	6,438	17,317	19.3X
AGI	Atlantica (AT 808)	RS II	4	331,056	4,036,330,519	74,575	12,373	9,881	20,414	1.0X
DNASC	Eriantha (BYU 132)	Sequel	54	28,257,346	276,554,530,583	151,576	8,699	6,479	15,102	70.9X

Illumina	Species	Technology	Average Read Length (bp)	Number of Reads	Total Size (bp)	Genome Coverage*
BGI	Atlantica (AT 808)	HiSeq	142	183,480,412	26,116,814,272	6.7X
Aberystwyth University	Atlantica (AT 808)	HiSeq	143	1,156,806,386	165,681,504,195	42.5
BGI	Eriantha (BYU 132)	HiSeq	142	267,489,998	40,123,499,700	10.03X

*Coverage is based on 3.7 Gb for *A. atlantica* and 4.0 Gb for *A. eriantha*

2.2.3: RNA-Seq and Transcriptome Assembly

For *A. atlantica*, RNA-Seq data consisted 2 X 100 bp PE Illumina reads derived from 11 different plant tissue types including, stem, mature leaf, stressed mature leaf, seed

(2 days old), hypocotyl (4-5 day old), root (4-5 days old), vegetative meristem, green grain, yellow grain, young flower (meiotic) and green anthers. For *A. eriantha*, 2 X 150 PE RNA-Seq data was generated by BGI from six tissue sources, including young leaf, mature leaf, crown, roots, immature panicle, and whole seedling, harvested from plants grown hydroponically in 1X Maxigro™ (GH Inc., Sebastopol, CA, USA) in growth chambers maintained at 21°C under broad-spectrum halogen lamps, with a 12-hour photoperiod at BYU (Table 2.2). The resulting reads were trimmed using Trimmomatic (Bolger, et al., 2014), then aligned to the either the *A. atlantica* or *A. eriantha* reference using HiSat2 v2.0.4 (Kim, et al., 2015) with default parameters except that the max intron length was set to 50,000 bp. Following alignment, the resulting SAM file was sorted and indexed using SAMtools v1.6 (Li, et al., 2009) and assembled into putative transcripts using StringTie v1.3.4 (Pertea, et al., 2016). The quality of the assembled transcriptome was assessed relative to completeness using BLAST comparisons to the reference *Brachypodium distachyon* L. (ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/brachypodium_distachyon/pep/).

Table 2.2: Raw read statistics for RNASeq data for A. atlantica and A. eriantha. All reads were illumina pair-end reads from standard 500 bp insert libraries.

***A. atlantica* (AT 808)**				
Source	**Technology**	**Read length (bp)**	**Number of Reads**	**Total base pairs (bp)**
A1. Stem	Illumina	101	16,597,324	1,676,329,724
B1. Mature leaf	Illumina	101	23,773,832	2,401,157,032
C1. Stressed Mature leaf	Illumina	101	23,937,698	2,417,707,498
D1. Seed (2 days old)	Illumina	101	23,451,368	2,368,588,168
E1. Hypocotyl (4/5 day old)	Illumina	101	30,204,274	3,050,631,674
F1. Root (4/5 days old)	Illumina	101	24,402,346	2,464,636,946
G1. Vegetative meristem	Illumina	101	23,448,778	2,368,326,578
H1. Green grain	Illumina	101	28,358,800	2,864,238,800
A2. Yellow grain	Illumina	101	22,643,900	2,287,033,900
B2. Young flower (meiotic)	Illumina	101	22,817,776	2,304,595,376
C2. Green anthers	Illumina	101	25,725,050	2,598,230,050
		Total:	**265,361,146**	**26,801,475,746**
***A. eriantha* (BYU 132)**				
Source	**Technology**	**Read length (bp)**	**Number of Reads**	**Total base pairs (bp)**
Young leaf	Illumina	150	67,538,458	10,130,768,700
Mature leaf	Illumina	150	52,218,112	7,832,716,800
Crown tissue	Illumina	150	55,027,004	8,254,050,600
Roots	Illumina	150	60,794,380	9,119,157,000
Whole seedling	Illumina	150	64,350,142	9,652,521,300
Immature panicle	Illumina	150	68,747,494	10,312,124,100
		Total:	**368,675,590**	**55,301,338,500**

2.2.4: Genome Size, Assembly, Polishing, and Scaffolding

Genome size was estimated using Jellyfish (Marcais and Kingsford, 2011) and GenomeScope v1.0 (Vurture, et al., 2017) at k-mer length = 21 for each species. Initial assemblies were done using Canu v1.7 (Koren, et al., 2017) with default parameters except setting corMhapSensitivity = normal and corOutCoverage = 40. The resulting assemblies were polished using Arrow from the GenomicConsensus package in the Pacific BioSciences SMRT portal v5.1.0 and PILON v0.22 (Walker, et al., 2014) using Illumina short reads. Chicago and Hi-C proximity-guided assemblies were performed by Dovetail

Genomics LLC (Santa Cruz, CA, USA) to produce chromosome-scale assemblies. Fresh leaf tissue from a single dark treated (72 h), 3-week old plant, derived directly from selfing of the original *A. atlantica and A. eriantha* plants, was sent to Dovetail Genomics for Chicago and Hi-C library preparation as described by Putnam et al. (2016) and Lieberman-Aiden et al. (2009), respectively, using the *Dpn*II restriction endonuclease. The libraries were sequenced using a standard Illumina library prep followed by sequencing on an Illumina HiSeq X in rapid run mode. The HiRiSE scaffolder and the Chicago and Hi-C library-based reads pairs were used to a produce a likelihood model for genomic distance between read pairs, which was used to break putative miss-joins and to identify and make prospective joins in the *de novo* Canu assemblies.

2.2.5: Repeat Analysis, Genome Completeness and Annotation

RepeatModeler v1.0.11 (Smit, 2008-2015) and RepeatMasker v4.0.7 (Smit, et al., 1996-2010) were used to quantify and classify repetitive elements in the final assemblies, relative to RepBase libraries v20181026; www.girinst.org). Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (Simão, et al., 2015) was employed to assess the completeness of the assembly using the Embryophyta odb9 dataset and the --long argument, which applies Augustus (Stanke, et al., 2006) optimization for self-training.

MAKER2 v2.31.10 (Cantarel, et al., 2008; Holt and Yandell, 2011) was used to annotate the final genomes. Expressed sequence tag evidence for annotation included the *de novo* transcriptomes for each species as well as the cDNA models from *Brachypodium distachyon* L. (v 1.0; Ensembl genomes). Protein evidence included the uniprot-sprot database (downloaded September 25, 2018) as well as the peptide models from *B.I*

distachyon (v 1.0; Ensembl genomes). Repeats were masked based on species-specific consensi.fa.classified files produced by RepeatModeler. For *ab initio* gene prediction, *A. atlantica* and *A. eriantha* species-specific AUGUSTUS gene prediction models were provided as well as rice (*Oryza sativa*)-based SNAP models.

2.2.6: Variant Identification and Tree Creation

Single Nucleotide Polymorphisms (SNPs) for the diversity panel were identified from the Illumina reads by mapping the A-subgenome and C-subgenome diploid accessions against the *A. atlantica* and *A. eriantha* reference genome assemblies, respectively, using BWA-mem v0.7.17 (Li, 2013). Output SAM files were converted to BAM files and sorted using SAMtools v1.6 (Li, et al., 2009), and indexed using Sambamba v0.6.8 (Tarasov, et al., 2015). InterSnp, an analysis tool from the BamBam v1.4 package (Page, et al., 2014) was used to call SNPs with the arguments -m 2 and -f 0.35. Bash scripting was used to removed SNPs with less than 100% genotypic calls across all accessions or where 5% or more of the accessions were called as heterozygotes. SNPs on unscaffolded contigs were also removed prior to phylogenetic analysis. SNPhylo v20160204 (Lee, et al., 2014), which uses MUSCLE (Edgar, 2004) for sequence alignments and linkage disequilibrium to down sample the SNP dataset, was used to build Phylogenies with the bootstrapping parameter set to 1000. The resulting tree was visualized using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree).

2.2.7: Genome Comparison

Genomic comparisons, including calculations of synonymous substitutions per synonymous sites (Ks) and homology searches for syntenic gene-sets with *Hordeum*

vulgare L., *Oryza sativa* L., *Zea mays* L., and *B. distachyon* were accomplished using the DAGchainer output file from the CoGe (https://genomevolution.org/coge/) SynMap tool.

2.2.8: Data availability

The raw sequences used for *A. atlantica* genome assembly are deposited in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) database under the BioProject PRJNA546592 with the following accession numbers: SRR9720684 (PacBio reads), SRR9841448 - SRR9841455 (Hi-C reads), SRR9841587 - SRR9841597 (Transcriptome). Similarly, the raw sequences for the *A. eriantha* genome assembly are found in BioProject PRJNA546595 with the following accession numbers: SRR9720373 (PacBio reads), SRR9833273 - SRR9833276 (Hi-C reads), SRR9722223 (Polishing short reads), SRR9722219 - SRR9722222, SRR9722225, SRR9722226 (Transcriptome). The raw reads for the resequencing panel of the diploid species (APPENDIX A) are found in BioProject PRJNA556219 with the following NCBI accession numbers: SRR9933122 - SRR9933198 (resequencing panel). Genome browsing and bulk data downloads, including annotations and BLAST analysis of the final proximity-guided assemblies are available at CoGe (https://genomevolution.org/coge/) with genome IDs: id53337 (*A. atlantica*) and id53381 (*A. eriantha*).

2.3: Results and Discussion

2.3.1: Whole Genome Sequencing and Assembly

We selected the *A. atlantica* accession Cc7277 (Institute of Biology, Environmental and Rural Sciences (IBERS) collection, Aberystwyth University, UK) and the *A. eriantha* accession CN 19328 (Plant Gene Resources of Canada, Saskatchewan, CN) for whole-

genome shotgun sequencing. Both accessions are highly inbred, phenotypically homogeneous and represent type accessions for their respective species. A total of 31,544,396 and 28,257,346 PacBio reads were generated across 122 (RSII and Sequel) and 54 (Sequel) SMRT cells generating a total of 325.9 (~84X coverage) and 276.6 (~71X coverage) Gb of sequence data for *A. atlantica* and *A. eriantha*, respectively. The longest reads for each species, 194,884 and 151,576 bp, came from the Sequel instrument. The N50 read length for *A. atlantica* and *A. eriantha* was 18,658 and 15,102 bp, respectively. In addition to PacBio sequencing, a total of 192 Gb for *A. atlantica* and 40 Gb for *A. eriantha* of 2 x 150 bp Illumina sequences were generated. A k-mer analysis (at k = 21 scale) using Genoscope (Vurture et al. 2017) predicted a genome size of 3.72 Gb with 0.07% heterozygosity and a repeat fraction of 78% for *A. atlantica* and a genome size of 4.17 Gb with a 0.12% heterozygosity and a repeat fraction of 76% for *A. eriantha* (Table 2.3). The relative magnitude of these values agree well with reported values by Bennett and Smith (1976) and Yan, et al. (2016), both of which report that the genomes of the A-genome diploid are ~15% smaller than the C-genome diploids However, former estimates determined by replicated flow cytometry measurements ranged in size from 4.1-4.6 Gb for A-genomes, and from 5.0 to 5.1 Gb for C genomes (Yan, et al., 2016). The differences in genome size predicted by k-mer vs. flow cytometry analyses is likely a reflection of the significant repeat fraction in the oat genome that is difficult to account for using a k-mer analysis.

Table 2.3: Estimation of genome heterozygosity, repeat content and genome size using GenomeScope (Vurture et al. 2017) for A. atlantica and A. eriantha at k=21.

	Avena atlantica		***Avena eriantha***	
	min	max	min	max
Heterozygosity (%)	0.06	0.07	0.12	0.12
Genome Haploid Length (bp)	3,722,497,502	3,722,912,267	4,164,092,415	4,165,797,510
Genome Repeat Length (bp)	2,891,980,461	2,892,302,688	3,175,053,164	3,176,353,272
Genome Unique Length (bp)	830,517,042	830,609,579	989,039,251	989,444,238
Model Fit (%)	86.63	97.01	89.95	99.06
Read Error Rate (%)	0.05	0.05	0.07	0.07

Prior to Hi-C scaffolding, Canu was used to assemble the *A. atlantica* and *A. eriantha* PacBio long reads into 3,914 and 8,067 contigs with an N50 of 5,544,947 and 1,385,002 bp, spanning a total of 3.68 and 3.77 Gb of total length, respectively (Table 2.4). The L50 of the assemblies were 196 and 797 and the longest contigs spanned 25,143,700 and 10,103,775 bp, respectively. The average G+C content of the assemblies were 44.4% and 43.9%, which is similar to most monocotyledonous cereals (e.g., *Sorghum bicolor*, 43.9%; *Oryza sativa*, 43.6% G+C) but significantly higher than G+C content predicted for dicots, which typically range between 33 - 36% (e.g., *Carica papaya*, 34%; *Arabidopsis thaliana*, 36%) (Singh, et al., 2016). As these were PacBio read based assemblies, no "N" gaps were present in the Canu assemblies.

Table 2.4: Summary statistics for the Canu (Koren et al. 2017) and Hi-C assemblies for A. atlantica and A. eriantha.

Assembly	***A. atlantica***		***A. eriantha***	
	Canu	Hi-C	Canu	Hi-C
Number of scaffolds	3,941	2,195	8,067	2,652
Total size of scaffolds (bp)	368,352,214	368,505,449	377,353,911	377,778,748
Longest scaffold (bp)	25,143,700	577,845,554	10,103,775	588,203,704
Shortest scaffold (bp)	1,010	1,010	1,020	1,020
Number of scaffolds > 1M nucleotides	768	9	1,203	7
N50 scaffold length	5,544,947	513,237,590	1,385,002	534,821,622
L50 scaffold count	196	4	797	4
Scaffold % A	27.81	27.81	28.06	28.04
Scaffold % C	22.2	22.19	21.94	21.91
Scaffold % G	22.19	22.18	21.93	21.91
Scaffold % T	27.8	27.79	28.07	28.05
Scaffold % N	0	0.03	0	0.09
Scaffold N nt	0	1,250,201	0	3,223,400
Scaffold % non-ACGTN	0	0	0	0
Percentage of assembly in scaffolded contigs	0.00%	97.00%	0.00%	97.80%
Average number of contigs per scaffold	1	1.9	1	3.1
Average length of breaks (20 or more Ns) between	0	601	0	578
Number of contigs	3,941	4,275	8,067	8,228
Number of contigs in scaffolds	0	2244	0	5740
Number of contigs not in scaffolds	3,941	2,031	8,067	2,488
Total size of contigs	3,683,522,149	3,683,804,291	3,773,539,112	3,774,564,081
Longest contig	25,143,700	21,736,085	10,103,775	10,106,525
Shortest contig	1,010	120	1,020	198
Number of contigs > 1M nt	768	868	1,203	1,202
N50 contig length	5,544,947	4,310,367	1385002	1,314,218
L50 contig count	196	245	797	838
Contig % A	27.81	27.81	28.06	28.07
Contig % C	22.2	22.2	21.94	21.93
Contig % G	22.19	22.19	21.93	21.93
Contig % T	27.8	27.8	28.07	28.07
Contig % N	0	0	0	0
Contig %non-ACGTN	0	0	0	0

To improve the Canu assembly, contigs were further scaffolded using chromatin contact maps using DoveTail Chicago® and Hi-C libraries. Chicago® library contact maps are based on purified DNA that is reconstituted *in vitro* and thus limited to chromatin associations no larger than the size of the purified input DNA fragments (< 100 kb).

Nonetheless they are ideal for detecting and correcting miss-joins in *de novo* assemblies as well as short range scaffolding (Putnam, et al., 2016). Approximately 73X coverage of 1-100 kb read pairs (2 X 150) were generated from Chicago® libraries for each *Avena* species and used to scaffold the Canu assemblies using the HiRiSE™ scaffolder. In total, 334 and 158 breaks were made, while 1,157 and 2,962 joins were made in the *A. atlantica* and *A. eriantha* assemblies, respectively. The net effect of these changes was to decrease the number of total scaffolds to 3,118 in the *A. atlantica* assembly and to 5,263 in the *A. eriantha* assembly, which was accompanied by a slight decrease in N50 (4,310 and 1.314 kb, respectively) for each assembly. Whenever a join was made between contigs, an "N" gap, consisting of 100 Ns was created. The total percent of the genome in gaps, for both species, was less than 0.1%.

The Chicago assemblies were then further scaffolded using *in vivo* Hi-C libraries, created from native chromatin to produce ultra-long-range mate pairs. Mate pair reads (10 - 10,000 kb) representing a physical coverage of 2,749X and 513X were generated for the *A. atlantica* and *A. eriantha* genomes and scaffolded using the HiRiSE™ scaffolder. In total 922 joins and 2,614 joins (plus three breaks) in the *A. atlantica* and *A. eriantha* were made, respectively, producing ultra-long scaffolds, putatively representing full length chromosomes and/or chromosome arms. The HiRiSE assembly for *A. atlantica* had a scaffold N50 of 513.2 Mb, and an L50 of 4, spanning a total sequence length of 3.685 Gb. The longest scaffold spanned 577.8 Mb. Similarly, the *A. eriantha* assembly had a scaffold N50 of 534.8 Mb, an L50 of 4, and spanned a total sequence length of 3.778 Gb with the longest scaffold reaching 588.3 Mb. Scaffold joins produced by the Hi-C mate pairs

introduced new "N" gaps in the assembly (each consisting of 1000 Ns) thereby increased the number of gaps in the assembly to 2,079 and 5,576 for *A. atlantica* and *A. eriantha*, respectively. The final percentage of "N" nucleotides in the final assemblies was less than 0.1%, with the average gap size of 600 and 578 bp, respectively (Table 2.4).

The longest eight scaffolds of the *A. atlantica* assembly, presumably representing two chromosome arms (205 and 278 Mb) and six full length chromosomes (448 - 577 Mb), consisted of > 96% of the total sequence length from the Canu assembly. Similarly, the longest seven scaffolds, ranging in size from 455 - 588 Mb, presumably represent each of the seven haploid *A. eriantha* chromosomes consisted of > 97% of the total Canu assembly sequence. For simplicity, scaffolds representing each of the seven chromosomes from each species are referred to forthwith by size (longest to shortest) as AA1-AA7 and AE1-AE7. The scaffolds in the *A. atlantica* and *A. eriantha* assemblies that remain unintegrated into one of the chromosome-scale pseudomolecules are relatively small and repetitive, with an average size of 61 and 38 kb, which likely contributed to the inability of the proximity-guided assembler to confidently place the contigs within the framework of the chromosomes – specifically the low number of interactions on a short fragment as well as the inability to discern interaction distance differences over the short molecule.

2.3.2: Chromosome Arm Merging

We compared the *A. atlantica* assembly with a recently published genetic linkage map, constructed from a $F_{6:8}$ recombinant inbred line population from a cross of *A. strigosa* x *A. wiestii*, both A_sA_s *Avena* diploid species (Latta, et al., 2019). This map was based on 11,455 ordered, codominant 64-base tag-level haplotypes on seven linkage groups

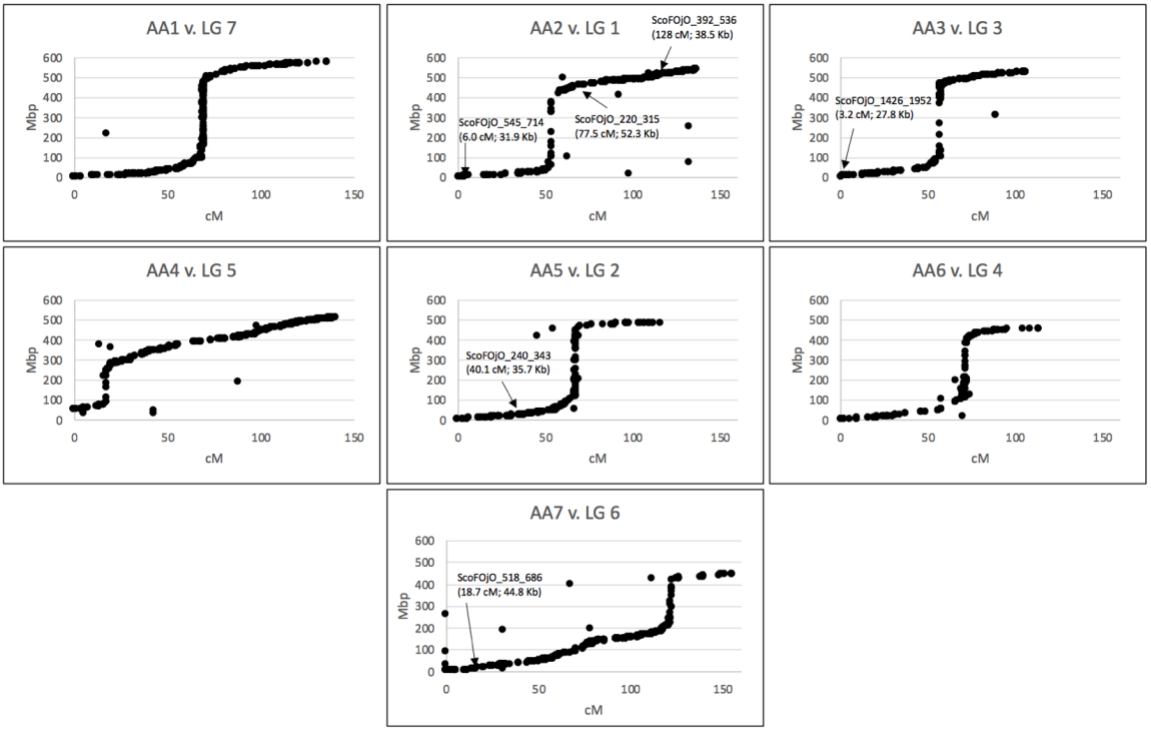
generated using the Haplotag pipeline (Tinker, et al., 2016). Of these, 4,551 haplotypes had perfect matches to single sites on the eight largest scaffolds. A clear one-to-one correspondence between linkage groups (LG) and physical assembly scaffolds was observed (Figure 2.1), with greater than 97% of the tag-level haplotypes mapping to a specific scaffold derived from a single LG. For example, of the 846 tag-level haplotypes mapping to scaffold ScoFOjO_324_449 (AA1), 838 (>99%) were derived from LG 7 (Table 2). Of the 464 tag-level haplotypes derived from LG 2, 378 mapped to scaffold ScoFOjO_1310 (278 Mb) and 85 mapped to scaffold ScoFOjO_1577 (205 Mb), indicating that these two smaller scaffolds should be merged to produce a single, full-length pseudo chromosome (AA5; 485 Mb), thus completing the assembly of seven full-length haploid chromosomes for *A. atlantica*. A head-to-tail merging of these chromosome arms (separated by 1000 Ns) was determined based on the collinearity of the tag-level haplotypes with respect to their orientation within the linkage group. A near perfect collinear relationship was observed between the linkage map and the physical map for all chromosome-linkage group comparisons, with the exceptions being the anticipated reductions of linkage distances relative to physical distances observed at the pericentromeric regions of each chromosome (Figure 2.1). It is well documented that recombination is suppressed in centromeres at rates ranging from fivefold to greater than 200-fold, depending on the species (Haupt, et al., 2001; Talbert and Henikoff, 2010). Of the 2,188 contigs that were unintegrated into an *A. atlantica* chromosome using the Hi-C data, we identified segregating haplotypes linked to 22, spanning a total length of 1.07 Mb,

which could tentatively place them into the context of the seven haploid chromosomes based on their linkage position (Figure 2.1; See Table on Figure 2.1).

Table 2.5: Physical map and Linkage map assignment. Haplotag markers from the consensus map of Latta et al. 2019 were used to assign scaffold assemblies to linkage groups. Two scaffolds mapped to LG 2 and were merged.

LG	Total markers	Miss-matches	% Miss-matches	% Matches	*A. atlantica Hi-C* scaffold	*A. atlantica* chromosome
1	705	12	1.7%	98.3%	ScoFOjO_1702_2338	AA2
2A	85	2	2.4%	97.6%	ScoFOjO_1577	AA5
2B	378	10	2.6%	97.4%	ScoFOjO_1310	
3	546	11	2.0%	98.0%	ScoFOjO_2069_2732	AA3
4	370	16	4.3%	95.7%	ScoFOjO_2050_2712	AA6
5	872	24	2.8%	97.2%	ScoFOjO_350_483	AA4
6	749	36	4.8%	95.2%	ScoFOjO_1760_2399	AA7
7	846	8	0.9%	99.1%	ScoFOjO_324_449	AA1
Total:	4551	119	2.7%	97.3%	-	-

2A = Markers spans 205.8 on the physical map corresponding to linkage positions 0 - 48 cM on the consensus linkage map.
2B = Markers spans 278.2 Mb on the physical map corresponding to linkage position 49-116 cM on the consensus linkage map.



Linkage group	Haplotag marker	Contig ID	pct_identity	Start Position	End Position	E-value	Linkage Position (cM)	Contig size (bp)
1	Oat34640.1184.331Wiestii	ScoFOjO_1246_1686	100	41246	41183	1.66E-25	84.331	63959
1	Oat44215.1160.002Wiestii	ScoFOjO_1356_1854	100	12726	12663	1.66E-25	60.002	22955
1	SW_sltn_61181104.122Strigosa	ScoFOjO_1417_1939	100	20320	20380	7.72E-24	104.122	28046
1	Oat18796.1171.726Wiestii	ScoFOjO_1664_2286	100	25615	25678	1.66E-25	71.726	38815
1	SW_sltn_126431136.153Strigosa	ScoFOjO_1809_2450	100	69758	69703	4.65E-21	136.153	76929
1	SW_sltn_9768177.489Strigosa	ScoFOjO_220_315	100	12535	12476	2.78E-23	77.489	52317
1	SW_sltn_9768177.489Strigosa	ScoFOjO_220_315	100	32604	32545	2.78E-23	77.489	-
1	Oat17132.11128.994Strigosa	ScoFOjO_392_536	100	17317	17259	9.99E-23	128.994	38492
1	Oat37886.11128.994Strigosa	ScoFOjO_392_536	100	32208	32271	1.66E-25	128.994	-
1	Oat42735.11128.994Wiestii	ScoFOjO_392_536	100	3810	3873	1.66E-25	128.994	-
1	Oat4203.112.701Strigosa	ScoFOjO_545_714	100	23590	23653	1.66E-25	2.701	31820
1	SW_sltn_66415.958Strigosa	ScoFOjO_545_714	100	8694	8757	1.66E-25	5.958	-
1	SW_sltn_759515.958Strigosa	ScoFOjO_545_714	100	8697	8634	1.66E-25	5.958	-
1	Oat5311.1142.02Strigosa	ScoFOjO_604_776	100	35826	35769	3.59E-22	42.02	56760
1	SW_sltn_22511136.153Strigosa	ScoFOjO_994_1270	100	61461	61524	1.66E-25	136.153	141876
2	Oat38756.1275.14Strigosa	ScoFOjO_240_343	100	24871	24934	1.66E-25	75.14	35746
2	Oat40687.1275.14Wiestii	ScoFOjO_240_343	100	20824	20887	1.66E-25	75.14	-
2	Oat40687.1275.14Wiestii	ScoFOjO_240_343	100	22869	22932	1.66E-25	75.14	-
2	SW_sltn_9630275.14Strigosa	ScoFOjO_240_343	100	23315	23378	1.66E-25	75.14	-
2	Oat42524.1249.219Wiestii	ScoFOjO_53_70	100	35278	35341	1.66E-25	49.219	47274
2	Oat33700.1248.155Wiestii	ScoFOjO_890_1121	100	12730	12667	1.66E-25	48.155	27651
3	Oat43306.133.238Strigosa	ScoFOjO_1426_1952	100	9019	9082	1.66E-25	3.238	27851
3	Oat43306.133.238Strigosa	ScoFOjO_1426_1952	100	21685	21748	1.66E-25	3.238	-
4	Oat10724.1430.873Wiestii	ScoFOjO_1041_1346	100	19153	19091	5.97E-25	30.873	48913
4	SW_sltn_11194444.126Strigosa	ScoFOjO_1331_1817	100	41284	41347	1.66E-25	44.126	54338
5	Oat33645.15106.458Wiestii	ScoFOjO_1387_1897	100	41049	41112	1.66E-25	106.458	72676
6	Oat36457.1618.715Strigosa	ScoFOjO_518_686	100	9421	9358	1.66E-25	18.715	44764
6	Oat36457.1618.715Strigosa	ScoFOjO_518_686	100	25490	25427	1.66E-25	18.715	-
7	Oat11582.1766.691Strigosa	ScoFOjO_1058_1373	100	23381	23318	1.66E-25	66.691	31421
7	SW_sltn_3462766.136Strigosa	ScoFOjO_2061_2724	100	8684	8626	9.99E-23	66.136	41662
7	SW_sltn_8105768.927Wiestii	ScoFOjO_226_322	100	17975	18033	9.99E-23	68.927	60925
7	Oat3765.1766.691Wiestii	ScoFOjO_963_1222	100	11381	11318	1.66E-25	66.691	22560
							Total:	1067750

Figure 2.1: Validation of the A. atlantica assembly. The genetic position of mapped markers is plotted as a function of physical distance relative to the A. atlantica genome assembly. The linkage position of six unassigned scaffolds with multiple mapping markers is shown.

2.3.3: Analysis of Repetitive Elements

The repeat fraction of the *Avena* genome assemblies was identified and annotated using RepeatModeler and RepeatMasker. In total, ~83% of each genome was classified as repetitive, with the most commonly identified repetitive elements being classified as long terminal repeat retrotransposons (LTR-RTs); LTR-RTs are the most abundant genomic components in flowering plants (Du, et al., 2010; Galindo-Gonzalez, et al., 2017) and their abundance is strongly correlated with genome size (Tenaillon, et al., 2010). Within published plant genomes, repeat content varies widely, ranging from 3% for the minute 82 Mb genome of *Utricularia gibba* L. (Ibarra-Laclette, et al., 2013) to 85% for maize (Schnable, et al., 2009). Given the large size of these genomes, it is not surprising that < 20% of the genome is classified as non-repetitive.

