Annotating and making use of the Avena sativa cv. Sang reference genome

Tsardakas Renhuldt, Nikos

2023

Document Version:
Publisher's PDF, also known as Version of record

Annotating and making use of the *Avena sativa* cv. Sang reference genome

NIKOS TSARDAKAS RENHULDT
DEPARTMENT OF CHEMISTRY | FACULTY OF ENGINEERING | LUND UNIVERSITY

Annotating and making use of the *Avena sativa* cv. Sang reference genome

Annotating and making use of the *Avena sativa* cv. Sang reference genome

Nikos Tsardakas Renhuldt

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Engineering, Lund University, Sweden.
To be defended at Lecture Hall A, Kemicentrum, Naturvetarvägen 14, Lund.
1st of September 2023 at 13.00.

Faculty opponent
Prof. Pär K. Ingvarsson

Department of Plant Biology, Swedish University of Agricultural Sciences, Uppsala

Organization: LUND UNIVERSITY

Document name: DOCTORAL DISSERTATION **Date of issue:** 1st of September 2023

Author: Nikos Tsardakas Renhuldt

Title and subtitle: Annotating and making use of the *Avena sativa* cv. Sang reference genome

Abstract:

Oats is an important cereal used for both food and feed. The topic of this thesis is the annotation of the genome of oat (*Avena sativa*) cv. Sang, as well as some of the things this genome and its annotation have been used for.

The first part of the thesis provides a short background to oat genomic resources, genomic resources in other plant species, and assembly of cereal genomes. Following this, it goes through the pipeline used to annotate the oat genome, covering various tools used, mentioning annotation pipelines used for other plant genomes, and comparing the results of cv. Sang annotation to annotations of other released oat genomes. It also briefly discusses a couple of tools used for functional annotation and identification of homologous genes.

The following chapter looks at how this annotation may be used. It describes the pipeline used to identify homologous genes and provides an overview of how this was used to identify genes involved in epicuticular wax biosynthesis and in detoxification of *Fusarium* mycotoxins.

Use of genetic markers, including how they have already been used both to identify breeding barriers in oats, and to establish that this oat reference genome corresponds to cv. Sang are brought up in the next chapter. How the markers may be aligned to the genome and how they may be visualized are also discussed.

Next, mapping-by-sequencing is discussed, providing more details regarding the work done to identify the genes *AsCer-q* and *AsGSK2.1*. The method is explained, including selection of the number of individuals to include in the sequenced pools as well as the choices made to filter variants and genes. A background on the identified genes is also provided, before concluding with some thoughts regarding future work.

Key words: oats, *Avena sativa*, genomics, genome annotation, epicuticular wax biosynthesis, brassinosteroids, GSK3, GFViz, visualization, bioinformatics, mapping-by-sequencing, gene mapping

Classification system and/or index terms (if any) Supplementary bibliographical information

Language: English **ISSN**:

ISBN: 978-91-7422-966-0

Recipient's notes **Number of pages**: 63

Price Security classification

Signature Date 2023-05-16

Annotating and making use of the *Avena sativa* cv. Sang reference genome

Nikos Tsardakas Renhuldt

LUND
UNIVERSITY

Faculty of Engineering
Department of Chemistry
Division of Pure and Applied Biochemistry

ISBN 978-91-7422-966-0 (print)
ISBN 978-91-7422-967-7 (digital)

Printed in Sweden by Media-Tryck, Lund University
Lund 2023

När man ser på hur barna
växer upp och står i
kan man undra om barna
nånsin får det som vi.
Om det finns jobb
om det finns mat
om det är drägligt där dom bor?
Finns det får och kor
och vatten och luft?

- *Hasse och Tage, Ett glas öl*

Contents

Popular science summary

Oats is an important crop used to feed both humans and animals. Many people eat oat products daily, and regular consumption of oat beta-glucans is known to lower the risk of heart disease. In spite of this, the full genetic sequence of oats has been unknown until recently, and a comprehensive overview of oat genes has been missing. This has made life harder for both researchers and plant breeders. In order to ensure that people have food to eat, we need to be able to improve our crops and make them more resilient to a changing climate while minimizing their climate impacts.

The first part of this thesis focuses on the creation of a comprehensive list of oat genes and how we identified where in the oat genetic sequence they are located. Not only does this list identify the location of these genes, but it also provides an idea about gene function, and a way of relating oat genes to similar genes in other species. This will help researchers use what is known about other plants when doing research on oats going forward. We used this connection to genes in other species, along with a barley gene known to be involved in resistance against *Fusarium* infections, to identify oat genes with similar function.

Another part of the thesis focuses on so called 'genetic markers', short pieces of genetic sequence with some known properties. When studying large populations, or when the full genetic sequence is not known, we can get the relative position of markers, and this may help us identify approximate locations of genetic regions controlling different plant traits. This thesis looks at different ways of connecting existing markers to the oat genetic sequence, how they may be visualized, and how we can combine them with information about oat genes. The combination of previous research identifying genetic regions controlling traits, and new lists of oat genes open doors for researchers and plant breeders to understand which gene variants cause different traits, study the exact function of these variants further, or select for variants that are beneficial.

Not only does this research allow us to connect to what has already been studied, but it also lets us study the genes underlying traits seen in mutant oats. By crossing mutants with non-mutant oats and looking at which parts of the genetic sequence is identical among siblings that share a trait, we were able to identify the function of two genes in oats, relating to oat surface wax, and oat hormones and development respectively. Plant surface wax helps plant conserve water and also ensures that water does not stick to the plant surface, and may also play a role in defense against pathogens. Plant hormones control innumerable aspects of plant development, and understanding this interplay may help produce oats with

higher yields or improved stress resistance.

This thesis provides a look at how to find genes in plants with large genetic sequence, as well as a first peek of what researchers will be able to achieve with access to the oat genes and genetic sequence. Hopefully they will prove a valuable resource to researchers and breeders alike, and hopefully it will let them create better oats, helping to ensure that people have food to eat in spite of the ongoing climate crisis.

Abstract

Oats is an important cereal used for both food and feed. The topic of this thesis is the annotation of the genome of oat (*Avena sativa*) cv. Sang, as well as some of the things this genome and its annotation have been used for.

The first part of the thesis provides a short background to oat genomic resources, genomic resources in other plant species, and assembly of cereal genomes. Following this, it goes through the pipeline used to annotate the oat genome, covering various tools used, mentioning annotation pipelines used for other plant genomes, and comparing the results of cv. Sang annotation to annotations of other released oat genomes. It also briefly discusses a couple of tools used for functional annotation and identification of homologous genes.

The following chapter looks at how this annotation may be used. It describes the pipeline used to identify homologous genes and provides an overview of how this was used to identify genes involved in epicuticular wax biosynthesis and in detoxification of *Fusarium* mycotoxins.

Use of genetic markers, including how they have already been used both to identify breeding barriers in oats, and to establish that this oat reference genome corresponds to cv. Sang are brought up in the next chapter. How the markers may be aligned to the genome and how they may be visualized are also discussed.

Next, mapping-by-sequencing is discussed, providing more details regarding the work done to identify the genes *AsCer-q* and *AsGSK2.1*. The method is explained, including selection of the number of individuals to include in the sequenced pools as well as the choices made to filter variants and genes. A background on the identified genes is also provided, before concluding with some thoughts regarding future work.

List of papers

Papers included in the thesis

Paper I: Kamal*, N., **Tsardakas Renhuldt***, N., Bentzer, J., Gundlach, H., Haberer, G., Juhász, A., Lux, T., Bose, U., Tye-Din, J. A., Lang, D., van Gessel, N., Reski, R., Fu, Y.-B., Spégel, P., Ceplitis, A., Himmelbach, A., Waters, A. J., Bekele, W. A., Colgrave, M. L., Hansson, M., Stein, N., Mayer, K. F. X., Jellen, E. N., Maughan, P. J., Tinker, N. A., Mascher, M., Olsson, O., Spannagl, M., and Sirijovski, N. (2022). "The Mosaic Oat Genome Gives Insights into a Uniquely Healthy Cereal Crop". In: *Nature* 606.7912 (7912), pp. 113–119. DOI: 10.1038/s41586-022-04732-y

*: These authors contributed equally to the publication.

Paper II: Khairullina, A., **Tsardakas Renhuldt**, N., Wiesenberger, G., Bentzer, J., Collinge, D. B., Adam, G., and Bülow, L. (2022). "Identification and Functional Characterisation of Two Oat UDP-Glucosyltransferases Involved in Deoxynivalenol Detoxification". In: *Toxins* 14.7 (7), p. 446. DOI: 10.3390/toxins14070446

Paper III: Tsardakas Renhuldt, N., Bentzer, J., Ahrén, D., Marmon, S., and Sirijovski, N. (2023). "Mutation in SHAGGY-like kinase AsGSK2.1 causes a short kernel phenotype in oat (*Avena sativa*)". *Manuscript*

Paper IV: Tsardakas Renhuldt, N. (2023). "GFViz: A tutorial on creating interactive visualization of genomic features using R Tidyverse and plotly". *Manuscript*

Papers not included in the thesis

Tinker, N. A., Wight, C. P., Bekele, W. A., Yan, W., Jellen, E. N., **Tsardakas Renhuldt**, N., Sirijovski, N., Lux, T., Spannagl, M., and Mascher, M. (2022). "Genome Analysis in Avena Sativa Reveals Hidden Breeding Barriers and Opportunities for Oat Improvement". In: *Communications Biology* 5.1 (1), pp. 1–11. DOI: 10.1038/s42003-022-03256-5

Sardari, R. R. R., Jasilionis, A., **Tsardakas Renhuldt**, N., Adlercreutz, P., and Karlsson, E. N. (2023). "HPAEC-PAD Analysis for Determination of the Amino Acid Profiles in Protein Fractions from Oat Flour Combined with Correction of Amino Acid Loss during Hydrolysis". In: *Journal of Cereal Science* 109, p. 103589. DOI: 10.1016/j.jcs.2022.103589

Darwish, E., Ghosh, R., Bentzer, J., **Tsardakas Renhuldt**, N., Proux-Wera, E., Kamal, N., Spannagl, M., Hause, B., Sirijovski, N., and Van Aken, O. (2023). "The Dynamics of Touch-Responsive Gene Expression in Cereals". In: *The Plant Journal*. DOI: 10.1111/tpj.16269

My contribution to the papers

Paper I: I was involved in the annotation (training and running *ab initio* prediction with AUGUSTUS; running PASApipeline; merging predictions using EVidenceModeler; prediction of UTRs using PASApipeline; confidence classification; production of BUSCO stats), marker alignment (aligning of the 6K chip markers to the reference genome), mapping of the *AsCer-q* gene (mapped reads; called and filtered variants; produced the mutant allele frequency visualization along with candidate gene lists including multiple sequence alignments and phylogenetic trees and their respective visualizations; identified the gene as being mutated in a separate line sharing the phenotype by establishing a database of TILLING population variants), and in writing the paper, including the production of final figures.

Paper II: I used bioinformatics to identify potential orthologs of UGT genes known from barley. This included sequence searches using DIAMOND, multiple sequence alignments using muscle, and production of trees using fasttree, along with analysis and interpretation of these results, guiding the choice of which oat genes to validate experimentally.

Paper III: I participated in study planning, ranging from discussions regarding pool sizes for sequencing to which experiments to use in validating brassinosteroid insensitivity. I also performed the bioinformatics analyses involved in the mapping-by-sequencing (mapped reads; called and filtered variants; produced the mutant allele frequency visualization along with candidate gene lists including multiple sequence alignments and phylogenetic trees and their respective visualizations), as well as the statistical analyses. Further, I identified the oat homologs of previously published seed shape genes, identified the locations of previously published QTL markers in cv. Sang, and created the visualization of oat seed shape genes and QTL markers. I did most of the writing as well as figure production and formatting.

Paper IV: I built an interactive visualization tool for genomic markers and other data, including gene locations, and wrote the tutorial on how to recreate it using R.

List of abbreviations

Abbreviation	Meaning
AHRD	Assignment of Human Readable Descriptions
As	*Avena sativa*
At	*Arabidopsis thaliana*
BAC	Bacterial Artificial Chromosome
BES1	BRI1-EMS-SUPPRESSOR 1
BIN2	BRASSINOSTEROID-INSENSITIVE 2
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
Bp	Basepair
BR	Brassinosteroid
BUSCO	Benchmarking Universal Single-Copy Orthologue
BZR	BRASSINAZOLE RESISTANT
cDNA	Complementary DNA
Cer	*Eceriferum*, waxless
CORE	Collaborative Oat Research Enterprise
Csl	Cellulose Synthase-Like
Cv.	Cultivar
CYP	Cytochrome P450
DKS	Diketone Synthase
DMC	Diketone Metabolism CYP
DMH	Diketone Metabolism Hydrolase
DMP	Diketone Metabolism PKS
EFSA	European Food Safety Authority
EMS	Ethyl Methanesulfonate
EVM	EVidenceModeler
FAO	Food and Agriculture Organization of the United Nations
Gb	Gigabases
GBS	Genotyping-By-Sequencing
GCMS	Gas Chromatography–Mass Spectrometry
GFP	Green Fluorescent Protein
GFViz	Genomic Feature Visualization
GL3.3	GRAIN LENGTH 3.3

Abbreviation	Meaning
GMAP	Genomic Mapping and Alignment Program
GO	Gene Ontology
GSK	GSK3/SHAGGY-like kinase
GSK3	Glycogen Synthase Kinase 3
GWAS	Genome-Wide Association Study
HC	High Confidence
HISAT	Hierarchical Indexing for Spliced Alignment of Transcripts
HMM	Hidden Markov Model
Hv	*Hordeum vulgare*
IONC	International Oat Nomenclature Committee
IWGSC	The International Wheat Genome Sequencing Consortium
KIB1	KINK SUPPRESSED IN BZR1-1D 1
LC	Low Confidence
Mb	Megabases
MCL	Markov Cluster Algorithm
Mrg	Merge Group
MSA	Multiple Sequence Alignment
ORF	Open Reading Frame
PASA	Program to Assemble Spliced Alignments
PGSB	The Plant Genome and Systems Biology group at Helmholtz Munich
PKS	Polyketide Synthase
QTL	Quantitative Trait Loci
RBNH	Reciprocal Best Normalized Hit
SEM	Scanning Electron Microscopy
SG	Semispherical Grain
SK	SHAGGY-like KINASE
SNP	Single-Nucleotide Polymorphism
SnpEff	SNP Effect
STAR	Spliced Transcripts Alignment to a Reference
T-DNA	Transfer DNA
Ta	*Triticum aestivum*
TAIR	The Arabidopsis Information Resource
TE	Transposable Element
TGW3	THOUSAND GRAIN WEIGHT 3
TILLING	Targeting Induced Local Lesions IN Genomes
TREP	TRansposable Elements Platform
UDP	Uridine Diphosphate
UGT	UDP Glycosyltransferase
UTR	Untranslated Region
WES	Wax Ester Synthase

Preface

I came into this project during the fall of 2018. I wanted to do bioinformatics and something related to plant breeding, and the research center ScanOats – an oat research collaboration involving Lund University, the Swedish University of Agricultural Sciences, Lantmännen and Oatly – was looking for a PhD student interested in oat genomics. Initially I was not stoked about oats, but Nick Sirijovski, who would be my supervisor during these years, sold me on the project by telling me about this impressive collaboration that was set up, and how there would be a bunch of experts in genome assembly and annotation involved. This, combined with promises of access to solid bioinformatics supervision from Dag Ahrén and Björn Canbäck[1], access to impressive computational resources, and a promise regarding not having to touch a pipette for the coming five years, sealed the deal.

The first wheat reference genome was released in August of that year – a 16Gbp short-read assembly, an impressive step forward for one of the most important food crops in the world. As I was getting started, Martin Mascher at IPK Gatersleben was in the final stages of scaffolding the cv. Sang oat (*Avena sativa*) genome into pseudomolecules using lessons learned from the wheat genome assembly process. In the spring of 2019, the assembly had been finalized and handed over to Manuel Spannagl and the Plant Genome and Systems Biology (PGSB) group at Helmholtz Munich, and they began the laborious process of annotating this 11Gb genome. In June of 2019 I took a train to Munich, and spent a couple of weeks learning from them, working closely with Thomas Lux and Nadia Kamal in particular, getting an initial understanding of what their annotation pipeline looked like, and discussing what I might be able to add to this. It was settled that I would add an additional set of predicted transcripts to the annotation, make sure that some type of prediction of untranslated regions (UTRs) was added, and do a final confidence classification. The work took the bulk of that fall, but in early 2020 an initial gene annotation had been performed. This was once again handed back to Nadia, Manuel, Thomas, and the other brilliant people at PGSB, where functional annotations were added along with information on gene homology.

With functional annotation and homology information in place, a new world of possibilities opened up. Together with my colleague Johan Bentzer (responsible for many of the expression analyses in this project, along with invaluable work in making the genome and

[1]Björn did a brief stint as my supervisor before continuing his career outside of academia. During my interview he stressed the importance of regular coffee breaks and a sensible work-life balance, which ended up a major selling point for me.

its annotation accessible by providing both a genome browser and custom data analysis interfaces, as well as numerous other contributions), and my supervisors Nick Sirijovski, Dag Ahrén, and Sofia Marmon, my work has spanned a number of different projects, integrating multiple different types of data, allowing us to do science that would have otherwise been impossible. Together with Alfia Khairullina we have identified proteins involved in the detoxification of *Fusarium* mycotoxins (**paper II**), we have worked with Alfredo Zambrano on identifying genes involved in the biosynthesis of arabinoxylans (not yet published), and – as a part of the reference genome publication (**paper I**) – with Nadia Kamal in identifying cellulose synthase genes. Having had access to the functional annotation, and homology information in particular, has been invaluable in this process.

In parallel with work using the functional annotation, Nick Tinker and Wubishet Bekele did work on mapping previous genotyping-by-sequencing (GBS) marker data sets to the new reference genome, with me doing the same thing for the older 6K BeadChip markers. Nick and Wubi provided analyses of recombination rates, and also identified breeding barriers resulting from reorganizations across the genomes of different oat varieties. Having these marker locations anchored to our reference genome gives us the ability to connect previous work from genome-wide association studies (GWAS) and work on identifying quantitative trait loci (QTL) to the reference genome and its annotation. My hope is that the visualization tool that I developed to study these QTLs and candidate genes (**paper IV**) more closely will be useful to other researchers, and that it may help guide others in establishing gene function in oat as well as in other species.

Mapping-by-sequencing is made possible by the above, and in particular by access to the reference genome and to the functional annotation. I have found mapping-by-sequencing to be a very effective tool in establishing gene function. This effectiveness has in part been driven by access to a TILLING population based on cv. Belinda, which is closely related to the reference genome cv. Sang. Mapping-by-sequencing in combination with the functional annotation has helped us identify gene function in a way that would not otherwise have been possible, and in a way that has been very resource efficient, requiring only a single cross, few generations, and small number of phenotype plants. It has allowed us to start peeking at how oats work, it is a method I am very enthusiastic about, and I hope to see others continuing to use it to shed light on gene function.

Chapter 1

Introduction

This thesis deals with how the oat (*Avena sativa*) cv. Sang reference genome was annotated. It also covers some of the things this genome has been used for so far.