Of the various LTR-RT present (APPENDIX B), *Gypsy*-like and *Copia*-like elements represent > 60% of each genome, in a ratio of 2.3:1 and 3.5:1 for the *A. atlantica* and *A. eriantha* genomes, respectively, which is similar to the ratios reported for other Poaceae species (e.g., rice, 4.9:1 (Tian, et al., 2009); sorghum, 3.7:1; (Paterson, et al., 2009) and maize, 1.6:1, (Baucom, et al., 2009)). The next most common element was identified as class II CMC-EnSpm DNA transposons, representing ~5% of each genome. Interestingly, a significant proportion (*A. atlantica*: 10.6% and *A. eriantha*: 14.3%) of the interspersed repeat fraction for each genome was classified as "unknown". Given the extensive investigations of repeat elements in the grasses (Bilinski, et al., 2017; Feschotte, et al., 2003; Minaya, et al., 2013), this unknown fraction likely represents repeat elements unique to *Avena* and could be invaluable in differentiating the A-, C- and D-subgenomes

in hexaploid oat. For example, Solano, et al. (1992) reported the identification of a tandem repeat sequence clone (pAm1; GenBank X83958) from *Avena murphyi* L., an AACC tetraploid, which selectively hybridized to the C-subgenome. A repeat that was highly homologous (E-value: 2E-82) to pAm1 was identified by RepeatModler in *A. eriantha*, but is missing in the *A. atlantica* genome (Figure 2.2A; Tracks 4 & 5). Similarly, Katsiotis, et al. (2000) reported the identification of an interspersed repeat (pAvKB26; GenBank AJ297385.1) that selectively hybridized to the A- and D-subgenomes. This repeat was identified in the unknown repeat fraction of *A. atlantica* but is missing in the *A. eriantha* genome (Figure 2.2B; Tracks 4 & 5). Repeat content is believed to be an important driver of genome organization and evolution (Michael, 2014), and these data will be important for understanding the overall evolution of common hexaploid oats.

In addition to the interspersed repeat elements identified, ~0.5% of the genome was classified as low complexity, satellite, telomeric repeat (see genomic feature section below) or microsatellite. Indeed, 5,217 and 3,404 putative microsatellite loci were identified, with the most common di-, tri- and tetranucleotide repeat motif identified being $(AT)_n$, $(AAC)_n$ or $(GGC)_n$ and $(TTTA)_n$, in *A. atlantica* and *A. eriantha* respectively. To date, no microsatellites have been generated specifically for the *Avena* diploid species – thus these new putative microsatellite loci represent important genetic tools for studying diversity and can also be used for advancing breeding in the A-genome diploids.

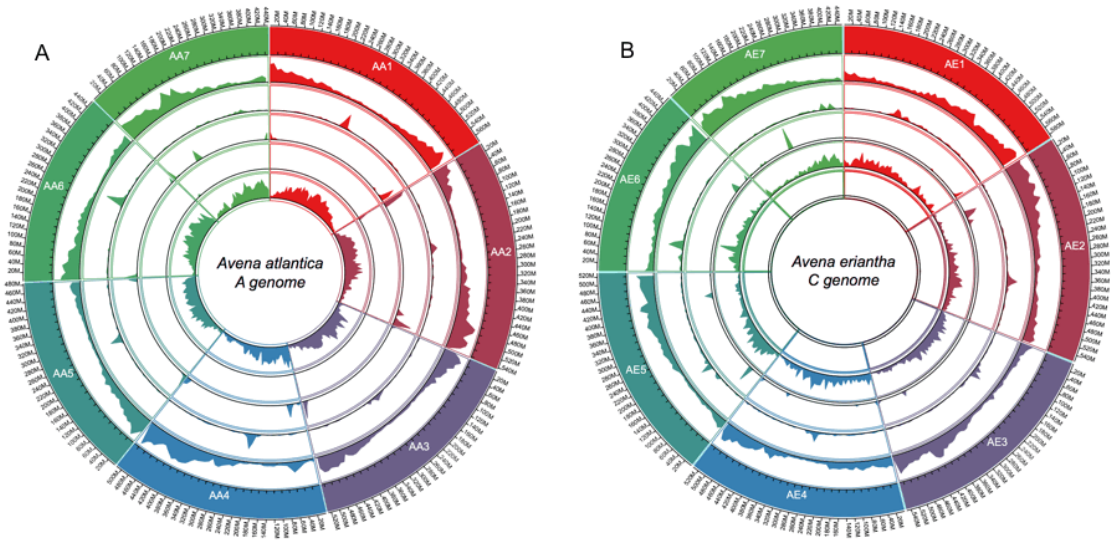


Figure 2.2: Genome overview of (A) A. atlantica and (B) A. eriantha. Track 1: Chromosome and sizes; Tracks 2: Annotated gene density; Track 3: Centromeric repeat density; Track 4: Telomeric sub-repeat density; Track 5: C-genome specific repeat (pAm1) density; Track 6: A-genome specific repeat (pAvKB26) density.

2.3.4: Transcriptome Assembly and Functional Annotation

The *A. atlantica* and *A. eriantha* transcriptomes, which consisted of 51,223 and 47,361 scaffolded isoforms, the *Brachypodium* cDNA and peptide models (v 1.0; Ensembl genomes) and the uniprot-sprot database were provided as primary evidence for annotation in the MAKER pipeline (Cantarel, et al., 2008). The RNAseq data mapped with high efficiency to the assemblies, with > 96% of the reads mapping to their respective genome with 93.1% concordance for pair alignment rates, suggestive of high-quality genome assemblies for both species. The MAKER pipeline identified a total of 51,100 and 49,105 gene predictions, with mean transcript lengths of 3,018 and 3,153 bp, with 70% and 66% of the annotations have annotation edit distance (AED) measures < 0.25 for *A. atlantica* and *A. eriantha* genomes, respectively. AED integrates sensitivity, specificity, and accuracy measurement to calculate annotation quality, where AED values < 0.25 are

indicative of high-quality annotations (Holt and Yandell, 2011). The mean G+C content of the transcripts in both species was ~52%. The increase in G+C content within coding regions relative to the overall G+C content of the genome (~44%) is a well-known phenomenon and is hypothesized to be the result of GC-biased gene conversion – a process by which the G+C content of DNA increases due to gene conversion during recombination (Duret and Galtier, 2009).

The completeness of the gene space defined by the genome and the annotations were quantified using BUSCO which provides a quantitative measure for genome and transcriptome completeness based on large core set of highly conserved plant-specific single-copy orthologs, (Simão, et al., 2015). Of the 1,440 plant-specific orthologs, 1,387 (96.3%) were identified in the *A. atlantica* genome assembly, while 1,395 (96.9%) were identified in the *A. eriantha* assembly, suggesting high-quality and complete genome assemblies. As expected for diploid species, the level of gene duplication, as identified by BUSCO for the conserved orthologous genes was low for both species (2.2% and 2.3%). Similarly, a BUSCO analysis of the transcript and protein annotation sets produced by MAKER, identified similar numbers of conserved orthologs for both species which is indicative of a successful annotation process.

2.3.5: Genomic Features

Pericentromeric regions, associated with reduced recombination relative to physical distance, were evident from the linkage and physical map comparison in *A. atlantica* (Figure 2.1). The observation that gene density is substantially reduced provided further evidence that that these regions represented centromeric regions in both species, as

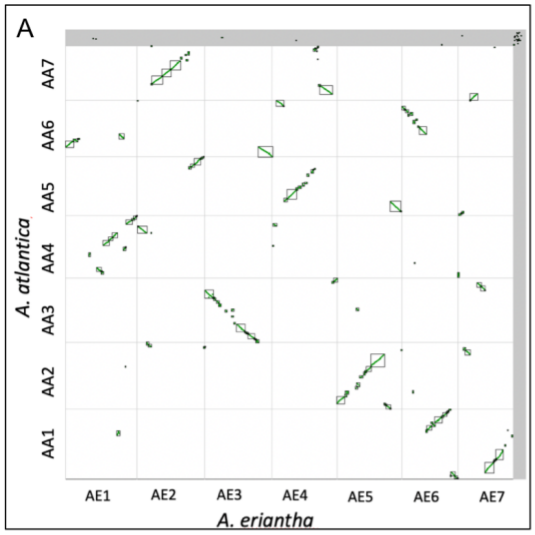
has been well-documented previously in other eukaryotes (Mizuno, et al., 2011; Philippe, et al., 2013). Centromeres in most plant species are complex but are dominated by megabase-sized arrays of tandemly arranged monomeric satellite repeats. While complex and highly diverse among plant species, they commonly share a unit length ranging between 150 - 180 bp, which is close to the size of the nucleosome unit (Henikoff, et al., 2001). Melters et al. (Melters, et al., 2013) showed that due to the relative size of the centromere the most common repeat found in whole genome sequencing data is the putative centromeric repeat. Using the output of RepeatModeler from *A. eriantha*, we identified a high-copy-number 159 bp tandem repeat that aligned specifically with the putative centromere location in each of the *A. atlantica* and *A. eriantha* chromosomes (Figure 2.2; APPENDIX B). Although the 159 bp repeat is similar in size to the putative centromeric repeats found in other grass species (e.g., *B. distachyon*, 156 bp; *H. vulgare*, 139 bp; *Oryza brachyantha* A.Chev. & Roehr., 154 bp; *Z. mays*, 156 bp), not surprisingly it shares little sequence similarity with the centromeric repeats of those species. Indeed, centromeric repeats exhibit little to no evidence of sequence similarity beyond ~50 million years of divergence (Melters, et al., 2013). As has been documented in other plant species, these putative centromeric repeats span a large portion (often > 50 Mb) of the *A. atlantica* and *A. eriantha* chromosomes, suggesting the presence of large pericentromeric heterochromatic regions (Gan, et al., 2016; Willing, et al., 2015). Moreover, the positioning of the centromeres, as defined by this putative repeat and the gene density plots, is consistent with the cytological positioning of the centromere, which suggests that the centromeres in *A. atlantica* are almost all metacentric to submetacentric, while the

centromeres in *A. eriantha* are almost all sub-telocentric (Baum and Fedak, 1985; Rajhathy and Dyck, 1963). Indeed, per our analyses, we identified three metacentric, two sub-metacentric and two sub-telocentric pseudochromosomes in *A. atlantica*, and five sub-telocentric, one submetacentric and one metacentric pseudochromosomes in *A. eriantha* (Figure 2.2).

RepeatModeler annotated a putative telomeric satellite sequence (665 bp) for *A. eriantha* (APPENDIX B). A homologous sequence (639 bp) with significant homology (E-value = 0.0) and alignment identity (Identity = 80%; Gap = 6%) was identified from the repeat sequences identified by RepeatModeler (APPENDIX B). BLAST searches of the assemblies with their respective satellite telomeric repeat sequence identified enriched regions of the telomeric repeat on all seven of the chromosomes for each species (Figure 2.2). In *A. atlantica*, telomeric satellite sequences were located toward the distal end of each chromosome; however, in *A. eriantha* the location of the sequence is more dispersed, being found primarily at the end of the chromosomes on ten of the 14 chromosome arms, while in a few instances being interspersed interstitially. Interstitial telomere-like repeats have been reported in several plant species including *Anthurium*, *Vicia*, *Sideritis*, *Typhonium* and *Pinus* where they were implicated in chromosome rearrangements, including inversions, translocations, and chromosome fusions (Islam-Faridi, et al., 2007; Raskina, et al., 2008; Schubert, et al., 1992; Sousa, et al., 2014). While chromosomal rearrangements are common in *Avena*, we caution that the very repetitive nature of telomeric sequences makes them susceptible to collapse during the assembly process and are thus inherently difficult to order and orient in the Hi-C scaffolding process.

2.3.6: Comparative genomics

We calculated the rate of synonymous nucleotide substitutions per synonymous site (K_s) for orthologous gene-pairs between the *A. atlantica* and *A. eriantha* assemblies using the CodeML (Yang, 2007) tool on the CoGe platform (genomevolution.org/coge). A total of 18,002 duplicate pairs were identified with a clear peak seen at $K_s = 0.0875$, suggesting that speciation between *A. atlantica* and *A. eriantha* occurred between 2.9–5.4 million years ago (MYA), depending on whether an *Arabidopsis*- or core eukaryotic-based synonymous mutation rate was used in the calculation (Koch, et al., 2000; Lynch and Conery, 2000) (Figure 2.4). As seen in the SynMap dot-plot alignment (Figure 2.3), significant synteny was observed between the *Avena* chromosomes consisting of 187 syntenic blocks with 21,021 collinear genes pairs (112 genes/block) with 98.2% coverage across both the *A. atlantica* and *A. eriantha* genomes. As expected, given the relatively close ancestry of the species, the size (bp) of the syntenic blocks between species was highly correlated (R^2=0.88; Figure 2.3C). The large blocks of syntenic genes are suggestive of homoeologous relationships between the chromosomes of the species (Figure 2.3A). For example, slightly more than 77% (349 Mb) of the syntenic sequence found on AA2 is derived from AE5, suggesting that they are homoeologs. Indeed, using a majority rule (>50% syntenic sequence) we identified the following homoeologous chromosome pairs: AA1 + AE6 (61%; 248 Mb); AA2 + AE5 (77%; 349 Mb); AA3 + AE3 (74%; 318 Mb); AA4 + AA1 (71%; 271 Mb); AA7 + AE2 (57%; 274 Mb); with AA5 and AA6 sharing homoeology with several *A. eriantha* chromosomes (Figure 2.3B).



B

A. atlantica Hi-C scaffold/Chromosome name		*A. eriantha* Hi-C scaffold/Chromosome Name												Total syntenic sequence (bp)	Homoelogs†
		1585_3068		4_4		2055_3936		430_889		1285_2510		670_1352		2331_4482			
		AE6		AE5		AE3		AE1		AE4		AE7		AE2			
324_449	AA1	61%	248,113,104	0%	-	0%	-	2%	8,964,879	0%	-	37%	150,069,527	0%	-	407,147,510	AE6 (60%); AE7 (36%)
1702_2338	AA2	1%	6,268,489	77%	349,869,485	0%	-	0%	-	0%	-	11%	50,572,926	10%	46,620,358	453,331,258	AE5 (77%)
2069_2732	AA3	0%	-	3%	14,026,198	74%	317,581,481	0%	-	10%	43,128,960	13%	55,964,464	0%	-	430,701,103	AE3 (74%)
350_483	AA4	1%	2,448,066	0%	-	0%	-	71%	271,866,573	8%	28,837,464	2%	7,950,910	19%	72,738,363	383,841,376	AE1 (71%)
1310_1788	AA5	0%	-	22%	85,602,020	0%	-	0%	-	41%	158,055,695	9%	34,177,532	28%	109,867,904	387,703,151	AE4 (41%); AE2 (28%)
2050_2712	AA6	29%	128,665,157	0%	-	25%	111,842,443	33%	146,006,081	13%	56,972,331	0%	-	0%	1,603,520	445,089,532	AE1 (32%); AE6 (29%); AE3 (25%)
1760_2399	AA7	0%	-	0%	-	0%	-	0%	-	29%	139,831,412	13%	64,550,926	57%	274,094,733	478,477,071	AE2 (57%); AE4 (29%)

†Subgenome homoeologous chromosomes representing > 50% syntenic blocks.

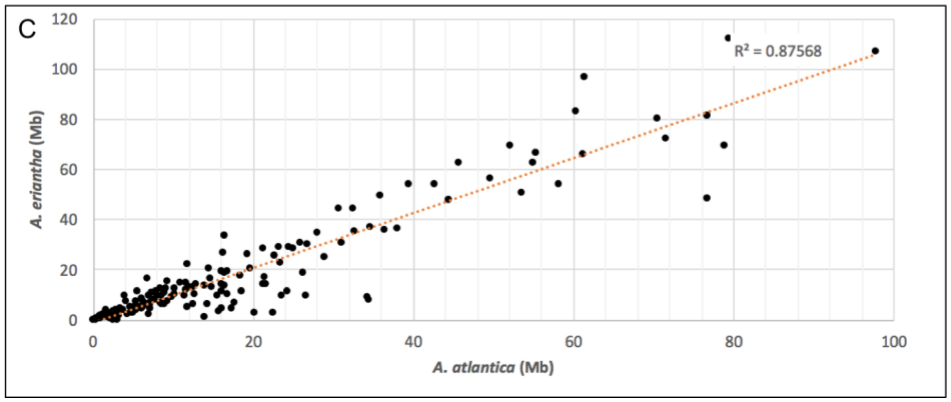


Figure 2.3: Homoeologous genes were identified between A. atlantica and A. eriantha genomes to detect homoeologous chromosome relationships. Genome synteny was (A) visualized by dotplot analysis, with boxes drawn around syntenic regions, (B) quantified, where the chromosome pairs with the highest amount of syntenic block connections, expressed as a percentage of the total syntenic bases, are colored red and transition to white as the number of connections decreases and (C) correlation between syntenic block sizes between A. atlantica and A. eriantha.

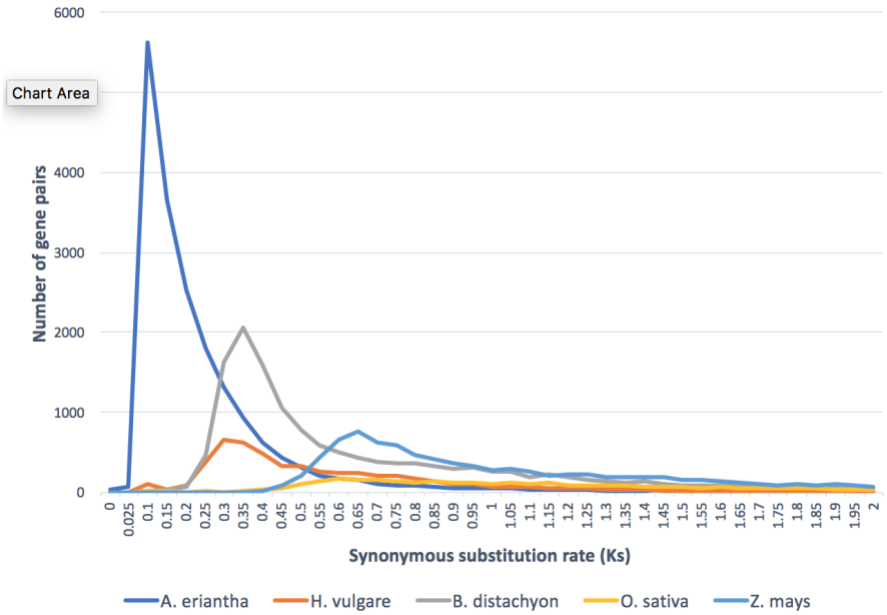
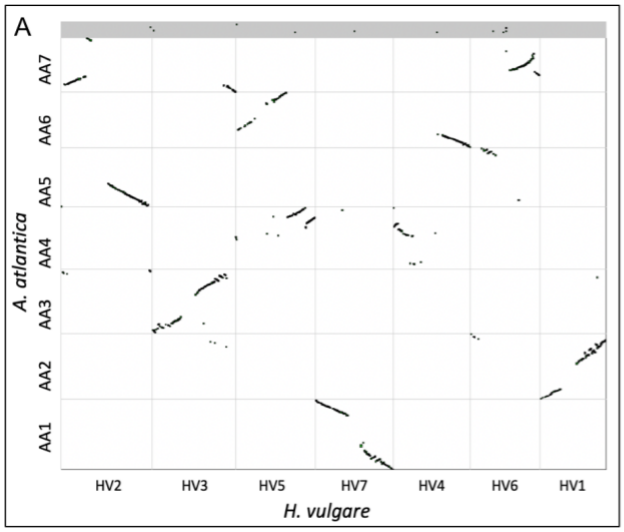


Figure 2.4: Rate of synonymous substitutions per synonymous sites (Ks) within duplicated gene pairs from coding sequences predicted from A. atlantica comparisons with A. eriantha, H. vulgare, B. distachyon, O. sativa, and Z. mays.

The paleobotanical fossil record indicates that the Poaceae family appeared approximately 50–70 MYA (Bouchenak-Khelladi, et al., 2010; Bremer, 1992). The family consists of many agronomically important species, commonly referred to as cereals, that are found in three subfamilies: Oryzoideae (rice), Panicoideae (maize, sorghum) and Pooideae (wheat, barley, oat, and rye). Pooideae forms 14 tribes, including the tribes Brachypodieae, Poeae (syn Aveneae, including oat) and Triticeae (barley, rye, and wheat), with Poeae and Triticeae tribes having separated ~20 MYA (Huang, et al., 2002). This agrees well with the K_S analyses presented here for *A. atlantica* and *A. eriantha* and with the published *Hordeum vulgare* genome (Mascher, et al., 2017), which both show a clear peak at 0.3 – suggestive of a divergence time of 19 MYA (per the core eukayotic-base synonymous substitution rate). As expected, the K_S analyses from the *Avena* comparisons with the *B. distachyon* genome (International *Brachypodium* Initiative, 2010) suggested a

more distant divergence of 20–22 MYA for the split of the *Avena–Brachypodium* lineages (Figure 2.4).

SynMap was also used to investigate syntenic relationships between the *Avena* and *Hordeum* chromosomes (Figure 2.5 & 2.6). Although more syntenic blocks (719 and 714), were identified in the *Avena–Hordeum* comparisons, they were smaller – consisting of ~8.5 genes/block, accompanied by a lower syntenic block size correlation (R_2=0.35; Figure 2.5C & Figure 2.6C). The decrease in block size and correlation is reflective of the more distant evolutionary relationship between the species. Nonetheless, the shared ancestry between the two Pooideae species was evident as substantial synteny was observed across all seven *Avena–Hordeum* chromosomes comparisons Figure 2.5A & Figure 2.6A). As expected, large, proximal, non-syntenic regions were observed in regions corresponding to putative centromeres where gene density is substantially reduced (Mizuno, et al., 2011; Philippe, et al., 2013). The synteny observed among the *Avena* and *Hordeum* chromosomes suggests several homeologous relationships. For example, *Hordeum* chromosome 1H is clearly homoeologous with *Avena* chromosome AA2 and AE5. Indeed, of the syntenic sequence on 1H, 99% (116 Mb) was syntenic to AA2 and 85% (105 Mb) syntenic to AE5 – which is not surprising since we previously showed that AA2 and AE5 are homoeologs (see above). Using a simple majority rule (>50% syntenic sequence) the following are putative *Hordeum–Avena* homoeologs: 1H + AA2/AE5; 2H + AA5/AE4; 3H + AA3/AE3; 6H + AA7/AE2; and 7H + AA1/AE6. The specific *A. atlantica* homoeologs of 4H and 5H are likely AA6 and AA4, respectively; however, rearrangements in the *A. eriantha* genome obscure the likely homoeologs for *A. eriantha* (Figure 2.5B & Figure 2.6B).



B

H. vulgare chromosomes	*A. atlantica* chromosome													Total syntenic sequence (bp)
	AA1		AA2		AA3		AA4		AA5		AA6		AA7	
HV1	0%	-	99%	116,552,573	1%	827,377	0%	-	0%	-	0%	-	0%	-	117,379,950
HV2	0%	-	0%	-	3%	3,970,402	0%	-	69%	93,678,206	0%	-	29%	38,970,738	136,619,346
HV3	0%	-	2%	2,583,749	88%	131,236,323	0%	-	0%	-	0%	-	10%	15,378,574	149,198,646
HV4	0%	-	0%	-	0%	-	41%	42,236,801	0%	-	59%	61,862,029	0%	-	104,098,830
HV5	0%	-	0%	-	0%	-	54%	72,599,961	0%	-	46%	62,333,988	0%	-	134,933,949
HV6	0%	-	5%	5,476,359	0%	-	0%	-	22%	25,316,901	0%	-	73%	82,857,940	113,651,200
HV7	99.7%	125,071,153	0%	-	0%	-	0%	409,703	0%	-	0%	-	0%	-	125,480,856
Homoelogs†	HV7 (99.7%)		HV1 (99%)		HV3 (88%)		HV5 (54%); HV4 (41%)		HV2 (69%)		HV4 (59%); HV5 (46%)		HV6 (73%)		

†Subgenome homoeologous chromosomes representing > 50% syntenic blocks.

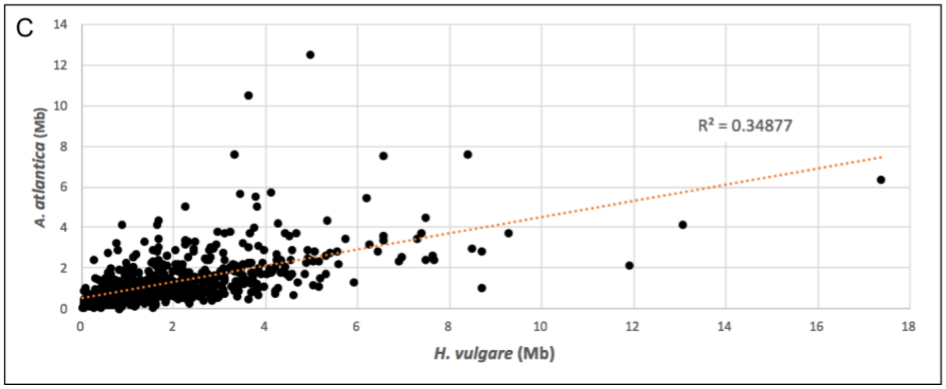
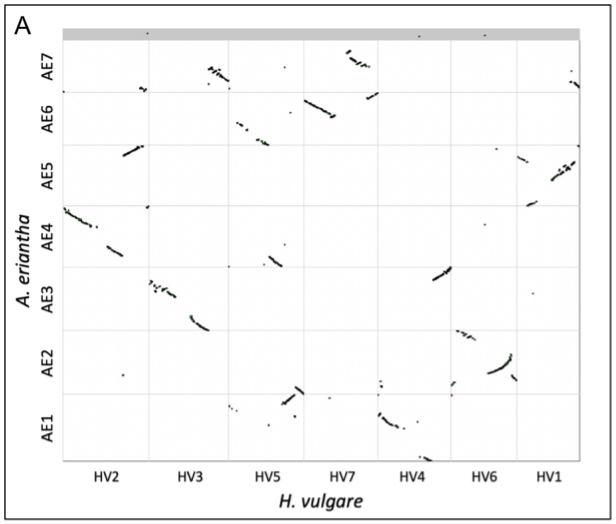


Figure 2.5: Homoeologous genes were identified between A. atlantica and H. vulgare genomes to detect homoeologous chromosome relationships. Genome synteny was (A) visualized by dotplot analysis, with boxes drawn around syntenic regions, (B) quantified, where the chromosome pairs with the highest amount of syntenic block connections, expressed as a percentage of the total syntenic bases, are colored red and transition to white as the number of connections decreases and (C) correlation between syntenic block sizes between A. atlantica and H. vulgare (Hv_IBSC_PGSB_v2; Ensembl Release 36).



B

H. vulgare chromosomes	*A. eriantha* chromosome												Total syntenic sequence (bp)
	AE1		AE2		AE3		AE4		AE5		AE6		AE7		
HV1	0%	-	0%	-	0%	155,397	0%	-	85%	105,383,642	0%	-	15%	18,862,612	124,401,651
HV2	0%	-	0%	701,399	0%	-	65%	113,211,874	28%	49,394,789	0%	-	6%	10,493,315	173,801,377
HV3	0%	-	0%	-	64%	106,439,135	0%	-	0%	-	0%	-	36%	58,708,358	165,147,493
HV4	52%	78,704,051	4%	6,618,174	44%	66,297,455	0%	-	0%	-	0%	-	0%	-	151,619,680
HV5	30%	44,906,428	15%	22,615,266	0%	-	28%	42,982,878	0%	-	27%	41,512,321	0%	36,184	152,053,077
HV6	1%	1,002,263	98%	139,586,141	0%	-	0%	44,332	1%	1,150,999	0%	-	0%	-	141,783,735
HV7	0%	603,748	0%	-	0%	-	0%	-	0%	-	53%	71,356,066	47%	63,613,086	135,572,900
Homoelogs†	HV4 (52%)		HV6 (98%)		HV3 (64%)		HV2 (65%)		HV1 (85%)		HV7 (53%)		-		

†Subgenome homoeologous chromosomes representing > 50% syntenic blocks.

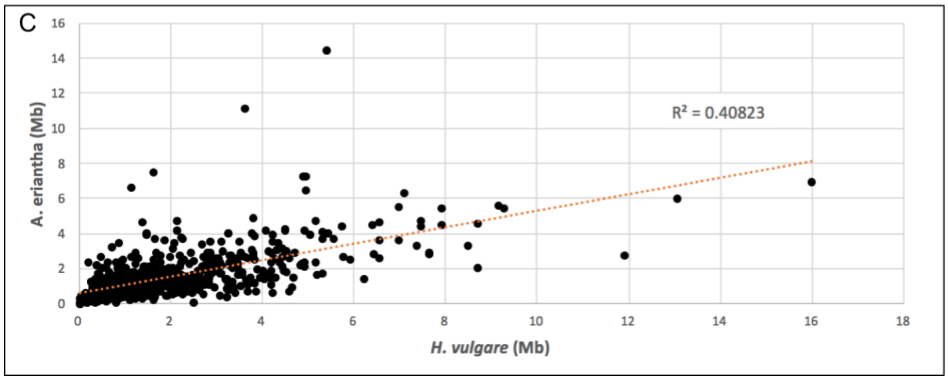


Figure 2.6 Homoeologous genes were identified between A. eriantha and H. vulgare genomes to detect homoeologous chromosome relationships. Genome synteny was (A) visualized by dotplot analysis, with boxes drawn around syntenic regions, (B) quantified, where the chromosome pairs with the highest amount of syntenic block connections, expressed as a percentage of the total syntenic bases, are colored red and transition to white as the number of connections decreases and (C) correlation between syntenic block sizes between A. eriantha and H. vulgare (Hv_IBSC_PGSB_v2; Ensembl Release 36).