This introduction provides some background on oat genome research, and an overview of some of the previous research in some other plants and cereals. It also discusses genome assembly in cereal genomes. In the chapters following this introduction I discuss the genome annotation process (**paper I**), followed by the work done in identifying oat homologs of barley proteins involved in wax biosynthesis (**paper I**) and mycotoxin detoxification (**paper II**). Following this is a chapter on the use of markers and how to make previous marker studies useful (**paper III** and **paper IV**), and a chapter about mapping-by-sequencing and how we used that to identify both the previously mentioned wax biosynthesis proteins and proteins involved in kernel size (**paper I** and **paper III**). Finally there is a chapter on the biochemistry connected to these proteins, before concluding with a discussion on some promising future work.

1.1 Why bother with an oat genome?

As I am typing this, I am eating oat porridge and drinking my oat milk frappé – in Sweden, oats is a very significant crop culturally. Globally, oats was grown on 9.8m hectares with a total production of 25m tonnes in 2020, making it the 7th most produced cereal according to the Food and Agriculture Organization of the United Nations, FAO (2022). A large proportion of oats is used for animal feed – within the EU, 77% of oats were used for this in 2019-2020 (European Commission, Directorate-General for Agriculture and Rural Development 2023). The past decades have seen a more diverse set of oat products being developed, ranging from oat milk to cosmetics. Oat food products in particular have seen an interest in part because of their health claims, with oats being known to decrease risk of heart disease, as well as reduce glycemic responses after meals (European Food Safety Authority, EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA) 2010, 2021).

A lot of research has been done on oats prior to this thesis. The chromosome number

and ploidy of *A. sativa* ($2n = 6x = 42$) was established more than a hundred years ago, with speculation regarding its evolutionary origins going back further than that (Kihara 1919). The subgenomes of the hexaploid were established to be A, C and D in Rajhathy and Morrison (1959), and the subgenome origins has continued to be an active research question (Peng, Yan, Guo, et al. 2022). It has also been shown that chromosome structure varies across oat species (Rajhathy and Thomas 1974), as well as that different varieties of *A. sativa* have different sets of translocations (Jellen and Beard 2000; Jellen, Gill, and Cox 1994; Jellen, Rines, et al. 1997).

Early genetic maps were published 30 years ago (O'Donoughue et al. 1992), and a number of different marker sets have been published since (e.g. Jannink and Gardner 2005; Tinker, Kilian, et al. 2009). The 6K marker chip and GBS marker datasets described in Tinker, Chao, et al. (2014) and Huang et al. (2014), respectively, were based on work done as a part of the Collaborative Oat Research Enterprise (CORE). They were both used as inputs to different parts of the work done in this thesis. The 6K marker chip was a BeadChip using both Infinium I and Infinium II designs. It was able to assay 4975 single-nucleotide polymorphisms (SNPs), and integrated previously published SNPs with new ones into an assay that was easily commercially available. These combined with the Huang et al. (2014) GBS markers were used in the initial characterization of the genetic diversity of the 635 lines in the CORE diversity panel. The CORE diversity panel consists mainly of spring oat varieties, but also has lines included to capture global diversity, and specifically to capture diversity in the southern US and South America (Esvelt Klos et al. 2016). The protocol used in Huang et al. (2014) has been the basis for a number of subsequent oat GBS papers, including updated analyses of the genetic diversity of the CORE collection, and in construction of high-density consensus linkage maps (Bekele et al. 2018; Chaffin et al. 2016). These maps have in turn been used in various studies of quantitative trait loci (QTLs), including studies on oat beta-glucans and on seed shape (Fogarty et al. 2020; Zimmer et al. 2021).

Another useful resource has been the oat Targeting Induced Local Lesions IN Genomes (TILLING) population (Chawade et al. 2010). It consists of approximately 2500 oat lines from cv. Belinda mutagenized using ethyl methanesulfonate (EMS). EMS mainly causes GC to AT point mutations, which have the benefit of being easy to identify using short-read sequencing data. These induced mutations and the phenotypes they cause have allowed us to study gene function in oats, but the usefulness of both the TILLING population and the QTL studies have been hampered by the lack of a high quality, annotated reference genomes – the main topic of this thesis.

Beyond the needs of researchers, there is a more general argument to be made for why this research is worth doing relating to the importance of plant breeding and crop improvement. We are facing an urgent climate crisis along with an increasing global population: people need to eat, and in order to do that we need crops that are capable of handling increasingly extreme weather, as well as crops that do not further contribute to climate change[1]. Oats is interesting in this context not only because of its health claims and generally lower climate impacts of plant based foods, but also because it has received relatively little attention

[1]This is of course mainly a political issue, not a technological one. I despair at the thought of trying to breed crops capable of feeding all of us in the climate that the current emissions trajectory is rapidly hurling us toward.

from researchers and breeders, with lower yields and profitability than e.g. wheat or barley (Gorash et al. 2017).

All of the above – oats being an important but under-researched crop, the lack of a reference genome making both research and breeding difficult, the necessity of plant breeding to ensure that people have food to eat that does not unnecessarily contribute to further changing the climate – are reasons why producing an annotated oat genome has been a worthwhile undertaking.

1.2 Previous plant genomic resources

More than 100 crops have had their genomes sequenced, with the improvement of sequencing technologies allowing the sequencing of progressively larger and more complex genomes (Purugganan and Jackson 2021). *Arabidopsis thaliana* was the first plant to have its genome (115 megabases, Mb) sequenced (The Arabidopsis Genome Initiative 2000), and in the decades since The *Arabidopsis* Information Resource (TAIR, Berardini et al. 2015) has served as a valuable resource for the plant community. Among other things, TAIR provides functional and Gene Ontology (GO) annotations for genes, orthology information, and a comprehensive database of previously published *Arabidopsis* literature, all of which may be useful for doing work in *Arabidopsis*, as well as in studying gene function in other plant species (Reiser et al. 2022).

Rice was the first cereal to have its genome (420 Mb) sequenced, with draft genomes released in 2002 (Goff et al. 2002; Yu et al. 2002). The maize genome (2.3 gigabases, Gb) was released in 2009 (Schnable et al. 2009). In the years since, a number of new versions of these genomes have been released along with genomes of different genotypes of each of these species (Hufford et al. 2021; Wang et al. 2018). These genomes have allowed researchers to identify the genetic architecture of many traits, investigate issues ranging from domestication to genetic diversity and heterosis, and provide a foundation to studies of gene function (Song, Tian, et al. 2018; Yang and Yan 2021).

1.3 Assembling cereal genomes

Our oat sequencing project was heavily influenced by the recent work done in barley and wheat (Mascher, Gundlach, et al. 2017; The International Wheat Genome Sequencing Consortium (IWGSC) et al. 2018). The barley genome (4.8 Gb) was constructed by sequencing bacterial artificial chromosomes (BACs) using short-read pair-end and mate-pair Illumina technology, which were assembled and scaffolded. Genetic markers and maps were used to group scaffolds by chromosome, and Hi-C sequencing data was used to order and orient these into pseudomolecules corresponding to the barley chromosomes. BACs were a part of the wheat genome (17 Gb) sequencing project as well, but there the whole genome assembly relied to a larger extent on Illumina libraries with differing insert sizes (450 basepairs (bp) – 10 kbp), which were assembled by the company NRGene.

Our allohexaploid (AACCDD) oat assembly of cv. Sang avoided the laborious and expensive BACs entirely. Scaffolds were assembled by NRGene using pair-end and mate-pair

Illumina libraries (mate-pair insert sizes ranging from 2 kbp – 10 kbp), as well as 10X Chromium data. The Bekele et al. (2018) consensus map was used as a guide map, and merge (Mrg) groups – together with the 10X and Hi-C data – were used to construct the pseudomolecules from the NRGene scaffolds using the TRITEX short-read pipeline (Monat et al. 2019). This resulted in an 11 Gbp reference sequence, split across 21 pseudochromosomes (**paper I**). As had been indicated by previous research (e.g. Chaffin et al. 2016; Jellen and Beard 2000; Jellen, Gill, and Cox 1994; Jellen, Rines, et al. 1997), the genome had large rearrangements relative to its progenitors, including translocations across the subgenomes.

Gradually, long reads have been replacing short reads as the tool of choice for *de novo* genome assemblies. For the *A. longiglumis* (AA) and *A. insularis* (CCDD) genomes – relatives of the A and CD subgenome progenitors of *A. sativa*, respectively, and published as a part of **paper I** – PacBio continuous long reads sequencing was used, with Illumina short reads used for polishing. Similarly, the recently released *A. sativa* ssp. *nuda* cv. Sanfensan used Oxford Nanopore long-read data, again with Illumina short reads for polishing (Peng, Yan, Guo, et al. 2022). The OT3098 reference was initiated later than the cv. Sang sequencing project but was released without an accompanying publication of its own, prior to the publication of **paper I**. In **paper I** we took the opportunity to include a number of comparisons of cv. Sang and OT3098. The OT3098 assembly was constructed using PacBio HiFi reads, and scaffolded using BioNano (Waters 2022a). In all of the above cases, Hi-C data was used for pseudochromosome assembly. The oat pangenome project PanOat (*PanOat: The Oat Pangenome Project | GrainGenes* 2023) is an ongoing research collaboration that aims at assembling and annotating 29 oat genotypes. Many of the PanOat genomes combine PacBio HiFi reads and Hi-C data, as this has been shown to be a cost-effective way of generating high-quality assemblies for these big genomes (Mascher, Wicker, et al. 2021).

To summarize, this introduction has provided a background on some previous oat research, and talked briefly about genomic resources present in other plants and what they have been used for. It has also mentioned how the cv. Sang genome was assembled, and described the assembly of cereal genomes broadly. This introduction also explained why creating an annotated oat genome is worthwhile. The next chapter discusses the annotation of plant genomes in general, with a particular focus on how we annotated the cv. Sang genome along with some of the results of that work.

Chapter 2

Genome annotation

With the reference sequence in place, the cv. Sang genome was ready for annotations, i.e., marking which sections of the genome correspond to e.g. protein coding genes or transposable elements (TEs). This chapter digs into the strategy and specific tools that were used for the annotation of protein coding genes performed as a part of **paper I**, and also discusses strategies used in other annotation projects. It provides some results from the annotation of cv. Sang, and compares the final annotation results of cv. Sang, cv. Sanfensan, and OT3098. It also discusses functional annotation, and how homologs in other species were identified as a part of the functional annotation of cv. Sang. This chapter will not discuss the significant work done by others in annotating TEs, also as a part of **paper I**[1]. Before we dig into the annotation of protein coding genes, however, a short note regarding oat nomenclature in general, and gene and protein identifiers in this thesis.

2.1 Oat nomenclature

A direct result of the work done in **paper I** has been a new internationally recognized nomenclature for oat chromosomes and genes, agreed upon by the International Oat Nomenclature Committee (IONC). Oat chromosomes are now designated 1-7 along with a letter to indicate the subgenome (e.g., 1C, 4A), with numbering based on the chromosome's phylogenetic relationship to other *Triticeae*, and chromosome orientation determined by the orientation of a conserved core region. Oat genes are identified using a string consisting of five dot-separated fields, such as AVESA.00010b.r2.3AG0419820.1. The first field (AVESA) is a five-character species indicator; the second field (00010b) is a six-character germplasm identifier where the letter indicates germplasm version; the third field (r2) indicates the annotation version; the fourth field (3AG0419820) starts with the chromosome (3A), followed by G for gene, and a numeric identifier that increments along the chromosomes. The final field (1) corresponds to the isoform/transcript id (International Oat Nomenclature Committee 2021a,b). Hopefully this shared nomenclature will

[1] TE annotation relies on different sets of tools and databases, and has been necessary in producing a high-quality gene annotation, but it is not work that I have been involved in.

make oat research easier going forward: naming will be synchronized across the genomes released as a part of the oat pangenome project PanOat, identifying which annotation a publication uses will be easier since this information is embedded in the gene identifier, and studies of synteny may be simplified by a consistent chromosome numbering both within the *Avenas* but also connected to other *Triticeae*. As a part of the work done in establishing the new chromosome nomenclature, IONC has also released a table connecting previous nomenclatures to the most recent one, facilitating the integration of previous studies with these new genomic resources (International Oat Nomenclature Committee 2021a).

In this thesis, gene names will be italicized, while proteins will be capitalized. This means that *AsCer-q* corresponds to the gene sequence of the *A. sativa* (*As*) *eceriferum* (waxless, *cer*) gene, where the *-q* suffix is given due to its homology to *HvCer-q*, the corresponding gene in *Hordeum vulgare* (*Hv*, barley). AsCER-Q refers to the protein that *AsCer-q* codes for. *AsGSK2.1* is capitalized since GSK is an acronomy for glycogen synthase kinase 3/SHAGGY-like kinase (GSK3/SHAGGY-like kinase), but refers to the gene, while AsGSK2.1 refers to the protein.

2.2 Annotating the genome

This section will first briefly discuss how we evaluate genome annotations, and will bring up the tool BUSCO (Benchmarking Universal Single-Copy Orthologue, Manni, Berkeley, Seppey, and Zdobnov 2021) in particular. This is followed by a discussion of how we annotated the oat genome in **paper I**, going through the PGSB (Plant Genome and Systems Biology group at Helmholtz Munich) pipeline, confidence classification of genes, protein-coding gene prediction using AUGUSTUS (Stanke et al. 2008), and integration of additional predicted transcripts and prediction of untranslated regions (UTRs) using PASApipeline (Program to Assemble Spliced Alignments, Haas, Delcher, et al. 2003), with EVidenceModeler (EVM, Haas, Salzberg, et al. 2008) used to merge everything together. An overview of the full genome annotation pipeline is shown in figure 2.1. This section also covers approaches taken in a few other genome annotation projects, along with some results from the annotation of cv. Sang, and comparisons to cv. Sanfensan and OT3098. My contributions here are mainly in running AUGUSTUS, PASA, EVM, and performing evaluations and generating summary results for this thesis. I was not involved in running the PGSB pipeline.

2.2.1 Evaluating genome annotations

A number of different measures are used in evaluating the genome annotation process. These include gene numbers, overlap between genes, the proportion of predicted genes that have both start and stop codons present, the distribution of exons per gene and the number of single exon genes, the number of different splice variants predicted, gene length distributions, how the genes are distributed across the genome, RNA-seq coverage of gene regions, RNA-seq mapping rates to a transcriptome derived from the annotation, etc. All of these may help diagnose issues in the annotation process, and help us trust the annotation product, but as with many things, there are few absolute rules in doing these evaluations, and it is hard to rank the quality of annotations that fall within some acceptable range

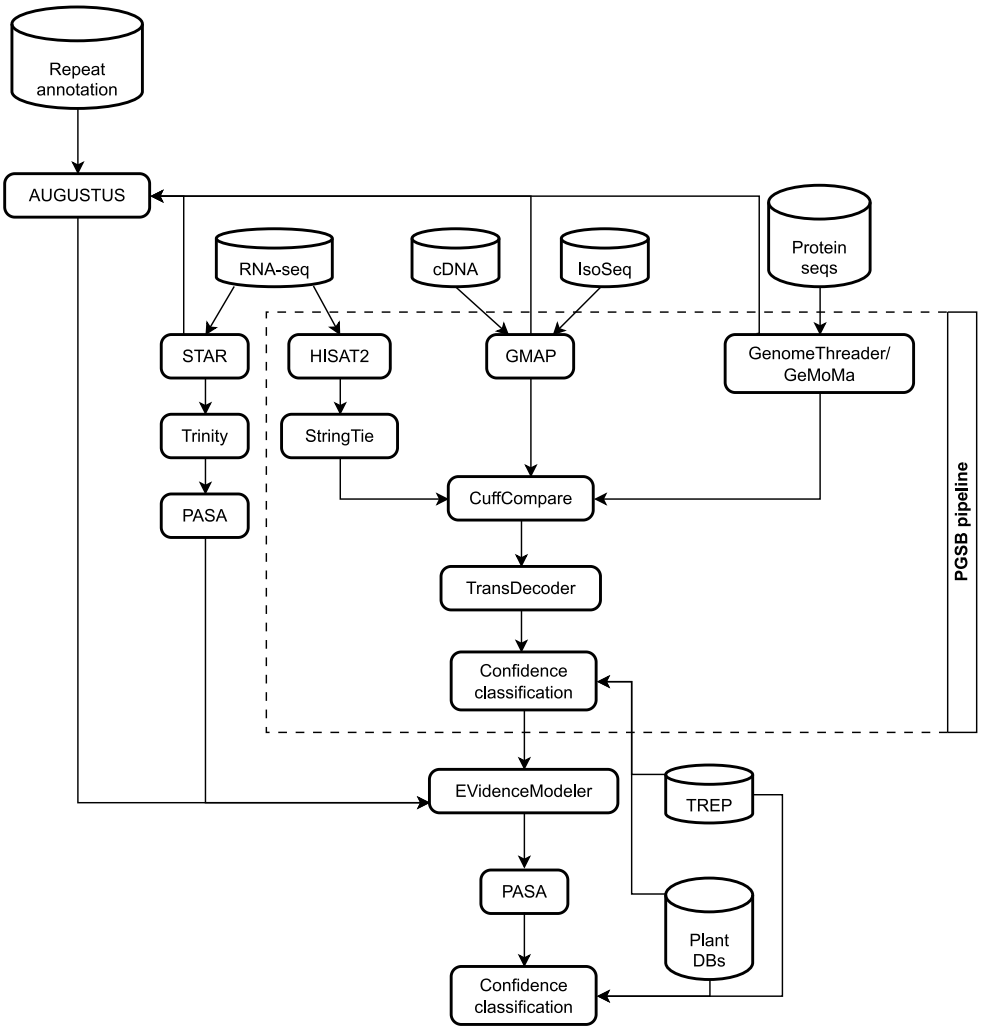

Figure 2.1: Overview of the annotation pipeline used in **paper I**, with the PGSB pipeline highlighted.

of these measures. Here, BUSCO (Manni, Berkeley, Seppey, Simão, et al. 2021; Manni, Berkeley, Seppey, and Zdobnov 2021; Simão et al. 2015) may be of some help.

The basic idea of BUSCO is that there is a set of single-copy genes that are expected to be present within (almost) all organisms belonging to a clade, and that presence or absence of these may help indicate the level of completeness of an assembly or annotation. An annotation containing all the expected BUSCO genes is likely more complete than an annotation that is lacking a large proportion of said genes. The lists of single-copy genes is curated based on OrthoDB (Zdobnov et al. 2021) and consist of genes with single-copy orthologs present in at least 90% of the clade. BUSCO may be used to evaluate genomes, transcriptomes or proteins. In the latest versions (v4, v5) of BUSCO, protein sequences are extracted from eukaryotic genomes and transcriptomes using MetaEuk (Levy Karin, Mirdita, and Söding 2020), with Prodigal (Hyatt et al. 2010) used to extract proteins from prokaryote genomes and tBLASTn[2] (Basic Local Alignment Search Tool, Camacho et al. 2009) used for prokaryote transcriptomes. In all cases, HMMER 3 (Eddy 2011) is used to identify BUSCOs among the predicted proteins. For the BUSCO statistics to make sense, only one representative transcript per gene should be evaluated. Each BUSCO is represented by a Hidden Markov Model (HMM) profile. If a single predicted protein matches the profile, i.e., has a score above the cutoff and is within the expected length interval, that BUSCO is considered 'complete'; with multiple matches, the BUSCO is considered 'duplicated'; a BUSCO with a match that is too short but has a score above the threshold is considered 'fragmented'; and a BUSCO that lacks matches passing the threshold is considered 'missing'.