Bekele, et al. (2018) recently published a high-density, tag-level, haplotype linkage map of hexaploid oat (*A. sativa*). This consensus linkage map consisted of 21 well-formed linkage groups, putatively corresponding to each of the 21 hexaploid oat chromosomes. To

identify the ancestral subgenome groups (A-, C- and D-) for each of the 21 linkages groups we mapped the haplotag markers to both the *A. atlantica* and *A. eriantha* genomes. To avoid false hits, which is particularly problematic due to the highly repetitive nature of the oat genomes, only BLAST hits with perfect identity across the entirety of the marker sequence (e.g., zero gap openings and mismatches) were retained for downstream analyses. In total, 2,119 and 969 haplotags were mapped to the *A. atlantica* and *A. eriantha* genomes, respectively. The increased number (~2X) of haplotags mapping successfully against the *A. atlantica* genome was not unexpected since we anticipated that both A- and some D-subgenome derived haplotags would potentially map against the A-genome diploid given the close phylogenetic relationship of these two subgenomes (Yan, et al., 2016). Indeed, close inspection of the mapping showed that in nearly all cases, haplotags mapping to a specific *A. atlantica* chromosome were derived from two separate consensus linkage groups - presumably corresponding to homoeologs derived from A- and D-subgenomes. For example, 322 haplotags mapped to chromosome AA1, with 153 (48%) derived from linkage group Mrg12 and 111 (35%) derived from Mrg02, which were previously identified as being derived from the A-, and D- subgenomes (Yan, et al., 2016) (Table 2.6A). Other homoeologous chromosome pairs between the A- and D-subgenome included: Mrg33/Mrg08, Mrg18/Mrg01, Mrg05/Mrg04, Mrg24/Mrg06, Mrg23/Mrg11, Mrg12/Mrg02, and Mrg20/Mrg21. Similar mapping of the haplotags against the *A. eriantha* genome elucidated linkage groups Mrg13, Mrg03, Mrg15, Mrg17, Mrg19, Mrg09 and Mrg 11 as being derived from C-subgenome (Table 2.6B). Interestingly, Mrg18, which we previously designated as an A-subgenome derived linkage group also showed

substantial mapping to the C-genome chromosome AE7 – suggesting that this Mrg18 is actually derived from an A/C- intergenomic reciprocal translocation. This is a well-documented reciprocal translocation, first reported by Jellen et al. (Jellen, 1996) where it was identified as 7C-17A. Other identifiable rearrangements include D/C-intergenomic exchanges on Mrg06/Mrg13, Mrg08/Mrg03, and Mrg19/Mrg28 (Table 2.6).

Table 2.6: Ancestral subgenome groups (A-, C- and D-) designation for each of the 21 consensus linkages groups reported for A. sativa (Bekele et al.). Haplotag markers mapping to (A) A. atlantica and (B) A. eriantha chromosomes, where highest haplotag mapping are colored red and transition to white as the number of haplotags mapping decreases. Subgenome designation for each linkage group are as previously reported by Chaffin et al. and/or Yan et al.

A.

A. atlantica chromosomes	*A. sativa* consensus linkage groups																						Total Markers
	Mrg12	Mrg18	Mrg23	Mrg20	Mrg33	Mrg24	Mrg05	Mrg02	Mrg01	Mrg11	Mrg21	Mrg08	Mrg06	Mrg04	Mrg13	Mrg19	Mrg28	Mrg17	Mrg15	Mrg09	Mrg03		
	Subgenome A*							Subgenome D*							Subgenome C*								
	A	C/A	A	A	A	A	A	D	D	D/C**	D/C	D/C	D/C	D	C	C/D	C/D***	C	C	C	C		
AA1	153	3	5		37			111			9	1	1	1				1					322
AA2	68	134	1	1		3			124			22	1	1		5	7			3			370
AA3		1	189	15	1	1	8	2	1	15	3	3				15	19		6		1		280
AA4	1		1	164	1	77			2	1	73		78				1			11	3		413
AA5	4	2			62	1	39		1		1	43	2	2	3			2					162
AA6	2		1	18		92		1	26	1	1		22										164
AA7				149	8	1	142	1	2	13	33	3		40	2								394
Total:	228	140	197	347	109	175	189	115	156	30	120	72	104	44	5	20	27	3	6	14	4		2105
Average mapping	197.9							91.6							11.3								

*Chromosomal designation as previously reported by Chaffin et al. and Yan et al.
**Chaffin designated Mrg11 as C/A, while Yan et al. designated it as C. Here we assign it as D/C (See Figure X).
***Yan et al. designated Mrg28 as D/C, while Chaffin et al. 2016 designated it as C. Here we re-designate it as D/C (See Figure X).

B.

A. eriantha chromosomes	*A. sativa* consensus linkage groups																						Total Markers
	Mrg09	Mrg17	Mrg15	Mrg13	Mrg19	Mrg03	Mrg11	Mrg04	Mrg02	Mrg01	Mrg06	Mrg28	Mrg08	Mrg21	Mrg33	Mrg18	Mrg05	Mrg24	Mrg23	Mrg12	Mrg20		
	Subgenome C*							Subgenome D*							Subgenome A*								
	C*	C	C	C	C/D	C	D/C**	D	D	D	D/C	C/D***	D/C	D/C	A	C/A	A	A	A	A	A		
AE1	55	2	2			55	16				4		2			2		1			1		140
AE2	49	92	1		1	1	4	3			1	2	2	3		1	2				2		164
AE3	7	2	69	3		1	25			2	1	3				1			5				119
AE4		1		25	2	14	5				37	1			1								86
AE5		1		21	31	5	3		2		1	38	1			8				2			113
AE6	2	18		2	1	62	2		11			2	16		2					2			120
AE7	5	1	44	30	2	1	27		1	2		6	2	59	1	36				1	3		221
Total:	118	117	116	81	37	139	82	3	14	4	44	52	23	62	4	48	2	1	5	5	6		963
Average mapping	98.6							28.9							10.1								

*Chromosomal designation as previously reported by Chaffin et al. and Yan et al.
**Chaffin designated Mrg11 as C/A, while Yan et al. designated it as C. Here we assign it as D/C (See Figure X).
***Yan et al. designated Mrg28 as D/C, while Chaffin et al. 2016 designated it as C. Here we re-designate it as D/C (See Figure X).

2.3.7: Utility of the genome assemblies

Given the genetic complexity of polyploid species, diploid species have frequently been used as simplified genetic models (Bertioli, et al., 2016; Du, et al., 2018; Jarvis, et al., 2017). We show the value of these diploid assemblies using published genome wide association studies (GWAS) for heading date and crown rust resistance - both major breeding targets for common oat. Heading date (flowering time) is critically important for regional adaptation, photosynthetic efficiency, stress avoidance; and through these factors

it strongly influences overall yield (Mathan, et al., 2016). A haplotag-based GWAS of heading date in the CORE diversity panel (*n* = 635) of common hexaploid (AACCDD) oat identified two major associations on linkage groups Mrg02, at position 34 cM in eight field trails and on Mrg12 at position 40–42 cM in seven field trials (Bekele, et al., 2018). Interestingly, our comparative analysis (see above) suggests that Mrg02 and Mrg12 are homoeologous (Table 2.6), with Mrg12 and Mrg02 being derived from the A-subgenome and D-subgenome, respectively. BLAST searches against the *A. atlantica* genome assembly using the maker sequences associated with heading date on Mrg12 localized the heading date QTL to chromosome AA1, at an interval spanning bases 548,905,448–553,755,648. A total of 175 annotated gene sequence are found within this region, including a likely candidate gene at the center of this interval, AA006173 (Figure 2.7A; 550,704,569–555,704,964), which is annotated as being homologous to HD3A (Heading Date 3A) from *O. sativa*, and is homologous (E-value = 9e-125, Identity = 97%) to the flowering time protein (FT-like; AAZ38709.1). Yan et al. (Yan, et al., 2006) described HD3A as the vernalization gene, VRN3, in wheat and barley. Interestingly, while Mrg02 is likely of a D-subgenome origin, BLAST search of markers associated with heading date from the Mrg02 linkage mapped significantly to the A genome chromosome AA1 at an interval spanning 550,053,072–550,947,435 bp, only 242,471 bp from the aforementioned HD3A gene, suggesting that the candidate gene for both major QTLs for flowering time are functional homoelogs of the flowering time (FT) HD3A gene in the A and D subgenomes (Figure 2.7B).

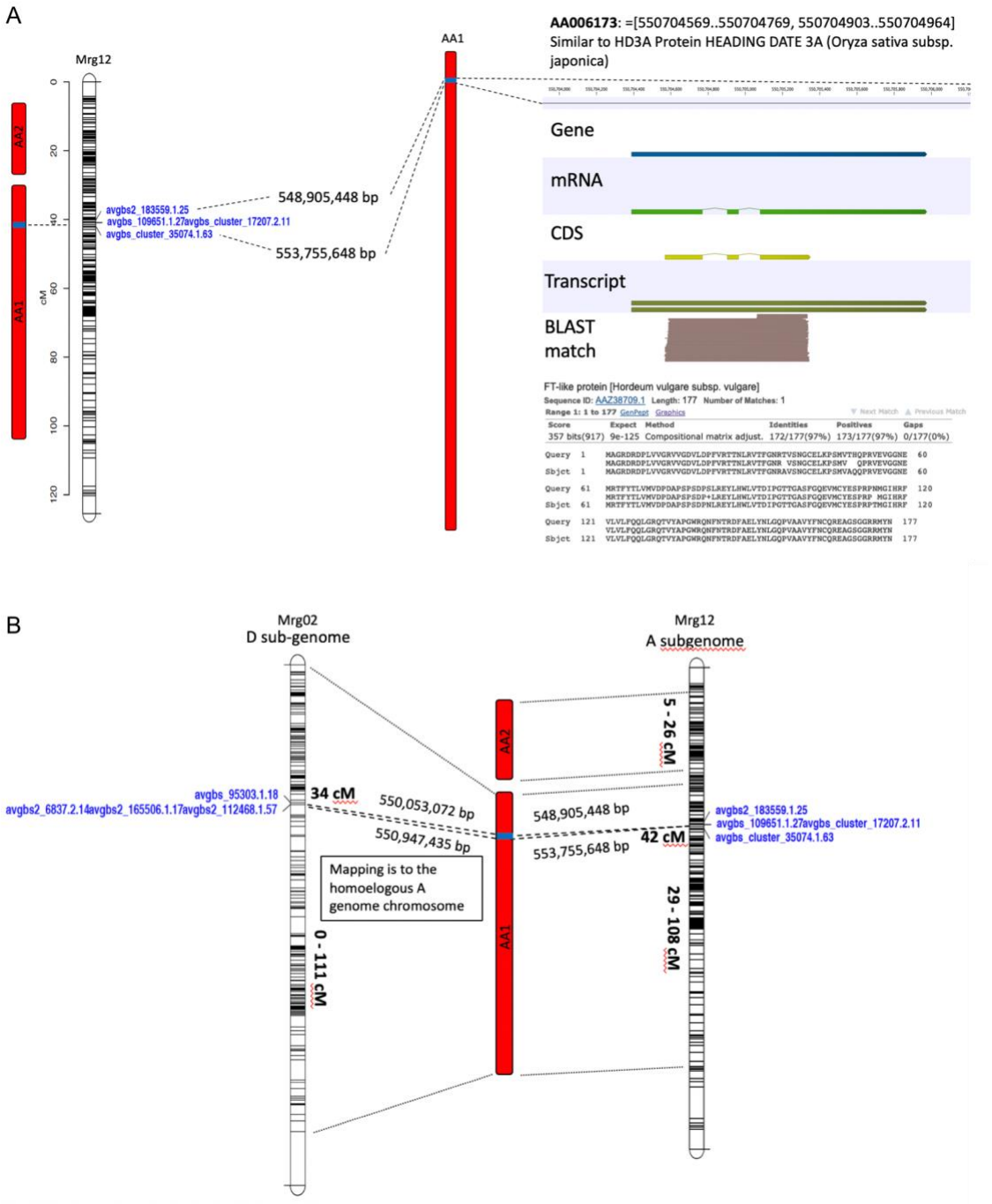


Figure 2.7: Identification of candidate genes putatively underlying heading date in oats. Candidate gene loci were identified using BLAST searches against the A. atlantica genome assembly using makers sequences associated with heading date QTLs located on the homoelogous linkage groups (A) Mrg12 and (B) Mrg02 (Bekele et al.). Markers from both QTLs mapped to the same physical position on chromosome AA1, within an interval containing an FT-like protein (HD3A), suggesting that heading date in modern oat is controlled by two functional homeologs of the flowering time gene.

Crown rust, caused by *Puccinia coronata f. Sp. avenae*, is the most damaging and widespread disease of oat worldwide (Carson, 2011). Moderate to severe outbreaks can reduce yield by 10–40% (Simmons, 1985). Klos, et al. (2017) performed a GWAS of crown rust resistance on elite common oat accessions challenged with multiple *P. coronata* isolates and identified multiple QTL on 12 linkage groups that were associated with crown rust resistance, several of which were associated with known resistance genes (e.g., *Pc91*). Resistance gene analogs (RGAs) contain specific conserved domains and motifs that can be used to identify and classify R-genes in to four main RGA families, specifically NBS-encoding proteins, receptor-like protein kinases (RLKs), receptor-like proteins (RLPs), and transmembrane coiled-coil (TM-CC). We employed the RGAugury pipeline (Li, et al., 2016) to predict RGAs in the *A. atlantica* and *A. eriantha* genomes which we then associated with BLAST searches of the markers linked to crown rust resistance associated with *Pc*91. The RGAugury pipeline annotated a total of 1,563 (511 NBS, 722 RLK; 120 RLP; 160 TM-CC) and 1,402 (459 NBS; 654 RLK; 135 RLP; 154 TM-CC) RGAs in the *A. atlantica* and *A. eriantha* genomes, respectively (APPENDIX E). As has been observed in other monocots, no Toll/Interleukin-1 receptor-NBS-LRR R-genes were predicted in either genome, supporting the hypothesis that this class of R-gene never evolved in monocots (Akita and Valkonen, 2002) or were lost during the evolution of monocots (Bai, et al., 2002). The RGAs, specifically the NBS encoding RGAs, cluster primarily in subtelomeric regions (APPENDIX E), with clusters identified on almost all chromosomes and often correlated with the mapping position of crown rust QTLs. For example, the *Pc91* gene, a known seedling resistance gene previously associated with QTL QPc.CORE.18.3

(Klos, et al., 2017) maps, via two SNPs, to the *A. atlantica* chromosome AA2 at positions 510,519,361 and 533,475,317, co-locating with a predicted disease gene cluster (APPENDIX F). We note that the diagnostic SCAR and DART markers developed for *Pc91* also map to this same location (527,126,948; (Gnanesh, et al., 2013; McCartney, et al., 2011)). The closest annotated disease resistance genes to these markers are AA013376, similar to RPH8A (a nonfunctional homolog of rpp8 in Arabidopsis (McDowell, et al., 1998)) located at position 510,828,316 and AA014151, similar to RPM1 (a well-documented resistance gene in *Arabidopsis* (Grant, et al., 1995)), is located at 533,698,614 (APPENDIX F). Both candidate RGAs were identified by the RGAugury pipeline as CC-NBS-LRR containing R-genes. We caution that while these two candidate RGAs are positioned in the immediate vicinity of the associated markers, at least 87 RGA are present at the RGA cluster defined by the QTL. We note that the diagnostic SCAR and DART markers developed for Pc91 also map to this same location (527,126,948; (Gnanesh, et al., 2013; McCartney, et al., 2011)).

2.3.8: SNP discovery and Genetic Diversity

To characterize the diversity and phylogenetic relationships among *Avena* A- and C-genome diploid species, we resequenced at 10X coverage 61 A-genome diploid accessions (including *A. atlantica, A. brevis, A. canariensis, A. damascena, A. hirtula, A. longiglumis, A. lusitanica, A. strigosa, A. strigosa-brevis, A. strigosa-hispanica, A. strigosa-nuda,* and *A. wiestii*) and 10 C-genome diploids (*A. clauda, A. eriantha, A. ventricosa*; APPENDIX A). The resequencing produced ~40 Gb sequence data per accession (APPENDIX A). The A-genome accession reads were mapped against the *A.*

atlantica genome, while the C-genome species were mapped against the *A. eriantha* genome for SNP discovery using InterSnp (Page, et al., 2014). InterSnp uses BAM files to calls SNPs between samples based on consensus alleles called at each genomic position, filtered to produce a dataset with 0% missing data across all lines. Considering the cleistogamous nature of the accessions included, any SNPs with > 5% heterozygous calls were deemed likely to result from spurious read-mapping and were removed from the dataset. Using a minimum allele frequency threshold of < 0.1, a total of 286,567 and 3,185,959 putative SNPs were identified within the A-genome and C-genome diploid datasets, respectively, and used by SNPhylo (Lee, et al., 2014) to investigate phylogenetic relationships. SNPhylo reduces oversampling effects at linked SNPs using an LD threshold (0.1) with a sliding window of 500,000 base pairs. Thus, a total of 7,221 and 11,530 SNPs, with an average of 1,032 and 1,647 SNPs per chromosome, were selected prior to tree construction for the A-genome and C-genome diploids dataset, respectively (Table 2.7).

Table 2.7: SNPs per chromosome use for maximum likelihood phylogeny produced using SNPhylo (Lee et al. 2014).

A. atlantica chromsome	Total Starting SNPs identified	Number of SNPs after pruning*	% of SNPs after pruning	A. eriantha chromsome	Total Starting SNPs identified	Number of SNPs after pruning	% of SNPs after pruning
AA1	1,036,245	1,073	0.10%	AE1	3,725,125	1,724	0.05%
AA2	1,018,956	1,117	0.11%	AE2	4,392,742	1,756	0.04%
AA3	1,034,920	1,089	0.11%	AE3	3,891,851	1,726	0.04%
AA4	1,082,592	1,046	0.10%	AE4	3,526,546	1,690	0.05%
AA5	1,018,296	972	0.10%	AE5	3,826,401	1,666	0.04%
AA6	952,804	938	0.10%	AE6	3,493,915	1,479	0.04%
AA7	1,013,017	986	0.10%	AE7	4,259,350	1,489	0.03%
Total:	7,156,830	7,221		Total:	27,115,930	11,530	

*SNPs were pruned using a linkage disequilbrium of 0.1, a missing rate > 0.1 using a 500,000 base pair sliding window.

The bootstrapped maximum likelihood phylogenetic trees were rooted with either the *A. eriantha* accession CN 19328 for the A-genome accessions tree (Figure 2.8A) or with the *A. atlantica* accession Cc 7277 for the C-genome accessions tree (Figure 2.8B). The A-genome diploids formed two distinct clades: one of these consisted primarily of accessions classified in taxa having the A_sA_s subgenome, which had previously been described by Rajhathy and Morrison (Rajhathy and Morrison, 1959) and Leggett (Leggett, 1987) as including *A. atlantica*, *A. hirtula*, the domesticated forms of *A. strigosa*, and *A. wiestii*; and a second clade comprised mostly of *A. canariensis* (A_cA_c), Syrian accessions of *A. damascena* (A_dA_d), *A. longiglumis* (A_lA_l), and three floret-shattering accessions that were possibly misidentified as *A. hirtula* and *A. lusitanica*. As expected, the spikelet-shattering *A. atlantica* occupied the basal position on the A_sA_s branch of the tree, while the *A. strigosa* (domesticated A_sA_s) genotypes formed a clade at the top of the tree and included a single accession of weedy *A. wiestii* (Clav 1994) that, upon inspection of the panicles, more closely resembled a long-awned, semi-shattering *A. strigosa* genotype.

The *A. strigosa* branch shows clearly the effect of a domestication bottleneck. This branch of the tree is subdivided into two distinct sub-branches. The upper sub-branch consists predominantly of genotypes of Iberian origin (i.e., 824, 670, 815, 117, etc.) and includes seven homogeneous accessions that are derivates of the Brazilian ‘Saia’ variety of forage oat (i.e. Clav 7010, PI 291990, etc.). Interestingly, the *A. hispanica* genotypes form a unitary subclade within this branch that is strongly supported by the bootstrap value. The lower sub-branch is comprised of accessions from outside the Iberian Peninsula (PI 83721, PI 287314, PI 304557, Clav 9022, etc.) and includes all the *A. strigosa-nuda*

varieties. Since *A. brevis* strains are distributed in both branches, there is no evidence to confirm its identity as a distinct taxon within or apart from *A. strigosa*. The presence of branches containing multiple, genetically indistinct accessions indicates there is a high degree of duplication being curated within the USDA and PGR-Canada gene banks for *A. strigosa*.

The remainder of the A-genome tree consists of entirely free-living genotypes. *Avena lusitanica* is not a universally accepted taxon, and the presence of these strains on various branches of the tree confirms that this is not a valid independent taxonomic entity; instead, it is part of the floret-dispersing *A. hirtula-wiestii* complex of semi-desert and Mediterranean scrub ecotypes of the A_sA_s biological species complex. The presence of three floret-shattering accessions from Morocco that were previously identified as *A. damascena* (A_dA_d) on the A_sA_s branch (PI 657468, PI 657471, PI 657472) and the two Syrian *A. damascena* genotypes on the other branch (CN 19457, CN 19459) suggests that the Moroccan group are mis-identified and are therefore, like *A. lusitanica*, either members of the A_sA_s *A. hirtula-wiestii-atlantica-strigosa* complex or, possibly, misclassified accessions of tetraploid *A. barbata*.

The rooted C-genome tree had the lone *A. ventricosa* (C_vC_v) accession at the base of the C-genome branch that consisted of two subclades. The more basal branch consisted of accessions of spikelet-shattering *A. eriantha* (C_pC_p) from Algeria (CN 24022) and four samples of *A. eriantha* from an extended population growing in the Middle Atlas Mountains of Morocco (PI 657575-8). The other branch included Algerian (CN 24040)

and Turkish (CN 19238) accessions of floret-shattering *A. clauda* (C_pC_p) along with Iranian (CN 19256) and Algerian (CN 19328, the reference genome) *A. eriantha* genotypes.

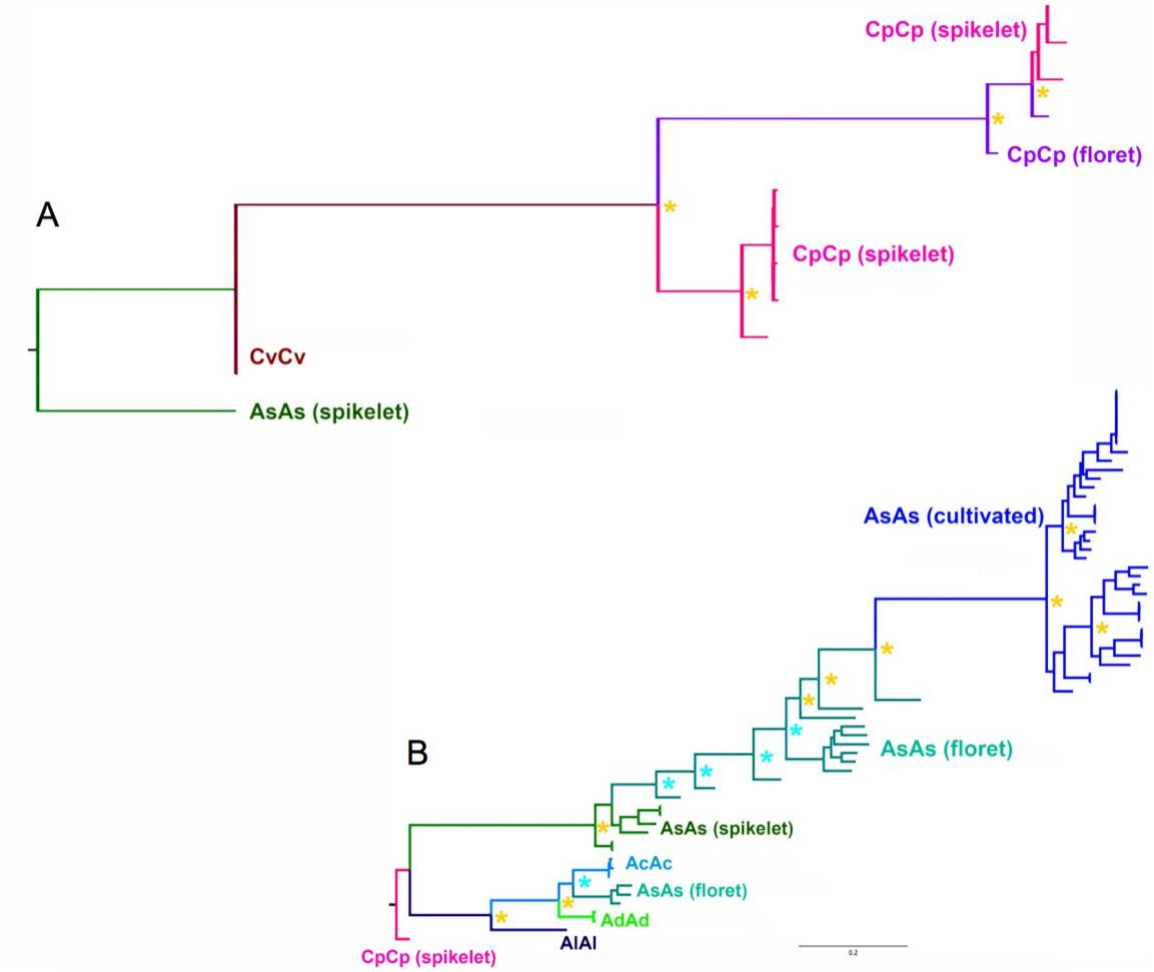


Figure 2.8: Abbreviated maximum-likelihood tree generated using (A) 10,894 SNPs for C-genome diploids rooted to the A. atlantica (AT_Cc7277) reference; and (B) 7,221 SNPs for A-genome diploids rooted to the A. eriantha reference (ER_CN 19238). Asterisks denote percentage of 1,000 bootstrap replicates that support the topology at 90-100% (gold) and 75-89% (blue). Scale bar represents substitutions per site. Branch labels are based on subgenomes composition and, in some cases, diaspore morphology ("floret-shattering", "spikelet-shattering", or "cultivated"). Unabbreviated trees are provided as APPENDIX BAPPENDIX C.

2.4: Conclusions

Reference-quality, *de novo* whole-genome sequence assemblies for two highly repetitive ~4 Gb *Avena* diploid species were produced using a hybrid approach involving PacBio long-reads, Illumina short-reads, and both *in vitro* and *in vivo* chromatin-contact

mapping. The whole-genome reference assemblies for A_s- and C_p-genome oat diploids provided for the first time in this paper represent powerful tools for identifying genes that underlie adaptive, disease resistance, and grain-quality traits critical for oat improvement. The utility of these wholegenome references was demonstrated first by analyzing sequences homologous to headingdate QTL-containing regions that were previously identified via GWAS in common hexaploid oat (*A. sativa*) to find linked candidate-genes in *A. atlantica* and *A. eriantha*.

Additionally, we used these references in successfully identifying RGAs homologous to oat crown rust resistance genes. *Avena atlantica* retains a remarkable degree of synteny with barley while *A. eriantha* has undergone a relatively greater degree of chromosomal rearrangement, suggesting the presence of an underlying genomic instability in the C-genome diploids. This might be related to the abundant heterochromatin, including the underlying pAm1 repeat motif, distributed throughout the chromosomes of this genome (Figure 2.2B, Track 5; (Sanz, et al., 2010)). Their genome sequences shed enormous insight into the complex evolutionary processes that have led to the appearance of cultivated diploid, tetraploid, and hexaploid oat going back millions of years. These processes included responses to natural selective events such as the Zanclean Cataclysm ~5 Mya and repeated cycles of global climate change characterized by boreal glacial maxima interspersed with humid periods and desertification due to northerly expansions of the Saharan and Arabian Deserts (Blanc, 2002; Garcia-Castellanos, et al., 2009; Kröpelin, et al., 2008).

We demonstrate that *A. atlantica*, *A. strigosa*, and *A. wiestii* represent multiple ecotypes or subspecies of a single biological species complex sharing the subgenome designation A_sA_s. The phylogeny presented here, which was generated by analyzing thousands of SNPs identified via resequencing of dozens of geographically diverse accessions, clearly illustrates a monophyletic relationship with *A. atlantica* accessions at the root of the A-genome clade (Figure 2.8A). This is further seen in the high degree of synteny and collinearity between the *A. atlantica* chromosomes and *A. strigosa* X *wiestii* linkage groups reported by Kremer, et al. (2001) (APPENDIX G). This result is remarkable, given the high degree of chromosomal rearrangement previously observed among different species and genomes within *Avena* (Sanz, et al., 2010).

The oat community has struggled without a reference genome for decades. Finally, we have complete references for what are, essentially, all three component genomes of cultivated hexaploid oat and the four known subgenomes of the genus, given the close correspondence between *Avena* subgenomes A, B, and D. Moreover, once a complete hexaploid reference is available, the utility of these component genomes will increase further, as they will provide a precise road-map of the structural and functional evolutionary steps that took place in the formation of this unique and important polyploid species.

CHAPTER 3: IDENTIFICATION OF *AVENA*-SPECIFIC GENES

3.1: Introduction

Identifying genes that are specific to grasses and, more specifically, the *Avena* genus is a key step in identifying the genes associated with *Avena*-specific protein production. Knowing which genes are unique to *Avena* and function in novel pathways will allow breeders and researchers to breed a better oat. Not only can breeders ensure that these important genes are not lost during selection, but it is also possible to identify oat germplasm with extra copies or natural or induced mutations in these target genes producing desirable effects that can be selected for during variety development. For example, if we identify an *Avena*-specific protein with benefits to human health, breeders can ensure that this trait is retained and not lost due to physical linkage with less desirable traits.

As the number of sequenced plant genomes and transcriptomes have increased, studies such as Graham et al. (2004) were performed to identify species-specific genes, in this case in legumes. The goal of this study was to incrementally reduce the dataset through a series of BLAST analyses to identify unique gene annotations found only in legumes. Each step in this process removed legume annotations with matches to conserved sequences found in an existing database from other non-legume species. Once conserved sequences were identified, they were removed from the set of sequences before proceeding to the next iteration of BLAST analysis (Graham, et al., 2004). The methods and results from this study are summarized in Table 3.1.