In our allohexaploid, we are expecting most BUSCOs to be present in triplicates. BUSCO has been used to evaluate the assembly quality, but it has also been extensively used in the ongoing evaluations of different parts of the genome annotation pipeline, as well as in evaluating thresholds used for confidence classification of gene models. It has been a valuable tool in assuring quality of the results and in guiding the annotation process. Figure 2.2 – shown in the results section of this chapter – provides an overview of BUSCO results for the different steps of the annotation process used to annotate cv. Sang in **paper I**, as well as a comparison of the final cv. Sang annotation to other oat genome annotations.

2.2.2 The PGSB annotation pipeline

In this thesis 'the PGSB annotation pipeline' refers to the workflows for genome annotation used by PGSB over the past several years. There are some differences in regards to what exactly is included as a part of the pipeline – to my knowledge, the work I did with AUGUSTUS as a part of **paper I** was one of the earlier attempts at including *ab initio* gene predictions as a part of the pipeline, and it was not used in PGSB's work for e.g. the first barley or wheat annotations (IWGSC et al. 2018; Mascher, Gundlach, et al. 2017). This section will discuss the PGSB pipeline as used for oat (see figure 2.1 for an overview), and covers the annotation process up until the first confidence classification, with AUGUSTUS, PASApipeline, EVidenceModeler and confidence classification being discussed in later sections.

[2]BLAST is the famous sequence search tool, used to find sequences that are similar to each other. tBLASTn matches protein sequences against nucleotide sequences translated in all six reading frames.

The PGSB pipeline is a way of integrating diverse types of data that are available to a project into a single annotation – essentially it boils down to aligning these data to the genome, merging the alignments, identifying the best transcripts, and assigning a confidence class to each transcript. In practice, this means identifying available RNA-seq data present in public repositories and within the project, and mapping these using HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts, Kim, Paggi, et al. 2019), and assembling these into transcripts using StringTie (Pertea et al. 2015). GMAP (Genomic Mapping and Alignment Program, Wu and Watanabe 2005) is used to map full length cDNA sequences and IsoSeq data if these are available, and GenomeThreader (Gremme et al. 2005) and GeMoMa (Keilwagen, Hartung, et al. 2018; Keilwagen, Wenk, et al. 2016) may be used to align known protein sequences to the genome. Cuffcompare from Cufflinks (Trapnell et al. 2010) is used to merge these into a set of consensus transcripts. TransDecoder (Haas 2021) is then used to identify the best open reading frames (ORFs) and transcripts, based on homology searches of candidate transcripts against databases of known proteins. This output is filtered to ensure that it is non-redundant, before being assigned a confidence class.

2.2.3 Confidence classification

The confidence classification step of the PGSB pipeline assigns confidence based on three criteria. These are completeness, i.e., the gene model having both a start and a stop codon; known homologous genes, i.e., the gene model having matches in public databases; and whether the gene model matches the repeat database TREP (The TRansposable Elements Platform, Schlagenhauf and Wicker 2016). Complete gene models that match the non-repeat databases are considered high confidence (HC), while genes that are either incomplete or do not match any of the non-repeat databases are considered low confidence (LC). Gene models with matches in TREP are clasified as repetitive, unless they are complete and have a match in the non-repeat databases.

In short: a HC gene model by this definition is a gene model that both has a start and a stop codon (i.e., is complete), and that is similar to other known genes. This may seem straightforward, but opens other questions: which are the appropriate databases to compare to here? How similar is similar enough, i.e., what cutoffs do we use when comparing similarity to other databases? How do we evaluate the cutoffs that we set?

During the final confidence classification of the oat annotation, BUSCO was used to evaluate the similarity cutoffs. This allowed us to consider a trade-off between the number of genes classified as HC relative to the number of BUSCO genes included in the LC set. It is worth noting that this is only one measure of quality, and that doing a comprehensive evaluation and ranking of multiple cutoffs is a time-consuming analysis.

There is no universally agreed upon definition of what constitutes a HC gene model, and there is no obviously correct way to evaluate this – an example of a different way of doing classification is mentioned later in this chapter. This means that comparing HC gene models across annotations is hard, since definitions and ways of evaluating this vary across the annotations. When working with or filtering on gene model confidence class, be aware of which definition of HC you are dealing with. If you are working with multiple annotations, be aware that the confidence classifications are unlikely to be comparable. This incompa-

rability may be due to entirely different methods of evaluating confidence being used, but even if the overall framework is the same, databases change over time, and the cutoffs used are likely different. In many cases, ignoring the confidence classification entirely is the way to go, and if some type of filtering on gene model quality is necessary, it may be better to focus on measures such as presence of RNA-seq evidence or similar.

2.2.4 AUGUSTUS

AUGUSTUS (Stanke et al. 2008) is a tool for protein-coding gene prediction. It uses statistical modeling (HMMs, more recently also conditional random fields) to predict whether genomic sequences are protein coding or not. These models are trained on datasets of known coding and non-coding genomic regions, and are then evaluated on genomic sequences to predict genes. These predictions may be done *ab initio*, i.e., without any additional information beyond the model parameters, or using additional data (RNA-seq, protein alignments, repeat annotations, etc.) as hints.[3]

During training of the AUGUSTUS model for oat, the paper by Hoff and Stanke (2019) was used. It provided valuable guidance and the concrete steps to take to train the AUGUSTUS model, including optimization of meta parameters. Even so, this process was not entirely straightforward: identifying and formatting suitable training data was non-trivial, as was evaluation of model results. The model we ended up training predicted a very large number of false-positives, with more than 900,000 predicted gene models, most of which lacked any type of transcript or protein evidence. In the end, the AUGUSTUS output was used for recovering incomplete exons with transcript support that would otherwise have been discarded by EVidenceModeler, but presence of an AUGUSTUS prediction on its own was not deemed sufficiently reliable to be included in the final annotation.

2.2.5 PASApipeline

PASApipeline (Haas, Delcher, et al. 2003) had two uses within this project: as an additional way to integrate RNA-seq data by aligning transcripts to the genome, and as a way to add UTRs and alternative splicing to the gene models.

For the first use case, RNA-seq data was mapped using STAR (Spliced Transcripts Alignment to a Reference, Dobin et al. 2013) and transcripts were assembled using Trinity (Grabherr et al. 2011). What PASA does is clean these transcripts; aligning them to the genome using BLAT (BLAST-like alignment tool, Kent 2002), GMAP or minimap2 (Li 2018); and filtering for alignments that both meet high identity thresholds along a large proportion of the transcript length, and that have appropriate splice sites. Using their mapping locations, these are clustered and gene structures are assembled such that the the number of compatible transcript alignments in each assembly is maximized. Overlapping assemblies are grouped to form clusters of assemblies, which may in turn be used to output a new annotation, or update an existing one (Haas 2023).

[3] It is worth noting that both '*ab initio*' and '*de novo*' are used to describe these predictions. It is also worth noting that '*ab initio*' may also be used to describe predictions by AUGUSTUS or similar tools even when run using additional evidence.

For adding UTRs and alternative splicing isoforms, the annotation produced by EVM was loaded into PASA's database, and updated using the PASA transcript alignments. PASA transcripts overlapping the annotation gene models are used to update these. In order for these updates to occur, the PASA transcripts must share orientation and meet a number of additional requirements. PASA assumes that the existing annotation has a high quality and integrates a lot of existing evidence. The criteria for updating a gene model are configurable, but by default these include the PASA transcript having e.g., at least 50% overlap with the annotation, and the updated protein sequence having at least 70% identity to the previous protein sequence along at least 70% of the annotated protein length, as well as a number of additional criteria.

2.2.6 EVidenceModeler

EVM (Haas, Salzberg, et al. 2008) is a tool used to merge evidence from different annotations of protein-coding genes. In our case, this meant merging the outputs of the PGSB pipeline with the results produced by PASApipeline and AUGUSTUS into one unified annotation. EVM merges annotations by building a graph of exons, introns and intergenic regions and finding the highest scoring path through the graph, with different scores assigned based on the weights of different sources of evidence. The weights may be set intuitively, with e.g. *ab initio* gene predictors having lower weights than protein alignments, in turn having lower weight than e.g. IsoSeq alignments. In our case, this meant that an exon from a HC PGSB gene model based on transcript evidence would be preferred to e.g. an exon from a PASA model or from protein alignment evidence. Beyond the weights, EVM also takes evidence types, which determine which parts of the annotation the evidence is allowed to affect and how scores are calculated. EVM produces neither UTRs nor alternative splicing isoforms, but running a PASA update following EVM adds these to the annotation.

It is worth noting that weights are added as a part of the scoring of paths through the graph, i.e., multiple predicted gene models from a source may add up to calculate the support for an exon. This means that evidence with lower quality (e.g., *ab initio* predictions) may overrule higher-quality evidence (e.g., aligned IsoSeq data) if the number of lower quality predictions is high or if weights are not sufficiently different. If all input data is of similar quality and output similar number of gene models this might not be a problem, but if you have reason to be more distrustful of some of your input or if some of your methods produce a large number of models relative to the other methods, do make sure that that your weight settings account for this. As mentioned previously, our AUGUSTUS results contained a large number of false positive gene predictions, which meant that we needed to be careful in our choice of EVM weights.

Also worth noting is that only exons that have a start and a stop codon; a start codon and a donor splice junction; an acceptor splice junction and a donor splice junction; or an acceptor splice junction and a stop codon can be considered as a candidate exon by EVM, but incomplete exons may be used as evidence in selecting among complete exons. This is where AUGUSTUS came into our annotation process: we did not deem the AUGUSTUS results sufficiently reliable on their own, but an incomplete gene model from PASA or PGSB, lacking e.g. a start codon, could be rescued if it overlapped a complete AUGUSTUS

exon.

For **paper I**, several sources of evidence were provided as both `OTHER_PREDICTION` and `PROTEIN` evidence types – this allowed them to contribute all types of exons, while also allowing incomplete exons to contribute in the predictions. Weights were set manually to ensure that more reliable evidence (e.g. high confidence transcript alignments) were preferred to less reliable evidence (e.g. AUGUSTUS predictions). A more detailed discussion regarding this is available at *Turn Non-Recoverable Predictions into Alignments?* (2023). Multiple sets of weights were used, and the resulting annotations were mainly evaluated using BUSCO scores.

2.2.7 Protein coding gene annotation in other plant genomes

The tools described above were the ones used in annotation of the oat genome as a part of **paper I**. The PGSB pipeline has also been used in various other species, including barley, rye, wild emmer, several species of *Aegilops*, and wheat (Avni, Lux, et al. 2022; Avni, Nave, et al. 2017; IWGSC et al. 2018; Mascher, Gundlach, et al. 2017; Rabanus-Wallace et al. 2021). The barley genome published in Mascher, Gundlach, et al. (2017) relied entirely on the PGSB pipeline, with the pangenome (Jayakodi et al. 2020) combining using this pipeline for some accessions with projection of the gene models onto all accessions using BLAT (Kent 2002) and exonerate (Slater and Birney 2005).

The wheat cv. Chinese Spring genome combined the PGSB pipeline with the TriAnnot pipeline (Leroy et al. 2012), merging the results into one annotation (IWGSC et al. 2018). In brief, the TriAnnot version used for annotation of the wheat genome works by first masking TEs; aligning transcript and protein sequences to the genome; gene finding using both *ab initio* methods and methods relying on alignment evidence; and quality filtering of the gene models. As of writing this thesis, TriAnnot does not seem to have a publicly available implementation (*DECODAGE – Communauté d'Annotation Des Génomes - TriAnnot* 2023). A later publication of 15 wheat genomes (Walkowiak et al. 2020) relied on the cv. Chinese Spring annotations and projected these onto the new genomes using BLAT (Kent 2002) and exonerate (Slater and Birney 2005).

Other released oat genomes used different annotation tools. The OT3098 v2 release used AUGUSTUS with wheat parameters and IsoSeq hints to predict genes. BLASTing against the National Center for Biotechnology Information BLAST database of non-redundant protein sequences (National Library of Medicine (US), National Center for Biotechnology Information 2004–2023) was used to identify genes with evidence and provide gene identities. Only predicted genes with at least 20% overlap to IsoSeq alignments were kept (Waters 2022b). The cv. Sanfensan annotation (Peng, Yan, Guo, et al. 2022) used IsoSeq data to predict genes using GeneMarkS-T (Tang, Lomsadze, and Borodovsky 2015). Data from protein databases was mapped using GeMoMa. Both of these were combined with *ab initio* gene predictions from AUGUSTUS trained using GeneMark-ET (Lomsadze, Burns, and Borodovsky 2014) using EVM following filtering by TransposonPSI (Haas 2010).

For the 3,000 Rice Genomes Project, they used the annotation pipeline MAKER (Holt and Yandell 2011; Wang et al. 2018). MAKER masks the genome using RepeatRunner (Smith et al. 2007) and RepeatMasker (Smit, Hubley, and Green 2013–2023); produces

Table 2.1: Number of genes and transcripts present following different steps of the annotation of cv. Sang, as well cv. Sang v1.1, cv. Sanfensan, and OT3098 v2.

Annotation	Genes	Transcripts
First PASA run	245,438	245,438
PGSB results	117,494	320,853
AUGUSTUS	960,475	960,475
EVM results	153,594	153,594
PASA update	152,836	175,444
Sang v1.1	152,335	174,318
Sanfensan	120,769	120,769
OT3098 v2	68,572	273,045

ab initio predictions using AUGUSTUS and SNAP (Korf 2004); and aligns protein and transcript evidence using BLAST and exonerate. MAKER also has additional capabilities for selection of gene models, but it seems as if Wang et al. (2018) preferred to use EVM for merging the *ab initio* predictions and the alignments, followed by filtering out of incomplete or redundant gene models. Novel genes were evaluated by verifying presence using whole-genome sequencing data and expression using RNA-seq data.

For 26 maize genomes published in Hufford et al. (2021), yet another approach was taken. Transcripts were assembled using five different programs, and merged using Mikado (Venturini et al. 2018). Mikado used splice sites generated by Portcullis (Mapleson et al. 2018), TransDecoder ORFs, and BLAST results against Swiss-Prot to select among transcripts. BRAKER (Hoff, Lomsadze, et al. 2019) was used to generate *ab initio* predictions of protein coding genes. PASA was used to add UTRs, additional isoforms and genes, and TEs were filtered out using TEsorter (Zhang, Wang, et al. 2019). Confidence classification of genes and annotation accuracy was based on Annotation Edit Distance scores calculated using MAKER-P (Campbell et al. 2014).

2.2.8 Annotation results

Table 2.1 gives an overview of the number of genes and transcripts that each of the pipeline steps resulted in, and figure 2.2 shows the BUSCO (v5.1.2, database poales_odb10 created 2020-08-05, Manni, Berkeley, Seppey, Simão, et al. 2021) distribution following each step – cv. Sanfensan (downloaded from *OatBioDB* 2023) and OT3098 v2 (Waters 2022b) are included as comparisons. Note that these numbers are based on all transcripts, i.e. not only representative transcripts, inflating the number of duplicated BUSCOs in annotations containing multiple isoforms per gene.

Paper I provides the detailed settings of the annotation pipeline, as well as the resulting gene numbers and gene distribution across the genome. In total, more than 150,000 gene models were predicted, with about 80,000 of these being HC. The genes are less abundant in the C subgenome, in spite of the C subgenome being slightly larger than the A and D subgenomes. The difference in gene abundance is explained by gene-rich regions of the C subgenome being translocated into the A and D subgenomes, while a likely explanation of

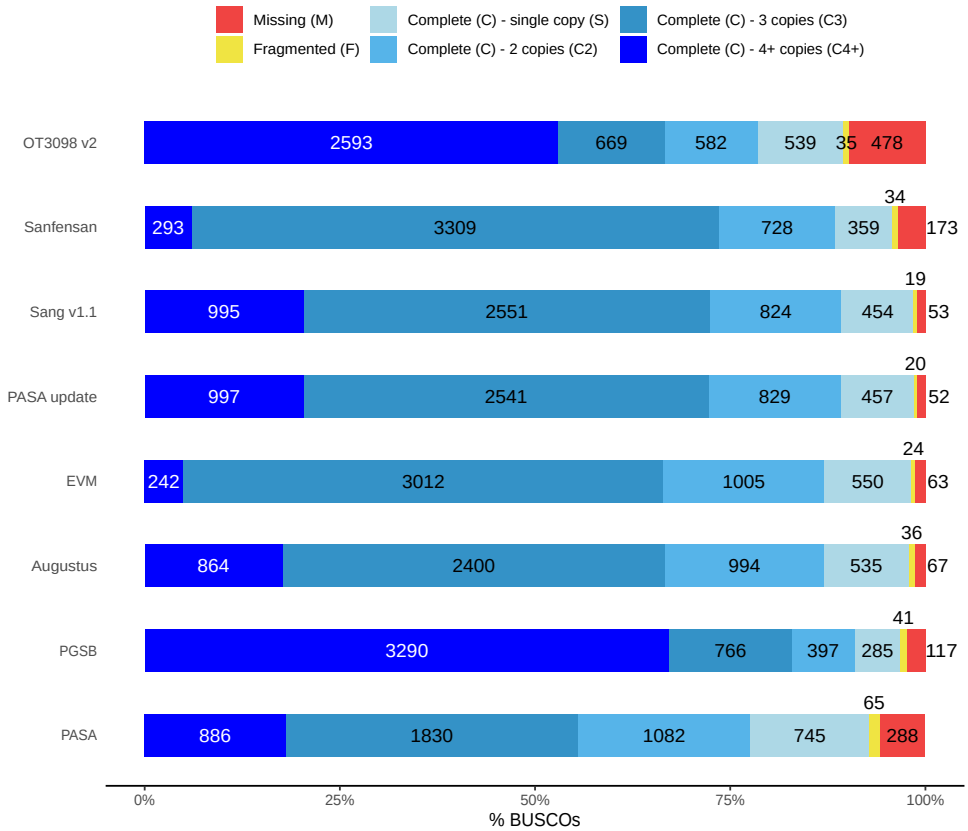

Figure 2.2: Number of BUSCOs (v5.1.2, database poales_odb10 created 2020-08-05, Manni, Berkeley, Seppey, Simão, et al. 2021) present following different steps of the annotation of cv. Sang, as well cv. Sang v1.1, cv. Sanfensan, and OT3098 v2. Note that BUSCO was run on all transcripts for each annotation, not only representative transcripts, making it hard to compare duplication rates.

the size difference is transposon activity in the C subgenome ancestor.

2.2.9 Discussion of the annotation of protein coding genes

Above I have provided some background on the genome annotation pipeline used in **paper I**, briefly presented some annotation approaches taken in other projects, and presented some results from the annotation process, including BUSCO scores for intermediate results in the cv. Sang annotation as well as for final annotations of cv. Sang, cv. Sanfensan and OT3098.