Table 3.1: Computational Identification of Legume-specific genes

BLAST Algorithm, E-Value Cutoff	**Data Set**	**Total Sequences in Data Set**	**Legume-Specific *G. max/soja* TCs retained**	**Legume-specific *M. truncatula* retained**	**Legume-specific *L. japonicus* retained**
	Legume TCs	45,783	24,750	17,243	3,790
BLASTN, 10^{-4}	TIGR AtGI, LeGI, OsGI, ZmGI	142,492	7,938	5,309	886
TBLASTX, 10^{-4}	TIGR AtGI, LeGI, OsGI, ZmGI	142,492	2,412	1,417	144
BLASTX, 10^{-4}	GenBank nonredundnant database	1,335,905	2,230	1,267	136
TBLASTX, 10^{-4}	Remaining TIGR Plant GIs	334,347	2,101	1,141	110
TBLASTX, 10^{-4}	Arabidopsis and rice genomes	-	2,081	1,128	106
Tera-BLASTN, 10^{-4}	EST_others	6,985,891	2,020	1,046	103
Tera-BLASTN, 10^{-4}	EST_others	6,985,891	1,997	1,028	101
TBLASTX, 10^{-20}	Eliminated Legume TCs	42,657	1,572	861	92

This table was taken from Computational Identification and Characterization of Novel Genes from Legumes, which appeared in Plant Physiology in July 2004.

Of the legume gene sets, Graham's group identified 1,572 genes from *Glycine max/soja*, 861 from *Medicago truncatula*, and 92 from *Lotus japonicus* that were determined to be legume-specific. After the identification of these legume-specific genes, InterProScan (Jones, et al., 2014) was used to identify conserved motifs. The conserved motifs included matches to a cyclin-like F-box motif, pectinesterase inhibitors, a zinc finger, and a nodulin. Additional sequences were identified that were rich in specific amino acids, primarily proline and cysteine. A subset of these cysteine rich proteins had similarity to motifs found in scorpion toxins and were identified to have strong similarities to plant defensins.

Additional studies have been performed using similar methods, including in *Oryza sativa* (Campbell, et al., 2007), *Arabidopsis thaliana* (Lin, et al., 2010), and sweet orange

(Xu, et al., 2015). The rice study identified a diversified subset of genes that were found primarily in grasses, which have a shorter mean gene length and a smaller number of exons than genes found in other families. They also identified a large number of F-box domain containing proteins in this grass-specific set (Campbell, et al., 2007). The *Arabidopsis* study identified *Arabidopsis* Lineage Specific Genes (ALSGs) and Conserved Brassica Specific Genes (CBSGs). Many of the CBSGs are involved in binding function, which could play a role to prevent inbreeding or breeding across species (Lin, et al., 2010). The study into sweet orange lineage specific genes (LSGs) identified 12 LSGs that are stimulated by abiotic stresses, including cold, heat, and UV treatments, though those gene functions are still unknown. Three genes were stimulated by more than one abiotic stress. Additionally, it was determined that not all LSGs are protein-coding and perhaps function as non-coding RNA at a transcription-level (Xu, et al., 2015). The methods for the identification of *Avena*-specific genes are loosely based on the methods shown in Table 3.1.

3.2: Materials and Methods

Annotations for *Avena atlantica* and *Avena eriantha* were created using the MAKER pipeline (Cantarel, et al., 2008) in Aim I. These annotations were used as input for the *Avena*-specific genes pipelines.

Assembly, gene, cds, protein, transcript, and annotation data for *Arabidopsis thaliana*, *Brachypodium distachyon*, and *Triticum aestivum* (common wheat) was downloaded from the Joint Genome Institute's Phytozome database on February 4, 2019 (Cheng, et al., 2017; Goodstein, et al., 2011; Rokhsar, et al., 2011; The International

Brachypodium Initiative, et al., 2010; The International Wheat Genome Sequencing Consortium, 2014). *A. thaliana* is a plant model organism, *B. distachyon* is a model grass organism, and *T. aestivum* or bread wheat is in the *Poaceae* (grass) family with oat. These three plants were chosen as the database for the first two steps of the BLAST pipeline in efforts to quickly eliminate common plant genes. Transcript and protein sequences from these three organisms were concatenated into transcript and protein files representing these three major species, and BLAST databases were created from them.

NCBI's nonredundant protein (nr) database was downloaded on February 2, 2019 (NCBI Resource Coordinators, 2016). A list of GI numbers, used to identify organisms, associated with *Avena* species was created for use in BLAST options on February 13, 2019 to ensure BLAST does not compare other *Avena* genes to our input sequences. Additionally, 27 Phytozome pre-release species transcripts were downloaded on April 13, 2019. A list of these species can be found in APPENDIX H.

For this project, we interrogated our datasets with multiple BLAST algorithms available from the National Center for Biotechnology Information (NCBI Resource Coordinators, 2016). Using multiple algorithms ensures that conserved protein sequences are identified, even if the nucleotide sequences have diverged. Step 1 was a nucleotide BLAST (BLASTN) with default parameters, output format 6 (tab delimited), and an e-value cutoff of 10^{-4} to compare the annotation of *Avena atlantica* (49,539 predicted genes) and *Avena eriantha* (49,203 predicted genes) to the transcripts of our three subject organisms. *Avena* nucleotide sequences with no matches to the dataset in Step 1 were used as query for a BLASTX search against protein sequences of the three subject organisms

(Step 2) using default parameters, output format 6 (tab delimited), and an e-value cutoff of 10^{-4}. The remaining *Avena* transcript sequences were compared to the non-redundant protein database using BLASTX (Step 3) with default parameters, output format 6 (tab delimited), the –negative_gi option with a list of *Avena* GIs, and an e-value cutoff of 10^{-4}. The remaining *Avena* transcript sequences were compared to the pre-release Phytozome species using BLASTN. After these four steps were completed, the remaining *Avena* genes were compared to a database of *Avena* genes previously eliminated from the dataset using BLASTX, which translates both query and subject from nucleotide to protein. Between each step, a Python script was run to create new query files and append eliminated sequences to the eliminated file. The entirety of this pipeline can be found in APPENDIX Iand the Python parser script can be found in APPENDIX J. A visualization of these methods is shown in Figure 3.1.

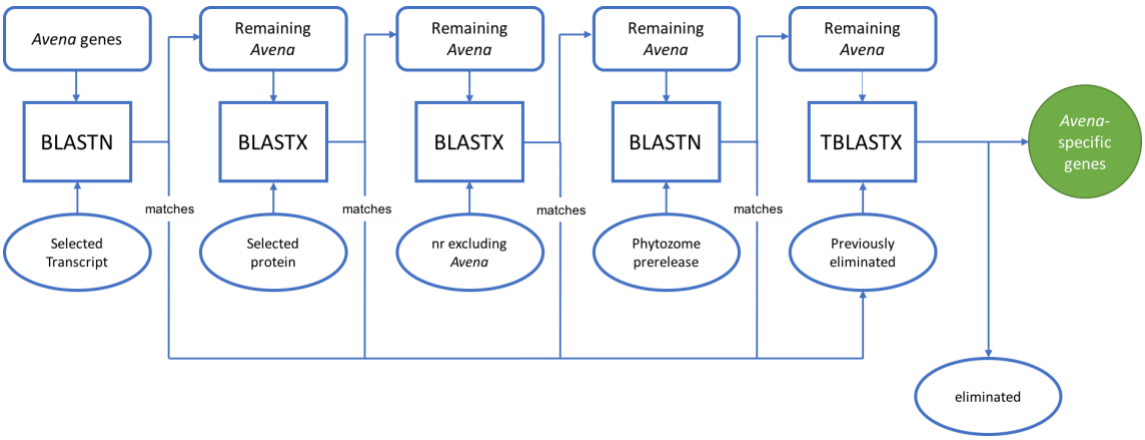


Figure 3.1: A flowchart of the methods for identification of Avena-specific genes.

Following identification of *Avena atlantica* and *Avena eriantha* specific genes, the genes common between the two species were determined using a TBLASTX search to compare the two, with the remaining *A. atlantica* annotations as the database and remaining

A. eriantha annotations as the query. The *Avena*-specific gene sets were clustered using CD-HIT-EST with default parameters from CD-HIT v. 4.8.1 (Fu, et al., 2012; Li and Godzik, 2006). Custom Python scripts were written to determine if the identified candidate resistance genes from Aim I were found in either of the *Avena*-specific genes lists.

Using the 11 *A. atlantica* and six *A. eriantha* RNA-seq datasets from Aim I, the Cufflinks and Cuffdiff pipeline (Trapnell, et al., 2012; Trapnell, et al., 2010) was used with default parameters and 8 threads (-p 8) to determine if any of these genes were differentially expressed in various tissues.

Domain analysis was performed on each output dataset using InterProScan (v. 5.34-73.0) with options -t n (indicating nucleotide input) and -f tsv for a tab separated output file (Jones, et al., 2014). The input files were split into multiple FASTA files to speed up the process and the results from each species were concatenated. Multiple sequence alignments for selected domains were performed using ClustalW (Larkin, et al., 2007; Sievers, et al., 2011) with default parameters. The nucleotide sequences for the selected genes were downloaded from the UniProt (The UniProt Consortium, 2019). These alignments were viewed using AliView (Larsson, 2014).

3.3: Results and Discussion

The *Avena atlantica* annotated genome dataset contained 49,539 predicted genes. After the initial BLAST comparisons to the transcript and protein sequences from the three plant species downloaded from Phytozome, 5,695 putative genes remained in the dataset. An additional 2,024 putative genes were eliminated from the dataset through a BLASTX against the NCBI non-redundant protein database, leaving 3,671 sequences. Phytozome

pre-release species transcripts were downloaded on April 13, 2019 and remaining *A. atlantica* sequences were compared to them using BLASTN, leaving 3,566 unique *A. atlantica* sequences. These remaining sequences were compared to the previously eliminated sequences, for a final 2,511 putative *Avena atlantica*-specific genes. A summary of the results of each step can be seen in Table 3.2.

Table 3.2: Computational identification of Avena atlantica*-specific genes*

Step	BLAST Algorithm	E-value Cutoff	Query	Subject	Total Sequences in Dataset After BLAST iteration	Sequences Eliminated from Dataset
Starting point before any BLAST analyses					49,539	
1	BLASTN	10^{-4}	*Avena atlantica*	*A. thaliana*, *B. distachyon*, *T. aestivum* transcript sequences	18,784	30,755
2	BLASTX	10^{-4}	Remaining *A. atlantica*	*A. thaliana*, *B. distachyon*, *T. aestivum* protein sequences	5,695	13,089
3	BLASTX	10^{-4}	Remaining *A. atlantica*	NCBI's Non-Redundant Protein Database (nr), excluding *Avena* GIs	3,671	2,024
4	BLASTN	10^{-4}	Remaining *A. atlantica*	Pre-release Phytozome Species transcripts	3,566	105
5	TBLASTX	10^{-4}	Remaining *A. atlantica*	Previously eliminated *A. atlantica*	**2,511**	1,055

The same procedure was followed for the *Avena eriantha* predicted genes. The *Avena eriantha* annotated genome dataset contained 49,203 predicted genes. After the initial BLAST comparisons to the transcript and protein sequences from the three plant species downloaded from Phytozome, 5,532 putative genes remained in the dataset. An additional 1,136 putative genes were eliminated from the dataset through a BLASTX against the NCBI non-redundant protein database leaving 4,396 sequences. Remaining *A.*

eriantha sequences were compared to the Phytozome pre-release species transcripts using BLASTN, leaving 4,129 unique *A. eriantha* sequences. These remaining sequences were compared to the previously eliminated sequences, for a final 3,043 putative *Avena eriantha*-specific genes. A summary of the results of each step can be seen in Table 3.3.

Table 3.3: Computational identification of Avena eriantha-specific genes

Step	BLAST Algorithm	E-value Cutoff	Query	Subject	Total Sequences in Dataset After BLAST iteration	Sequences Eliminated from Dataset
Starting point before any BLAST analyses					49,203	
1	BLASTN	10^{-4}	*Avena eriantha*	*A. thaliana*, *B. distachyon*, *T. aestivum* transcript sequences	17,569	31,634
2	BLASTX	10^{-4}	Remaining *A. eriantha*	*A. thaliana*, *B. distachyon*, *T. aestivum* protein sequences	5,532	12,037
3	BLASTX	10^{-4}	Remaining *A. eriantha*	NCBI's Non-Redundant Protein Database (nr), excluding *Avena* GIs	4,396	1,136
4	BLASTN	10^{-4}	Remaining *A. eriantha*	Pre-release Phytozome Species transcripts	4,129	267
5	TBLASTX	10^{-4}	Remaining *A. eriantha*	Previously eliminated *A. eriantha*	**3,043**	1,086

The 2,511 remaining *A. atlantica* genes were compared against the remaining 3,043 *A. eriantha* genes utilizing a TBLASTX search with *A. atlantica*-specific genes as the database and *A. eriantha*-specific genes as the query. 715 genes from *A. eriantha* and 687 genes from *A. atlantica* were similar across datasets. This translates into 28.47% of the *A. atlantica*-specific genes in common with 22.58% of the *A. eriantha*-specific genes. What this means is roughly 75% of the species-specific genes identified in each of these genomes

from the same Genus are unique to just that genome. This information is visualized in Figure 3.2.

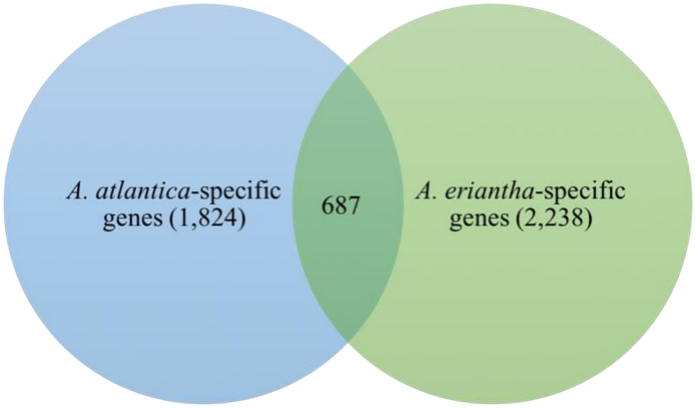


Figure 3.2: Venn diagram showing comparison between Avena eriantha *and* Avena atlantica *specific genes.*

To identify potential gene families within these species-specific genes, clustering was performed. No substantially large groups of *Avena*-specific genes were identified. The *A. atlantica* genes clustered into 2,392 groups. 73 of these clusters contain two genes, 13 contain three, four contain four genes, and one contains five genes. The *A. eriantha* genes clustered into 2,886 groups. Of these clusters, 73 have two genes, 11 have three, five clusters have four, one has six, one has 8 genes, one has 13 genes, and one cluster contains 19 genes. The clusters containing 13 and 19 genes contain only tRNAs. While it is perhaps unexpected to see tRNAs in the *Avena*-specific genes, tRNAs experience high levels of mutation due to their high levels of transcription (Thornlow, et al., 2018). In addition, the shorter length of tRNA sequences makes it more difficult to find significant BLAST hits using the selected parameters. This lack of clustering suggests that the genes identified as Avena-specific are not part of larger gene families such as resistance genes. In fact, when we compare the Avena-specific gene set to the putative resistance genes identified in Aim

I, we find that of the 1,563 and 1,402 resistance gene analogs identified in *A. atlantica* and *A. eriantha*, respectively, none of these candidate resistance genes were identified in the *Avena*-specific genes sets.

It is of note that β-glucan and avenanthramides synthesis genes are not found in the *Avena*-specific dataset. These compounds are often thought to be oat-specific, but that is not the case. In fact, β-glucans are found in many cereals and other organisms. Analyses of β-glucans have been performed for wheat (Pritchard, et al., 2011), barley (Delaney, et al., 2003; Islamovic, et al., 2013), and bamboo (Kweon, et al., 2003). Analysis of various compounds in cereal grains was performed by Mattila, et al. and avenanthramides were only detected in oat grains. However, the methods described seem to indicate that only oat was analyzed for avenanthramides (Mattila, et al., 2005). Annotations for genes involved in avenanthramide biosynthesis from *Avena sativa* were used as queries for online BLASTN search against nr with default parameters. These *A. sativa* annotations can be found in Table 3.4. Results included significant matches (e-value = 0.0) to genes from other cereal grains including Chinese Spring wheat, *Oryza sativa*, *Brachypodium distachyon*, *Panicum hallii*, *Zea mays*, and *Hordeum vulgare* showing that genes involved in the production of avenanthramides are found in other grasses.

Table 3.4: Avena sativa genes & accession numbers involved in avenanthramide biosynthesis, which were used as queries for BLASTN analyses.

Accession	Definition
MH397065	*Avena sativa* hydroxycinnamoyl-CoA:hydroxyanthranilate N-hydroxycinnamoyltransferase (HHT5) mRNA
MH397064	*Avena sativa* hydroxycinnamoyl-CoA:hydroxyanthranilate N-hydroxycinnamoyltransferase (HHT4) mRNA
MH397066	*Avena sativa* hydroxycinnamoyl-CoA:hydroxyanthranilate N-hydroxycinnamoyltransferase (HHT6) mRNA
MH397063	*Avena sativa* 4-coumarate:CoA ligase (4CL) mRNA

Each of the Avena-specific genes was queried in the RNA-seq datasets from Aim I to identify potentially interesting expression patterns. The results of the Cufflinks and Cuffdiff pipeline indicated that there were no significantly differentially expressed *Avena*-specific genes in the dataset. Though there were genes in each dataset that were expressed more frequently in certain tissues, no statistical significance was identified. The gene expression profiles for particular *Avena*-specific genes containing domains of interest is further explored below.

Given that each of these *Avena*-specific genes is unique to the *Avena* genus, identifying potential functions is not possible with a traditional annotation technique of matching sequence similarity to genes of known function. Much as was done in the Graham et al. (2004) publication, each Avena-specific gene was queried for protein domains using InterProScan (Jones, et al., 2014). What this approach aims to do is to narrow down putative functions based upon domains found in these genes. It is important to note that while we are discussing where these domains are found in other species (and in some cases other plants), the full genes are significantly different enough to indicate being *Avena*-specific. Results from the domain analysis of the *Avena atlantica*- and *Avena eriantha*-specific genes are summarized in Table 3.5 and Table 3.6, respectively.

Table 3.5: A summary of domains identified in Avena atlantica-specific genes, showing the number of genes containing at least one of each domain. The number of genes identified as shared with Avena eriantha in parentheses in first column.

Number of genes containing domain (number in shared set)	Domain name
139	[undefined coil domain]
1	AAI_LTSS
1	AP2/ERF domain profile.
1	Animal heme peroxidase superfamily profile.
1	B3 DNA-binding domain profile.
1	CHAP domain profile.
1	Chitin recognition or binding domain signature.
1	Chitin recognition protein
1	Chitin-binding type-1 domain profile.
1	ChtBD1
1	Contryphan family signature.
1 (1)	Cysteine-rich TM module stress tolerance
1	Death domain profile.
3 (2)	Eggshell protein signature
1	F-box-like
1	Histone H2A signature
1	Hydroxymethylglutaryl-coenzyme A reductases signature 1.
2	Metallothionein
1	Methylated-DNA--protein-cysteine methyltransferase active site.
1 (1)	PIF1-like helicase
1	Pentraxin domain signature.
1	Phospholipase A2 histidine active site.
1	Plant invertase/pectin methylesterase inhibitor
1	Plant transposon protein
1	Probable lipid transfer
15 (6)	Prokaryotic membrane lipoprotein lipid attachment site profile.
20 (18)	Proline rich extensin signature
2 (2)	Protein of unknown function (DUF1110)
1	Proteolipid membrane potential modulator
1	PsbL protein
1	PseudoU_synth
1 (1)	Putative S-adenosyl-L-methionine-dependent methyltransferase
1	REJ domain profile.
1	Reelin domain profile.
3 (1)	Reverse transcriptase-like (Ribonuclease H domain)
1	Rieske [2Fe-2S] iron-sulfur domain profile.
1 (1)	S4
1	Serine proteases, subtilase family, serine active site.
1 (1)	Twin arginine translocation (Tat) signal profile.
1 (1)	Vinculin signature
1	YDG domain profile.
1 (1)	Zinc finger SWIM-type profile.
10130	consensus disorder prediction

Table 3.6: A summary of domains identified in Avena eriantha-specific genes, showing the number of genes containing at least one of each domain. The number of genes identified as shared with Avena atlantica in parentheses in first column.

Number of genes containing domain (number in shared set)	Domain name
121	[undefined coil domain]
1	Alpha-carbonic anhydrases profile.
1 (1)	Arterivirus papain-like cysteine protease alpha (PCPalpha) domain profile.
1	Basic-leucine zipper (bZIP) domain profile.
1	Basic-leucine zipper (bZIP) domain signature.
1 (1)	CBL proto-oncogene N-terminus, EF hand-like domain
1	CD80-like C2-set immunoglobulin domain
1	Chlorovirus glycoprotein repeat
1	Domain of unknown function (DUF588)
1	Domain of unknown function DUF223
6 (5)	Eggshell protein signature
2	IQ motif profile.
1	LSD1: zinc finger domain, LSD1 subclass
1	Metallothionein
1	NUP C-terminal domain profile.
2 (1)	PLAT domain profile.
2	PLAT/LH2 domain
1 (1)	PTS HPR domain histidine phosphorylation site signature.
1	Pentraxin domain signature.
20 (6)	Prokaryotic membrane lipoprotein lipid attachment site profile.
22 (10)	Proline rich extensin signature
1 (1)	Protein of unknown function (DUF1110)
1	Protein of unknown function (DUF1668)
2	Protein of unknown function (DUF3681)
1	Putative AtpZ or ATP-synthase-associated
2	Reverse transcriptase-like (Ribonuclease H domain)
1	Sema domain profile.
1	Ulp1 protease family, C-terminal catalytic domain
1	Zinc finger PHD-type signature.
2 (1)	Zinc knuckle
1	bZIP transcription factor
12288	consensus disorder prediction

Across the two *Avena* datasets, seven domains are found in both sets of *Avena*-specific genes. These include eggshell protein signature, metallothionein, pentraxin domain signature, prokaryotic membrane lipoprotein lipid attachment site signature, Proline rich extensin signature, a protein of unknown function (DUF1110), and reverse transcriptase-like (Ribonuclease H) domains.

3.3.1: Eggshell Protein Signature

Perhaps the most intriguing and surprising of these shared *Avena*-specific gene domains is the eggshell protein signature. This signature is found in many different organisms, including several *Mycobacterium*, viruses, blood flukes, shrimp, and several plants including barley, rice, and kidney beans. In these plants, this signature is found in various amino-acid rich cell-wall structural proteins. Other proteins associated with this signature in plants are rich in glycine as well. The alfalfa gene with this signature is associated with draught and cold regulation (Hunter, et al., 2009; PR01228).

Many plant genes that contain the eggshell protein signature fall into the glycine rich protein (GRP) family of proteins. GRP proteins are often associated with structure of cell walls, plant defense, and stress mediation. This stress mitigation is made possible by the presence of cold-shock domains in Class IVc RNA binding GRPs, however, other RNA binding GRPs also play a role in cold tolerance through various mechanisms (Mangeon, et al., 2010). An alignment of eggshell domain-containing GRPs and four of the eggshell-protein signature-containing *Avena* proteins can be seen in Figure 3.3. While the *Avena*-specific proteins align with the other GRPs and the eggshell protein signatures can be found in them, they contain many mismatches and gaps that keep them different from the known proteins in other species. The difference in sequence could be due to the different climates that other species are grown in, and their different needs for temperature response. Due to the cool temperate climate that oat grows in, it is logical that it would contain proteins that mediate cold response, which would help the plants to survive. Additionally, it is logical that a plant such as oat would have cell wall structural proteins.

Figure 3.3: Portion of a multiple sequence alignment of select Glycine-Rich eggshell domain proteins and the eggshell domain-containing proteins of the Avena species investigated. From top down: 1) Medicago sativa Cold and drought-regulated protein, 2) Hordeum vulgare (barley) Glycine-rich cell wall structural protein, 3) Phaseolus vulgaris (kidney bean) Glycine-rich cell wall structural protein, 4) Oryza sativa Glycine-rich cell wall structural protein 2, 5) Oryza sativa Putative glycine-rich cell wall structural protein 1, 6) Oryza sativa Putative glycine-rich cell wall structural protein 1, 7) Acanthoscurria gomesiana (tarantula) Acanthoscurrin-1, 8) Arabidopsis thaliana Glycine-rich protein 5, 9) Avena atlantica AT039540, 10) Avena atlantica AT033627, 11) Avena eriantha AE002675, and 12) Avena eriantha AE007819. Eggshell domains enclosed in red boxes.

Avena atlantica AT033627 was expressed in stressed mature leaf (FPKM = 47.8587), mature leaf (28.3166), hypocotyl (8.76709), and stem (6.21579) tissues. *A. atlantica* AT039540 was expressed in stem (7.52493) mature leaf (4.67218), stressed mature leaf (4.0189), hypocotyl (1.37163), and root (2.16442) tissues. *A. eriantha* AE002675 was expressed in root (51.1699) and whole seedling (0.16158) tissues. *A. eriantha* AE007819 was expressed in root (63.7129), whole seedling (27.307), crown (3.44563), and mature leaf (0.213781) tissues. These expression profiles may point toward the function of these genes. Given that the *A. atlantica* genes are more highly expressed in tissues that are more exposed to the elements, they may be playing a role in cold response. However, as the *A. eriantha* genes are more highly expressed in root and seedling tissues, they may play more of a role in development and the interaction between plant and soil environment.

3.3.2: Metallothionein Family 15

Three of the *Avena*-specific genes identified in these two species are associated with metallothionein family 15, which is found mostly in plants. Members of this family are cysteine rich proteins and bind to heavy metals. Genes in this family have been identified in many plants including apple, banana, rice, barley, and maize (Hunter, et al., 2009; Metallothionein, family 15, plant (IPR000347)). Metallothioneins occur in many plants and the expression levels of different metallothionein genes vary greatly between tissues, making determining their function a bit more difficult. Different classes of metallothioneins have higher expression levels in different tissues (Joshi, et al., 2016). In *Arabidopsis*, MT1a is expressed in roots, MT2a and MT3 in leaves, and MT4a and MT4b in seeds. In addition, MT1a and MT2b are expressed in phloem and MT2a and MT3 are expressed in mesophyll (Guo, et al., 2003). When exposed to excessive heavy metals, metallothioneins can be found primarily in roots, shoots, and seedlings (Hassinen, et al., 2009). It has been observed that some abiotic stresses can affect expression levels of metallothioneins in plants. Draught and salt stresses in rice and barley have caused significant upregulation of metallothioneins (Kawasaki, et al., 2001; Ozturk, et al., 2002; Rabbani, et al., 2003). In cotton, draught, salt stress, and low temperatures all contributed to upregulation (Xue, et al., 2008). Additionally, some plant functions are regulated by metallothioneins including cell growth, homeostasis, and protection from oxidative stress (Joshi, et al., 2016).

It is not uncommon for soil to be contaminated with metals, such as arsenic, lead, copper, zinc, cadmium, and nickel (Burges, et al., 2018). The ability of plants to

phytoextract these metals is commonly used as a growing means to decrease these levels of contamination. *Avena strigosa* has shown superior abilities regarding cadmium accumulation, specifically in its leaves (Uraguchi, et al., 2006). It is possible that these identified metallothionein family proteins in *A. atlantica* and *A. eriantha* play roles in the phytoextraction of heavy metals allowing them to grow in poor soil conditions. An alignment of several metallothionein family 15 genes, seen in Figure 3.4, shows that those found in the *Poaceae* family of plants align closely together and that the two found in *Avena atlantica* are closely related to each other with a relatively small number of mismatches.

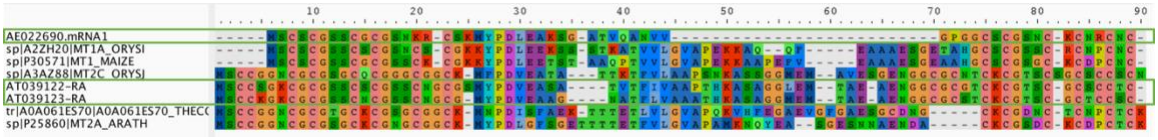


Figure 3.4: A multiple sequence alignment of select metallothionein family proteins and the metallothionein family genes of the Avena species investigated. From top down: 1) Avena eriantha AE022690, 2) Oryza sativa subsp. indica Metallothionein-like protein 1A, 3) Zea mays Metallothionein-like protein 1, 4) Oryza sativa subsp. japonica Metallothionein-like protein 2C, 5) Avena atlantica AT039122, 6) Avena atlantica AT039123, 7) Theobroma cacao Metallothionein 2A, and 8) Arabidopsis thaliana Metallothionein-like protein 2A.

Avena atlantica gene AT039122 is expressed in roots (FPKM = 30.908), seed (9.45), young flower (1.5), stressed mature leaf (1.306), green grain (1.0), and green anthers (0.305) but not in mature leaf, stem, or hypocotyl. While these levels were not considered significant, it is noteworthy that a hypothetical metallothionein protein is expressed in a stressed mature leaf, but not a non-stressed mature leaf. In addition, the high levels of expression in seed and root tissues are consistent with expression patterns reported in other plants. It is possible that this gene is a type I metallothionein as the higher expression levels in roots are consistent with higher expression of type I MTs in other plants (Yang, et al.,

2009). *Avena atlantica* AT039123 and *A. eriantha* AE022690 were not expressed in any of the tissues examined in the RNA-Seq.

3.3.3: Pentaxins

Both *A. atlantica* and *A. eriantha* contained one *Avena*-specific gene with the pentaxin domain signature. Proteins in the pentaxin family contain five subunits that have been shown to be disc-shaped and noncovalently bound. There are currently 118 known architectures contained in this gene family with various gene functions. This domain signature is found mostly in animals, but rice has one annotated gene with the signature as well, though the function of that protein is still unknown (Hunter, et al., 2009; Pentaxin, conserved site (IPR030476)). Structures of some pentaxin domain proteins have shown similarity to proteins in the lectin fold superfamily. Proteins in this superfamily include plant and animal carbohydrate-binding proteins (Pepys, 1998).

In mammals, pentaxin family proteins are involved in apoptosis. These proteins mark apoptotic cells for phagocytosis (Roos, et al., 2004). In plants, programmed cell death occurs in developmental regulated forms, as well as environmentally induced forms. Environmentally induced programmed cell death can be induced by stressors including heat or cold stress, salinity, and water availability (Locato and De Gara, 2018). It is possible that these pentaxin-related proteins are involved in programmed cell death in plants. An alignment of several selected pentaxin domain-containing genes from plants and animals can be seen in Figure 3.5.

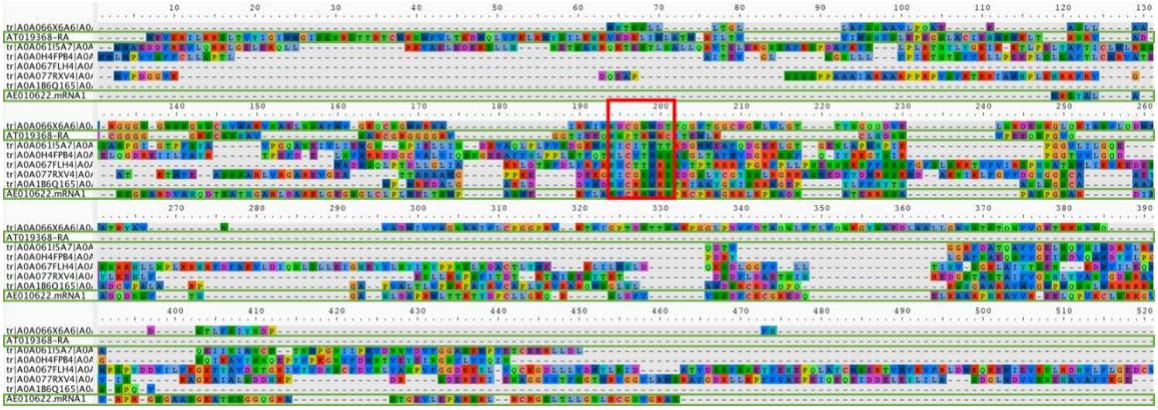


Figure 3.5: Portion of a multiple sequence alignment of select pentaxin domain proteins and the pentaxin domain-containing proteins of the Avena species investigated. From top down: 1) Colletotrichum sublineola (causal agent of sorghum anthracnose) Peroxidase, 2) Avena atlantica AT019368, 3) Cricetulus griseus (Chinese hamster) Neuronal pentraxin-2-like protein, 4) Carassius auratus (goldfish)Pentaxin, 5) Citrus sinensis (sweet orange) F-box domain-containing protein, 6) Triticum aestivum (wheat) PPM-type phosphatase domain-containing protein, 7) Sorghum bicolor Uncharacterized protein, 8) Avena eriantha AE010622. Pentaxin domain shown in red box.

The identified *Avena atlantica* pentaxin family protein AT019368 was not expressed in any tissue in our RNA-Seq samples. *A. eriantha* AE010622 was very lowly expressed in whole seedling (FPKM = 0.704), crown tissue (0.557), roots (0.101), and immature panicle (0.0798), but not in young or mature leaf tissues. If these genes are in fact involved in apoptosis, the low levels of expression could indicate a low level of programmed cell death occurring within these tissues.

3.3.4: Prokaryotic Lipoprotein Lipid Attachment Site Signatures

Between the two *Avena* species studied, there are 35 prokaryotic membrane lipoprotein lipid attachment site signatures. These signatures are found in prokaryotic cell membranes and have been identified in many bacteria as well as bacteriophages (PROSITE documentation PDOC00013; PROSITE Entry: PS51257; Sigrist, et al., 2013). Though the signatures are primarily characterized in archaea, they have been identified in various plant and animal proteins as well. These proteins have various annotations including

carboxypeptidase, ceramidase, and sodium channel proteins in chimpanzees and Indole-3-pyruvate monooxygenase in rice (Hunter, et al., 2009; PS51257).

An alignment of several prokaryotic lipoprotein lipid attachment site signature-containing genes can be seen in Figure 3.6. The selected *Avena atlantica* genes align most closely with the head blight pathogen (which can affect oat), grapevine downy mildew, and an *Oryza sativa* gene. The selected *A. eriantha* genes align most closely with a secreted RxLR effector protein from grape downy mildew and the regulator of itaconic acid biosynthesis 1 in corn smut. Considering the large number of gaps in the alignments, it is possible that these signatures have been transferred into oat and other plant species via horizontal gene transfer from the pathogens they are found in and have been incorporated into other plant genes.

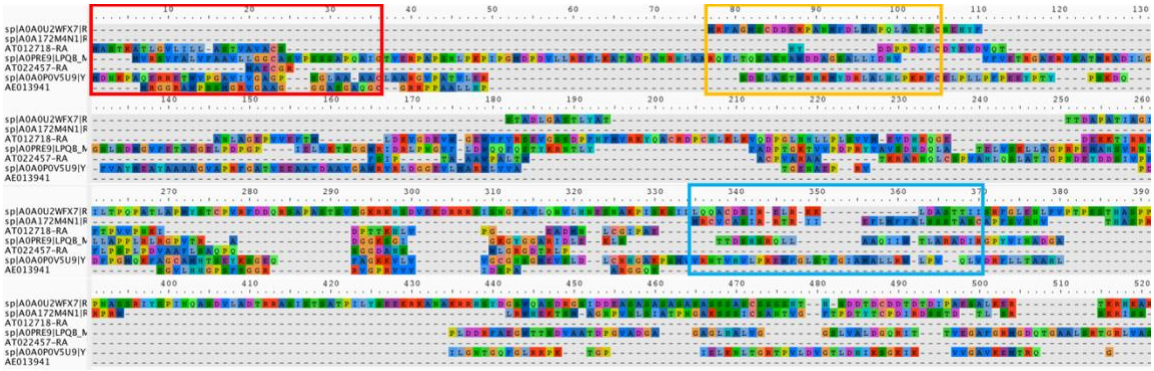


Figure 3.6: Portion of a multiple sequence alignment of genes containing Prokaryotic Lipoprotein Lipid Attachment Site Signatures and the Prokaryotic Lipoprotein Lipid Attachment Site Signature-containing proteins of the Avena species investigated. From top down: 1) Ustilago maydis (corn smut) Regulator of itaconic acid biosynthesis (domain in orange box), 2) Plasmopara viticola (grapevine downy mildew) Secreted RxLR effector protein 68 (domain in blue box), 3) Avena atlantica AT012718, 4) Mycobacterium ulcerans Lipoprotein LpqB, 5) Avena atlantica AT022457, 6) Oryza sativa Indole-3-pyruvate monooxygenase YUCCA1 (domain in red box), and 7) Avena eriantha AE013941.

Avena atlantica AT012718 is only expressed in green anther (FPKM = 6.3088) tissue. *A. atlantica* AT022457 is expressed in mature leaf (30.8171), vegetative meristem (12.83), stressed mature root (11.6404), seed (11.4918), leaf (9.64208), hypocotyl (9.66154), green anthers (7.77495), stem (6.21997), and young flower (5.88788) tissues.

A. eriantha AE013941 is expressed in crown (8.7995), whole seedling (8.4059), root (5.76189), and young leaf (2.27902) tissues. *A. eriantha* AE043222 is expressed in young leaf (2.11179), mature leaf (1.91578), crown (1.87052), root (0.849157), and whole seedling (0.308415) tissues. There does not seem to be any discernable pattern to the expression profiles of these prokaryotic lipoprotein lipid attachment site signature-containing genes.

3.3.5: DUF1110

Interestingly, a protein of unknown function (DUF1110) was also identified by InterProScan in both *Avena* species analyzed. This protein family is associated with the *Poaceae* family (grasses), but the proteins in this family have an unidentified function. According to InterPro, the proteins in this family are found in many grasses, and not yet in other organisms. The 190 proteins within the family occur in *Aegilops tauschii* (Tausch's goatgrass), various *Oryza* (rice) species, *Triticum aestivum* (wheat), *Triticum monococcum* (Einkorn wheat), multiple varieties of *Hordeum vulgare* (Barley), *Zea mays* (Maize), and *Panicum miliaceum* (Proso millet) (Hunter, et al., 2009; Protein of unknown function DUF1110 (IPR010535)). Several genes in the DUF1110 family have been identified as BTR1 genes, which play roles in spike morphology and grain shattering (Civáň and Brown, 2017; Zhao, et al., 2019). Both *A. atlantica* and *A. eriantha* are spikelet shattering. An alignment of proteins identified to be in the DUF1110 family can be seen in Figure 3.7. Due to the occurrence of genes in organisms they seem to not belong in (such as prokaryotic membrane signatures in eukaryotes), it is very interesting that these genes have only been identified in grasses. Though these proteins do align, there are many gaps and mismatches

keeping the *Avena* genes distinct from the other grass genes. Further studies on these genes, along with others, could lead to greater understanding of what makes grasses distinct from other plant species as well as insight into the domestication process.

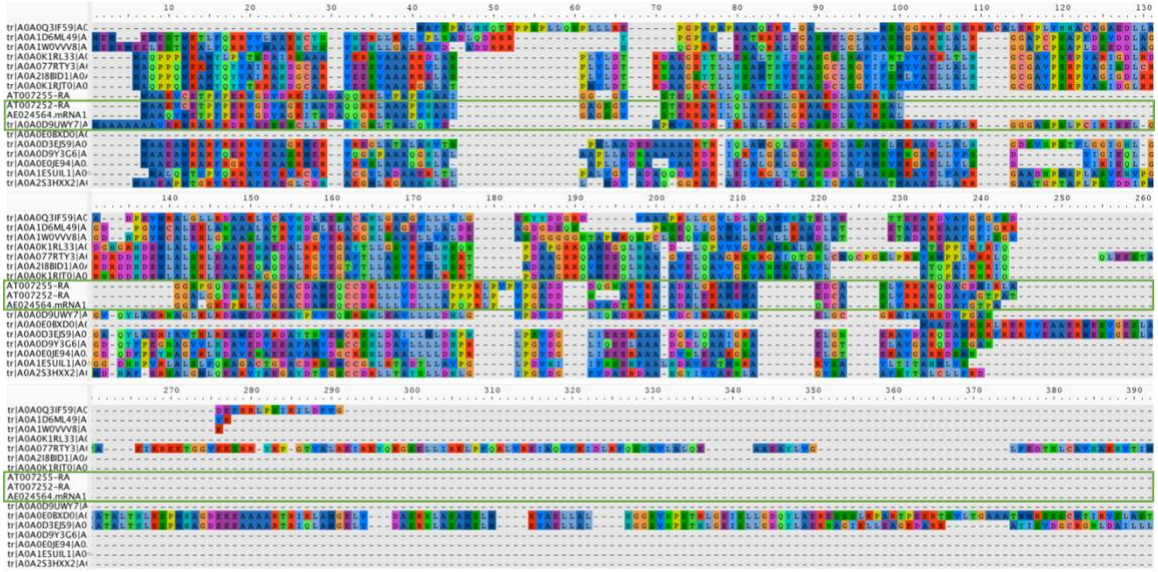


Figure 3.7: A multiple sequence alignment between several DUF1110 protein sequences. From top to bottom: 1) Brachypodium distachyon uncharacterized protein, 2) Zea mays uncharacterized protein, 3) Sorghum bicolor uncharacterized protein, 4) Triticum aestivum BTR1-A-like protein, 5) Triticum aestivum Histone domain-containing protein, 6) Triticum monococcum subsp. aegilopoides, 7) Hordeum vulgare subsp. vulgare 8) Avena atlantica AT007255, 9) Avena atlantica AT007252, 10) Avena eriantha AE024564, 11) Leersia perrieri uncharacterized protein, 12) Oryza meridionalis uncharacterized protein, 13) Oryza barthii uncharacterized protein, 14) Oryza glumipatula uncharacterized protein, 15) Oryza punctata uncharacterized protein, 16) Dichanthelium oligosanthes uncharacterized protein, 17) Panicum hallii uncharacterized protein

Avena atlantica AT007252 is only expressed in green anthers (3.62675). *A. atlantica* AT007255 is expressed in green grain (2830.82), yellow grain (789.689), stem (16.7948), and seed (5.50066) tissues. *A. eriantha* AE024564 is expressed in root (17.0673), whole seedling (6.322), crown (1.7275), young leaf (1.27942), and mature leaf (1.26946) tissues. High expression in grain tissues suggests that these genes may be playing a role in grain morphology as in other grasses.

3.3.6: Reverse Transcriptase Domains

Reverse transcriptase-like domains are usually associated with viruses, as they are required for converting RNA to double-stranded DNA, which helps viruses replicate within their host cells. Ribonuclease H (RNase H) domains are responsible for breaking down the RNA portion of the RNA-DNA hybrids formed during reverse transcription. These domains have been found in many species, including plants, animals, and bacteria. RNase H domains have been identified in over 900 rice genes and over 200 *Arabidopsis* genes (Hunter, et al., 2009; Ribonuclease H domain (IPR002156)). Many plants contain RNase H proteins and many are also infected by RNA virus-like viroids. It is possible that RNase H domains could serve as a protection for plants when these viroids infect them (Moelling, et al., 2017). There are several known oat diseases that are caused by viruses, including oat mosaic virus, oat blue dwarf virus, oat soil-borne stripe virus, oat sterile dwarf virus, barley yellow dwarf virus, and cereal tillering virus (Clifford, 1995). These virus diseases could be a reason for oat to contain genes with RNase H domains.

A multiple sequence alignment for protein sequences containing the Ribonuclease H domains can be seen in Figure 3.8. The four *Avena* sequences represented align together, then with *Anaeromyxobacter* sp. (strain K), a bacteria commonly found in soil (Sanford, et al., 2002), *Erwinia tasmaniensis,* a bacteria which infects apple and pear trees (Geider, et al., 2006), and *Serratia proteamaculans*, a bacteria found in plant roots, shown to promote growth (https://genome.jgi.doe.gov/portal/serpr/serpr.home.html). While the sequences do align, the *Avena* sequences maintain enough dissimilarity from the bacterial sequences to still be *Avena*-specific.

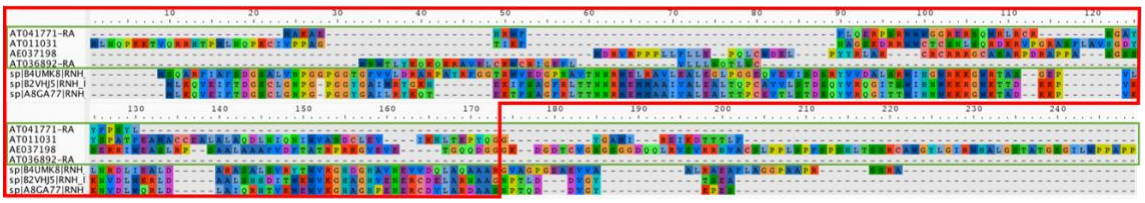


Figure 3.8: A multiple sequence alignment of Ribonuclease H domain proteins. From top to bottom: 1) Avena atlantica AT041771, 2) Avena atlantica AT011031, 3) Avena eriantha AE037198, 4) Avena atlantica AT036892, 5) Anaeromyxobacter sp. (strain K) Ribonuclease H, 6) Erwinia tasmaniensis Ribonuclease H, 7) Serratia proteamaculans Ribonuclease H. Domain shown in red box.

Avena atlantica AT011031 is only expressed in the stem (FPKM = 2.74518) and young flower (0.94987) tissues. *A. atlantica* AT036892 is expressed in root (87.16), green grain (84.126), seed (60.419), yellow grain (54.205), vegetative meristem (52.464), mature leaf (51.436), stressed mature leaf (50.143), hypocotyl (48.914), young flower (47.441), stem (35.754), and green anther (33.006) tissues. *A. eriantha* AE037198 is expressed in mature leaf (58.125), young leaf (56.772), root (35.399), crown (24.921), immature panicle (17.713), and whole seed (13.136) tissues. There does not seem to be any clearly discernable pattern to these expression values.

3.4: Conclusions

The study of *Avena*-specific genes has identified 2,511 and 3,043 *Avena atlantica* and *Avena eriantha*-specific genes, respectively. 687 of these genes were similar between the two species. Many of the genes one would expect to find in a list of oat-specific genes, such as β-glucan and avenanthramides production genes, were not present due to their previously described identification in other species.

While the function of many of the *Avena*-specific genes from both *Avena atlantica* and *Avena eriantha* cannot be predicted, their presence in *Avena* and absence in other genera may make them worth investigating in the future. This further knowledge will help inform future breeding efforts in oat. This information could help to ensure that the

beneficial properties of oat are not diminished and perhaps even enhanced. Additionally, further research could indicate some of these properties are negative and should be removed or reduced if possible.

CHAPTER 4: NUTRITIONAL PATHWAYS

4.1: Introduction

The identification of *Avena*-specific genes is an important step in understanding the unique benefits of oat consumption, but the roles of these genes and their resulting proteins have on human health are perhaps more important. Disease and health are often affected by a multitude of factors. Some of these factors are diet (He, et al., 2007), exposure to air pollution and cigarette smoke (Miller, et al., 2007; Pope, et al., 2011), access to clean water (Fink, et al., 2011), and obesity (Grundy, 2002; Wolin, et al., 2010). Oats continue to be considered a key component in a healthy diet. In efforts to identify how these genes encoding important proteins may affect those who consume oat, we seek to utilize a Knowledge Base to computationally build linkages in the literature between genes/proteins that are either unique to oat or are known to be beneficial and their potential health benefits.

Scientific papers are being published at a higher rate than ever before leading researchers to constantly struggle with not only remaining on top of literature in their own field of interest, but struggling to branch out into areas that are potentially relevant but that linkage may not be easily identifiable through a traditional Google Scholar search engine. In efforts to keep up with the vast amount of information being published, students within the Plant Pathways Elucidation Project (P2EP) in Kannapolis, NC have been developing a knowledge base (Linchangco, 2018). A knowledge base is a specialized database, which stores complex information for use by computer system. The difference between standard databases and knowledge bases is that knowledge base data is structured, with pointers between objects. It can then be queried in efforts to find information stored within. The

P2EP Knowledge Base utilizes text-mining technologies in efforts to minimize the amount of reading and searching that researchers must do to find papers relevant to their research.

The P2EP Knowledge Base compiles information from multiple sources including NCBI, USDA, EMBL-EBI, OMIM, OBO Ontologies, Comparative Toxicogenomics Database, and the Therapeutics Target Database. Text mining was used to aggregate information from scientific literature and other sources. For example, an abstract from a journal article in PubMed could be mined with the resulting output being agricultural and biomedical entities, as well as the relationships between them. The information is organized and stored in an information retrieval system, allowing it to be quickly and easily searchable by users utilizing keyword mapping and document indexing. Information extraction then takes the unstructured information and identifies the entities and relationships from it. This is done by identifying the named entities of interest and extracting the relevant relationships from the data. There are multiple techniques for this named entity recognition that are based on dictionary, rule, or machine learning approaches. The ultimate goal of the text mining is to use existing associations to discover possible new associations (Linchangco, 2018).

The P2EP Knowledge Base contains over 7 million entities and over 9,000 unique relationship types between these entities. There are several meta paths that are possible between Plants and Disease, starting with Plant $\rightarrow$ Chemical $\rightarrow$ Gene $\rightarrow$ Pathway $\rightarrow$ Disease (Linchangco, 2018). This vast knowledge base is used to identify potential health benefits related to genes and proteins that are characteristic of *Avena*.

The pathways and chemicals studied in this Aim will provide a starting point for future research. Identification of all genes affecting human pathways is important for oat breeding as it will help ensure that nutritional content of oat can be optimized for human consumption.

4.2: Materials and Methods

The P2EP Knowledge Base and neo4j (https://neo4j.com/) were installed locally and the graph.db file from the knowledge base was added to neo4j data sources. The knowledge base was then accessed via a web browser by navigating to http://localhost:7474/browser/, where SQL-like queries were input and graphs and tables were produced as output (Linchangco, 2018). These graphs were manipulated only to ease readability and downloaded from the local host webpage.

Gene products identified in the *Avena*-specific genes project (Aim II) were interrogated with the P2EP-KB with no returned relationships. Several diseases were then selected as input using wildcard characters to allow for partial matches. Results were limited to 100 matches to minimize the number of edges between nodes for readability. These input queries are outlined in Table 4.1. Once results from the Knowledge Base were obtained, the literature was examined more thoroughly to investigate these possible relationships.

Table 4.1: List of queried used in the P2EP KB for identification of diet-disease networks.

Plant queried	Phenotype queried	Query
.**Avena sativa*	.*cardi.*	match (a:Plant)-[]-(b:Chemical)-[]-(c:Gene)-[]-(d:Pathway)-[]-(e:Phenotype) where a.name =~ ".*Avena sativa.*" and e.definition =~ ".*cardi.*" return a,b,c,d,e limit 100
.**Avena sativa*.*	.*diabetes.*	match (a:Plant)-[]-(b:Chemical)-[]-(c:Gene)-[]-(d:Pathway)-[]-(e:Phenotype) where a.name =~ ".*Avena sativa.*" and e.definition =~ ".*diabetes.*" return a,b,c,d,e limit 100
.**Avena sativa*.*	.*dermatitis.*	match (a:Plant)-[]-(b:Chemical)-[]-(c:Gene)-[]-(d:Pathway)-[]-(e:Phenotype) where a.name =~ ".*Avena sativa.*" and e.definition =~ ".*dermatitis.*" return a,b,c,d,e limit 100
.**Avena sativa*.*	.*eczema.*	match (a:Plant)-[]-(b:Chemical)-[]-(c:Gene)-[]-(d:Pathway)-[]-(e:Phenotype) where a.name =~ ".*Avena sativa.*" and e.definition =~ ".*eczema.*" return a,b,c,d,e limit 100
.**Avena sativa*.*	.*cancer.*	match (a:Plant)-[]-(b:Chemical)-[]-(c:Gene)-[]-(d:Pathway)-[]-(e:Phenotype) where a.name =~ ".*Avena sativa.*" and e.definition =~ ".*cancer.*" return a,b,c,d,e limit 100

4.3: Results and Discussion

4.3.1: Oat and Heart Health

The relationship between oat consumption and it's positive effects on heart health has been well documented (Wolever, et al., 2016). Figure 4.1 shows several relationships between *Avena sativa* and cardiac diseases. *Avena sativa*, shown in dark green on the left, affects lipoprotein, sterol, and cholesterol levels (shown in orange in Figure 4.1). These relationships are largely due to the presence of β-glucan in oat grains. It has been shown that consumption of β-glucan-containing oat can help lower LDL cholesterol (Wolever, et al., 2016). A recent study showed that the cholesterol lowering effects of oat can also be attributed to the presence of certain lipids and proteins as well. The proteins in oat with low lysine-arginine and methionine-glycine ratios contributed to lower total cholesterol and LDL cholesterol levels. The study concluded that the hypocholesterolemic properties of oat cannot simply be attributed to one factor, but a combination of many, including oleic acid, vitamin E, and plant sterols (Guo, et al., 2014).

The meta path between oat and heart conditions shown in Figure 4.1 identified one gene (HSD11B1, shown in red) and one pathway (lipid metabolic process, shown in light green). HSD11B1, or hydroxysteroid 11-beta dehydrogenase 1, an enzyme which is responsible for the conversion between cortisol and cortisone. Cortisone is the metabolically inactive form of cortisol, the human stress hormone. HSD11B1 expression is increased in adipose tissues of obese individuals (Paulsen, et al., 2007). Dysregulation of HSD11B1 is associated with an imbalance of glucocorticoid in adipose tissues, glucose imbalance, and visceral fat accumulation, though it is unclear what the role of HSD11B1 is in this association (Dammann, et al., 2019). These factors contribute to metabolic syndrome, which puts patients at a higher risk for cardiac diseases (Turek, et al., 2014). Various SNPs in HSD11B1 have associations with type II diabetes (Nair, et al., 2004), metabolic syndrome (Gambineri, et al., 2011), and hypertension (Freedman, et al., 2001; Goff, et al., 2005).

The four cardiac diseases shown in Figure 4.1 (blue) are left ventricular hypertrophy, coronary disease, cardiomegaly, and myocardial ischemia. Left ventricular hypertrophy is the enlargement of the left ventricle of the heart and can be caused by obesity, diabetes, and older age (https://www.mayoclinic.org/diseases-conditions/left-ventricular-hypertrophy/symptoms-causes/syc-20374314). Myocardial ischemia and coronary disease are the decrease of blood flow to the heart, which affects the oxygen flow to the heart. Risk factors for coronary disease include diabetes, high blood pressure, high cholesterol levels, obesity, and more (https://www.mayoclinic.org/diseases-conditions/myocardial-ischemia/symptoms-causes/syc-20375417). Cardiomegaly is the

name for an enlarged heart, which is often a symptom of another condition. It can be caused by heart valve disease, high blood pressure, and coronary artery disease (https://www.mayoclinic.org/diseases-conditions/enlarged-heart/symptoms-causes/syc-20355436). These diseases are very intertwined, as an enlarged heart can be caused by coronary disease.

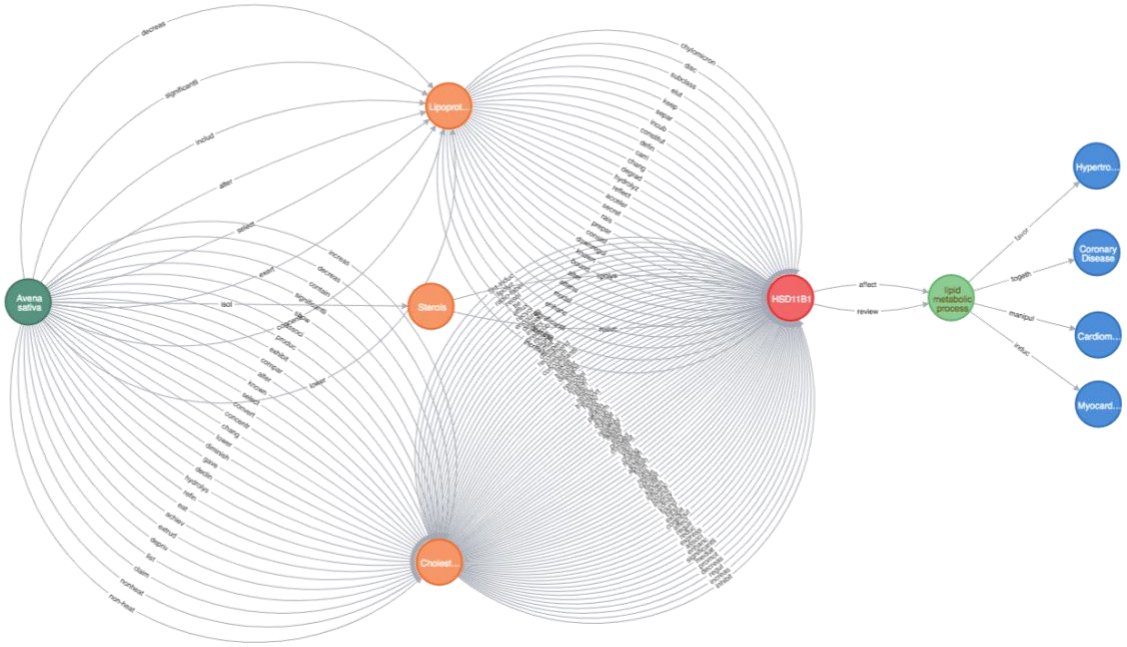


Figure 4.1: Figure depicting the meta path between Avena sativa and cardiac diseases with multiple hypothetical explanations for the relationships.

Further examination of the cholesterol – HSD11B1 relationship could be very informative in exploring this connection further. Due to the relationship between oat β-glucan and cholesterol and weight, the connection to heart health is logical. Decreased weight, specifically visceral fat in the abdomen, would lower the expression levels of HSD11B1, which would improve regulation of cortisol. Decreased weight also lowers the risk for type II diabetes, which is a risk factor for many heart conditions.

4.3.2: Oat and Diabetes

A meta path was created depicting the diet-disease network between *Avena sativa* and diabetes and is shown in Figure 4.2. As with the previous example exploring the relationship between Avena sativa and cardiac disease (Figure 4.1), this path shows that consumption of *Avena sativa* affects cholesterol levels in the body, which in turn is associated with the gene HSD11B1, which affects lipid metabolic processes having both positive and negative impacts on the incidence of diabetes.

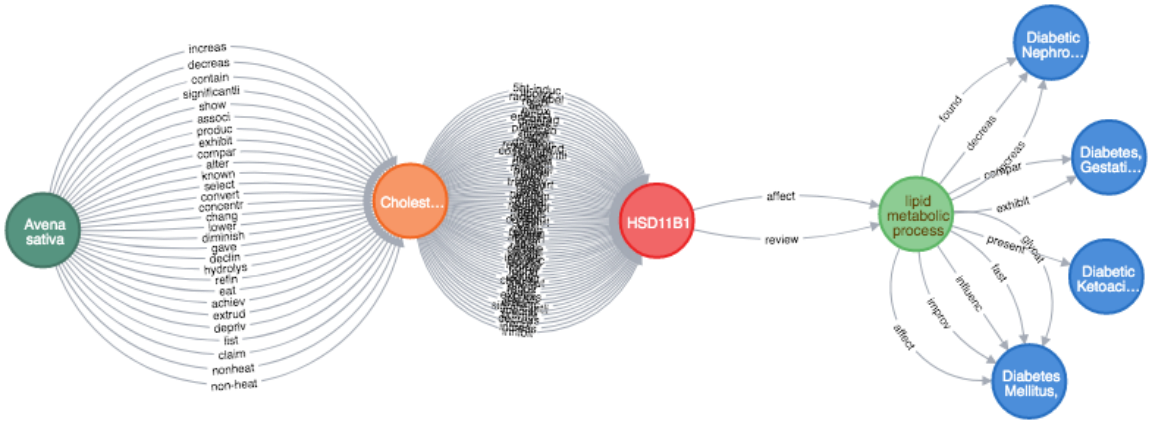


Figure 4.2: Figure depicting the meta path between Avena sativa and diabetes conditions with multiple hypothetical explanations for the relationships.