As shown above, there are multiple approaches to annotation of protein coding genes in plant genomes. Common to the approaches described here is the combination of *ab initio* predictions with alignment evidence based on RNA-seq, IsoSeq or previously known protein sequences. Many of the pipelines combine this with some type of filtering for TEs. The details of how these evidence types are generated differ even among the few annotations described above – even within the various genomes annotated in part using the PGSB pipeline the implementations differ.

Figure 2.2 shows the BUSCO scores of the currently available annotations of protein coding genes in oats, along with the BUSCO scores of intermediate steps in the process of annotating cv. Sang. Since the figure is based on total transcripts, comparison of duplicated genes is not possible, but it is possible to say something about missingness. Among the inputs to EVM we note PASA having a higher level of missingness than PGSB and AUGUSTUS. This is expected, not least since the PASA run relied only on the RNA-seq data aligned using STAR whereas the PGSB pipeline also relied on cDNA, IsoSeq data, and aligned protein sequences. AUGUSTUS catches a large proportion of BUSCOs, but also predicts close to one million gene models (see table 2.1). EVM reduces the number of genes relative to both AUGUSTUS and PASA while reducing the number of missing BUSCOs. The PASA update step manages to reduce both the number of fragmented and the number of missing BUSCOs.

Comparing the gene numbers and BUSCO scores of cv. Sang, cv. Sanfensan and OT3098 shows that Sang has both higher numbers of annotated genes, and lower number of missing BUSCOs than the other two genomes, likely meaning that it is more complete than the other two annotations. The OT3098 v2 annotation does not rely on alignment evidence from protein sequence databases, which may be one explanation for it missing many genes. The IsoSeq data used in annotating OT3098 was also used as a part of the cv. Sang annotation. It is not obvious why cv. Sanfensan has more missing genes, but it seems that the annotation process did not make use of publicly available RNA-seq and IsoSeq data used as a part of the cv. Sang annotation – this could be one explanation for the cv. Sang annotation covering a larger number of BUSCOs.

I have touched on some issues with genome annotation processes in this section, and I will discuss some of them here. Evaluating annotation quality is hard, and relies largely on the reliability of various other databases, which in turn are often also based on other annotation pipelines. Errors present in these other databases may propagate, e.g. through incorrect proteins being aligned to the genome and later being verified by presence in that same database. Relying heavily on BUSCO scores is tempting, as these provide a clear way of evaluating

quality, but relying on and repeatedly evaluating primarily based on BUSCO may in the long term lead to what researchers in machine learning might call 'overfitting', i.e., the tools we use to identify proteins become disproportionately good at identifying BUSCOs, degrading BUSCOs usefulness as a proxy for annotation completeness. Meanwhile, weighting multiple measures of annotation quality is inherently hard – many measures may help identify a bad annotation, but comparing decent annotations is tricky.

This difficulty evaluating annotation quality also brings with it the difficulty evaluating pipeline settings and pipelines as a whole. As shown above, there are a large number of potential softwares that may be combined, and infinite possible combinations of software parameters and data sources. The problem of evaluating these is made worse by these pipelines being time-consuming and computationally expensive to run. Occasionally there are also issues with running these for large plant genomes – some tools require modification to run, while some simply refuse to run at all.

A pointed way to put this is that annotations are a tangle of decisions that are considered 'good enough' during the annotation process, but that genome annotations, in spite of this messiness, end up being very useful. My own take-away from this work has been to be skeptical of genome annotations – in oats I routinely look at transcript evidence as well as multiple sequence alignments when evaluating the quality of the gene models I work more closely with. I have learned that any cross-annotation comparisons, in particular between different species, require some thinking, and a wariness regarding whether the comparisons we are attempting make any sense. Bearing this in mind, we can start doing useful science with these annotations. Functional annotations and homology information are useful tools for this, and they are the topic of the next section of this thesis.

2.3 Functional annotation and homology information

Annotation of protein coding genes present within a genome is a great start, and may be very helpful in narrowing the search space if you have known genes from other species that you are interested in identifying. More useful is if we can also identify likely functions of our annotated proteins, along with homologous genes from other species. This section will cover the tools Automated Assignment of Human Readable Descriptions (AHRD, Hallab et al. 2022) as well OrthoFinder (Emms and Kelly 2019) that were used as a part of **paper I**. I was not involved in running these, but the summary results of the orthogroups shown here are generated by me.

2.3.1 AHRD

Functional annotation of annotated protein coding genes was performed using AHRD (Hallab et al. 2022). AHRD creates human-readable descriptions by searching databases of annotated proteins and parsing their annotations. In short, this involves BLASTing different databases containing annotated proteins, parsing fasta headers containing protein descriptions, and weighing different parts of the description in order to create a representative description of the protein to be annotated. When Gene Ontology (GO, The Gene Ontology Consortium 2019) annotation databases are available, these may also be used to add

a functional annotation. The functional annotation released as a part of **paper I** was generated using AHRD with the UniProt databases Swiss-Prot and TrEMBL (The UniProt Consortium 2023) as well as TAIR (Berardini et al. 2015) and included GO annotations.

2.3.2 OrthoFinder

OrthoFinder (Emms and Kelly 2015, 2019) is a tool used to, among other things, infer orthogroups and orthologs among sets of protein coding genes. Orthogroups consist of sets of genes that all originate from the same gene in the last common ancestor of the species being analyzed. Orthogroups may encompass both orthologs and paralogs, and are a useful tool when trying to infer gene function based on homology. We have used orthogroups extensively in our work, both for identifying homologous genes within oat and to identify related genes in other species.

OrthoFinder takes one fasta file per species as input, containing protein or DNA sequences of representative transcripts for each gene. To construct orthogroups, OrthoFinder does an all-against-all BLAST search of the inputs, normalizes the BLAST bit scores based on sequence length and phylogenetic distance, groups genes based on the reciprocal best normalized hits (RBNHs), and uses these gene groups as basis for clustering with the Markov Cluster Algorithm (MCL, Van Dongen 2008).

The length normalization is done to avoid biases with longer sequences achieving higher bit scores and lower e-values than shorter sequences. Briefly, this is done by binning BLAST hits by the sequence length, and picking the top 5% of bit scores among each bin. These are in turn used to model the relationship between sequence length and bit scores, with this model in turn used to normalize all bit scores. This de-couples sequence length from the bit score, allowing the normalized bit score to be used as a measure of sequence similarity. This normalization is done independently for every combination of species, which also provides a normalization for phylogenetic distance. These normalized bit scores are used to identify the RBNHs among the species. To group genes into orthogroups, the normalized bit scores of all the RBNHs of a gene is identified, and all genes with a normalized bit score exceeding the lowest normalized bit score among the RBNHs are used as orthogroup candidates, and passed on to MCL for generation of the final orthogroups.

The orthogroups presented in **paper I** are based on protein sequences from *A. sativa*, *A. insularis*, *A. longiglumis*, *A. atlantica*, *A. eriantha*, *Brachypodium distachyon*, wheat, barley, rye, maize and rice. A total of 74,304 orthogroups were generated, with a median of 3 and mean of 11 proteins per orthogroup. The distribution is skewed with the top 5,533 largest orthogroups containing half of all included proteins, with 54,390 orthogroups containing less than 11 proteins, i.e. less than one protein per species on average. Figure 2.3 provides an overview of the distribution of proteins and species in the orthogroups.

The resulting orthogroups have been very useful to us, both in finding oat proteins homologous to proteins from other species and vice versa, but also in identifying oat proteins that are similar to each other. It is important to remember that these groups are not perfect. As outlined above, the algorithm used for orthogroup generation is based on a set of clever heuristics, but this does not represent a ground truth, and tuning e.g. MCL parameters will result in a different set of orthogroups being generated. In our work with oat we have reg-

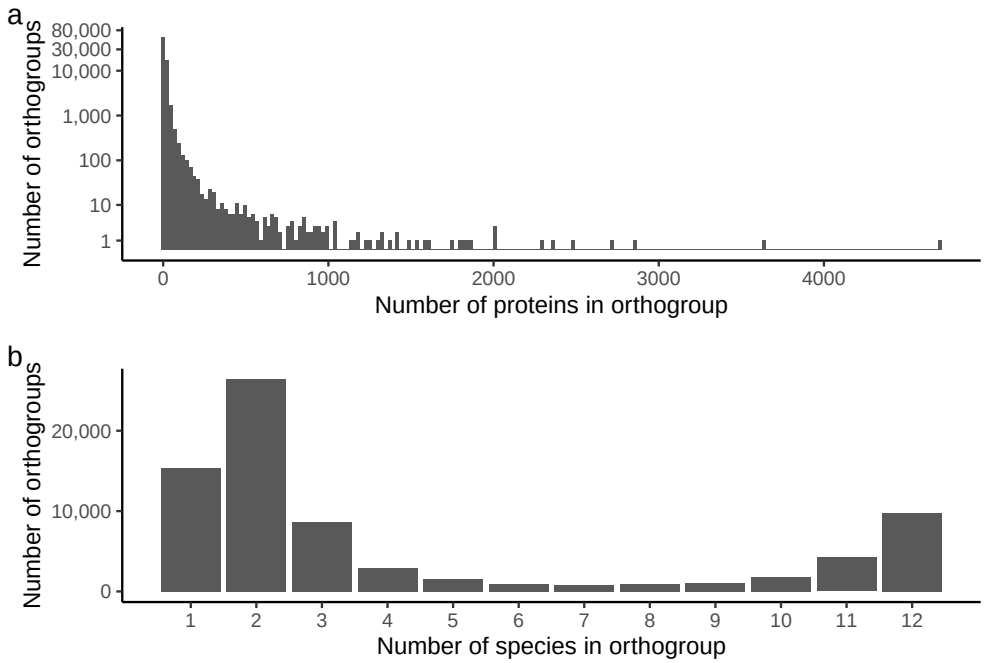

Figure 2.3: Protein and species numbers in the orthogroups generated in **paper I**. a) Distribution of the number of proteins per orthogroup. b) Distribution of the number of species per orthogroup.

ularly encountered orthogroups where relevant homologous proteins have been placed in separate orthogroups – it is important to not blindly trust that the orthogroup containing some interesting protein covers all other interesting homologous proteins.

2.4 Chapter summary

Annotation of large plant genomes is a non-trivial task with a large number of potential moving parts. Outcomes depend on the available data, both newly generated experimental data such as IsoSeq and RNA-seq, but also data present in various other databases; as well as a large number of programs used to make sense of this data, along with the parameters used to run these tools. BUSCO and other annotation statistics are useful for evaluating the quality of annotations, but it is inherently hard to evaluate full genome annotations. In spite of all of this, these resources are very useful. For cv. Sang, the annotation, along with the predicted gene functions, orthogroups, lists of genes syntenic to genes in the OT3098 annotation, as well as expression data that were published as a part of **paper I** may be found at doi:10.5447/ipk/2022/2.

In the next chapter I will discuss how we may use these annotations in spite of their limitations, with the help of the pipeline I developed to more completely capture homologous proteins.

Chapter 3

Identifying homologous proteins

The previous chapter outlined how we generated a genome annotation of protein-coding genes, including functional annotation and orthogroups. This section will consider how we use these to identify relevant homologs of interesting proteins. We are interested in doing this because a lot of research has been done in species other than oats, and for homologous proteins, function is often conserved across species: being able to piggyback on this previous research is of course valuable. Proteins identified in other species – plants and other cereals in particular – may provide a good starting point for research in oats. This chapter will go through the pipeline I developed for identifying homologous proteins, and look at how I did this in practice for the proteins identified in **paper I** and **paper II**.

There are a number of different challenges in identifying proteins of interest. One is the issue of polyploidization: in hexaploid oat we may expect three homoeologs being present in the haploid assembly, but beyond this we are also dealing with the introduction of possible paralogs and pseudogenization, along with different expression levels among the homoeologs. As pointed out earlier, OrthoFinder does some neat adjustments of BLAST results to ensure that a good set of homologous proteins are identified, but the generated orthogroups are parameter-dependent, and do not constitute a ground truth. Beyond this we are also dealing with assembly and annotation errors, including unresolved or misassembled regions of the genome; genes located in such regions; and genes that are incorrectly annotated or entirely missing. The pipeline described in the next section is an attempt at a workflow for capturing relevant homologous proteins.

3.1 The homology pipeline

Since a number of our projects involved identifying homologs in oats of proteins that had been previously investigated in other species, I set up a pipeline for this, to help automate some of the evaluation steps I consider mandatory for this work – an overview of this is show in figure 3.1. The pipeline starts with DIAMONDing[1] (Buchfink, Reuter, and

[1]Like BLASTing, but for protein sequences and faster.

Drost 2021) the known protein sequence against the proteins present in the orthogroups. This is done separately for each species, using one database for each of the species used to construct the orthogroups. The orthogroup of the top match of each species is selected, and the proteins of the identified orthogroups are used to construct a multiple sequence alignment (MSA) using muscle (Edgar 2004) and a phylogenetic tree using fasttree (Price, Dehal, and Arkin 2010). Often the identified proteins will belong to the same orthogroup across all species, but occasionally they will be spread out across separate orthogroups, and identifying these is useful when attempting to get a complete picture of the homologs of a protein.

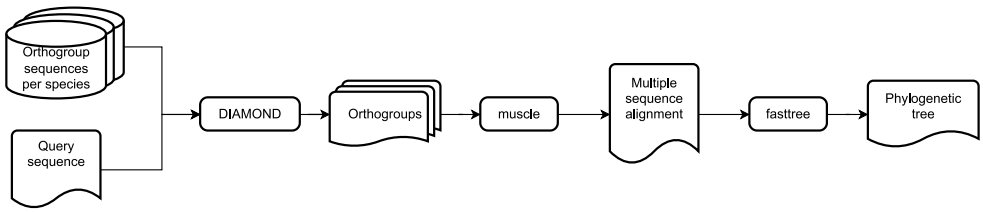

Figure 3.1: Overview of the pipeline used to identify homologous proteins.

3.2 What does this look like in practice?

AsCER-Q (AVESA.00010b.r2.UnG1403470.1) – identified as a part of the work done in **paper I** and discussed further in the chapters on mapping-by-sequencing and gene function – is one such case where identifying additional orthogroups was very useful. OrthoFinder assigned AsCER-Q to orthogroup OG0028142, together with two additional proteins from *A. sativa* and one protein each from *A. atlantia* and *A. insularis*. DIAMONDing AsCER-Q against the other orthogroup species shows that the best hits from the other oat species, barley, wheat, rye, *Brachypodium* and rice all belong to orthogroup OG0001044 which includes a total of 71 proteins, including HvCER-Q, known to be involved in epicuticular wax biosynthesis in barley (Schneider et al. 2016). Inspecting MSAs and the phylogenetic tree containing both orthogroups showed that AVESA.00010b.r2.UnG1403470.1 was indeed very similar to HvCER-Q. The mutation in AVESA.00010b.r2.UnG1403470.1 was located in the residue adjacent to a known loss-of-function mutation in HvCER-Q, known to cause the same lack of epicuticular beta-diketone wax in barley as was observed in the oat mutant, making AVESA.00010b.r2.UnG1403470.1 a strong AsCER-Q candidate.

The homology pipeline was also used as a part of the reverse genetics approach in **paper II**. Here, the barley uridine diphosphate-glycosyltransferase (UDP-glycosyltransferase, UGT) HvUGT13248 was already known from the literature (Schweiger et al. 2010), and we were tasked with identifying homologous proteins in oats. Here, the best hits among all the grasses belonged to the same orthogroup, OG0000783, containing a total of 89 proteins, shown in figure 3.2. The two oat proteins most closely related to HvUGT13248 were experimentally validated and named AsUGT1 (AVESA.00010b.r2.6AG1068650.1) and AsUGT2 (AVESA.00010b.r2.6AG1068570.1) respectively. Looking at the tree, AVESA.00010b.r2.4CG1255890.1 is located in the same clade as the functionally

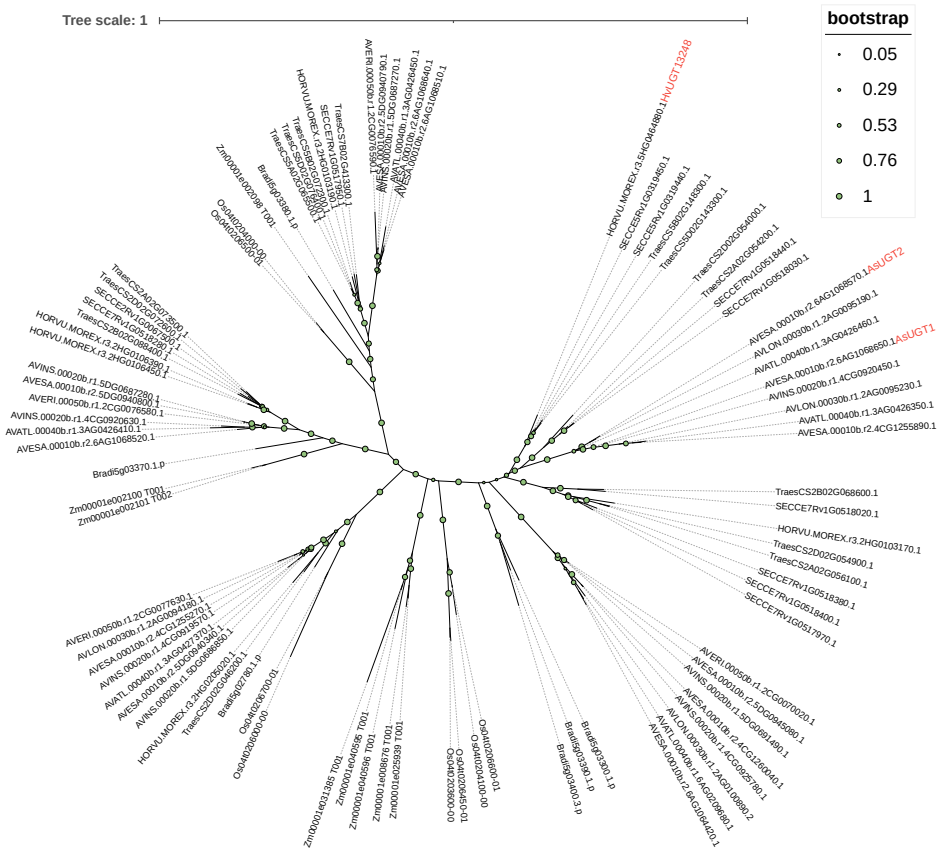

Figure 3.2: Phylogenetic tree of orthogroup OG0000783 with HvUGT13248, AsUGT1 and AsUGT2 highlighted. Bootstrap values are fasttree local support values.

validated AsUGT proteins. Inspecting the MSA it turns out that the protein is heavily truncated, and when inspecting the gene model it turns out that it is located in a poorly assembled region of the cv. Sang genome. Digging further and DIAMONDing HvUGT13248 against the OT3098 annotation (Waters 2022b), the top hit is a protein located on chromosome 4C, AVESA.00001b.r3.4Cg0000153, with another strong hit located nearby (AVESA.00001b.r3.4Cg0000162). Looking closer at MSAs and trees, it seems that these proteins are missing in the cv. Sang annotation, but that the truncated protein has a high similarity to AVESA.00001b.r3.4Cg0000162. The function of these proteins remains to be confirmed, but given the high similarity to the two verified AsUGT proteins, it seems likely that these too may be UGTs. In addition to these there are a number of additional proteins in oats as well as other species present in this group, which may potentially also be functional UGT proteins.

In summary, the orthogroups are a very useful tool in identifying relevant homologous

proteins, but if we want to avoid missing out on relevant homologs we need to look carefully and not settle on only the first orthogroup we come across. The pipeline outlined here is a good starting point for identifying a comprehensive set of homologous proteins.

Chapter 4

Making use of marker studies

Studies identifying genes in other species provide one way of investigating traits in oats. Another source of information is provided by marker based studies, including studies looking at QTLs and genome-wide association studies (GWASs). These marker studies can tell us things about genetic diversity as well as help us identify genes underlying phenotypes of interest. The introduction has mentioned the 6K chip (Tinker, Chao, et al. 2014) as well as GBS markers developed for oats (Huang et al. 2014). This chapter goes through how the 6K markers have been aligned to the cv. Sang reference genome (**paper I**), how they have been used in establishing that it is indeed the cv. Sang that has been sequenced, how they may be visualized (**paper IV**), and how they have been used to map genes and identify breeding barriers in oats. I have been responsible for the marker mapping discussed here, as well as for identifying the genome assembly as being cv. Sang, and for developing the marker visualization in **paper IV**. My contribution to Tinker, Wight, Bekele, Yan, Jellen, Renhuldt, et al. (2022) was in annotating candidate genes identified in the GWAS – I was not involved in the marker analyses of that paper.

4.1 Aligning markers

There are multiple tools that could be used to align markers to the reference genome. In my experience, the marker sequences may be formatted in a number of different ways, and be delivered in a number of different file formats, ranging from nicely formatted fasta files to xlsx-files with unreliably formatted columns: parsing this tends to be a large part of the work. There could probably be many ways to parse these into alignable formats, but in the pipeline I set up, markers are parsed into fasta files with one entry per allele, with the fasta header indicating the original name of the marker, along with an identifier to indicate which allele the entry corresponds to. These fasta files are easily alignable using standard tools – I use GMAP (Wu and Watanabe 2005) in part as it provides nice gff output that is very easy to do additional filtering on using R or Python, along with being easy to visualize in standard genome browsers.

The marker filtering done to align the 6K markers to cv. Sang in **paper I** starts off with removing any non-perfect alignments from the GMAP output. In cases where a single allele was matching in a single place across the genome, this is used as the marker position in the genome. If multiple matches are present, but on separate chromosomes, this is disambiguated by going for the one that matches the previously published consensus maps, but if there are multiple matches on the same chromosome, or if no match is present on the chromosome expected by the consensus map, or if a match is present among the unanchored scaffolds ('unknown chromosome', chrUn), it is not possible to tell the true location, and the marker is discarded. Once markers are aligned we are able to connect our genome to previous research, enabling a number of new analyses.

4.2 Establishing the reference genome as cv. Sang

One of the challenges of this project turned out to be identifying what starting material the genome assembly was based on. It was initially thought that material came from cv. Belinda, but comparing it to cv. Belinda material from the TILLING population (Chawade et al. 2010), it turned out that these mutants consistently shared a large number of variants relative to the assembly. This led us to want to verify which cultivar we had actually assembled. Preliminary analyses using proprietary marker data pointed to cv. Sang, which was known to have been present in the lab at the time of initiating the sequencing project.

Analyzing markers was one line of evidence in this investigation. Comparing the mapped markers to those released as a part of the CORE collection showed that the genome assembly shared the highest number of markers with cv. Sang. We also obtained a cv. Sang reference accession from the Swedish Board of Agriculture. When resequencing this and performing variant calling on the results, we found no markers with the other allele present in the resequenced sample when compared to the assembled genome.

Another strong argument in favor of the assembly being cv. Sang was comparing variant calling results when resequencing the material used to construct the assembly, cv. Belinda and cv. Sang. The resequenced cv. Belinda sample showed millions of high-quality variants, compared to tens-of-thousands for both cv. Sang and the resequenced material used for the assembly. In the case of cv. Sang and the resequenced assembly material, more variants were shared across the samples than were found to be unique for either of the samples. The tens-of-thousands of intracultivar variants identified through whole-genome resequencing is in line with previous work in maize, where exome sequencing showed thousands of variants within cv. Williams 82 (Haun et al. 2011).

The combination of the high similarity of the resequenced assembly material to cv. Sang; that the 6K marker results show that the assembled genome is most similar to cv. Sang among all of the CORE collection accessions; and the presence of cv. Sang in the lab at the time of the start of the project were enough to convince us of our assembly being cv. Sang.

4.3 GFViz

Another use for these markers is in going back to previous work with GWAS or QTLs, in order to try to pinpoint which genes are underlying significant markers. For these to be useful we need the marker locations – depending on which reference genome you are working with, you may need to align these yourself, e.g. using the workflow I outlined above. In the case of the cv. Sang reference genome the marker locations of the Bekele et al. (2018) consensus map have been published as a part of Tinker, Wight, Bekele, Yan, Jellen, Renhuldt, et al. (2022). Visualizing these together with genes identified from a literature search is one way to help identify potential genes of interest, and is done as a part of **paper III**. **Paper IV** is a tutorial for creating this type of visualization, and names it *GFViz*, short for *Genomic Feature Visualization*.

GFViz combines multiple types of data into one interactive plot, inspired by a visualization in Fogarty et al. (2020). It shows significant QTL markers, along with genes identified as potentially relevant in the literature, along with e.g., expression data or homology information presented for the identified genes. A screenshot of this is shown in figure 4.1. An interactive version of this plot may be found at doi:10.5281/zenodo.7924641 together with the code and data necessary to generate it. This visualization is obviously not sufficient to pinpoint gene function, but it may provide a good starting point for identifying genes suitable for experimental validation, by indicating which genes are located close to QTLs on the chromosome.

There are multiple reasons for the decision to publish GFViz as a tutorial paper as opposed to e.g. an R package. Having made public a Snakemake (Mölder et al. 2021) workflow for variant calling (Tsardakas Renhuldt 2021), it has become apparent to me that the ongoing maintenance of software is a ton of work. A steady stream of package updates ensures that software that is not actively maintained has a very limited life span, and finding time, resources and interest(!) for ongoing software maintenance as a PhD student or academic is not easy. Add to this the inherent limits of software packages: by wrapping things into a neat box, customizations require digging into package internals – beyond the scope of many users, potentially saddling the maintainer (me!) with requests for customizations. A tutorial avoids these issues: the user builds the visualization with public tools. They learn what makes up the visualization, and hopefully this also provides for easy customization. If software breaks – a function becomes deprecated or similar – it will hopefully be obvious to the user where this happens, and how to troubleshoot it.

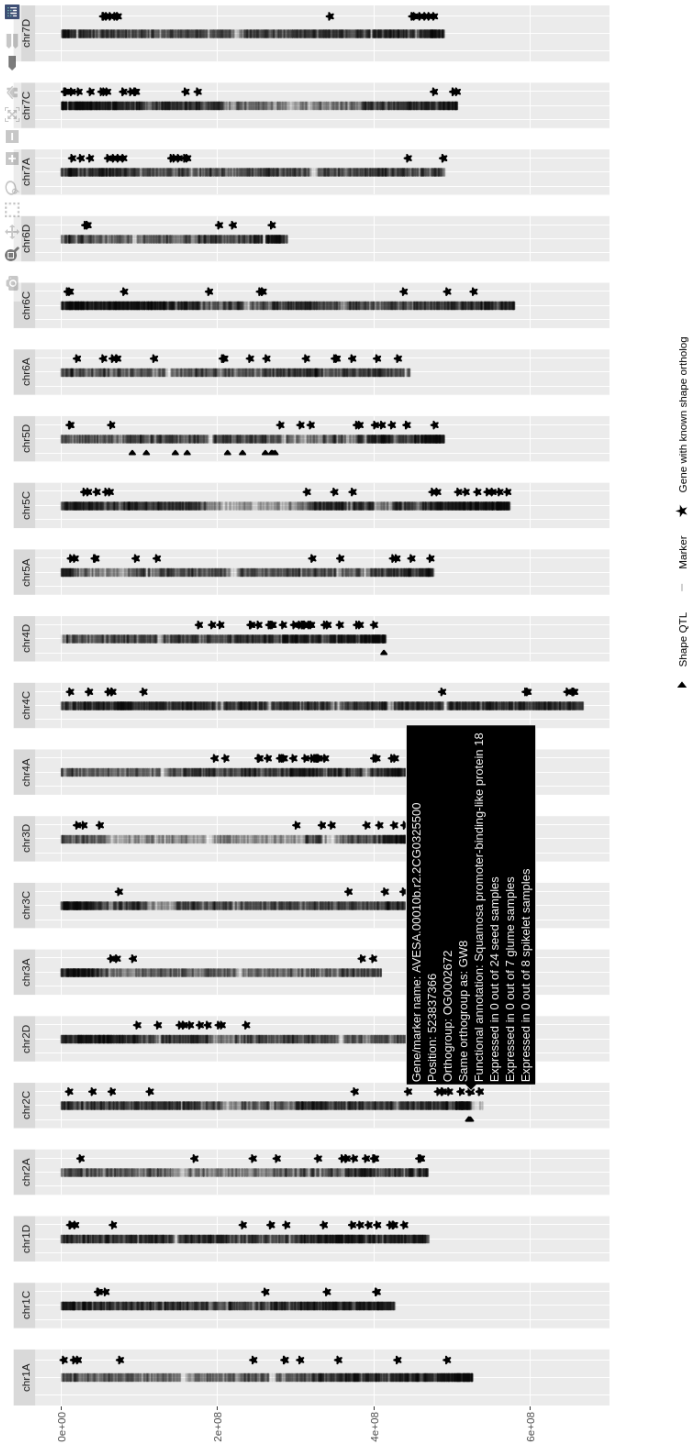

Figure 4.1: Screenshot of the GFViz tool.

4.4 Breeding barriers

In **paper I** we compared *A. sativa* to descendants of its subgenome progenitors, and looked at the distribution of different types of transposable elements and confirmed that there had been large translocations relative to the progenitors – this was previously known, and similar findings had been produced using marker data in e.g. Chaffin et al. (2016). Both Chaffin et al. (2016) and previous research (Jellen and Beard 2000; Jellen, Gill, and Cox 1994; Jellen, Rines, et al. 1997) had also shown that there were significant translocations, not only in *A. sativa* relative to its progenitors, but also across different varieties of *A. sativa*. With reference genomes and robust marker datasets in place, this has now become possible to study further.

Tinker, Wight, Bekele, Yan, Jellen, Renhuldt, et al. (2022) digs into this, among other things. Using GBS markers – de-duplicated and ordered with the help of the cv. Sang reference genome – and five recombinant inbred line populations, they establish QTLs for beta-glucan content, lipid content and heading date, and identify candidate genes for these traits in oats with the help of the cv. Sang annotation. By looking at recombination rates in the populations they also confirm a previously known translocation from chr1C to chr1A: a cross between HiFi (lacking the translocation) and Goslin (containing the translocation) shows a reduced recombination rate (i.e., markers not mixing randomly; pseudo-linkage) between parts of chromosome 1A and 1C. In another population, recombination is completely suppressed along almost all of chr7D, which could potentially be explained by some combination of a large pericentric inversion and other rearrangements preventing pairing of the chromosomes.

Given the seeming abundance of various rearrangements within oat genomes, studies such as these potentially become an important tool for plant breeders. Charting out the rearrangement landscape of oats, and helping to identify incompatible crosses may help improve breeding results going forward.

This chapter has provided an introduction to how existing markers may be aligned to genomes, and showed how markers have been used to identify the genome assembly as being cv. Sang, how we may visualize markers and genes across the genome, and how markers have already been used to identify both genomic translocations and genes in oats. The following chapter discusses mapping-by-sequencing, and how we may use mutations to help identify gene function.

Chapter 5

Mapping-by-sequencing

This chapter provides a background on mapping-by-sequencing, and how this was used to identify genes in **paper I** and **paper III**. The full work done with *AsCer-q* and *AsGSK2.1* is shown in **paper I** and **paper III** respectively, but a short summary is provided here, followed by a more in-depth analysis of the approach we have taken in identifying these genes, with the next chapter providing more information about the genes themselves. This chapter discusses how mapping-by-sequencing results were combined with other types of data to narrow in on the genes we identified, along with some things to keep in mind when designing these experiments. My work on mapping-by-sequencing has been focused mainly on the bioinformatics, ranging from read mapping and variant calling, into gene and variant filtering etc., but also involved reviewing the literature connected to candidate genes.

Mapping-by-sequencing is a useful tool for identifying a single causative gene affecting some trait of interest. In short, the process involves crossing parents that differ in the phenotype, pooling the offspring based on the phenotype, and sequencing the pools. The idea is that recombination ensures that only parts of the genome close to the causative mutation are common to all of the plants in the sequenced pool. The first published tool for this type of analysis, SHOREmap (Schneeberger et al. 2009), crossed an *Arabidopsis* mutant of one strain with a different strain in their example application. Using known markers and SNPs from the strains, they calculated the parental allele frequency across the genome, and used this to identify regions that were common to the pooled plants. Even in the first version of SHOREmap it was possible to do this analysis without markers from both parents being known, with later versions of SHOREmap providing analyses adapted to both outcrossed and backcrossed experiments (Sun and Schneeberger 2015). In the time since SHOREmap was first released, a number of other tools have come out, using a number of different methods to identify regions common to the sequenced pool (see e.g. Abe et al. 2012; Lup et al. 2021; Takagi et al. 2013; Zhang and Panthee 2020).

At the time of starting the experiments that led up to the identification of the genes presented in **paper I** and **paper III**, we were simply not aware of SHOREmap existing, nor of the concept 'mapping-by-sequencing'. We were familiar with bulk segregant analysis

(genotyping pools of siblings sharing a phenotype in order to identify differences in the pool, Michelmore, Paran, and Kesseli 1991) and initially this is what we set out to do. We had previously done some variant calling on different lines from the TILLING population, as well as on different unmutagenized cv. Belinda material, and I had experience working directly with vcf files to visualize variant distributions and variants shared across samples – we did not stop to consider that there might have been tools available to provide these visualizations already, and building them myself felt both fun and straightforward.

The mutants we used to identify the genes were first identified as a part of the cv. Belinda TILLING population (Chawade et al. 2010). As a part of **paper I**, we used a mutant lacking epicuticular wax to demonstrate the usefulness of the reference genome. The mutant has a glossy look due to the lack of wax, and using scanning electron microscopy (SEM) and gas chromatography–mass spectrometry (GCMS) we were able to identify that it was lacking in a beta-diketone wax known to form wax tubules. *AsCer-q* was identified as an homolog of the barley *HvCer-q* by using mapping-by-sequencing to identify the mutation as located on chr1C, followed by filtering of genes based on variant support and gene expression. A candidate gene was identified based on homology to *HvCer-q*, in turn known to be involved in epicuticular wax biosynthesis in barley. The gene was confirmed to be *AsCer-q* by identifying an independent mutant. In **paper III**, *AsGSK2.1* was identified using a mutant with short kernels, shorter stature and altered plant architecture. Again, mapping-by-sequencing was used to narrow in on chr3A, gene candidates were filtered similarly to the process used to identify *AsCer-q*, and a gene candidate was identified based on similarity to GSK proteins with mutants known to cause similar phenotypes in barley and rice. These identified homologs indicated that the mutant would be partially insensitive to brassinosteroids, which we confirmed experimentally.

For both of the genes we have worked with, we have taken mutants from the TILLING population, backcrossed them to cv. Belinda, and pooled homozygous F_2 plants descended from independent F_1 seeds, with zygosity determined by F_3 plant phenotypes. Following this, we sequenced the pools, took the necessary steps to call variants, and filtered these to remove variants shared between the pool and other sequenced samples, including cv. Belinda samples, to account for differences between cv. Belinda and the reference genome cv. Sang. Additional quality filtering was also performed, to ensure acceptable coverage etc. of the called variants. Following this we visualized the mutant allele frequency across the genome. The EMS mutations themselves were used as markers in this approach, and these were detected directly by sequencing and variant calling. In both cases we have studied, we have been able to pinpoint the causative mutation itself, as it has been among the called variants. I have found it remarkable how powerful this is, and how we have been able to resolve these questions with minimal sequencing efforts: a reference genome, a set of variants to account for differences between the unmutagenized parent and the reference genome, and a sequenced pool of back-crossed siblings sharing the phenotype has been sufficient to identify the chromosome of both *AsCer-q* and *AsGSK2.1*.

Doing large crosses and phenotyping many plants is expensive and time-consuming, even in the easy phenotypes we have studied so far. For pool sizes, we used 11 plants to identify *AsCer-q* and 15 plants to identify *AsGSK2.1*, far lower than the 500 plants that were pooled as a part of the original SHOREmap publication (Schneeberger et al. 2009). Our smaller

pools provide enough resolution to identify which chromosome the gene is located on, but does not enable us to reliably narrow in on a chromosome region. To put these numbers into additional context, assuming both parents are homozygous and ignoring recombination, the probability of a chromosome being randomly shared by a pool of n offspring with c chromosomes is $1 - (1 - 2 \cdot 0.25^n)^c$. This gives that a pool size of 11 plants means a 1×10^{-5} probability that a chromosome is common to all plants in the pool by chance: i.e., for 11 siblings in the glossy mutant, the probability that they all end up homozygous for chr1C from the mutant parent by chance is 1×10^{-5}. Five or more plants are required to have less than 5% probability of a chromosome being randomly shared in the pool – in the case of the 15 siblings used to identify *AsGSK2.1* the probability was 4×10^{-8}. To estimate the pool size necessary to narrow in on a specific region some idea of the recombination rate would be useful – Tinker, Wight, Bekele, Yan, Jellen, Renhuldt, et al. (2022) shows that this varies wildly across crosses and chromosome regions, being close to zero for many regions, and sometimes for entire chromosomes. Research in *Arabidopsis* points to hybrids and inbred lines having similar recombination landscapes, but points out that crossovers are suppressed in regions with large rearrangements (Lian et al. 2022) – with oats having multiple large rearrangements, and the recombination rate in **paper I** and Tinker, Wight, Bekele, Yan, Jellen, Renhuldt, et al. (2022) estimated from hybrids, it might make sense to estimate the recombination rate in backcrosses and other inbred lines directly.

With a good idea about which chromosome our gene is located on, but with no idea about the exact region, we ended up combining our different sources of data to narrow in on which mutated gene is causing the observed phenotype. The filtering has relied on the causative mutation being called, but also that it is located in an annotated gene and that the mutation is classified as having high or moderate impact by SnpEff (SNP effect, Cingolani et al. 2012). It is worth noting that this excludes all mutations outside of genes, including mutations potentially affecting promoters or enhancers, making this assumption fairly bold. It also very effectively narrows the search space into a manageable number of genes. In the case of the short kernel mutant, there was something like 156,000 possible homozygous variants on chr3A, with more than ten thousand of these passing sensible variant filtering, but only 123 genes on chr3A were affected by potential high or moderate impact mutations even without accounting for variant quality measures. Filtering candidate genes based on gene expression – i.e., only considering genes expressed in the relevant tissues – and only considering mutations with sufficient read depth etc. provides another way of narrowing in on likely causative genes. In doing further analysis of candidates, we also use the orthogroups to check whether the gene seems to have multiple expressed homoeologs or paralogs present – this has not been used to filter directly, but the results we find here need to make sense. In the case of *AsGSK2.1*, with multiple genes in the orthogroup having similar levels of expression for all candidates, it would be hard to explain why a mutation affecting one of these matters, had it not been for known gain-of-function mutations in the corresponding residue of homologous genes from other species (Pérez-Pérez, Ponce, and Micol 2002).