The four diabetes conditions that appear in the meta path in Figure 4.2, from top to bottom, are diabetic nephropathies, gestational diabetes, diabetic ketoacidosis, and diabetes mellitus, type 2. Type 2 diabetes is a condition in which the body does not use insulin properly, and eventually the body cannot create enough insulin to keep up with the metabolism of sugar (http://www.diabetes.org/diabetes-basics/type-2/). The diagnosis of full type 2 diabetes is often proceeded by what is now called insulin resistance, and in many cases patients who have insulin resistance may experience diabetic ketoacidosis before they are diagnosed with type 2 diabetes. Both diabetic ketoacidosis and diabetic nephropathies are complications of diabetes. Diabetic ketoacidosis is a condition which occurs when the

body cannot create enough insulin and no outside source is provided. As the body breaks down fats for energy, ketone acids are released into the bloodstream negatively affecting the kidneys causing a situation that can quickly become life-threatening if not treated (https://www.mayoclinic.org/diseases-conditions/diabetic-ketoacidosis/symptoms-causes/syc-20371551). Diabetic nephropathy is when the kidneys are damaged from diabetes, type 1 or 2. Many patients with diabetes eventually acquire kidney disease (https://www.mayoclinic.org/diseases-conditions/diabetic-nephropathy/symptoms-causes/syc-20354556). Gestational diabetes is a type of diabetes that occurs in pregnant women who have not previously had diabetes. It can lead to many complications throughout the pregnancy and birth, including preeclampsia and high birth weight of their babies. Though gestational diabetes typically resolves itself shortly after the birth of the child, it occasionally turns into type 2 diabetes (https://www.cdc.gov/pregnancy/diabetes-gestational.html).

Diabetes patients often have abnormal levels of many different lipids, as well as abnormal qualities to these lipids. For example, LDL cholesterol levels are normal or slightly high, but they experience higher levels of LDL oxidation and glycation (Vergès, 2005). Multiple studies have examined dyslipidemia in type 2 diabetes patients and the relationship between type 2 diabetes and cardiovascular disease (Pokharel, et al., 2017; Shahwan, et al., 2019). Patients with type 2 diabetes are at elevated risk for cardiovascular diseases including atherosclerosis, and this dyslipidemia may play a role in these risks (Shahwan, et al., 2019).

As in the diet-disease network for *Avena sativa* and cardiac diseases shown in Figure 4.1, HSD11B1 is shown to be the human gene connecting this relationship. Due to the association between HSD11B1 and glucose imbalance (Dammann, et al., 2019), the connection to diabetes is clear. As above, the relationship between HSD11B1, cholesterol and the role that consuming oats has in helping to prevent diabetes should be examined further.

4.3.1: Oat and Dermatitis

Two meta path graphs were created by the P2EP Knowledge Base describing the possible relationships between *Avena sativa* and dermatitis. Figure 4.3 depicts the possible relationship between *A. sativa* and atopic dermatitis, commonly known as eczema. Eczema is a condition that causes red and itchy skin, common in children but possible at any age. Eczema is chronic, and patients often experience periodic flare-ups. Patches, when scratched too much, can bleed and scab, and skin can appear scaly and cracked. It is usually spot-treated with medicated creams and moisturizers, and patients are instructed to avoid scented or harsh soaps or detergents (https://www.mayoclinic.org/diseases-conditions/atopic-dermatitis-eczema/symptoms-causes/syc-20353273). Figure 4.4 depicts the possible relationships between *A. sativa* and phototoxic dermatitis and contact dermatitis. Phototoxic dermatitis describes a wide range of abnormal reactions to sunlight. Often, prevention is the best treatment for these types of reactions (Lehmann and Schwarz, 2011). Contact dermatitis is a rash caused by contact with an allergen or irritant substance. These substances could be in the form of cosmetics, metals or jewelry, or plants. As with phototoxic dermatitis, the best recommendation is to avoid the problem substance

altogether. If scratched too much, a rash caused by contact dermatitis can become an open wound and cause an infection (https://www.mayoclinic.org/diseases-conditions/contact-dermatitis/symptoms-causes/syc-20352742). Both graphs identify the same two human genes, TNFSF11 and PARK7 as playing a role in apoptotic process and then the dermatitis conditions. Several chemicals, shown in orange in each graph, are hypothesized to play a role in these relationships as well. These chemicals include interleukins, dinoprostone, vitamin A, AMP-Activated protein kinases, NF-kappa B, collagen, adiponectin, and Prostaglandins E.

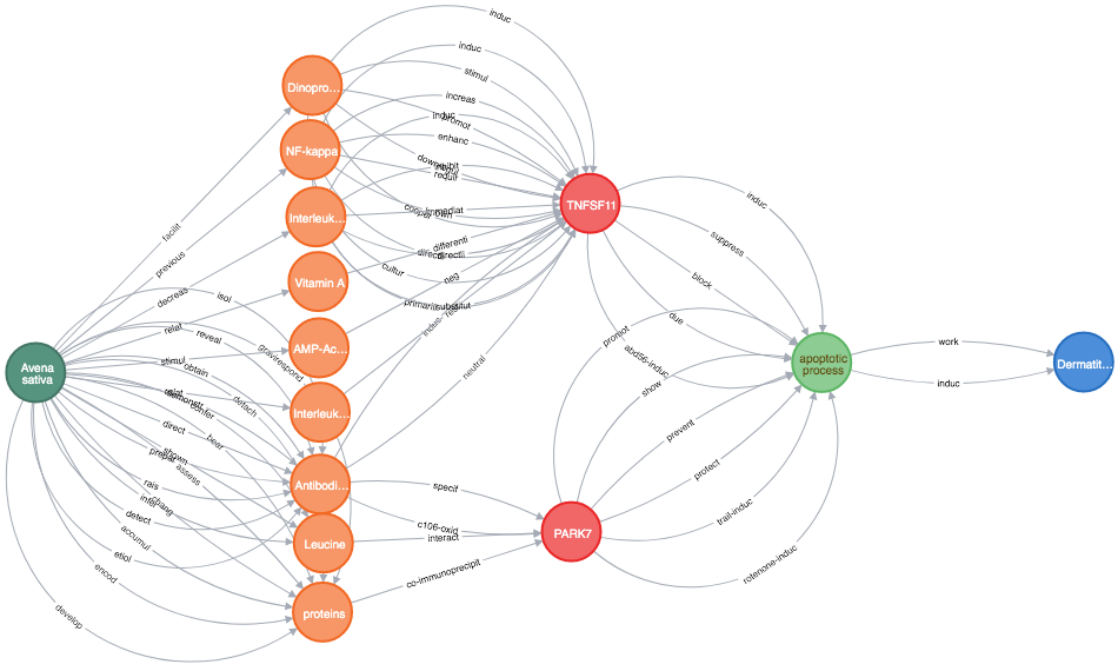


Figure 4.3: Figure depicting the meta path between Avena sativa and atopic dermatitis with multiple hypothetical explanations for the relationships.

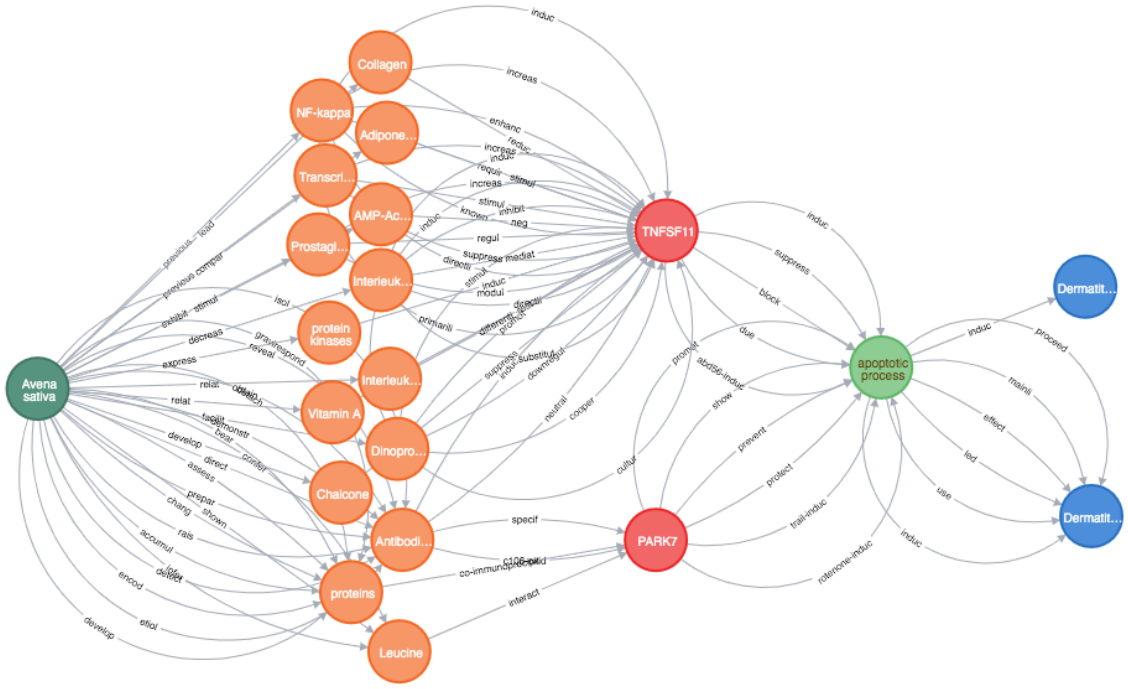


Figure 4.4: Figure depicting the meta path between Avena sativa and phototoxic dermatitis and contact dermatitis with multiple hypothetical explanations for the relationships.

TNFSF11 (also known as RANKL) is an immune response membrane protein in the tumor necrosis factor (TNF) family. The gene has a regulatory role in apoptosis and plays a role in osteopetrosis, or abnormally dense bones (https://ghr.nlm.nih.gov/gene/TNFSF11). Denosumab is an antibody drug used in the treatment of osteoporosis that blocks binding to RANKL. Patients on denosumab saw higher occurrences of eczema and infection (Cummings, et al., 2009). The other gene identified in both meta paths for dermatitis is PARK7, which codes for DJ-1. DJ-1 plays a role in antioxidative stress, and eczema and other conditions are associated with oxidative stress (Kim, et al., 2013).

Studies have identified that retinoic acid, an active metabolite of Vitamin A, inhibits the expression of RANKL (Hu, et al., 2010). A study into the effects of oat phytochemicals on skin showed that even small amounts of avenanthramides can be associated with a

decreased phosphorylation of NF-kappa B. In addition, the study also found that there was a reduction of interleukin-8 release, showing that the oat phytochemicals can help inhibit the inflammatory cytokines, which may contribute to patients feeling the need to scratch (Sur, et al., 2008).

Several mechanisms for atopic dermatitis being affected by dysregulated apoptosis (Trautmann, et al., 2001). By blocking or preventing apoptotic process, as indicated in both graphs, dysregulation of RANKL or PARK7 could be affecting dermatitis conditions. Therefore, oat as an ingredient in skin care products should continue to be explored as a natural treatment for dermatitis conditions.

4.3.2: Oat and Cancer

A meta path was created depicting hypothetical relationships between *Avena sativa* and diabetes and is shown in Figure 4.5. There are many chemicals (orange) shown which connect oat (dark green) to the mTOR, ARID4B, and KMT5C proteins (red) in humans. These proteins are related to polyamine catabolic process and histone H4-K20 trimethylation (light green), which in turn are connected to carcinogenic process and carcinoma: non-small-cell lung, breast neoplasms, and urinary bladder neoplasms (blue).

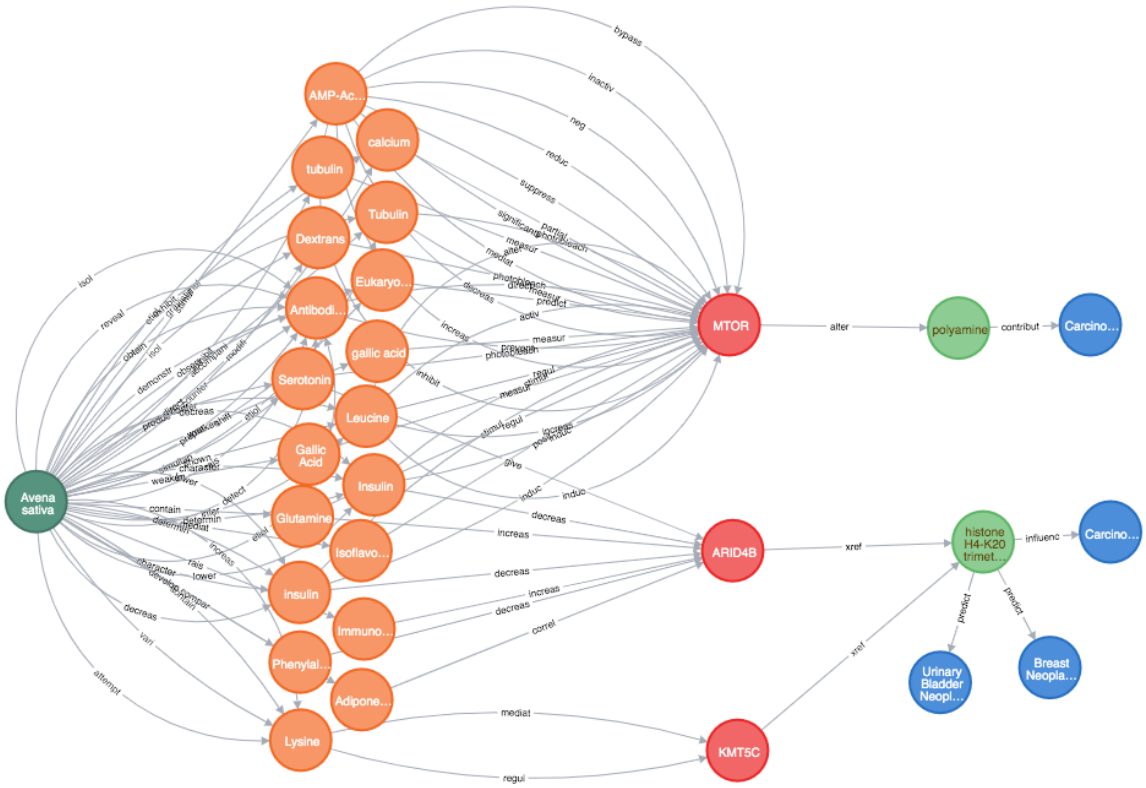


Figure 4.5: Figure depicting the meta path between Avena sativa various cancers with multiple hypothetical explanations for the relationships.

The histone H4-K20 trimethylation pathway is associated in the diet-disease network with non-small-cell lung carcinomas, breast neoplasms, and urinary bladder neoplasms. The histone H4-K20 trimethylation pathway (H4K20me3) plays a role in chromatin integrity due to its preferential targeting of repetitive elements in pericentromeric and telomeric heterochromatin. It has been suggested that the effects of disrupted H4K20me3 in tumors are due to genomic instability. (Shinchi, et al., 2015). H4K20me3 levels are significantly lowered in breast, urinary, and non-small-cell lung cancers (Fraga, et al., 2005; Schneider, et al., 2011; Van Den Broeck, et al., 2008).

The human gene KTM5C, or lysine methyltransferase 5C, is also known as SUV420H2. It is one of the genes responsible for H4K20me3 and could be a target for

disruption in cancer cells (Fraga, et al., 2005). When SUV420H2 is delivered exogenously, invasion of cancer cells is suppressed (Shinchi, et al., 2015). While oat products contain significant amounts of lysine – one cup of old fashioned rolled oats contain 516 mg (https://nutritiondata.self.com/facts/breakfast-cereals/1597/2) – there is currently no evidence to suggest that oat consumption plays a role in the KTM5C gene expression or function. Additionally, the human gene ARID4B, or RBBP1L1, is a tumor suppressor that regulates epigenetic markers including H4K20me3 (Gong, et al., 2012), however there is currently no evidence to suggest that oat consumption plays a role in ARID4B gene expression or function.

The polyamine catabolic process is related in the diet-disease network to the carcinogenic process. It is well-documented that the dysregulation of the polyamine metabolic process is frequent in cancer cells (Battaglia, et al., 2014). In fact, clinical trials are ongoing for the treatment and prevention of cancer by targeting polyamines (Casero, et al., 2018). The mTOR protein is involved with regulating the polyamine metabolic route that is necessary to the formation of tumors (Zabala-Letona, et al., 2017).

Gallic acid is one of the chemicals connecting *Avena sativa* to mTOR and therefore the carcinogenic process. Gallic acid is an antioxidant compound found in many plants. It has been shown to inhibit growth of cancer cells and reduce the phosphorylation of mTOR in some cancer cells (Tan, et al., 2014). Gallic acid is found in oat groats and hulls (Emmons and Peterson, 1999), which could contribute to the anti-carcinogenic properties of oat consumption. Further exploration of the anti-carcinogenic effects of consuming gallic-acid-containing plants could be useful in helping to prevent cancer through diet.

Additionally, mTOR can be inhibited by AMP-activated protein kinase (AMPK), as shown in the diet-disease network. AMPK can be activated by the calcium-calmodulin dependent kinase β in response to elevated calcium levels (Kania, et al., 2015). Thus, increased calcium consumption could play a role in the autophagy of cancer cells.

The relationships between oat consumption and the mTOR, ARID4B, and KMT5C proteins should be further researched. Further studies regarding the anti-carcinogenic effect of consuming oat, and other plants, would be useful for not only preventing cancer, but also as a potential mechanism for treating existing cancers. Additionally, studies of the oat metabolome could identify compounds that could be used in new treatments for cancers.

4.4: Conclusions

Further exploration of the HSD11B1 – cholesterol relationship could help improve understanding of the relationship between oat and hearth health, as well as diabetes. Additionally, further studies into the effects of avenanthramides and other oat chemicals on PARK7 and RANKL pathways could further inform treatment of skin conditions such as dermatitises. While this work is starting to be more commonplace (Cancer Research and Cancer Prevention Collaborative, 2019), the effect of phytochemicals from oat and other plants on carcinogenesis and autophagy of cancer cells should be further explored as a mechanism for preventing and treating cancers.

CHAPTER 5: CONCLUSIONS & FUTURE DIRECTIONS

In this work, two diploid *Avena* genomes have been presented, fully sequenced and annotated. These genomics resources will be valuable to the oat research community, both for the assembly and anchoring of the hexaploid *Avena sativa* genome, and in the identification of important genes relating to disease and stress resistance and high-quality grain traits. Once the hexaploid genome assembly is refined, further studies can be performed regarding the rearrangement and evolution of the subgenomes over time.

From these annotations, *Avena*-specific genes were identified using a BLAST pipeline and custom Python scripts. We identified 2,511 and 3,043 genes from *A. atlantica* and *A. eriantha*, respectively, that are specific to the *Avena* genus. Within these *Avena*-specific genes, we identified potential metallothionein family 15 proteins which may be involved in stress response. Additionally, three genes were identified which fell into the DUF1110 family, a family of genes only found in *Poaceae*, with putative function in spike morphology and grain shattering. Further research and understanding of these genes could help breeders ensure the good qualities of oat are maintained and enhanced in further breeding efforts, as well as which qualities should be removed if possible.

The relationship between *Avena* and several diseases was studied using the P2EP Knowledge Base created by students in the Plant Pathways Elucidation Project. These relationships were explored more in-depth and several pathways and relationships that need further study were identified, including the HSD11B1 – cholesterol relationship for the connections to heart health and diabetes, and those between oat phytochemicals and PARK7 and RANKL for the connection to skin conditions including atopic dermatitis.

Additionally, the study of oat phytochemicals and their relationship to cancer pathways should be further explored to develop new cancer prevention or treatment drugs.

In addition to the studies mentioned above, a full hexaploid oat genome assembly and annotation were pursued early in this dissertation research. Unfortunately, delays in sequencing, computational challenges in assembly and ultimately the discovery that a significant collapse in sub-genomes occurred during error correction led to the presentation of a hexaploid genome as part of this dissertation being not possible. However, the ultimate refinement of the hexaploid assembly will have a huge impact on the oat research community. The diploid genomes presented herein will be used to anchor the hexaploid genome during the assembly process to ensure the best possible assembly. A hexaploid genome will be an invaluable tool when it comes to oat improvement.

REFERENCES

Ahmad, M., *et al.* A review on Oat (Avena sativa L.) as a dual-purpose crop. *Scientific Research and Essays* 2014;9(4):52-59.

Akita, M. and Valkonen, J.P.T. A novel gene family in moss (Physcomitrella patens) shows sequence homology and a phylogenetic relationship with the TIR-NBS class of plant disease resistance genes. *Journal of Molecular Evolution* 2002;55(5):595-605.

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 1997;25.

Andon, M.B. and Anderson, J.W. State of the Art Reviews: The Oatmeal-Cholesterol Connection: 10 Years Later. *American Journal of Lifestyle Medicine* 2008;2(1):51-57.

Au, K.F., *et al.* Improving PacBio long read accuracy by short read alignment. *PLoS ONE* 2012;7(10):e46679.

Aung T, C.J., Leggett M. The transfer of crown rust resistance gene Pc94 from a wild diploid to cultivated hexaploid oat. In, *9th European and Mediterranean Cereal Rust & Powdery Mildews Conference*. Luntern, The Netherlands; 1996. p. 167–171.

Badouin, H., *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 2017;546:148.

Bai, J., *et al.* Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res* 2002;12(12):1871-1884.

Bao, E. and Lan, L. HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics* 2017;18(1):204.

Battaglia, V., *et al.* Polyamine catabolism in carcinogenesis: potential targets for chemotherapy and chemoprevention. *Amino acids* 2014;46(3):511-519.

Baucom, R.S., *et al.* Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet* 2009;5(11):e1000732.

Baum, B.R. and Fedak, G. Avena-Atlantica, a New Diploid Species of the Oat Genus from Morocco. *Canadian Journal of Botany-Revue Canadienne De Botanique* 1985;63(6):1057-1060.

Beier, S., *et al.* Construction of a map-based reference genome sequence for barley, Hordeum vulgare L. *Scientific data* 2017;4:170044-170044.

Bekele, W.A., *et al.* Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnology Journal* 2018;16(8):1452-1463.

Bennett, M.D. and Smith, J.B. Nuclear dna amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci* 1976;274(933):227-274.

Berlin, K., *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotech* 2015;33(6):623-630.

Bertioli, D.J., *et al.* The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature genetics* 2016;48:438.

Bilinski, P., *et al.* Genomic abundance is not predictive of tandem repeat localization in grass genomes. *PLoS One* 2017;12(6):e0177896.

Blanc, P.-L. The opening of the Plio-Quaternary Gibraltar Strait: assessing the size of a cataclysm. *Geodinamica Acta* 2002;15(5):303-317.

Bolger, A.M., Lohse, M. and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114-2120.

Bolger, A.M., *et al.* Computational aspects underlying genome to phenome analysis in plants. *The Plant Journal* 2019;97(1):182-198.

Bouchenak-Khelladi, Y., *et al.* Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time. *Botanical Journal of the Linnean Society* 2010;162(4):543-557.

Brawley, S.H., *et al.* Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta). *Proceedings of the National Academy of Sciences of the United States of America* 2017;114(31):E6361-E6370.

Bremer, K. Ancestral Areas - a Cladistic Reinterpretation of the Center of Origin Concept. *Systematic Biology* 1992;41(4):436-445.

Burges, A., *et al.* From phytoremediation of soil contaminants to phytomanagement of ecosystem services in metal contaminated sites. *International Journal of Phytoremediation* 2018;20(4):384-397.

Butt, M.S., *et al.* Oat: unique among the cereals. *European Journal of Nutrition* 2008;47(2):68-79.

Campbell, M.A., *et al.* Identification and characterization of lineage-specific genes within the Poaceae. *Plant Physiology* 2007;145(4):1311-1322.

Cancer Research, U.K.L.C.R.N. and Cancer Prevention Collaborative, G. Current opportunities to catalyze research in nutrition and cancer prevention - an interdisciplinary perspective. *BMC medicine* 2019;17(1):148-148.

Cantarel, B.L., *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 2008;18(1):188-196.

Carson, M.L. Virulence in Oat Crown Rust (Puccinia coronata f. sp avenae) in the United States from 2006 through 2009. *Plant Disease* 2011;95(12):1528-1534.

Casero, R.A., Jr., Murray Stewart, T. and Pegg, A.E. Polyamine metabolism and cancer: treatments, challenges and opportunities. *Nature reviews. Cancer* 2018;18(11):681-695.

Chaffin, A.S., *et al.* A Consensus Map in Cultivated Hexaploid Oat Reveals Conserved Grass Synteny with Substantial Subgenome Rearrangement. *The Plant Genome* 2016;9(2).

Cheng, C.-Y., *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal* 2017;89(4):789-804.

Chin, C.-S., *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth* 2013;10(6):563-569.

Chin, C.-S., *et al.* Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *bioRxiv* 2016.

Civáň, P. and Brown, T.A. A novel mutation conferring the nonbrittle phenotype of cultivated barley. *The New phytologist* 2017;214(1):468-472.

Clifford, B.C. Diseases, pests and disorders of oats. In: Welch, R.W., editor, *The Oat Crop: Production and Utilization*. Dordrecht: Springer Netherlands; 1995. p. 252-278.

Coon, M.A. MS. Provo, Utah: Brigham Young University; 2012. Characterization and Variable Expression of the CslF6 Homologs in Oat (*Avena* sp.).

Cummings, S.R., *et al.* Denosumab for Prevention of Fractures in Postmenopausal Women with Osteoporosis. *New England Journal of Medicine* 2009;361(8):756-765.

Dammann, C., Stapelfeld, C. and Maser, E. Expression and activity of the cortisol-activating enzyme 11β-hydroxysteroid dehydrogenase type 1 is tissue and species-specific. *Chemico-Biological Interactions* 2019;303:57-61.

Daou, C. and Zhang, H. Oat Beta-Glucan: Its Role in Health Promotion and Prevention of Diseases. *Comprehensive Reviews in Food Science and Food Safety* 2012;11(4):355-365.

Darwish, O., *et al.* Re-annotation of the woodland strawberry (Fragaria vesca) genome. *BMC Genomics* 2015;16(1):29.

Delaney, B., *et al.* β-Glucan Fractions from Barley and Oats Are Similarly Antiatherogenic in Hypercholesterolemic Syrian Golden Hamsters. *The Journal of Nutrition* 2003;133(2):468-475.

Du, J., *et al.* Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J* 2010;63(4):584-598.

Du, X., *et al.* Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nature genetics* 2018;50(6):796-802.

Duret, L. and Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* 2009;10:285-311.

Dyck, P.L. and Zillinsky, F.J. Inheritance of Crown Rust Resistance Transferred from Diploid to Hexaploid Oats. *Canadian Journal of Genetics and Cytology* 1963;5(5):398-+.

Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32(5):1792-1797.

Emmons, C.L. and Peterson, D.M. Antioxidant activity and phenolic contents of oat groats and hulls. *Cereal Chemistry* 1999;76(6):902-906.

Esvelt Klos, K., *et al.* Genome-wide association mapping of crown rust resistance in oat elite germplasm. *The Plant Genome* 2017.

EU Register on Nutrition Health Claims. In.; 2018.

Fardet, A. New hypotheses for the health-protective mechanisms of whole-grain cereals: what is beyond fibre? *Nutrition Research Reviews* 2010;23:65–134.

Feschotte, C., Swamy, L. and Wessler, S.R. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* 2003;163(2):747-758.

Fink, G., Günther, I. and Hill, K. The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986–2007. *International Journal of Epidemiology* 2011;40(5):1196-1204.

Fominaya, A., Vega, C. and Ferrer, E. Giemsa C-Banded Karyotypes of Avena Species. *Genome* 1988;30(5):627-632.

A Food Labeling Guide. In: Services, U.S.D.o.H.a.H., editor.; 2013.

Fraga, M.F., *et al.* Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature genetics* 2005;37(4):391.

Francis, G., *et al.* The biological action of saponins in animal systems: a review. *British Journal of Nutrition* 2007;88(6):587-605.

Freedman, D.S., *et al.* Distribution and correlates of high-density lipoprotein subclasses among children and adolescents. *Metabolism - Clinical and Experimental* 2001;50(3):370-376.

Fu, L., *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150-3152.

Fu, S., Wang, A. and Au, K.F. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biology* 2019;20(1):26.

Galindo-Gonzalez, L., *et al.* LTR-retrotransposons in plants: Engines of evolution. *Gene* 2017;626:14-25.

Gambineri, A., *et al.* A combination of polymorphisms in HSD11B1 associates with in vivo 11β-HSD1 activity and metabolic syndrome in women with and without polycystic ovary syndrome. 2011;165(2):283.

Gan, X., *et al.* The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. *Nature Plants* 2016;2(11).

Garcia-Castellanos, D., *et al.* Catastrophic flood of the Mediterranean after the Messinian salinity crisis. *Nature* 2009;462:778.

Geider, K., *et al.* Erwinia tasmaniensis sp. nov., a non-phytopathogenic bacterium from apple and pear trees. *International Journal of Systematic and Evolutionary Microbiology* 2006;56(12):2937-2943.

Giordano, F., *et al.* De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports* 2017;7:3935.

Gnanesh, B.N., *et al.* Chromosome location and allele-specific PCR markers for marker-assisted selection of the oat crown rust resistance gene Pc91. *Molecular Breeding* 2013;32(3):679-686.

Goff, D.C., Jr., *et al.* Insulin resistance and adiposity influence lipoprotein size and subclass concentrations. Results from the Insulin Resistance Atherosclerosis Study. *Metabolism - Clinical and Experimental* 2005;54(2):264-270.

Gong, W., *et al.* Structural insight into recognition of methylated histone tails by retinoblastoma-binding protein 1. *The Journal of biological chemistry* 2012;287(11):8531-8540.

Goodstein, D.M., *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 2011;40(D1):D1178-D1186.

Goodwin, S., *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research* 2015;25(11):1750-1756.

Goodwin, S., McPherson, J.D. and McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 2016;17:333.

Graham, M.A., *et al.* Computational Identification and Characterization of Novel Genes from Legumes. *Plant Physiology* 2004;135(3):1179-1197.