This strategy is not enough by itself to identify gene function or determine that a mutated gene is in fact responsible for the observed phenotype. There is an obvious risk in filtering, both in excluding the causative gene e.g. due to the variant hitting the promoter, and in the associated risk of narrowing in on the wrong gene. For both *AsCer-q* and *AsGSK2.1* we re-

solved this by some additional experimental validation to help confirm our findings. In the case of *AsCer-q* we had access to sequencing data from multiple lines in the same TILLING population. Using this data we were able to identify an independent line with a mutation to the same gene. Using SEM and GCMS we could verify that this line also lacked the same type of wax as our other mutant (**paper I**). For *AsGSK2.1*, one of six candidates remaining after stringent filtering had homologous genes with mutants producing phenotypes similar to what we observed. With this family of genes known to be involved in brassinosteroid (BR) signaling – described more in depth in the next chapter – the candidate gene offered a testable prediction regarding the plant phenotype: that the mutant would be partially insensitive to BR. We confirmed this insensitivity, strengthening the case for this gene being *AsGSK2.1* and it being causative of the phenotype we observed (**paper III**).

This chapter has covered some basics of mapping-by-sequencing, and how we combined it with other data to identify the genes *AsCer-q* and *AsGSK2.1* in **paper I** and **paper III** respectively. It has discussed some things to keep in mind about this approach to gene mapping, and provided some calculations to consider when selecting pool sizes. The next chapter provides some more information about these genes and their homologs in other species.

Chapter 6

What about the genes?

The earlier parts of this thesis have largely been focused on the bioinformatics involved in annotating the genome, and various ways of identifying gene function. This has also been a very large part of my work – I have not set foot in a lab for the duration of my PhD work. In identifying the function of these genes I have however done a number of dives in the literature surrounding candidate genes. Some of this has made it into **paper I** and **paper III**: in this chapter I say something regarding the what is known about homologs of these genes in other species.

6.1 *AsCer-q*

The *AsCer-q* mutant we studied in **paper I** has a glossy look due to a lack of epicuticular wax. *AsCer-q* was identified as an ortholog to the barley *HvCer-q* gene and it was found to be located within a cluster of genes where some also had orthologs involved in wax biosynthesis in barley. This section discusses some of the previous research done on the genes involved in the synthesis of epicuticular wax in wheat and barley, focusing on the genes present in the clusters identified in those species. It also looks closer at the genes present in the clusters identified in oat as a part of **paper I**, and connects these to what is known from barley and wheat.

A lot of previous research on *HvCer-q* and the *HvCer-cqu* gene cluster has been done in barley, and a review of this and other work connected to these genes is to be found in von Wettstein-Knowles (2017). Genes closely related to *HvCer-cqu* genes are also found in the wheat gene cluster, with genes designated *Diketone Metabolism -Hydrolase* (*TaDMH*), *-polyketide synthase* (PKS, *TaDMP*), and *-cytochrome P450* (CYP, *TaDMC*) corresponding to *HvCer-q*, *HvCer-c*, and *HvCer-u* respectively (Hen-Avivi et al. 2016). The *HvCer-cqu* genes are involved in the biosynthesis of hentriacontane-14,16-dione, which forms wax tubules on the plant cuticle. In short, HvCER-Q is active upstream of HvCER-C which in turn is active upstream of HvCER-U in the pathway. Alkan-2-ol esters – also known to be found in epicuticular wax of both wheat and barley – are also a downstream product of

HvCER-Q, with *HvCer-q* mutants lacking these alkan-2-ol esters, and *HvCer-c* mutants instead seeing an increase of these (von Wettstein-Knowles 2017).

It has been proposed that the TaDMH protein functions as a thioesterase hydrolase/lipase, based in part on expression experiments in *E. coli* in which a C16 3-ketoacid was determined to be a likely product of TaDMH, with an ketoacyl intermediate of fatty acid biosynthesis likely acting as substrate, but also based on its sequence similarity to hydrolases and carboxylesterases (Hen-Avivi et al. 2016). von Wettstein-Knowles (2017), expressing HvCER-Q in *E. coli*, showed that HvCER-Q is able to cleave fatty acids. The exact substrate of HvCER-Q is still up for debate (von Wettstein-Knowles 2017).

HvCER-C is a type III PKS denoted diketone synthase (DKS) and is known to perform at least two elongations of the 3-ketoacid produced by HvCER-Q, forming tri- and tetraketide intermediates. An additional five or six 2C elongations are performed, and during these the beta-oxygens are removed, possibly by ketoacyl-CoA reductase, hydroxyacyl-CoA dehydratase and enoyl-CoA reductase, which are involved in normal fatty acid elongation. The exact enzymes involved in the additional elongations are unknown, but one potential explanation would be HvCER-C working together with these other enzymes responsible for the removal of the beta-oxygens (Schneider et al. 2016; von Wettstein-Knowles 2017).

The HvCER-U protein has been established to be a CYP functioning as an hydroxylase – it produces hydroxy-beta-diketones, and its mutants see a lack of these with a corresponding decrease in the beta-diketones (Schneider et al. 2016; von Wettstein-Knowles 1972).

Both the oat and wheat clusters contain genes belonging to orthogroup OG0001279, with similarity to wax ester synthases (WES, **paper I**, Hen-Avivi et al. 2016) – the barley genes of that orthogroup are located on chromosomes 5H and 7H. In wheat, the *WES* is highly co-expressed with the *TaDM* genes, and Hen-Avivi et al. (2016) speculate that it may be involved in the synthesis of the alkan-2-ol esters.

Several of the oat clusters also contains Myb factors and short-chain dehydrogenase/reductases. The Myb factors found by the clusters on chromosomes 1C and 3A in both cv. Sang and OT3098 as well as on 1C in *A. insularis* belong to orthogroup OG0073676 which only contains proteins from *A. sativa* and *A. insularis*, but DIAMONDing these against the other orthogroup species brings up orthogroup OG0004345 as containing other related proteins. Among the wheat proteins in this orthogroup is TaMYB31 (TraesCS5B02G226100.1) which is known to be an activator of cuticle biosynthesis genes and involved in drought response (Bi et al. 2016). The short-chain dehydrogenase/reductases found to be present in many of the oat clusters do not seem to have homologs located close to the clusters in wheat or barley, nor do they seem to have homologous proteins with a known relation to wax biosynthesis.

It is worth noting that HvCER-U and the TaDMC as well as a number of their most similar oat proteins belong to orthogroup OG0001150. All of these oat proteins correspond to genes located on entirely different chromosomes than the identified clusters in oats. The function – if any – of the oat cluster CYPs in (hydroxy-)beta-diketone biosynthesis remains to be determined, as does which proteins are responsible for synthesizing the hydroxy-beta-diketones known to be present in oats (Dierickx and Buffel 1972; Tulloch and Hoffman 1973).

6.2 *AsGSK2.1*

AsGSK2.1 was identified in **paper III** thanks to a mutation causing the mutant to be partially insensitive to BR, leading to a short stature, altered plant architecture, and short kernels. AsGSK2.1 belongs to orthogroup OG0000448, which it shares with a number of proteins previously known in the literature, with a more comprehensive overview of these provided in **paper III**. Common to these proteins is that they are members of the glycogen synthase kinase 3 (GSK3) family. The family consists of four subgroups, and subgroup membership is determined by sequence homology or phylogeny (Charrier et al. 2002; Jonak and Hirt 2002; Yoo et al. 2006; Youn and Kim 2015).

Several of the proteins belonging to this family are involved in BR signaling which has seen extensive research over the past decades, with multiple recent reviews detailing its involvement in things ranging from plant development, growth and stress response (Nolan et al. 2020) to immunity (Ortiz-Morea et al. 2020) to interactions with nutrients (Han et al. 2023). There have also been multiple recent reviews covering GSK3s in various species, including *Arabidopsis*, rice, wheat, barley, maize, and others (Li, Zhang, and Yu 2021; Song, Wang, et al. 2023; Zolkiewicz and Gruszka 2022). Providing an in-depth review of this is beyond the scope of this thesis, but I will provide a brief introduction to some of what is known about some of the GSK3s in *Arabidopsis*, wheat, rice, and barley.

Among the most well-studied of the proteins in the orthogroup and among the GSK3s is the *Arabidopsis* BRASSINOSTEROID-INSENSITIVE 2 (BIN2, also known as ULTRACURVATA1, UCU1, SHAGGY-LIKE KINASE 21, AtSK21, ASKη, DWARF12), a negative regulator of BR signaling with several known gain-of-function mutations (Choe et al. 2002; Li and Nam 2002; Li, Nam, et al. 2001; Pérez-Pérez, Ponce, and Micol 2002). BIN2 hinders BR signaling by phosphorylating the BRASSINAZOLE RESISTANT (BZR) transcription factors BZR1 and BZR2 (also known as BRI1-EMS-SUPPRESSOR 1, BES1), which in turn are known to directly bind to DNA and regulate the expression of BR responsive genes (He, Gendron, Sun, et al. 2005; He, Gendron, Yang, et al. 2002; Yin, Vafeados, et al. 2005; Yin, Wang, et al. 2002). BR normally induces BIN2 depletion, and the gain-of-function mutation *bin2-1* is thought to stabilize BIN2 by hindering KINK SUPPRESSED IN BZR1-1D 1 (KIB1) binding to BIN2, and the proteasome-mediated degradation that this results in (Peng, Yan, Zhu, et al. 2008; Zhu et al. 2017). Nolan et al. (2020) reviews multiple mechanisms regulating BIN2 activity, with known ones including the KIB1 proteasomal degradation; inactivation through dephosphorylation; sequestration at the plasma membrane stopping BIN2 from phosphorylating BZR1 and BZR2 in the nucleus; inactivation through deacetylation; and activation through oxidation by reactive oxygen species. BIN2 gain-of-function mutants are known to exhibit a BR-deficient phenotype, including dwarfing, changes in plant and leaf architecture, and stomatal clustering, with different gain-of-function mutations displaying different intensities of the phenotypes. Most of these gain-of-function mutations affect the highly conserved 'TREE domain', substituting one of these amino acids for some other amino acid (Choe et al. 2002; Kim, Michniewicz, et al. 2012; Pérez-Pérez, Ponce, and Micol 2002).

In relation to AsGSK2.1, the wheat protein TaSG-D1 (SG for 'semispherical grain', D for wheat subgenome) is interesting due to it having a gain-of-function mutation in its TREE

domain, showing that the effect of substitutions in this domain remains in wheat (Cheng et al. 2020), making a similar effect in oats plausible. Cheng et al. (2020) identified the gene by fine-mapping and showed that the phenotype of *Triticum sphaerococcum* is caused by substitutions in the TaSG-D1 TREE domain both by showing that the TREE domain substitutions were present in all surveyed accessions of *T. sphaerococcum* and by overexpressing the mutant protein in *T. aestivum* and showing that this caused a phenotype similar to that in *T. sphaerococcum*. They also showed that the mutants had an altered response to BR, and that these mutations cause TaSG-D1 to not be degraded by BR when expressed in *Arabidopsis*. Similarly, when the rice OsSK22 (also known as OsGSK2) protein with mutations to the TREE domain was overexpressed in rice it caused a BR loss-of-function phenotype with altered plant architecture and small round seeds – overexpression the wildtype OsSK22 did not suffice to achieve this effect (Tong et al. 2012). Several other rice GSK3s also have known mutants including OsSK23 (also known as OsGSK3) and OsSK41 (also known as OsGSK5; GRAIN LENGTH 3.3; GL3.3; THOUSAND GRAIN WEIGHT 3; TGW3). T-DNA insertion mutants knocking out OsSK23 leads to altered architecture and longer seeds and increased sensitivity to BR (Gao et al. 2019). Loss of function in OsSK41 leads to increased weight and length of grains, with no major changes to plant architecture (Hu et al. 2018; Xia et al. 2018; Ying et al. 2018). In contrast to many other GSK3s, OsSK41 does not seem to be obviously involved in BR signaling, but does seem to be a negative regulator of auxin responsive genes in rice grains: known auxin response genes are more highly expressed in the rice grains in OsSK41 knock-outs (Hu et al. 2018). In barley, Kloc et al. (2020) targeted *HvGSK1.1* for silencing using RNA interference and showed that this led to higher kernel weight in normal growing conditions, and increased biomass in both normal and salt stress conditions. They also showed that silencing *HvGSK1.1* modified expression of *HvGSK1.2*, *HvGSK2.1*, *HvGSK3.1* and *HvGSK4.1* as well, and that the phenotypic effects were correlated to expression of several of these. In a leaf inclination assay, effects of BR were greater in the plants where the *HvGSKs* were silenced.

The Pro303Leu substitution in AsGSK2.1 identified in **paper III** affects the same residue as the weak gain-of-function mutation *ucu1-3* identified in Pérez-Pérez, Ponce, and Micol (2002). This mutation affects a highly conserved residue located 20 residues downstream of the TREE domain, and has seen less research than e.g. the TREE domain mutation *bin2-1*, possibly due to its less extreme phenotype. For both *Arabidopsis* and wheat GSK3s it has been shown that the TREE domain mutations cause a higher abundance of the affected GSK3 protein, and hinders BR-induced degradation of the protein when it is fused to green fluorescent protein (GFP) and expressed in *Arabidopsis* (Cheng et al. 2020; Peng, Yan, Zhu, et al. 2008). As stated above, it is thought that *bin2-1* stabilizes BIN2 through hindering the proteasomal degradation that results from KIB1 binding. A similar mechanism in oats is one possible explanation for the phenotype caused by the mutation of AsGSK2.1.

This chapter has provided some background on the genes *AsCer-q* and *AsGSK2.1*, identified in **paper I** and **paper III** respectively. It has also provided a short summary of some of what is known about their homologous genes in other species. The next chapter concludes this thesis and briefly discusses some potential future work.

Chapter 7

Conclusions and future work

Above, I provide a brief background to this project, outlining some of the previous oat genetic resources that were available at the start of this project, along with a tiny peek of what had been done in other species precviously. This is followed by an overview of the genome annotation process and the tools we have used in that work (**paper I**). That is in turn followed by an example of combining these resources with genes known from other species, and how this was used to identify genes involved in *Fusarium* resistance (**paper II**). The chapter following that discusses using previously published marker datasets (**papers III, IV**), with the next one discussing how we integrate the results from mapping-by-sequencing with the other resources in order to help identify mutated genes underlying observed phenotypes (**papers I, III**). The chapter prior to this conclusion gives more background on the genes we identified, along with information on their homologs in other species.

In some ways this thesis is a showcase of the usefulness of the resources that have been generated as a part of the work by me and others over the past five years. In this thesis I have illustrated how we have integrated the annotated reference genome, including functional annotations, orthogroups and gene expression data with both existing and newly generated data in order to shine a light on how oats work. To my knowledge, this thesis is the first publication identifying a GSK in oats, with potential implications for oat breeding. I have highlighted some ways that these resources may mislead potential users and how I have worked around these limitations in order to do meaningful science. This work has moved oats and its unwieldy genome into a new era, and has laid a foundation for a number of ongoing research projects generating even more comprehensive oat resources.

There are a number of interesting things happening currently, building on top of our work. The PanOat project is working on releasing another set of fully annotated reference genomes, bringing the total number of annotated oat genomes up to 30 or so within the near future. Beyond comparisons of the genomes and various comparative analyses, these results will also include a more comprehensive expression atlas that will undoubtedly help in understanding gene function etc. There is also ongoing work on a global analysis of oat genetic diversity using marker data – as Nordic oats are underrepresented in this panel,

ScanOats is also working together with The Nordic Genetic Resource Center NordGen to analyze the diversity among Nordic oats specifically. We can expect this combination of more annotated genomes and more comprehensive marker datasets to start providing us with a global overview of oat genetic diversity, including identification of the seemingly abundant rearrangements present in the oat genome.

In-depth studies of beta-glucan biosynthesis should also be imminent, as its health claim makes this trait very interesting. As a part of **paper I** we published a phylogenetic analysis of the cellulose synthase and cellulose synthase-like (Csl) genes – *CslF6* belongs to these, and is known to be involved in beta-glucan biosynthesis in wheat and barley (Burton et al. 2011; Nemeth et al. 2010). With the resources that are now made available, identifying oat homologs of e.g. the barley *Hv.NST1/lys5* gene (HORVU.MOREX.r3.6HG0578050.1, orthogroup OG0016407) should be straightforward, opening up for oats with a beta-glucan content of 20%, similar to what has been observed in barley mutants (Munck et al. 2004; Patron et al. 2004).

In conclusion, we can expect that access to a high quality annotated reference genome for oats will benefit researchers and plant breeders alike. Long term, this will probably mean people having access to better oats: more nutritious, more resilient to a extreme weather, and with a smaller climate impact. My hope is that the work I have done here will contribute a tiny piece of the puzzle in helping ensure that people have food to eat in the decades to come.

Acknowledgements

I cannot overstate the collaborative nature of this work. None of this would have happened without help, support and collaboration with a number of excellent scientists. My supervisors in particular have all played important roles in making this happen. Nick Sirijovski brought me into this project, and our collaboration has been very valuable – the Nature paper would not have happened without his hard work. Dag Ahrén has been very supportive, and his insight into bioinformatics has been invaluable throughout my PhD work. Sofia Marmon has been an excellent guide to actually finishing this PhD, and central to bringing the paper on short kernel oats together. Leif Bülow has provided good insight into oftentimes complex university workings. Björn Canbäck's brief stint as my supervisor left a lasting and very positive impact on my attitude to work-life balance.

Johan Bentzer has been my constant companion in this work, and reasoning together regarding these questions has been key to many results in this thesis. Manuel Spannagl, Nadia Kamal and Thomas Lux, the time spent with them in Munich during 2019, and our collaboration since has taught me a lot about genome annotation: without them the Nature paper would not have happened.

The different collaborations I have been involved in with Alfia Khairullina, Alfredo Zambrano, Roya Sardari, Olivier Van Aken, Essam Darwish, Alf Ceplitis, Pernilla Vallenback, Jan Svensson, Lars Arvestad, and many others have all been rewarding. ScanOats has been an excellent setting to do this PhD in, and the informal fikas arranged for and by the junior researchers within ScanOats have been great for keeping in touch and keeping perspective on my work. The Cesky Krumlov Workshop on Genomics – to which I was introduced by Dag, and where I participated both as a student and a teaching assistant – turned out to be both an excellent way to stay up to date regarding the latest best practices for work in bioinformatics, and an opportunity to talk to scientists with different backgrounds about science.

Brian Haas has been very helpful with regards to PASApipeline and EVidenceModeler, Nick Tinker and Wubishet Bekele have provided excellent insights regarding the use of genetic markers, and Martin Mascher and David Stuart have provided good insights into mapping-by-sequencing. My mentor Alex Holmes has been a valuable support in keeping a sensible work-life balance, and in making sure that I did not burn out during these years.