Grant, M.R., *et al.* Structure of the Arabidopsis Rpm1 Gene Enabling Dual-Specificity Disease Resistance. *Science* 1995;269(5225):843-846.

Gremme, G., *et al.* Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* 2005;47(15):965-978.

Grigoriev, I., *et al.* The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* 2012;40(Database issue):D26-32.

Grimalt, R., Mengeaud, V. and Cambazard, F. The Steroid-Sparing Effect of an Emollient Therapy in Infants with Atopic Dermatitis: A Randomized Controlled Study. *Dermatology* 2007;214(1):61-67.

Grundy, S.M. Obesity, Metabolic Syndrome, and Coronary Atherosclerosis. *Circulation* 2002;105(23):2696-2698.

Guo, L., *et al.* The cholesterol-lowering effects of oat varieties based on their difference in the composition of proteins and lipids. *Lipids in Health and Disease* 2014;13(1):182.

Guo, W.-J., Bundithya, W. and Goldsbrough, P.B. Characterization of the Arabidopsis metallothionein gene family: tissue-specific expression and induction during senescence and in response to copper. *New Phytologist* 2003;159(2):369-381.

Hackl, T., *et al.* proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 2014;30(21):3004-3011.

Haghshenas, E., *et al.* Colormap: Correcting long reads by mapping short reads. *Bioinformatics* 2016;32(17):i545-i551.

Harkess, A., *et al.* The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature communications* 2017;8(1):1279-1279.

Hassinen, V.H., *et al.* Metallothioneins 2 and 3 contribute to the metal-adapted phenotype but are not directly linked to Zn accumulation in the metal hyperaccumulator, Thlaspi caerulescens. *Journal of experimental botany* 2009;60(1):187-196.

Haupt, W., *et al.* The centromere1 (CEN1) region of Arabidopsis thaliana: architecture and functional impact of chromatin. *Plant J* 2001;27(4):285-296.

He, F.J., *et al.* Increased consumption of fruit and vegetables is related to a reduced risk of coronary heart disease: meta-analysis of cohort studies. *Journal Of Human Hypertension* 2007;21:717.

Henikoff, S., Ahmad, K. and Malik, H.S. The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science* 2001;293(5532):1098-1102.

Henson, J., Tischler, G. and Ning, Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 2012;13(8):901-915.

Hoff, K.J., *et al.* BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 2016;32(5):767-769.

Holt, C. and Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 2011;12.

Hu, L., *et al.* Retinoic acid increases proliferation of human osteoclast progenitors and inhibits RANKL-stimulated osteoclast differentiation by suppressing RANK. *PLoS ONE* 2010;5(10):e13305-e13305.

Huang, S., *et al.* Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol Biol* 2002;48(5-6):805-820.

Hunter, S., *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Research* 2009;37(Database issue):D211-D215.

Ibarra-Laclette, E., *et al.* Architecture and evolution of a minute plant genome. *Nature* 2013;498(7452):94-98.

Islam-Faridi, M.N., Nelson, C.D. and Kubisiak, T.L. Reference karyotype and cytomolecular map for loblolly pine (Pinus taeda L.). *Genome* 2007;50(2):241-251.

Islamovic, E., *et al.* Genetic dissection of grain beta-glucan and amylose content in barley (Hordeum vulgare L.). *Molecular Breeding* 2013;31(1):15-25.

Jarvis, D.E., *et al.* The genome of Chenopodium quinoa. *Nature* 2017;542:307.

Jayakumar, V. and Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. 2017.

Jellen, E.N., B.S. Gill, H.W. Rines, S.L. Fox, W.A. Wilson, and M.S. McMullen. Translocations in current and ancestral spring and winter oat accessions. In, *1996 Agronomy Abstracts*. Madison, WI: Agronomy Society of America; 1996. p. 78.

Jellen, E.N. and Beard, J. Geographical distribution of a chromosome 7C and 17 intergenomic translocation in cultivated oat. *Crop Science* 2000;40(1):256-263.

Jellen, E.N., Gill, B.S. and Cox, T.S. Genomic in-Situ Hybridization Differentiates between a/D-Genome and C-Genome Chromatin and Detects Intergenomic Translocations in Polyploid Oat Species (Genus Avena). *Genome* 1994;37(4):613-618.

Jenkins, A.L., *et al.* Depression of the glycemic index by high levels of beta-glucan fiber in two functional foods tested in type 2 diabetes. *European journal of clinical nutrition* 2002;56(7):622-628.

Jiao, Y., *et al.* Improved maize reference genome with single-molecule technologies. *Nature* 2017;advance online publication.

Jones, P., *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* 2014;30(9):1236-1240.

Joshi, R., Pareek, A. and Singla-Pareek, S.L. Chapter 9 - Plant Metallothioneins: Classification, Distribution, Function, and Regulation. In: Ahmad, P., editor, *Plant Metal Interaction*. Elsevier; 2016. p. 239-261.

Kania, E., Pająk, B. and Orzechowski, A. Calcium Homeostasis and ER Stress in Control of Autophagy in Cancer Cells. *BioMed Research International* 2015;2015:12.

Katsiotis, A., Loukas, M. and Heslop-Harrison, J.S. Repetitive DNA, genome and species relationships in Avena and Arrhenatherum (Poaceae). *Annals of Botany* 2000;86(6):1135-1142.

Kawasaki, S., *et al.* Gene expression profiles during the initial phase of salt stress in rice. *The Plant Cell* 2001;13(4):889-905.

Kim, D., Langmead, B. and Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;12(4):357-360.

Kim, D.K., *et al.* DJ-1 regulates mast cell activation and IgE-mediated allergic responses. *Journal of Allergy and Clinical Immunology* 2013;131(6):1653-1662.e1651.

Klos, K.E., *et al.* Genome-Wide Association Mapping of Crown Rust Resistance in Oat Elite Germplasm. *Plant Genome* 2017;10(2).

Koch, M.A., Haubold, B. and Mitchell-Olds, T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). *Molecular Biology and Evolution* 2000;17(10):1483-1498.

Koren, S., *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* 2012;30.

Koren, S., *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 2017.

Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;5(1):59.

Kremer, C.A., Lee, M. and Holland, J.B. A restriction fragment length polymorphism based linkage map of a diploid Avena recombinant inbred line population. *Genome* 2001;44(2):192-204.

Kröpelin, S., *et al.* Climate-Driven Ecosystem Succession in the Sahara: The Past 6000 Years. *Science* 2008;320(5877):765.

Kweon, M.-H., Hwang, H.-J. and Sung, H.-C. Isolation and Characterization of Anticomplementary β-Glucans from the Shoots of Bamboo Phyllostachys edulis. *Planta Med* 2003;69(01):56-62.

Larkin, M.A., *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23(21):2947-2948.

Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 2014;30(22):3276-3278.

Latta, R., *et al.* Comparative linkage mapping of diploid, tetraploid, and hexaploid Avena species suggests extensive chromosome rearrangement in ancestral diploids (In Press). *Sci Rep* 2019.

Lee, H., *et al.* Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv* 2014:006395.

Lee, T.H., *et al.* SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 2014;15:162.

Left ventricular hypertrophy. In, *Diseases & Conditions*. Mayo Clinic.

Leggett, J.M. Interspecific Hybrids Involving the Recently Described Diploid Taxon Avena-Atlantica. *Genome* 1987;29(2):361-364.

Lehmann, P. and Schwarz, T. Photodermatoses: diagnosis and treatment. *Deutsches Arzteblatt international* 2011;108(9):135-141.

Li, G., *et al.* The Sequences of 1504 Mutants in the Model Rice Variety Kitaake Facilitate Rapid Functional Genomic Studies. *The Plant Cell* 2017;29(6):1218-1231.

Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 2013;preprint arXiv:1303.3997.

Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Li, P.C., *et al.* RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* 2016;17.

Li, R., *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* 2010;20(2):265-272.

Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22(13):1658-1659.

Lieberman-Aiden, E., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326(5950):289-293.

Lightfoot, D.J., *et al.* Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. *BMC Biology* 2017;15(1):74.

Lin, H., *et al.* Comparative analyses reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. *BMC evolutionary biology* 2010;10:41-41.

Linchangco, R.V. Doctor of Philosophy: UNC Charlotte; 2018. The Semantics of Diet and Health: Knowledge Based Discovery through Data Integration, Text Mining, and Network Analysis.

Liu, Q., *et al.* The repetitive DNA landscape in Avena (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. *BMC plant biology* 2019;19(1):226-226.

Liu, Q., *et al.* Unraveling the evolutionary dynamics of ancient and recent polyploidization events in Avena (Poaceae). *Scientific Reports* 2017;7:41944.

Liu, S., *et al.* A prospective study of whole-grain intake and risk of type 2 diabetes mellitus in US women. *American journal of public health* 2000;90(9):1409-1415.

Locato, V. and De Gara, L. Programmed Cell Death in Plants: An Overview. In: De Gara, L. and Locato, V., editors, *Plant Programmed Cell Death: Methods and Protocols*. New York, NY: Springer New York; 2018. p. 1-8.

Lomsadze, A., Burns, P.D. and Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* 2014;42(15):e119-e119.

Loskutov, I.G. and Rines, H.W. *Avena*. In: Kole, C., editor, *Wild Crop Relatives: Genomic and Breeding Resources.* . Heidelberg: Springer; 2011. p. 109-183.

Lynch, M. and Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* 2000;290(5494):1151-1155.

Maki, K.C., *et al.* Effects of consuming foods containing oat β-glucan on blood pressure, carbohydrate metabolism and biomarkers of oxidative stress in men and women with elevated blood pressure. *European journal of clinical nutrition* 2006;61:786.

Mangeon, A., Junqueira, R.M. and Sachetto-Martins, G. Functional diversity of the plant glycine-rich proteins superfamily. *Plant Signaling & Behavior* 2010;5(2):99-104.

Marcais, G. and Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27(6):764-770.

Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 2008;24(3):133-141.

Mascher, M., *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* 2017;544:427.

Mathan, J., Bhattacharya, J. and Ranjan, A. Enhancing crop yield by optimizing plant developmental features. *Development* 2016;143(18):3283-3294.

Mattila, P., Pihlava, J.-M. and Hellström, J. Contents of phenolic acids, alkyl- and alkenylresorcinols, and avenanthramides in commercial grain products. *Journal of Agricultural and Food Chemistry* 2005;53(21):8290-8295.

McCartney, C.A., *et al.* Mapping of the oat crown rust resistance gene Pc91. *Theoretical and Applied Genetics* 2011;122(2):317-325.

McDowell, J.M., *et al.* Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of arabidopsis. *Plant Cell* 1998;10(11):1861-1874.

Melters, D.P., *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* 2013;14(1).

Metallothionein, family 15, plant (IPR000347). In.: EMBL-EBI.

Michael, T.P. Plant genome size variation: bloating and purging DNA. *Brief Funct Genomics* 2014;13(4):308-317.

Miclotte, G., *et al.* Jabba: Hybrid error correction for long sequencing reads using maximal exact matches. In, *International Workshop on Algorithms in Bioinformatics*. Springer; 2015. p. 175-188.

Miller, J.R., *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 2008;24(24):2818-2824.

Miller, K.A., *et al.* Long-Term Exposure to Air Pollution and Incidence of Cardiovascular Events in Women. *New England Journal of Medicine* 2007;356(5):447-458.

Minaya, M., *et al.* Distribution and evolutionary dynamics of Stowaway Miniature Inverted repeat Transposable Elements (MITEs) in grasses. *Mol Phylogenet Evol* 2013;68(1):106-118.

Mizuno, H., *et al.* Asymmetric distribution of gene expression in the centromeric region of rice chromosome 5. *Frontiers in Plant Science* 2011;2.

Moelling, K., *et al.* RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. *Frontiers in microbiology* 2017;8:1745-1745.

Myers, G. https://github.com/thegenemyers/DALIGNER

Nair, S., *et al.* 11β-Hydroxysteroid dehydrogenase Type 1: genetic polymorphisms are associated with Type 2 diabetes in Pima Indians independently of obesity and expression in adipocyte and muscle. *Diabetologia* 2004;47(6):1088-1095.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 2016;44(D1):D7-D19.

Oliver, R.E., *et al.* New Diversity Arrays Technology (DArT) markers for tetraploid oat (Avena magna Murphy et Terrell) provide the first complete oat linkage map and markers linked to domestication genes from hexaploid A. sativa L. *Theor Appl Genet* 2011;123(7):1159-1171.

Oliver, R.E., *et al.* SNP Discovery and Chromosome Anchoring Provide the First Physically-Anchored Hexaploid Oat Map and Reveal Synteny with Model Species. *PLoS ONE* 2013;8(3):e58068.

Ozturk, Z.N., *et al.* Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. *Plant Molecular Biology* 2002;48(5-6):551-573.

PacBio Sequel System. In.

Page, J.T., *et al.* BamBam: genome sequence analysis tools for biologists. *BMC Res Notes* 2014;7:829.

Paterson, A.H., *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* 2009;457(7229):551-556.

Paulsen, S.K., *et al.* 11β-HSD Type 1 Expression in Human Adipose Tissue: Impact of Gender, Obesity, and Fat Localization. *Obesity* 2007;15(8):1954-1960.

Pentaxin, conserved site (IPR030476). In.: EMBL-EBI.

Pepys, M.B. C-Reactive Protein. In: Delves, P.J., editor, *Encyclopedia of Immunology (Second Edition)*. Oxford: Elsevier; 1998. p. 663-665.

Perrelli, A., *et al.* Biological Activities, Health Benefits, and Therapeutic Properties of Avenanthramides: From Skin Protection to Prevention and Treatment of Cerebrovascular Diseases. *Oxidative medicine and cellular longevity* 2018;2018:6015351-6015351.

Pertea, M., *et al.* Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;11(9):1650-1667.

Peterson, D.M. Oat Antioxidants. *Journal of Cereal Science* 2001;33(2):115-129.

Philippe, R., *et al.* A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat. *Genome Biology* 2013;14(6).

Pokharel, D.R., *et al.* Prevalence and pattern of dyslipidemia in Nepalese individuals with type 2 diabetes. *BMC research notes* 2017;10(1):146-146.

Pope, C.A., 3rd, *et al.* Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure-response relationships. *Environmental health perspectives* 2011;119(11):1616-1621.

Potter, R.C., Castro, J.M. and Moffatt, L.C. Oat oil compositions with useful cosmetic and dermatological properties. In.: Google Patents; 1997.

PR01228. In.: EMBL-EBI.

Pritchard, J.R., *et al.* A survey of β-glucan and arabinoxylan content in wheat. *Journal of the science of food and agriculture* 2011;91(7):1298.

PROSITE documentation PDOC00013. In.: PROSITE.

PROSITE Entry: PS51257. In.: PROSITE.

Protein of unknown function DUF1110 (IPR010535). In.: EMBL-EBI.

PS51257. In.: EMBL-EBI.

Putnam, N.H., *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* 2016;26(3):342-350.

Quail, M.A., *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;13(1):341.

Rabbani, M.A., *et al.* Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. *Plant Physiology* 2003;133(4):1755-1767.

Rajhathy, T. and Dyck, P.L. Chromosomal Differentiation and Speciation in Diploid Avena .2. Karyotype of A. Pilosa. *Canadian Journal of Genetics and Cytology* 1963;5(2):175-&.

Rajhathy, T. and Morrison, J.W. Chromosome morphology in the genus *Avena*. *Canadian Journal of Botany* 1959;37(3):331-337.

Raskina, O., *et al.* Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. *Cytogenet Genome Res* 2008;120(3-4):351-357.

Reis-Filho, J.S. Next-generation sequencing. *Breast Cancer Research* 2009;11(3):S12.

Reyes-Chin-Wo, S., *et al.* Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature communications* 2017;8:14953-14953.

Rhoads, A. and Au, K.F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* 2015;13(5):278-289.

Riaño-Pachón, D.M. and Mattiello, L. Draft genome sequencing of the sugarcane hybrid SP80-3280. *F1000Research* 2017;6:861.

Ribonuclease H domain (IPR002156). In.: EMBL-EBI.

Rines, H.W., *et al.* Identification, introgression, and molecular marker genetic analysis and selection of a highly effective novel oat crown rust resistance from diploid oat, Avena strigosa. *Theoretical and Applied Genetics* 2018;131(3):721-733.

Rokhsar, D.S., *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 2011;40(D1):D1178-D1186.

Roos, A., *et al.* Mini-review: A pivotal role for innate immunity in the clearance of apoptotic cells. *European Journal of Immunology* 2004;34(4):921-929.

Roth, M.S., *et al.* Chromosome-level genome assembly and transcriptome of the green alga Chromochloris zofingiensis illuminates astaxanthin production. *Proceedings of the National Academy of Sciences of the United States of America* 2017;114(21):E4296-E4305.

Ruan, J. 2017. wtdbg. https://github.com/ruanjue/wtdbg

Ruan, J. and Li, H. Fast and accurate long-read assembly with wtdbg2. *bioRxiv* 2019:530972.

Salmela, L. and Rivals, E. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 2014;30(24):3506-3514.

Sambrook, J., E .F. Fritsch, and T. Maniatis. Molecular cloning: A laboratory manual. . Cold Spring Harbor, NY.: Cold Spring Harbor Laboratory Press; 1989. .

Sanford, R.A., Cole, J.R. and Tiedje, J.M. Characterization and description of Anaeromyxobacter dehalogenans gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium. *Applied and environmental microbiology* 2002;68(2):893-900.

Sanz, M.J., *et al.* A new chromosome nomenclature system for oat (Avena sativa L. and A. byzantina C. Koch) based on FISH analysis of monosomic lines. *Theoretical and Applied Genetics* 2010;121(8):1541-1552.

Schmidt, M.H.-W., *et al.* Reconstructing The Gigabase Plant Genome Of Solanum pennellii Using Nanopore Sequencing. *bioRxiv* 2017.

Schmidt, M.H.-W., *et al.* De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant Cell* 2017;29(10):2336-2348.

Schnable, P.S., *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* 2009;326(5956):1112-1115.

Schneider, A.-C., *et al.* Global histone H4K20 trimethylation predicts cancer-specific survival in patients with muscle-invasive bladder cancer. *BJU International* 2011;108(8b):E290-E296.

Schubert, I., *et al.* Telomeric Signals in Robertsonian Fusion and Fission Chromosomes - Implications for the Origin of Pseudoaneuploidy. *Cytogenetics and Cell Genetics* 1992;59(1):6-9.

Sequencing power for every scale. In.: Illumina. p. Compare key specifications across the whole portfolio of Illumina sequencing systems. Understand the primary differences between the MiniSeq, MiSeq, NextSeq, HiSeq, and HiSeq X Series.

Shahwan, M.J., *et al.* Prevalence of dyslipidemia and factors affecting lipid profile in patients with type 2 diabetes. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 2019;13:2387-2392.

Shinchi, Y., *et al.* SUV420H2 suppresses breast cancer cell invasion through down regulation of the SH2 domain-containing focal adhesion protein tensin-3. *Experimental Cell Research* 2015;334(1):90-99.

Sievers, F., *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 2011;7(1):539.

Sigrist, C.J.A., *et al.* New and continuing developments at PROSITE. *Nucleic Acids Research* 2013;41(Database issue):D344-D347.

Simão, F.A., *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31(19):3210-3212.

Simmons, M.D. The Cereal Rusts Vol II: Diseases, distribution, epidemiology and control. In: Bushnell, A.P.R.a.W.R., editor. Orlando, FL: Academic Press; 1985. p. 132-172.

Simpson, J.T., *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Research* 2009;19(6):1117-1123.

Singh, R., De, S. and Belkheir, A. Avena sativa (Oat), A Potential Neutraceutical and Therapeutic Agent: An Overview. *Critical Reviews in Food Science and Nutrition* 2013;53(2):126-144.

Singh, R., Ming, R. and Yu, Q.Y. Comparative Analysis of GC Content Variations in Plant Genomes. *Tropical Plant Biology* 2016;9(3):136-149.

Slater, G.S.C. and Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 2005;6(1):31.

Smit, A., Hubley, R. and Green, P. RepeatMasker Open-3.0. In.; 1996-2010.

Smit, A., Hubley, R. . RepeatModeler Open-1.0.

. In.; 2008-2015. p. <http://www.repeatmasker.org/>. .

Solano, R., *et al.* Organization of repeated sequences in species of the genus Avena. *Theor Appl Genet* 1992;83(5):602-607.

Soreng, R.J., *et al.* A worldwide phylogenetic classification of the Poaceae (Gramineae). *Journal of Systematics and Evolution* 2015;53(2):117-137.

Sousa, A., Cusimano, N. and Renner, S.S. Combining FISH and model-based predictions to understand chromosome evolution in Typhonium (Araceae). *Ann Bot* 2014;113(4):669-680.

Stanke, M., *et al.* Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 2008;24(5):637-644.

Stanke, M. and Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* 2005;33(suppl_2):W465-W467.

Stanke, M., Tzvetkova, A. and Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* 2006;7 Suppl 1:S11 11-18.

Sur, R., *et al.* Avenanthramides, polyphenols from oats, exhibit anti-inflammatory and anti-itch activity. *Archives of Dermatological Research* 2008;300(10):569.

Talbert, P.B. and Henikoff, S. Centromeres convert but don't cross. *PLoS Biol* 2010;8(3):e1000326.

Tan, H.K., Moad, A. and Tan, M.L. The mTOR signalling pathway in cancer and the potential mTOR inhibitory activities of natural phytochemicals. *Asian Pac J Cancer Prev* 2014;15(16):6463-6475.

Tapola, N., *et al.* Glycemic responses of oat bran products in type 2 diabetic patients. *Nutrition, Metabolism and Cardiovascular Diseases* 2005;15(4):255-261.

Tarasov, A., *et al.* Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;31(12):2032-2034.

Tenaillon, M.I., Hollister, J.D. and Gaut, B.S. A triptych of the evolution of plant transposable elements. *Trends Plant Sci* 2010;15(8):471-478.

The International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* 2012;491(7426):711-716.

The International Brachypodium Initiative, *et al.* Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* 2010;463:763.

The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 2014;345(6194):1251788.

The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 2019;47(D1):D506-D515.

Thornlow, B.P., *et al.* Transfer RNA genes experience exceptionally elevated mutation rates. *Proceedings of the National Academy of Sciences* 2018;115(36):8996.

Tian, Z.X., *et al.* Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Research* 2009;19(12):2221-2230.

Tinker, N.A., Bekele, W.A. and Hattori, J. Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. *G3-Genes Genomes Genetics* 2016;6(4):857-863.

Todd, J.J. and Vodkin, L.O. Duplications That Suppress and Deletions That Restore Expression from a Chalcone Synthase Multigene Family. *Plant Cell* 1996;8(4):687-699.

Trapnell, C., *et al.* Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012;7.

Trapnell, C., *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;28.

Trapnell, C., *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* 2010;28(5):511-515.

Trautmann, A., *et al.* Role of Apoptosis in Atopic Dermatitis. *International Archives of Allergy and Immunology* 2001;124(1-3):230-232.

Turek, L.V., *et al.* Gender-dependent association of HSD11B1 single nucleotide polymorphisms with glucose and HDL-C levels. *Genetics and molecular biology* 2014;37(3):490-495.

Unver, T., *et al.* Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 2017;114(44):E9413-E9422.

Uraguchi, S., *et al.* Characteristics of cadmium accumulation and tolerance in novel Cd-accumulating crops, Avena strigosa and Crotalaria juncea. *Journal of experimental botany* 2006;57(12):2955-2965.

Van Den Broeck, A., *et al.* Loss of histone h4k20 trimethylation occurs in preneoplasia and influences prognosis of non–small cell lung cancer. *Clinical cancer research* 2008;14(22):7237-7245.

Vergès, B. New insight into the pathophysiology of lipid abnormalities in type 2 diabetes. *Diabetes & Metabolism* 2005;31(5):429-439.

Vurture, G.W., *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 2017;33(14):2202-2204.

Walker, B.J., *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* 2014;9(11):e112963.

Wang, J.R., *et al.* FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics* 2018;19(1):50.

Wang, Z., Chen, Y. and Li, Y. A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics* 2004;2(4):216-221.

Welch, R.W., Brown, J.C.W. and Leggett, J.M. Interspecific and intraspecific variation in grain and great characteristics of wild oat (Avena) species: Very high great (1 -> 3),(1 -> 4)-beta-D-glucan in an Avena atlantica genotype. *Journal of Cereal Science* 2000;31(3):273-279.

Willing, E.M., *et al.* Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants* 2015;1(2):1-7.

Wolever, T.M.S., Raederstorff, D. and Duss, R. Oat β-Glucan Reduces Serum LDL Cholesterol in Humans with Serum LDL Cholesterol <160mg/dL. *Immunology, Endocrine & Metabolic Agents - Medicinal ChemistryCurrent Medicinal Chemistry - Immunology, Endocrine & Metabolic Agents)* 2016;16(2):122-128.

Wolin, K.Y., Carson, K. and Colditz, G.A. Obesity and Cancer. *The Oncologist* 2010;15(6):556-565.

Wu, G.A., *et al.* Genomics of the origin and evolution of Citrus. *Nature* 2018;554:311.

Xu, Y., *et al.* Identification, characterization and expression analysis of lineage-specific genes within sweet orange (Citrus sinensis). *BMC Genomics* 2015;16:995-995.

Xue, T., *et al.* Cotton metallothionein GhMT3a, a reactive oxygen species scavenger, increased tolerance against abiotic stress in transgenic tobacco and yeast. *Journal of experimental botany* 2008;60(1):339-349.

Yan, H., *et al.* High-density marker profiling confirms ancestral genomes of Avena species and identifies D-genome chromosomes of hexaploid oat. *Theoretical and Applied Genetics* 2016:1-17.

Yan, H., *et al.* High-density marker profiling confirms ancestral genomes of Avena species and identifies D-genome chromosomes of hexaploid oat. *Theor Appl Genet* 2016;129(11):2133-2149.

Yan, H., *et al.* Genome size variation in the genus Avena. *Genome* 2016;59(3):209-220.

Yan, L., *et al.* The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proceedings of the National Academy of Sciences of the United States of America* 2006;103(51):19581-19586.

Yang, X., Chockalingam, S. and Aluru, S. A survey of error-correction methods for next-generation sequencing. 2012.

Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;24(8):1586-1591.

Yang, Z., *et al.* OsMT1a, a type 1 metallothionein, plays the pivotal role in zinc homeostasis and drought tolerance in rice. *Plant Molecular Biology* 2009;70(1):219-229.

Yarnell, E. and Abascal, K. Botanical Remedies for Nicotine Addiction. *Alternative & Complementary Therapies* 2001:337-340.

Yarnell, E. and Abascal, K. Botanical Treatments for Depression. *Alternative and Complementary Therapies* 2001;7(3).

Yasui, Y., *et al.* Draft genome sequence of an inbred line of Chenopodium quinoa, an allotetraploid crop with great environmental adaptability and outstanding nutritional properties. *DNA Research* 2016;23(6):535-546.

Yin, D., *et al.* Genome of an allotetraploid wild peanut Arachis monticola: a de novo assembly. *GigaScience* 2018;7(6):giy066.

Zabala-Letona, A., *et al.* mTORC1-dependent AMD1 regulation sustains polyamine metabolism in prostate cancer. *Nature* 2017;547(7661):109-113.

Zhao, Y., *et al.* Btr1-A Induces Grain Shattering and Affects Spike Morphology and Yield-Related Traits in Wheat. *Plant and Cell Physiology* 2019;60(6):1342-1353.

Zhou, X., Jellen, E.N. and Murphy, J.P. Progenitor Germplasm of Domisticated Hexaploid Oat. *Crop Science* 1999;39(4):1208-1214.

Zimin, A.V., *et al.* The MaSuRCA genome assembler. *Bioinformatics* 2013;29(21):2669-2677.

Zimin, A.V., *et al.* Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* 2017;27(5):787-792.

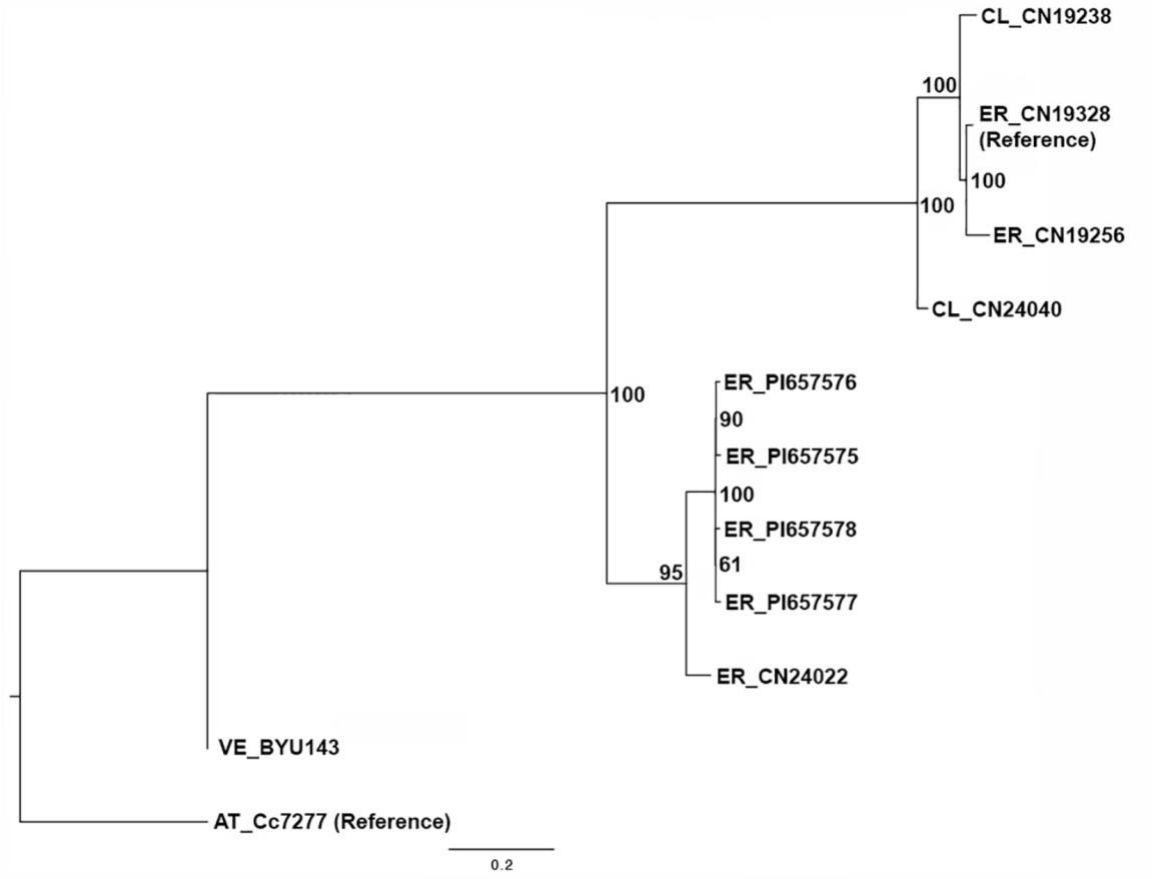
APPENDIX A: *AVENA* ACCESSIONS INCLUDED IN RESEQUENCING PANEL

Accession[1]	*Avena* Taxon	BYU	Genome	Origin
CN 19328	*eriantha*	132	C_pC_p	Boghari, Algeria
CN 19256	*eriantha*	689	C_pC_p	Iran
CN 24022	*eriantha*	690	C_pC_p	Saida, Algeria
PI 657575	*eriantha*	765	C_pC_p	Ifrane, Morocco
PI 657576	*eriantha*	766	C_pC_p	Ifrane, Morocco
PI 657577	*eriantha*	767	C_pC_p	Ifrane, Morocco
PI 657578	*eriantha*	768	C_pC_p	Ifrane, Morocco
CN 19238	*clauda*	685	C_pC_p	Cardak, Turkey
CN 24040	*clauda*	768	C_pC_p	Batna, Algeria
unknown	*ventricosa*	143	C_vC_v	unknown
CN 25888	*atlantica*	116	A_sA_s	Tiznit, Morocco
CN 25864	*atlantica*	139	A_sA_s	Tiznit, Morocco
CN 25849	*atlantica*	678	A_sA_s	Smimou, Morocco
CAV 6794	*atlantica*	680	A_sA_s	Tiznit, Morocco
Cc 7277	*atlantica*	803	A_sA_s	Wales
Cc 2043	*strigosa brevis*	202	A_sA_s	Unknown
CN 88826	*strigosa brevis*	681	A_sA_s	Portugal
PI 258545	*strigosa brevis*	804	A_sA_s	Portugal
PI 258544	*strigosa brevis*	807	A_sA_s	Portugal
PI 119009	*strigosa brevis*	808	A_sA_s	Rio Grande do Sul, Brazil
CIav 9036	*strigosa brevis*	815	A_sA_s	St. Petersburg, Russia
PI 573533	*strigosa brevis*	825	A_sA_s	Portugal
CN 25698	*strigosa hispanica*	117	A_sA_s	Santa Eulalia, Portugal
CN 25675	*strigosa hispanica*	126	A_sA_s	Elvora, Portugal
CN 25727	*strigosa hispanica*	127	A_sA_s	Ponte de Sor, Portugal
CN 25788	*strigosa hispanica*	699	A_sA_s	Fataca, Portugal
PGR 10012	*strigosa nuda*	164	A_sA_s	Germany
PI 401830	*strigosa nuda*	809	A_sA_s	Germany
CIav 9009	*strigosa nuda*	812	A_sA_s	Canada, Ontario
PI 287319	*strigosa nuda*	826	A_sA_s	Nordrhein-Westfalen, Germany
PI 292226	*strigosa*	666	A_sA_s	Tel Aviv, Israel
PI 291990	*strigosa*	667	A_sA_s	Israel
CIav 9011	*strigosa*	668	A_sA_s	Denmark
CIav 9012	*strigosa*	669	A_sA_s	Bulgaria
CIav 9019	*strigosa*	670	A_sA_s	Wales, United Kingdom
CIav 9020	*strigosa*	671	A_sA_s	Argentina
CIav 9022	*strigosa*	672	A_sA_s	The Netherlands
CIav 9038	*strigosa*	673	A_sA_s	N. Ireland, United Kingdom
PI 186606	*strigosa*	790	A_sA_s	Rio Grande do Sul, Brazil
CIav 7010	*strigosa*	810	A_sA_s	Rio Grande do Sul, Brazil
CIav 9035	*strigosa*	814	A_sA_s	St. Petersburg, Russia
PI 83719	*strigosa*	816	A_sA_s	New South Wales, Australia
PI 83721	*strigosa*	817	A_sA_s	New South Wales, Australia
PI 244471	*strigosa*	818	A_sA_s	Rio Grande do Sul, Brazil
PI 244472	*strigosa*	819	A_sA_s	Rio Grande do Sul, Brazil
PI 304557	*strigosa*	820	A_sA_s	Wales, United Kingdom
PI 401800	*strigosa*	821	A_sA_s	Germany
PI 436031	*strigosa*	823	A_sA_s	La Araucania, Chile
PI 436102	*strigosa*	824	A_sA_s	La Araucania, Chile

CN 19731	*hirtula*	136	A_sA_s	El Asnam, Algeria
CN 19530	*hirtula*	179	A_sA_s	Antalya, Turkey
CN 21478	*hirtula*	691	A_sA_s	Aghia Varvara, Crete, Greece
CN 21653	*hirtula*	692	A_sA_s	Uras, Sardinia, Italy
CN 22965	*hirtula*	693	A_sA_s	Thibo, Tunisia
CN 25414	*lusitanica*	137	A_sA_s	Cordoba, Spain
CN 25777	*lusitanica*	140	A_sA_s	Senera, Portugal
CN 25807	*lusitanica*	141	A_sA_s	Ben Slimane, Morocco
CN 25843	*lusitanica*	142	A_sA_s	Essaouira, Morocco
CN 19342	*wiestii*	138	A_sA_s	Chalus, Iran
CN 24352	*wiestii*	178	A_sA_s	Misaf Hanegev, Israel
CN 19348	*wiestii*	696	A_sA_s	East Azerbaijan, Iran
CIav 1994	*wiestii*	801	A_sA_s	Giza, Egypt
CN 58138	*longiglumis*	149	A_lA_l	Oran, Algeria
CN 23032	*canariensis*	118	A_cA_c	Fuerteventura, Canary Is
CN 23029	*canariensis*	151	A_cA_c	Fuerteventura, Canary Is
CN 25449	*canariensis*	684	A_cA_c	Fuerteventura, Canary Is
CN19457	*damascena*	687	A_dA_d	Syria
CN19459	*damascena*	688	A_dA_d	Syria
PI 657458	*damascena*$_2$	740	A_dA_d	Ait Kemara, Morocco
PI 657471	*damascena*$_2$	743	A_dA_d	Laassara, Morocco
PI 657472	*damascena*$_2$	744	A_dA_d	Nador, Morocco
PI 657587	*agadiriana*	772	$A_1A_1A_2A_2$	Tifnit, Morocco
PI 657588	*agadiriana*	773	$A_1A_1A_2A_2$	Tiznit, Morocco
PI 657589	*agadiriana*	774	$A_1A_1A_2A_2$	Tiznit, Morocco
CIav 9008	*sativa nuda*	811	AACCDD	Czechia
PI 401812	*sativa nuda*	822	AACCDD	Germany

APPENDIX B: UNABBREVIATED A-GENOME DIPLOIDS ROOTED TO THE *A. ERIANTHA* REFERENCE (ER_CN 19238)

APPENDIX C: UNABBREVIATED C-GENOME DIPLOIDS ROOTED TO THE *A. ATLANTICA* REFERENCE (AT_CC 7277)



APPENDIX D: SUMMARY OF THE REPEAT ELEMENT CONTENT IN THE AMARANTH GENOME ASSEMBLY AS IDENTIFIED BY REPEATMASKER RELATIVE TO THE REPBASE-DERIVED REPEATMASKER LIBRARIES.

Repeat Class‡	***A. atlantica*** (3,673,044,503 bp)			***A. eriantha*** (3,776,743,233 bp)		
	Count	Bases masked	Masked	Count	Bases masked	Masked
DNA	12343	2295245	0.06%	41071	13892671	0.37%
CMC-EnSpm	176565	183627714	5.00%	235692	181347535	4.80%
MULE-MuDR	16890	5211297	0.14%	13913	9147023	0.24%
MuLE-MuDR	6397	5618469	0.15%	13476	18601079	0.49%
Maverick	148	18931	0.00%	--	--	--
PIF-Harbinger	34425	10378659	0.28%	53235	28107529	0.74%
TcMar-Stowaway	82949	13125880	0.36%	108832	20258806	0.54%
hAT-Ac	2924	849064	0.02%	6590	2464996	0.07%
hAT-Tag1	878	410769	0.01%	5261	3946889	0.10%
hAT-Tip100	1561	651355	0.02%	1875	932660	0.02%
LINE	--	--	--	--	--	--
CR1	922	101045	0.00%	--	--	--
Jockey	145	30396	0.00%	5977	4017019	0.11%
L1	42540	33266727	0.91%	44555	36045210	0.95%
R1	1345	438210	0.01%	--	--	--
L2	--	--	--	573	326088	0.01%
RTE-X	--	--	--	1738	877096	0.02%
LTR	32110	49218294	1.34%	14612	5824958	0.15%
Copia	312901	641161159	17.46%	254114	522841719	13.84%
Gypsy	705163	1758990581	**47.89%**	715788	1829333860	**48.44%**
Viper	--	--	--	469	281891	0.01%
RC	--	--	--	--	--	--
Pao	519	285086	0.01%	--	--	--
Helitron	--	--	--	1695	568721	0.02%
L1	10983	4727496	0.13%	8142	1669307	0.04%
tRNA	3080	530660	0.01%	6237	4121298	0.11%
Unknown	533080	322656906	8.78%	693112	447222379	11.84%
Total interspersed	1977868	3033593943	82.59%	2226957	3131828734	82.92%
Low_complexity	22741	1212274	0.03%	21382	1166133	0.03%
Satellite	5217	2364614	0.06%	3404	943623	0.02%
Telo repeat‡	--	--	--	1815	14459837	0.38%
Simple repeat	176100	10467715	0.28%	162410	10363028	0.27%
Total	2181926	3047638546	82.97%	2415968	3158761355	83.64%

‡SINE, short interspersed nuclear elements; LINE, long interspersed nuclear elements; LTR, long terminal repeat; RC, Rolling circle

‡‡A. Eriantha telomeric satelllite repeat
CTCAAACNTGTATCGGGTCTTATGGNCGATGNAAATCGCNTGGAACCCCAAAACTGTGG
GCAATACCTCATGAAAACGGCCATAAAACGCGAAAACTACGAGTTTCGTGTCATAACAT
GTATCGGGTCTTACGGTCNTTGTAAATCNCCCTAGAACCCCAAAACCGTGAGCAATAGCT
CATGAAAACGGCCATAAAACGCGAAAAAGACGAGTTTTTGGTCATATCTCTCAAACATGT
ATCGGGTCTTACGNTCGTTGTAAATCGCCNTGGAACCCCAAAATTGTGGGCAATAGCTCA
TGAAAACGGCCATAAAACGCTGAAACATGTATCGGGTCTTACGGTCGTTGTAAATNNCCC
TAGAACCCCAAAACTGTGGGCAATAGCTCATGAAAACGGCCATAAAACGCGAAAACGAC
GAGTTTTTGGTCATATCTCTCAAACATGTATCGGGTCTTACGGTCGTTGTAAATCGCCCTG
AAACCCCAAAACTGTGGGCAATAGCTCATGAAAACGGCCATAAAGCGCGAAAACGACGA
GTTTTTGGTCATATCTCTCAAACNTGTATCGGGTCTTACGGTCNNTGNAAATCGCCTGGA
ACCCCAAAACTGTGGGCAATAGCTCATGAAAACGGCCATAAAACGCNAAAACTACGAGT
TTTTGTCAT
‡‡A. atlantica telomeric satelllite repeat
CTCAAACATGTATCGTGTCTTGCTGTCATTTTAAATCGCCCTGGAACACCAANANTATGG
GCAATAACTCATGAAAACGGCCATAAAACGCGAAAACGACGAGTTCTTGGTCATGACTC
TCAAACATGTAAATCGCCTTGGAACCCCAAAACTGTGGGCAATANCTCNTGAAAACGGC
CNTAAAACACGAAAATGGAGAGTTTTTGGTCATGCCNTCAAACATGTATCGGGTCTTACG
GTCATTTTAAATCGCCCTGGAACCCCAATATTATGGGCAATAACTCATGAAAACGGCCAT
AAAACGCGAAAACAACGAGTGCTTGGTCATAACTCTCAAACATGTAAATCGCCTTGGAA
CCCCAAAACTGTGGGCAATNGCTCATGAAAACGGCCATAAAACACGAAAATGGAGAGTT
TTTGGTCATGCCCTCAAACATGTATCGGGTCTTACGGTCATTTTAAATCGCCCTGGAACCC
CAATATTATGGGCAATAACTCATGAAAACGGCCATAAAACGCGAAAACGACGAGTTTTT
GGTCATAACTCTCAAACATGTAAATCGCCTTGGAACCCCAAAACTATGGGCAATNGCTCA
TGAAAACGGCCATAAAACGCGAAAACGGNGAGTTTTTGGTCAT

‡‡A. Eriantha cetnromeric satelllite repeat (rnd-1_family-822#unknown)
TGATGCAGCCCAACACATGGNAATCACCATTGGTCCACATCATGCTACGCCAACCATNNC
TACNAAGGTGAATTCCTTCCATCTCTTGCCTTCCTTTGGTCATCATGTGAATGGGATCCTA
CTTGAGACGAATNCCTATCGTATGACTAGGNCCATGAC

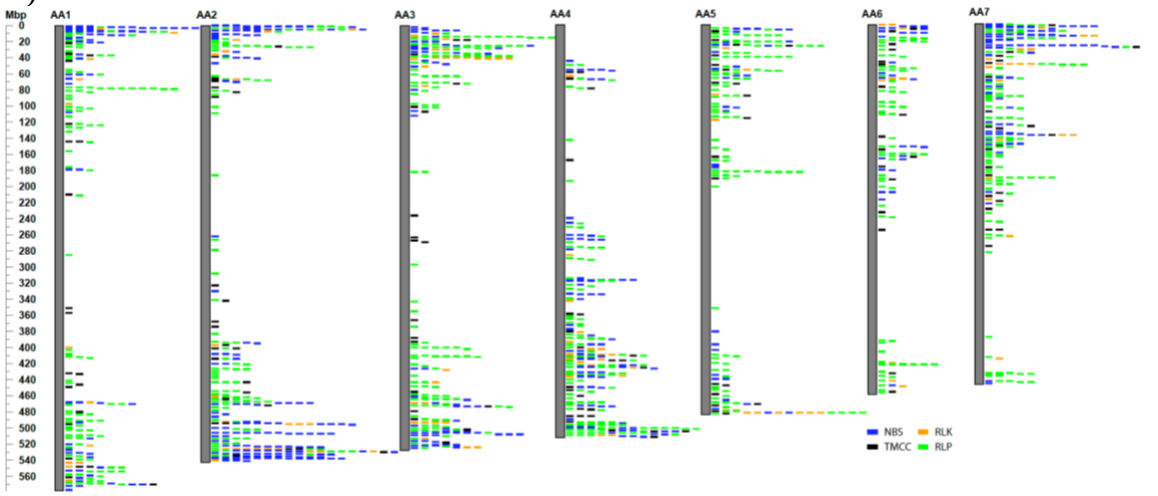
APPENDIX E: RESISTANCE GENE ANALOGS

Summary of resistance gene analog identification results using the RGAugury pipeline (Li et al. 2016) for A. atlantica and A. eriantha.

Species	Protein sequences	Mean Protein length (aa)	NBS encoding								RLP	RLK	TM-CC	Total
			NBS	CC-NBS-LRR	TIR-NBS-LRR	CC-NBS	TIR-NBS	NBS-LRR	TIR-unknown	Others				
A. atlantica	49,542	369	45	226	0	40	2	195	3	0	120	772	160	1563
A. eriantha	47,361	346	50	190	0	42	1	174	2	0	135	654	154	1402
CC: Coiled-coil; LRR: Leucine rich repeat; NBS: Nucleotide-binding site; RLP: Receptor like protein; RLK: Receptor like kinase; TIR: Toll/Interleukin-1 receptor; TM: Transmembrane														

Distribution of the resistance gene analogs (RGAs) encoding genes on the (A) A. atlantica and (B) A. eriantha genome. The scale is in megabases (Mb). NBS, RLK, RLP and TMCC are in blue, green, orange and black, respectively.

A)



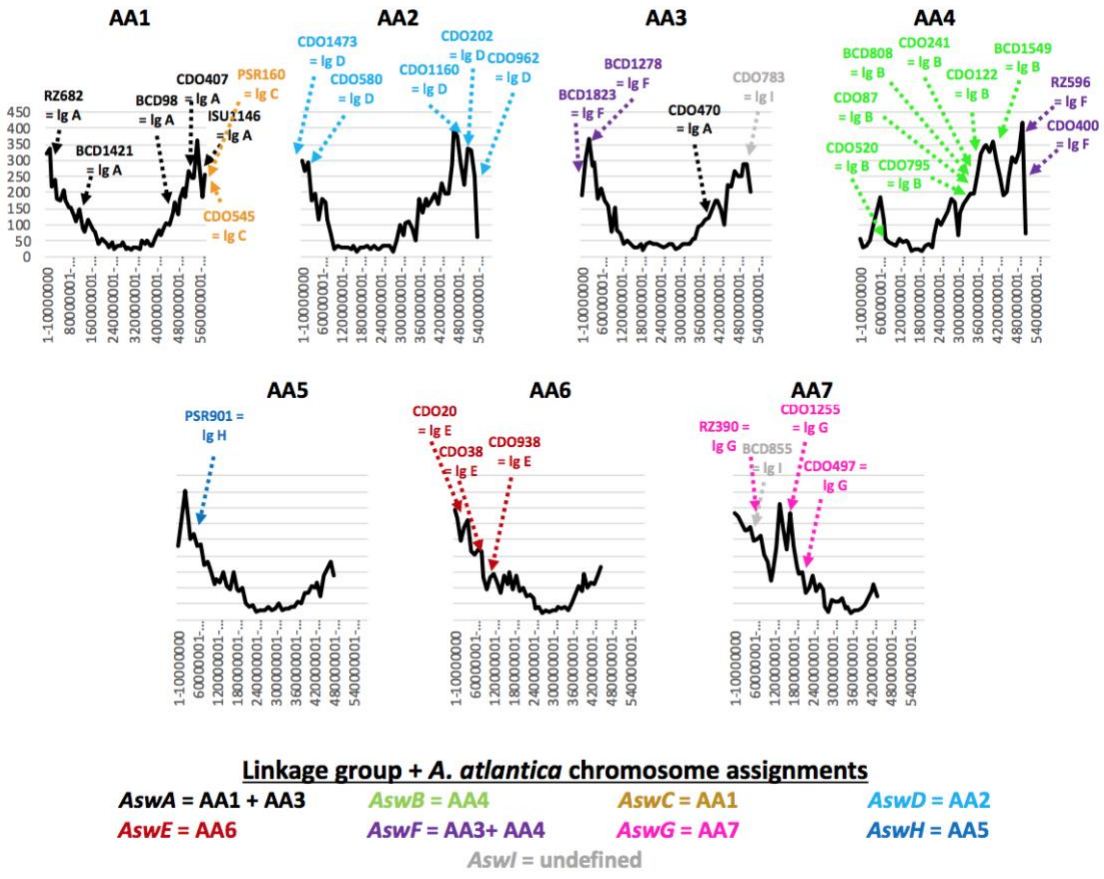
B)

APPENDIX F: CANDIDATE RESISTANCE GENES

Candidate resistance gene analogs associated with crown rust resistance on Mrg18 linkage group (Chaffin et al. 2016). Mrg18 was previously shown to be involved in a intergenomic translocation involving 7C and 17A, corresponding to A. eriantha chromosome AE7 and A. atlantica chromosome AA2. Klos et al. 2017 identified two QTLs associated P. coronata (crown rust) resistance on Mrg18, one of which determined to be Pc91. Candidate resistance gene analogs were identified using BLAST searches against the A. atlantica and A. eriantha genome assembly using makers sequences associated with the QTLs.

QTL	*P. coronata* resistance gene	SNP name	Linkage group	cM position	Chromosomal designation	Chromosome	Marker mapping position	E-value	Mapping species	Closest RGA annotation	Distance to closest RGA (bp)	RGA classification	Annotation ID	Closest Annotation Description
QPc.CORE.18.1	NA	GMI_D5_LB_2908	Mrg18	21.1	C/A	AE7	70,260,284	1E-103	Eriantha	70,245,021	(15,263)	NBS-NL	AE028733	Similar to RPM1: Disease resistance protein RPM1 (Arabidopsis thaliana)
QPc.CORE.18.2	*Pc91*	GMI_ES03_c2277_336	Mrg18	67.7	C/A	AA2	496,455,574	1E-57	Atlantica	496,480,229	24,655	NBS-NL	AA013068	Similar to RGA3: Putative disease resistance protein RGA3 (Solanum bulbocastanum)
QPc.CORE.18.3	*Pc91*	GMI_ES05_c11155_383	Mrg18	67.7	C/A	AA2	533,475,317	5E-53	Atlantica	533,698,614	223,297	NBS-CNL	AA014151	Similar to RPM1: Disease resistance protein RPM1 (Arabidopsis thaliana)
QPc.CORE.18.3	*Pc91*	GMI_GBS_24408	Mrg18	67.7	C/A	AA2	510,519,361	5E-25	Atlantica	510,828,316	308,955	NBS-CNL	AA013376	Similar to RPH8A: Disease resistance protein RPH8A (Arabidopsis thaliana)

APPENDIX G: POLYMORPHISM MARKERS



Corresponding location of restriction fragment length polymorphism markers mapped on a segregating A. strigosa X A. wiestii population developed by Kremer et al. on the A. atlantica chromosomes (shown on a gene density plot). Markers from each of their linkage groups (Asw A-I) are color-coded with approximate positions on A. atlantica scaffolds indicated with arrows.

APPENDIX H: LIST OF PHYTOZOME PRERELEASE GENOMES USED IN *AVENA*-SPECIFIC GENES STUDY

1. *Amaranthus hypochondriacus* v2.1 (Amaranth) (Lightfoot, et al., 2017)
2. *Anacardium occidentale* v0.9 (Cashew)
3. *Asparagus officinalis* V1.1 (Asparagus) (Harkess, et al., 2017)
4. *Arabidopsis thaliana* Araport11 (Thale cress) (Cheng, et al., 2017)
5. *Botryococcus braunii* v2.1
6. *Brachypodium distachyon* Bd21-3 v1.1 (Purple false brome)
7. *Brachypodium hybridum* v1.1
8. *Brachypodium sylvaticum* v1.1
9. *Chenopodium quinoa v1.0* (Quinoa) (Jarvis, et al., 2017)
10. *Chromochloris zofingiensis* v5.2.3.2 (Roth, et al., 2017)
11. *Citrus clementina* v1.0 (Clementine) (Wu, et al., 2018)
12. *Gossypium hirsutum* v1.1 (Upland cotton)
13. *Helianthus annuus* r1.2 (Sunflower) (Badouin, et al., 2017)
14. *Hordeum vulgare* r1 (Barley) (Beier, et al., 2017; Mascher, et al., 2017)
15. *Lactuca sativa* V8 (Lettuce) (Reyes-Chin-Wo, et al., 2017)
16. *Miscanthus sinensis* v7.1 (Chinese silvergrass)
17. *Olea europaea* var. sylvestris v1.0 (Wild olive) (Unver, et al., 2017)
18. *Oryza sativa* Kitaake v3.1 (Kitaake rice) (Li, et al., 2017)
19. *Panicum hallii* v3.1 (Hall's panicgrass)
20. *Panicum hallii* var. hallii v2.1 (Hall's panicgrass)
21. *Panicum virgatum* v1.1 (Switchgrass)
22. *Populus deltoides* WV94 v2.1 (Eastern cottonwood)
23. *Populus trichocarpa* v3.1 (Poplar)
24. *Porphyra umbilicalis* v1.5 (Laver) (Brawley, et al., 2017)
25. *Sorghum bicolor* Rio v2.1 (Sorghum Rio)
26. *Setaria viridis* v2.1 (Green foxtail)
27. *Vigna unguiculata* v1.1 (Cowpea)

APPENDIX I: BLAST ANALYSIS PIPELINE

```
#!/bin/bash

#PBS -N asg-eriantha
#PBS -l walltime=336:00:00
#PBS -l nodes=1:ppn=8
##PBS -l vmem=2000mb
#PBS -q copperhead

cd /nobackup/oat_genome/rachel/asg

module load blast
module list 2>&1

INFOFILE="eriantha.info.txt"
ELIMINATED="eriantha.eliminated.fasta"
MAKEELIMDB="eriantha.eliminated"

INPUT="eriantha.transcript.fa"
SPECIES="eriantha"


blastn -db step1transcript -query $INPUT -evalue 1e-4 -num_threads 8 -outfmt 6 -out $SPECIES.step1.out

cut -f 1 $SPECIES.step1.out | sort | uniq > $SPECIES.step1.names.out
python parseblastout.py $SPECIES.step1.names.out $INPUT $SPECIES.afterasg1.fasta $ELIMINATED $INFOFILE

blastx -db step2prot -query $SPECIES.afterasg1.fasta -evalue 1e-4 -num_threads 8 -outfmt 6 -out $SPECIES.step2.out

cut -f 1 $SPECIES.step2.out | sort | uniq > $SPECIES.step2.names.out
python parseblastout.py $SPECIES.step2.names.out $SPECIES.afterasg1.fasta $SPECIES.afterasg2.fasta $ELIMINATED $INFOFILE

blastx -db /projects/oat_genome/rachel/nr/nr -query $SPECIES.afterasg2.fasta -negative_gilist avena13Feb19.gi -evalue 1e-4 -num_threads 8 -outfmt 6 -out $SPECIES.step3.out

cut -f 1 $SPECIES.step3.out | sort | uniq > $SPECIES.step3.names.out
python parseblastout.py $SPECIES.step3.names.out $SPECIES.afterasg2.fasta $SPECIES.afterasg3.fasta $ELIMINATED $INFOFILE

blastn -db prerelease -query $SPECIES.afterasg3.fasta -evalue 1e-4 -num_threads 8 -outfmt 6 -out $SPECIES.step4.out
```

```
38.
39.    cut -f 1 $SPECIES.step4.out | sort | uniq >
    $SPECIES.step4.names.out
40.    python parseblastout.py $SPECIES.step4.names.out
     $SPECIES.afterasg3.fasta $SPECIES.afterasg4.fasta $ELIMINATED
     $INFOFILE
41.
42.    makeblastdb -in $MAKEELIMDB.fasta -dbtype nucl -out
     $MAKEELIMDB
43.    blastn -db $MAKEELIMDB -query $SPECIES.afterasg4.fasta
     -evalue 1e-4 -num_threads 8 -outfmt 6 -out $SPECIES.step5.out
44.
45.    cut -f 1 $SPECIES.step5.out | sort | uniq >
     $SPECIES.step5.names.out
46.    python parseblastout.py $SPECIES.step5.names.out
     $SPECIES.afterasg4.fasta $SPECIES.afterasg5.fasta $ELIMINATED
     $INFOFILE
```

APPENDIX J: BLAST PARSER – PYTHON SCRIPT

```
1.    #!/usr/bin/python
2.
3.    ########################################
4.    ## parseblastout.py                    ##
5.    ## Rachel Walstead                     ##
6.    ## February 8, 2019                    ##
7.
8.    ## USAGE:
9.    ## Run BLAST with -outfmt 6, then:
10.   ## cut -f 1 <blastout> > <blastoutnames>
11.   ## python3 parseblastout.py <blastoutnames>
      <blastinput fasta> <fasta out> <eliminated out> <info out>
12.
13.
14.   import sys
15.
16.   blastoutnames = sys.argv[1]
17.   fastain = sys.argv[2]
18.   fastaout = sys.argv(Pope, et al.)
19.   eliminateout = sys.argv[4]
20.   infoout = sys.argv[5]
21.
22.   def parsefasta(fastainfile):
23.    fasta = open(fastainfile, 'r')
24.    flag = 0
25.    fastadict = {}
26.    name, fullname = '', ''
27.
28.    for line in fasta:
29.       line = line.strip()
30.       if line.startswith('>'):
31.          if flag == 1:
32.             fastadict[name] = (fullname, seq)
33.          fullname = line
34.          name = line.split()[0].strip('>')
35.          seq = ''
36.       else:
37.          seq = seq + line
38.          flag = 1
39.    fastadict[name] = (fullname, seq)
40.    fasta.close()
41.    return fastadict
42.
43.
44.
45.   fastadict = parsefasta(fastain)
46.   #print(fastadict)
47.
48.   elimcount = 0
49.   outcount = 0
50.   outfile = open(fastaout, 'w')
51.   eliminatedfile = open(eliminateout, 'a')
```

```
52.
53.   namelist = []
54.
55.   with open(blastoutnames) as blastfile:
56.    for line in blastfile:
57.       line = line.strip() #preprocess line
58.       namelist.append(line)
59.   #print(namelist)
60.
61.   for key in fastadict:
62.    if key in namelist:
63.       elimcount += 1
64.       name = fastadict[key][0]
65.       seq = fastadict[key][1]
66.       eliminatedfile.write(name + "\n" + seq + "\n")
67.    else:
68.       outcount += 1
69.       name = fastadict[key][0]
70.       seq = fastadict[key][1]
71.       outfile.write(name + "\n" + seq + "\n")
72.
73.   outfile.close()
74.   eliminatedfile.close()
75.
76.   infofile = open(infoout, 'a')
77.   infofile.write(str(elimcount)+" added to "+eliminateout+"\n")
78.   infofile.write(str(outcount)+" written to "+fastaout+"\n\n")
79.   infofile.close()
80.
81.   print("done")
```