Matilda Berkell is a brilliant scientist and a dear friend: having followed her PhD process has been invaluable guidance in my own work. She also deserves a special thanks for proof-

reading this thesis – any remaining errors are entirely my own. My anchor partner L is more of a biologist than I am ever likely to be, and has kept me grounded during this work. I am incredibly grateful for her, and for all friends, lovers, comrades, queers and neighbors that have made Malmö my home over these past few years. In particular, a thanks to the people fighting for a livable climate and a juster world: none of the work I have done is worth anything unless they succeed, and we owe them our support and solidarity.

Stödet från min familj har varit en förutsättning för det här, för att det överhuvudtaget skulle vara möjligt att föreställa sig att doktorera. Att vara uppvuxen med och omgiven av människor som är bra på att bekräfta varandra är en ynnest, och hela livet hade varit helt knäppt utan det. Ett särskilt tack till mina föräldrar, Jenny och Kostas, och min bror, Rasmus: det här hade inte hänt utan er.

Bibliography

Abe, A. et al. (2012). "Genome Sequencing Reveals Agronomically Important Loci in Rice Using MutMap". In: *Nature Biotechnology* 30.2 (2), pp. 174–178. DOI: 10.1038/nbt.2095.

Avni, R., Lux, T., et al. (2022). "Genome Sequences of Three Aegilops Species of the Section Sitopsis Reveal Phylogenetic Relationships and Provide Resources for Wheat Improvement". In: *The Plant Journal* 110.1, pp. 179–192. DOI: 10.1111/tpj.15664.

Avni, R., Nave, M., et al. (2017). "Wild Emmer Genome Architecture and Diversity Elucidate Wheat Evolution and Domestication". In: *Science* 357.6346, pp. 93–97. DOI: 10.1126/science.aan0032.

Bekele, W. A. et al. (2018). "Haplotype-Based Genotyping-by-Sequencing in Oat Genome Research". In: *Plant Biotechnology Journal* 16.8, pp. 1452–1463. DOI: 10.1111/pbi.12888.

Berardini, T. Z. et al. (2015). "The Arabidopsis Information Resource: Making and Mining the "Gold Standard" Annotated Reference Plant Genome". In: *genesis* 53.8, pp. 474–485. DOI: 10.1002/dvg.22877.

Bi, H. et al. (2016). "Identification and Characterization of Wheat Drought-Responsive MYB Transcription Factors Involved in the Regulation of Cuticle Biosynthesis". In: *Journal of Experimental Botany* 67.18, pp. 5363–5380. DOI: 10.1093/jxb/erw298.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND". In: *Nature Methods* 18.4 (4), pp. 366–368. DOI: 10.1038/s41592-021-01101-x.

Burton, R. A. et al. (2011). "Over-Expression of Specific HvCslF Cellulose Synthase-like Genes in Transgenic Barley Increases the Levels of Cell Wall (1,3;1,4)-β-d-Glucans and Alters Their Fine Structure". In: *Plant Biotechnology Journal* 9.2, pp. 117–135. DOI: 10.1111/j.1467-7652.2010.00532.x.

Camacho, C. et al. (2009). "BLAST+: Architecture and Applications". In: *BMC bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421. pmid: 20003500.

Campbell, M. S. et al. (2014). "Genome Annotation and Curation Using MAKER and MAKER-P". In: *Current Protocols in Bioinformatics* 48.1, pp. 4.11.1–4.11.39. DOI: 10.1002/0471250953.bi0411s48.

Chaffin, A. S. et al. (2016). "A Consensus Map in Cultivated Hexaploid Oat Reveals Conserved Grass Synteny with Substantial Subgenome Rearrangement". In: *The Plant Genome* 9.2, plantgenome2015.10.0102. DOI: 10.3835/plantgenome2015.10.0102.

Charrier, B. et al. (2002). "Expression Profiling of the Whole Arabidopsis Shaggy-Like Kinase Multigene Family by Real-Time Reverse Transcriptase-Polymerase Chain Reaction". In: *Plant Physiology* 130.2, pp. 577–590. DOI: 10.1104/pp.009175.

Chawade, A. et al. (2010). "Development and Characterization of an Oat TILLING-population and Identification of Mutations in Lignin and β-Glucan Biosynthesis Genes". In: *BMC Plant Biology* 10.1, p. 86. DOI: 10.1186/1471-2229-10-86. pmid: 20459868.

Cheng, X. et al. (2020). "A Single Amino Acid Substitution in STKc_GSK3 Kinase Conferring Semispherical Grains and Its Implications for the Origin of Triticum Sphaerococcum". In: *The Plant Cell* 32.4, pp. 923–934. DOI: 10.1105/tpc.19.00580.

Choe, S. et al. (2002). "Arabidopsis Brassinosteroid-Insensitive dwarf12Mutants Are Semidominant and Defective in a Glycogen Synthase Kinase 3β-Like Kinase". In: *Plant Physiology* 130.3, pp. 1506–1515. DOI: 10.1104/pp.010496.

Cingolani, P. et al. (2012). "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3". In: *Fly* 6.2, pp. 80–92. DOI: 10.4161/fly.19695. pmid: 22728672.

DECODAGE – Communauté d'Annotation Des Génomes - TriAnnot (2023). URL: https://www6.inrae.fr/decodage/TriAnnot (visited on 05/07/2023).

Dierickx, P. J. and Buffel, K. (1972). "Hydroxy-Hentriacontanediones from Avena Sativa". In: *Phytochemistry* 11.8, pp. 2654–2655. DOI: 10.1016/S0031-9422(00)88587-8.

Dobin, A. et al. (2013). "STAR: Ultrafast Universal RNA-seq Aligner". In: *Bioinformatics (Oxford, England)* 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635. pmid: 23104886.

Eddy, S. R. (2011). "Accelerated Profile HMM Searches". In: *PLOS Computational Biology* 7.10, e1002195. DOI: 10.1371/journal.pcbi.1002195.

Edgar, R. C. (2004). "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput". In: *Nucleic Acids Research* 32.5, pp. 1792–1797. DOI: 10.1093/nar/gkh340. pmid: 15034147.

EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA) (2010). "Scientific Opinion on the Substantiation of a Health Claim Related to Oat Beta Glucan and Lowering Blood Cholesterol and Reduced Risk of (Coronary) Heart Disease Pursuant to Article 14 of Regulation (EC) No 1924/2006". In: *EFSA Journal* 8.12. DOI: 10.2903/j.efsa.2010.1885.

– (2021). "Scientific Opinion on the Substantiation of Health Claims Related to Beta-Glucans from Oats and Barley and Maintenance of Normal Blood LDL-cholesterol Concentrations (ID 1236, 1299), Increase in Satiety Leading to a Reduction in Energy Intake (ID 851, 852), Reduction of Post-Prandial Glycaemic Responses (ID 821, 824), and "Digestive Function" (ID 850) Pursuant to Article 13(1) of Regulation (EC) No 1924/2006". In: *EFSA Journal* (2011;9(6):2207). DOI: 10.2903/j.efsa.2011.2207.

Emms, D. M. and Kelly, S. (2015). "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy". In: *Genome Biology* 16.1, p. 157. DOI: 10.1186/s13059-015-0721-2.

– (2019). "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics". In: *Genome Biology* 20.1, p. 238. DOI: 10.1186/s13059-019-1832-y.

Esvelt Klos, K. et al. (2016). "Population Genomics Related to Adaptation in Elite Oat Germplasm". In: *The Plant Genome* 9.2, plantgenome2015.10.0103. DOI: 10.3835/plantgenome2015.10.0103.

European Commission, Directorate-General for Agriculture and Rural Development (2023). *Crops Market Observatory: EU Cereals Balance Sheets*. URL: https://circabc.europa.eu/sd/a/2f20cdb4-6113-48d8-9990-b1ac7edd2e2a/Cereals_bs_EUROPA_EU.xlsx (visited on 03/29/2023).

FAO (2022). *FAO.STAT. License: CC BY-NC-SA 3.0 IGO. Extracted From:* URL: https://www.fao.org/faostat/en/#data/QCL (visited on 12/13/2022).

Fogarty, M. C. et al. (2020). "Identification of Mixed Linkage β-Glucan Quantitative Trait Loci and Evaluation of *AsCslF6* Homoeologs in Hexaploid Oat". In: *Crop Science* 60.2, pp. 914–933. DOI: 10.1002/csc2.20015.

Gao, X. et al. (2019). "Rice qGL3/OsPPKL1 Functions with the GSK3/SHAGGY-Like Kinase OsGSK3 to Modulate Brassinosteroid Signaling". In: *The Plant Cell* 31.5, pp. 1077–1093. DOI: 10.1105/tpc.18.00836.

Goff, S. A. et al. (2002). "A Draft Sequence of the Rice Genome (Oryza Sativa L. Ssp. Japonica)". In: *Science* 296.5565, pp. 92–100. JSTOR: 3076391.

Gorash, A. et al. (2017). "Aspects in Oat Breeding: Nutrition Quality, Nakedness and Disease Resistance, Challenges and Perspectives". In: *Annals of Applied Biology* 171.3, pp. 281–302. DOI: 10.1111/aab.12375.

Grabherr, M. G. et al. (2011). "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome". In: *Nature Biotechnology* 29.7, pp. 644–652. DOI: 10.1038/nbt.1883.

Gremme, G. et al. (2005). "Engineering a Software Tool for Gene Structure Prediction in Higher Organisms". In: *Information and Software Technology* 47.15, pp. 965–978. DOI: 10.1016/j.infsof.2005.09.005.

Haas, B. J. (2010). *TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies*. URL: https : / / transposonpsi . sourceforge . net/ (visited on 05/07/2023).

– (2021). *TransDecoder*. URL: https://github.com/TransDecoder/TransDecoder (visited on 03/24/2021).

– (2023). *Gene Structure Annotation and Analysis Using PASA*. GitHub. URL: https://github.com/PASApipeline/PASApipeline/wiki/Home (visited on 05/15/2023).

Haas, B. J., Delcher, A. L., et al. (2003). "Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies". In: *Nucleic Acids Research* 31.19, pp. 5654–5666. DOI: 10.1093/nar/gkg770. pmid: 14500829.

Haas, B. J., Salzberg, S. L., et al. (2008). "Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments". In: *Genome Biology* 9.1, R7. DOI: 10.1186/gb-2008-9-1-r7.

Hallab, A. et al. (2022). *Automated Assignment of Human Readable Descriptions (AHRD).*

Han, C. et al. (2023). "Brassinosteroid Signaling and Molecular Crosstalk with Nutrients in Plants". In: *Journal of Genetics and Genomics*. DOI: 10.1016/j.jgg.2023.03.004.

Haun, W. J. et al. (2011). "The Composition and Origins of Genomic Variation among Individuals of the Soybean Reference Cultivar Williams 82". In: *Plant Physiology* 155.2, pp. 645–655. DOI: 10.1104/pp.110.166736.

He, J.-X., Gendron, J. M., Sun, Y., et al. (2005). "BZR1 Is a Transcriptional Repressor with Dual Roles in Brassinosteroid Homeostasis and Growth Responses". In: *Science* 307.5715, pp. 1634–1638. DOI: 10.1126/science.1107580.

He, J.-X., Gendron, J. M., Yang, Y., et al. (2002). "The GSK3-like Kinase BIN2 Phosphorylates and Destabilizes BZR1, a Positive Regulator of the Brassinosteroid Signaling Pathway in Arabidopsis". In: *Proceedings of the National Academy of Sciences* 99.15, pp. 10185–10190. DOI: 10.1073/pnas.152342599.

Hen-Avivi, S. et al. (2016). "A Metabolic Gene Cluster in the Wheat W1 and the Barley Cer-cqu Loci Determines β-Diketone Biosynthesis and Glaucousness". In: *The Plant Cell* 28.6, pp. 1440–1460. DOI: 10.1105/tpc.16.00197. pmid: 27225753.

Hoff, K. J., Lomsadze, A., et al. (2019). "Whole-Genome Annotation with BRAKER". In: *Gene Prediction: Methods and Protocols*. Ed. by M. Kollmar. Methods in Molecular Biology. New York, NY: Springer, pp. 65–95. DOI: 10.1007/978-1-4939-9173-0_5.

Hoff, K. J. and Stanke, M. (2019). "Predicting Genes in Single Genomes with AUGUSTUS". In: *Current Protocols in Bioinformatics* 65.1, e57. DOI: 10.1002/cpbi.57. pmid: 30466165.

Holt, C. and Yandell, M. (2011). "MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects". In: *BMC Bioinformatics* 12.1 (1), pp. 1–14. DOI: 10.1186/1471-2105-12-491.

Hu, Z. et al. (2018). "A Novel QTL qTGW3 Encodes the GSK3/SHAGGY-Like Kinase OsGSK5/OsSK41 That Interacts with OsARF4 to Negatively Regulate Grain Size and Weight in Rice". In: *Molecular Plant* 11.5, pp. 736–749. DOI: 10.1016/j.molp.2018.03.005.

Huang, Y.-F. et al. (2014). "Using Genotyping-By-Sequencing (GBS) for Genomic Discovery in Cultivated Oat". In: *PLOS ONE* 9.7, e102448. DOI: 10.1371/journal.pone.0102448.

Hufford, M. B. et al. (2021). "De Novo Assembly, Annotation, and Comparative Analysis of 26 Diverse Maize Genomes". In: *Science* 373.6555, pp. 655–662. DOI: 10.1126/science.abg5289.

Hyatt, D. et al. (2010). "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification". In: *BMC Bioinformatics* 11.1, p. 119. DOI: 10.1186/1471-2105-11-119.

International Oat Nomenclature Committee (2021a). *Meeting Notes 30 April 2021*.

– (2021b). *Meeting Notes 7 July 2021*.

Jannink, J.-L. and Gardner, S. W. (2005). "Expanding the Pool of PCR-Based Markers for Oat". In: *Crop Science* 45.6, pp. 2383–2387. DOI: 10.2135/cropsci2005.0285.

Jayakodi, M. et al. (2020). "The Barley Pan-Genome Reveals the Hidden Legacy of Mutation Breeding". In: *Nature* 588.7837 (7837), pp. 284–289. DOI: 10.1038/s41586-020-2947-8.

Jellen, E. N. and Beard, J. (2000). "Geographical Distribution of a Chromosome 7C and 17 Intergenomic Translocation in Cultivated Oat". In: *Crop Science* 40.1, pp. 256–263. DOI: 10.2135/cropsci2000.401256x.

Jellen, E. N., Gill, B. S., and Cox, T. S. (1994). "Genomic in Situ Hybridization Differentiates between A/D- and C-genome Chromatin and Detects Intergenomic Translocations in Polyploid Oat Species (Genus Avena)". In: *Genome* 37.4, pp. 613–618. DOI: 10.1139/g94-087.

Jellen, E. N., Rines, H. W., et al. (1997). "Characterization of 'Sun II' Oat Monosomics through C-banding and Identification of Eight New 'Sun II' Monosomics". In: *Theoretical and Applied Genetics* 95.8, pp. 1190–1195. DOI: 10.1007/s001220050680.

Jonak, C. and Hirt, H. (2002). "Glycogen Synthase Kinase 3/SHAGGY-like Kinases in Plants: An Emerging Family with Novel Functions". In: *Trends in Plant Science* 7.10, pp. 457–461. DOI: 10.1016/S1360-1385(02)02331-2.

Keilwagen, J., Hartung, F., et al. (2018). "Combining RNA-seq Data and Homology-Based Gene Prediction for Plants, Animals and Fungi". In: *BMC Bioinformatics* 19.1, p. 189. DOI: 10.1186/s12859-018-2203-5.

Keilwagen, J., Wenk, M., et al. (2016). "Using Intron Position Conservation for Homology-Based Gene Prediction". In: *Nucleic Acids Research* 44.9, e89. DOI: 10.1093/nar/gkw092.

Kent, W. J. (2002). "BLAT—The BLAST-Like Alignment Tool". In: *Genome Research* 12.4, pp. 656–664. DOI: 10.1101/gr.229202. pmid: 11932250.

Kihara, H. (1919). "Ueber Cytologische Studien Bei Einigen Getreidearten. Mitteilung II. Chromosomenzahlen Und Verwandtschaftsverhältnisse Unter Avena-Arten." In: *Shokubutsugaku Zasshi* 33.388, en94–en97. DOI: 10.15281/jplantres1887.33.388_94.

Kim, D., Paggi, J. M., et al. (2019). "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype". In: *Nature Biotechnology* 37.8 (8), pp. 907–915. DOI: 10.1038/s41587-019-0201-4.

Kim, T.-W., Michniewicz, M., et al. (2012). "Brassinosteroid Regulates Stomatal Development by GSK3-mediated Inhibition of a MAPK Pathway". In: *Nature* 482.7385 (7385), pp. 419–422. DOI: 10.1038/nature10794.

Kloc, Y. et al. (2020). "Silencing of HvGSK1.1—A GSK3/SHAGGY-Like Kinase–Enhances Barley (Hordeum Vulgare L.) Growth in Normal and in Salt Stress Conditions". In: *International Journal of Molecular Sciences* 21.18. DOI: 10.3390/ijms21186616. pmid: 32927724.

Korf, I. (2004). "Gene Finding in Novel Genomes". In: *BMC Bioinformatics* 5.1 (1), pp. 1–9. DOI: 10.1186/1471-2105-5-59.

Leroy, P. et al. (2012). "TriAnnot: A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes". In: *Frontiers in Plant Science* 3.

Levy Karin, E., Mirdita, M., and Söding, J. (2020). "MetaEuk—Sensitive, High-Throughput Gene Discovery, and Annotation for Large-Scale Eukaryotic Metagenomics". In: *Microbiome* 8.1, p. 48. DOI: 10.1186/s40168-020-00808-x.

Li, C., Zhang, B., and Yu, H. (2021). "GSK3s: Nodes of Multilayer Regulation of Plant Development and Stress Responses". In: *Trends in Plant Science* 26.12, pp. 1286–1300. DOI: 10.1016/j.tplants.2021.07.017.

Li, H. (2018). "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* 34.18, pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

Li, J. and Nam, K. H. (2002). "Regulation of Brassinosteroid Signaling by a GSK3/SHAGGY-like Kinase". In: *Science (New York, N.Y.)* 295.5558, pp. 1299–1301. DOI: 10.1126/science.1065769. pmid: 11847343.

Li, J., Nam, K. H., et al. (2001). "BIN2, a New Brassinosteroid-Insensitive Locus in Arabidopsis". In: *Plant Physiology* 127.1, pp. 14–22. DOI: 10.1104/pp.127.1.14.

Lian, Q. et al. (2022). "The Megabase-Scale Crossover Landscape Is Largely Independent of Sequence Divergence". In: *Nature Communications* 13.1 (1), p. 3828. DOI: 10.1038/s41467-022-31509-8.

Lomsadze, A., Burns, P. D., and Borodovsky, M. (2014). "Integration of Mapped RNA-Seq Reads into Automatic Training of Eukaryotic Gene Finding Algorithm". In: *Nucleic Acids Research* 42.15, e119. DOI: 10.1093/nar/gku557.

Lup, S. D. et al. (2021). "Easymap: A User-Friendly Software Package for Rapid Mapping-by-Sequencing of Point Mutations and Large Insertions". In: *Frontiers in Plant Science* 12, p. 601. DOI: 10.3389/fpls.2021.655286. pmid: 34040621.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., et al. (2021). "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes". In: *Molecular Biology and Evolution* 38.10, pp. 4647–4654. DOI: 10.1093/molbev/msab199.

Manni, M., Berkeley, M. R., Seppey, M., and Zdobnov, E. M. (2021). "BUSCO: Assessing Genomic Data Quality and Beyond". In: *Current Protocols* 1.12, e323. DOI: 10.1002/cpz1.323.

Mapleson, D. et al. (2018). "Efficient and Accurate Detection of Splice Junctions from RNA-seq with Portcullis". In: *GigaScience* 7.12, giy131. DOI: 10.1093/gigascience/giy131.

Mascher, M., Gundlach, H., et al. (2017). "A Chromosome Conformation Capture Ordered Sequence of the Barley Genome". In: *Nature* 544.7651 (7651), pp. 427–433. DOI: 10.1038/nature22043.

Mascher, M., Wicker, T., et al. (2021). "Long-Read Sequence Assembly: A Technical Evaluation in Barley". In: *The Plant Cell* 33.6, pp. 1888–1906. DOI: 10.1093/plcell/koab077.

Michelmore, R. W., Paran, I., and Kesseli, R. V. (1991). "Identification of Markers Linked to Disease-Resistance Genes by Bulked Segregant Analysis: A Rapid Method to Detect Markers in Specific Genomic Regions by Using Segregating Populations." In: *Proceedings of the National Academy of Sciences* 88.21, pp. 9828–9832. DOI: 10.1073/pnas.88.21.9828.

Mölder, F. et al. (2021). "Sustainable Data Analysis with Snakemake". In: *F1000Research* 10, p. 33. DOI: 10.12688/f1000research.29032.2.

Monat, C. et al. (2019). "TRITEX: Chromosome-Scale Sequence Assembly of Triticeae Genomes with Open-Source Tools". In: *Genome Biology* 20.1, p. 284. DOI: 10.1186/s13059-019-1899-5.

Munck, L. et al. (2004). "Near Infrared Spectra Indicate Specific Mutant Endosperm Genes and Reveal a New Mechanism for Substituting Starch with (1→3,1→4)-β-Glucan in Barley". In: *Journal of Cereal Science* 40.3, pp. 213–222. DOI: 10.1016/j.jcs.2004.07.006.

National Library of Medicine (US), National Center for Biotechnology Information (2004–2023). *Protein BLAST: Search Protein Databases Using a Protein Query*. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome (visited on 05/15/2023).

Nemeth, C. et al. (2010). "Down-Regulation of the CSLF6 Gene Results in Decreased (1,3;1,4)-β-d-Glucan in Endosperm of Wheat". In: *Plant Physiology* 152.3, pp. 1209–1218. DOI: 10.1104/pp.109.151712. pmid: 20089768.

Nolan, T. M. et al. (2020). "Brassinosteroids: Multidimensional Regulators of Plant Growth, Development, and Stress Responses[OPEN]". In: *The Plant Cell* 32.2, pp. 295–318. DOI: 10.1105/tpc.19.00335.

O'Donoughue, L. S. et al. (1992). "An RFLP-based Linkage Map of Oats Based on a Cross between Two Diploid Taxa (Avena Atlantica × A. Hirtula)". In: *Genome* 35.5, pp. 765–771. DOI: 10.1139/g92-117.

OatBioDB (2023). URL: http://waooat.cn/genome/1 (visited on 05/08/2023).

Ortiz-Morea, F. A. et al. (2020). "It Takes Two to Tango – Molecular Links between Plant Immunity and Brassinosteroid Signalling". In: *Journal of Cell Science* 133.22, jcs246728. DOI: 10.1242/jcs.246728.

PanOat: The Oat Pangenome Project | GrainGenes (2023). URL: https://wheat.pw.usda.gov/GG3/PanOat (visited on 04/26/2023).

Patron, N. J. et al. (2004). "The Lys5 Mutations of Barley Reveal the Nature and Importance of Plastidial ADP-Glc Transporters for Starch Synthesis in Cereal Endosperm". In: *Plant Physiology* 135.4, pp. 2088–2097. DOI: 10.1104/pp.104.045203. pmid: 15299120.

Peng, P., Yan, Z., Zhu, Y., et al. (2008). "Regulation of the Arabidopsis GSK3-like Kinase BRASSINOSTEROID-INSENSITIVE 2 through Proteasome-Mediated Protein Degradation". In: *Molecular Plant* 1.2, pp. 338–346. DOI: 10.1093/mp/ssn001.

Peng, Y., Yan, H., Guo, L., et al. (2022). "Reference Genome Assemblies Reveal the Origin and Evolution of Allohexaploid Oat". In: *Nature Genetics* 54.8 (8), pp. 1248–1258. DOI: 10.1038/s41588-022-01127-7.

Pérez-Pérez, J. M., Ponce, M. R., and Micol, J. L. (2002). "The UCU1 Arabidopsis Gene Encodes a SHAGGY/GSK3-like Kinase Required for Cell Expansion along the Proximodistal Axis". In: *Developmental Biology* 242.2, pp. 161–173. DOI: 10.1006/dbio.2001.0543.

Pertea, M. et al. (2015). "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-seq Reads". In: *Nature Biotechnology* 33.3, pp. 290–295. DOI: 10.1038/nbt.3122.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments". In: *PloS One* 5.3, e9490. DOI: 10.1371/journal.pone.0009490. pmid: 20224823.

Purugganan, M. D. and Jackson, S. A. (2021). "Advancing Crop Genomics from Lab to Field". In: *Nature Genetics* 53.5 (5), pp. 595–601. DOI: 10.1038/s41588-021-00866-3.

Rabanus-Wallace, M. T. et al. (2021). "Chromosome-Scale Genome Assembly Provides Insights into Rye Biology, Evolution and Agronomic Potential". In: *Nature Genetics* 53.4 (4), pp. 564–573. DOI: 10.1038/s41588-021-00807-0.

Rajhathy, T. and Morrison, J. W. (1959). "Chromosome Morphology in the Genus Avena". In: *Canadian Journal of Botany* 37.3, pp. 331–337. DOI: 10.1139/b59-024.

Rajhathy, T. and Thomas, H. (1974). *Cytogenetics of Oats (Avena L.)* In collab. with Internet Archive. Ottawa : Genetics Society of Canada. 98 pp.

Reiser, L. et al. (2022). "Using the Arabidopsis Information Resource (TAIR) to Find Information About Arabidopsis Genes". In: *Current Protocols* 2.10, e574. DOI: 10.1002/cpz1.574.

Schlagenhauf, E. and Wicker, T. (2016). *The TREP Platform: A Curated Database of Transposable Elements*. URL: https://trep-db.uzh.ch/ (visited on 02/24/2023).

Schnable, P. S. et al. (2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics". In: *Science (New York, N.Y.)* 326.5956, pp. 1112–1115. DOI: 10.1126/science.1178534. pmid: 19965430.

Schneeberger, K. et al. (2009). "SHOREmap: Simultaneous Mapping and Mutation Identification by Deep Sequencing". In: *Nature Methods* 6.8 (8), pp. 550–551. DOI: 10.1038/nmeth0809-550. pmid: 19644454.

Schneider, L. M. et al. (2016). "The Cer-cqu Gene Cluster Determines Three Key Players in a β-Diketone Synthase Polyketide Pathway Synthesizing Aliphatics in Epicuticular Waxes". In: *Journal of Experimental Botany* 67.9, pp. 2715–2730. DOI: 10.1093/jxb/erw105. pmid: 26962211.

Schweiger, W. et al. (2010). "Validation of a Candidate Deoxynivalenol-Inactivating UDP-Glucosyltransferase from Barley by Heterologous Expression in Yeast". In: *Molecular Plant-Microbe Interactions®* 23.7, pp. 977–986. DOI: 10.1094/MPMI-23-7-0977.

Simão, F. A. et al. (2015). "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs". In: *Bioinformatics* 31.19, pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351.

Slater, G. S. C. and Birney, E. (2005). "Automated Generation of Heuristics for Biological Sequence Comparison". In: *BMC Bioinformatics* 6.1, p. 31. DOI: 10.1186/1471-2105-6-31.

Smit, A. F. A., Hubley, R., and Green, P. (2013–2023). *RepeatMasker Open-4.0*.

Smith, C. D. et al. (2007). "Improved Repeat Identification and Masking in Dipterans". In: *Gene* 389.1, pp. 1–9. DOI: 10.1016/j.gene.2006.09.011.

Song, S., Tian, D., et al. (2018). "Rice Genomics: Over the Past Two Decades and into the Future". In: *Genomics, Proteomics & Bioinformatics* 16.6, pp. 397–404. DOI: 10.1016/j.gpb.2019.01.001.

Song, Y., Wang, Y., et al. (2023). "Regulatory Network of GSK3-like Kinases and Their Role in Plant Stress Response". In: *Frontiers in Plant Science* 14, p. 1123436. DOI: 10.3389/fpls.2023.1123436.

Stanke, M. et al. (2008). "Using Native and Syntenically Mapped cDNA Alignments to Improve *de Novo* Gene Finding". In: *Bioinformatics* 24.5, pp. 637–644. DOI: 10.1093/bioinformatics/btn013.

Sun, H. and Schneeberger, K. (2015). "SHOREmap v3.0: Fast and Accurate Identification of Causal Mutations from Forward Genetic Screens". In: *Plant Functional Genomics: Methods and Protocols*. Ed. by J. M. Alonso and A. N. Stepanova. Methods in Molecular Biology. New York, NY: Springer, pp. 381–395. DOI: 10.1007/978-1-4939-2444-8_19.

Takagi, H. et al. (2013). "QTL-seq: Rapid Mapping of Quantitative Trait Loci in Rice by Whole Genome Resequencing of DNA from Two Bulked Populations". In: *The Plant Journal: For Cell and Molecular Biology* 74.1, pp. 174–183. DOI: 10.1111/tpj.12105. pmid: 23289725.

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). "Identification of Protein Coding Regions in RNA Transcripts". In: *Nucleic Acids Research* 43.12, e78. DOI: 10.1093/nar/gkv227.

The Arabidopsis Genome Initiative (2000). "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana". In: *Nature* 408.6814 (6814), pp. 796–815. DOI: 10.1038/35048692.

The Gene Ontology Consortium (2019). "The Gene Ontology Resource: 20 Years and Still GOing Strong". In: *Nucleic Acids Research* 47.D1, pp. D330–D338. DOI: 10.1093/nar/gky1055.

The International Wheat Genome Sequencing Consortium (IWGSC) et al. (2018). "Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome". In: *Science* 361.6403. DOI: 10.1126/science.aar7191. pmid: 30115783.

The UniProt Consortium (2023). "UniProt: The Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1, pp. D523–D531. DOI: 10.1093/nar/gkac1052.

Tinker, N. A., Chao, S., et al. (2014). "A SNP Genotyping Array for Hexaploid Oat". In: *The Plant Genome* 7.3, plantgenome2014.03.0010. DOI: 10.3835/plantgenome2014.03.0010.

Tinker, N. A., Kilian, A., et al. (2009). "New DArT Markers for Oat Provide Enhanced Map Coverage and Global Germplasm Characterization". In: *BMC genomics* 10, p. 39. DOI: 10.1186/1471-2164-10-39. pmid: 19159465.

Tinker, N. A., Wight, C. P., Bekele, W. A., Yan, W., Jellen, E. N., Renhuldt, N. T., et al. (2022). "Genome Analysis in Avena Sativa Reveals Hidden Breeding Barriers and Opportunities for Oat Improvement". In: *Communications Biology* 5.1 (1), pp. 1–11. DOI: 10.1038/s42003-022-03256-5.

Tong, H. et al. (2012). "DWARF AND LOW-TILLERING Acts as a Direct Downstream Target of a GSK3/SHAGGY-Like Kinase to Mediate Brassinosteroid Responses in Rice". In: *The Plant Cell* 24.6, pp. 2562–2577. DOI: 10.1105/tpc.112.097394.

Trapnell, C. et al. (2010). "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation". In: *Nature Biotechnology* 28.5, pp. 511–515. DOI: 10.1038/nbt.1621.

Tsardakas Renhuldt, N. (2021). "Nikostr/Dna-Seq-Deepvariant-Glnexus-Variant-Calling: V0.3.1". In: DOI: 10.5281/ZENODO.5045703.

Tulloch, A. P. and Hoffman, L. L. (1973). "Leaf Wax of Oats". In: *Lipids* 8.11, pp. 617–622. DOI: 10.1007/BF02533144.

Turn Non-Recoverable Predictions into Alignments? (2023). *Turn Non-Recoverable Predictions into Alignments? · Issue #23 · EVidenceModeler/EVidenceModeler*. GitHub. URL: https://github.com/EVidenceModeler/EVidenceModeler/issues/23 (visited on 02/02/2023).

Van Dongen, S. (2008). "Graph Clustering Via a Discrete Uncoupling Process". In: *SIAM Journal on Matrix Analysis and Applications* 30.1, pp. 121–141. DOI: 10.1137/040608635.

Venturini, L. et al. (2018). "Leveraging Multiple Transcriptome Assembly Methods for Improved Gene Structure Annotation". In: *GigaScience* 7.8, giy093. DOI: 10.1093/gigascience/giy093.

Von Wettstein-Knowles, P. (1972). "Genetic Control of β-Diketone and Hydroxy-β-Diketone Synthesis in Epicuticular Waxes of Barley". In: *Planta* 106.2, pp. 113–130. DOI: 10.1007/BF00383991. pmid: 24477953.

Von Wettstein-Knowles, P. (2017). "The Polyketide Components of Waxes and the Cercqu Gene Cluster Encoding a Novel Polyketide Synthase, the β-Diketone Synthase, DKS". In: *Plants* 6.4, p. 28. DOI: 10.3390/plants6030028. pmid: 28698520.

Walkowiak, S. et al. (2020). "Multiple Wheat Genomes Reveal Global Variation in Modern Breeding". In: *Nature* 588.7837 (7837), pp. 277–283. DOI: 10.1038/s41586-020-2961-x.

Wang, W. et al. (2018). "Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice". In: *Nature* 557.7703 (7703), pp. 43–49. DOI: 10.1038/s41586-018-0063-9.

Waters, M. (2022a). *PepsiCo OT3098 Hexaploid Oat Version 2 Genome Assembly Release in Collaboration with GrainGenes | GrainGenes*. URL: https://graingenes.org/GG3/content/pepsico-ot3098-hexaploid-oat-version-2-genome-assembly-release-collaboration-graingenes (visited on 11/16/2022).

– (2022b). *PepsiCo Releases Annotated Gene Set and Associated Files for OT3098 v2 Genome in Partnership with GrainGenes | GrainGenes*. URL: https://graingenes.org/GG3/content/pepsico-releases-annotated-gene-set-and-associated-files-ot3098-v2-genome-partnership (visited on 11/16/2022).

Wu, T. D. and Watanabe, C. K. (2005). "GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences". In: *Bioinformatics* 21.9, pp. 1859–1875. DOI: 10.1093/bioinformatics/bti310.

Xia, D. et al. (2018). "GL3.3, a Novel QTL Encoding a GSK3/SHAGGY-like Kinase, Epistatically Interacts with GS3 to Produce Extra-long Grains in Rice". In: *Molecular Plant* 11.5, pp. 754–756. DOI: 10.1016/j.molp.2018.03.006. pmid: 29567448.

Yang, N. and Yan, J. (2021). "New Genomic Approaches for Enhancing Maize Genetic Improvement". In: *Current Opinion in Plant Biology*. Plant Biotechnology 60, p. 101977. DOI: 10.1016/j.pbi.2020.11.002.

Yin, Y., Vafeados, D., et al. (2005). "A New Class of Transcription Factors Mediates Brassinosteroid-Regulated Gene Expression in Arabidopsis". In: *Cell* 120.2, pp. 249–259. DOI: 10.1016/j.cell.2004.11.044.

Yin, Y., Wang, Z.-Y., et al. (2002). "BES1 Accumulates in the Nucleus in Response to Brassinosteroids to Regulate Gene Expression and Promote Stem Elongation". In: *Cell* 109.2, pp. 181–191. DOI: 10.1016/S0092-8674(02)00721-3.

Ying, J.-Z. et al. (2018). "TGW3, a Major QTL That Negatively Modulates Grain Length and Weight in Rice". In: *Molecular Plant* 11.5, pp. 750–753. DOI: 10.1016/j.molp.2018.03.007.

Yoo, M.-J. et al. (2006). "Phylogenetic Diversification of Glycogen Synthase Kinase 3/SHAGGY-like Kinase Genes in Plants". In: *BMC Plant Biology* 6.1, p. 3. DOI: 10.1186/1471-2229-6-3.

Youn, J.-H. and Kim, T.-W. (2015). "Functional Insights of Plant GSK3-like Kinases: Multi-Taskers in Diverse Cellular Signal Transduction Pathways". In: *Molecular Plant*. Cell Signaling 8.4, pp. 552–565. DOI: 10.1016/j.molp.2014.12.006.

Yu, J. et al. (2002). "A Draft Sequence of the Rice Genome (Oryza Sativa L. Ssp. Indica)". In: *Science* 296.5565, pp. 79–92. JSTOR: 3076390.

Zdobnov, E. M. et al. (2021). "OrthoDB in 2020: Evolutionary and Functional Annotations of Orthologs". In: *Nucleic Acids Research* 49.D1, pp. D389–D393. DOI: 10.1093/nar/gkaa1009.

Zhang, J. and Panthee, D. R. (2020). "PyBSASeq: A Simple and Effective Algorithm for Bulked Segregant Analysis with Whole-Genome Sequencing Data". In: *BMC Bioinformatics* 21.1, p. 99. DOI: 10.1186/s12859-020-3435-8.

Zhang, R.-G., Wang, Z.-X., et al. (2019). *TEsorter: Lineage-Level Classification of Transposable Elements Using Conserved Protein Domains*. DOI: 10.1101/800177. (Visited on 05/07/2023). preprint.

Zhu, J.-Y. et al. (2017). "The F-box Protein KIB1 Mediates Brassinosteroid-Induced Inactivation and Degradation of GSK3-like Kinases in Arabidopsis". In: *Molecular Cell* 66.5, 648–657.e4. DOI: 10.1016/j.molcel.2017.05.012.

Zimmer, C. M. et al. (2021). "Genome-Wide Association Mapping for Kernel Shape and Its Association with β-Glucan Content in Oats". In: *Crop Science* 61.6, pp. 3986–3999. DOI: 10.1002/csc2.20605.

Zolkiewicz, K. and Gruszka, D. (2022). "Glycogen Synthase Kinases in Model and Crop Plants – From Negative Regulators of Brassinosteroid Signaling to Multifaceted Hubs of Various Signaling Pathways and Modulators of Plant Reproduction and Yield". In: *Frontiers in Plant Science* 13.

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

Annotating and making use of the *Avena sativa* cv. Sang reference genome

Oats are not only filling and nutritious, they also have an absolutely massive genetic sequence. This thesis locates genes in that sequence, and shows some of the things we can do now that we have access to these genes. No spoilers, but it's some pretty neat stuff.

Division of Pure and Applied Biochemistry
Department of Chemistry
Faculty of Engineering
Lund University

ISBN 978-91-7422-966-0