Arabidopsis thaliana genome assemblies and their use in hybrid transcriptome analyses

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Maximilian Lothar Collenberg
aus Freudenberg

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	26.06.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Detlef Weigel
2. Berichterstatter/-in:	Prof. Dr. Gerd Jürgens

Acknowledgements

First and foremost, I'd like to thank Prof. Detlef Weigel, for his supervision, constant support, and especially for his continuous optimism throughout my PhD. I really learned a lot from you and I enjoyed my time working as a PhD student in the Weigelworld!

I'd also like to thank Prof. Gerd Jürgens for his input during my thesis advisory committee meetings as well as for being my supervisor at the University of Tübingen.

Moreover, I am thankful that Prof. Eric Kemen and Prof. Gerd Weber agreed to be part of my thesis defense committee, together with Prof. Detlef Weigel and Prof. Gerd Jürgens.

Special thanks go to Prof. Gerd Weber for his mentorship and guidance throughout my bachelors and masters as well as for everything that I have learned from him!

Moreover, I'd like to thank Rebecca Schwab for her continuous support, help, advice, chocolate (Milchschnitte) supply, and especially for encouraging me to finish writing my thesis. I also want to say thank you to Ilja Bezrukov, not only for helping me with any seemingly impossible informatics problem but also for being such a nice office mate and for supporting me especially at the end of my PhD. I'm grateful to Gautam Shirsekar for our collaboration, for all the fruitful discussions and for his critical questions. I am also thankful that Theresa Schlegel helped me with all the HMW DNA extractions. Further thanks go to Christa Lanz for everything that I have learned from her about DNA sequencing and for the good times in the genome center. Moreover, I'd like to thank my former officemates Christian Kubica, Travis Wrightsman, Anna-Lena Van de Weyer, Pablo Carbonell, Sebastian Vorbrugg, and Oliver Deusch for all the nice scientific and non-scientific discussions and everything beyond. I also want to say thank you to Cristina Barragan, Clemens Weiss, Julian Regalado, Sergio Latorre, and Miriam Lucke for their friendship, for being awesome colleagues and for the many good times we had in and outside the lab. I'd like to thank Felix Bemm for his guidance at the beginning of my PhD and for everything I have learned from him. Further thanks go to Moi Expósito Alonso, Patricia Lang, Adrian Contreras, Alejandra Duque, Sonja Kersten, Manuela Neumann, Kathrin Fritschi, Efthymia Symeonidi, Thanvi Srikant, Lisa Fuchs, Anja Barth, and Hülya Wicher for their help, all the good discussions and the nice lunch chats we had. Moreover, I am thankful that Katerina Romanova proofread my thesis.

I also want to say thank you to my parents and grandparents for being so supportive and for giving me the opportunity to study carefree. Last but certainly not least I want to thank my wife Franziska for her unconditional support, her endless patience, for being a stay-at-home mom and for taking care of our wonderful kids Charlotte and Carl Friedrich.

Table of contents

Summary

Arabidopsis thaliana exhibits tremendous phenotypic and genotypic variation while having a rather small, mostly homozygous genome. It was the first plant where a sequenced genome became available in 2000. However, this reference genome has its limitations especially when it comes to highly repetitive regions which currently are not well resolved. Moreover, it is widely recognized that the full spectrum of genetic diversity of a species cannot be represented by a single reference genome. Besides fundamental research, *A. thaliana* is also used to tackle questions that are relevant to plant breeding. Two important fields here are plant immune (R) genes and heterosis in F_1 hybrids.

Plant immune genes often encode intracellular nucleotide-binding leucine-rich repeat receptors (NLRs) that enable the plant to directly or indirectly sense the presence of a pathogen via its effector proteins. Thus, knowing the NLR gene repertoire of a species is key to breed novel, more pathogen resistant plants. However, NLR genes are highly diverse even within a species, they can exhibit extensive presence-absence or copy number variation, and are often located in repetitive clusters. These three factors make it unlikely that this gene family can be assessed using a single reference genome that in addition has a lower resolution in highly repetitive regions. In the first part of my work I established a robust workflow for processing the latest PacBio HiFi long-read data. This enabled me to generate high quality genome assemblies for 18 differential *A. thaliana* lines with a high resolution in repetitive regions. I found that the genomes differ in size which I can explain with length variation in highly repetitive centromeric regions that are absent in the current gold-standard reference. In these 18 differential lines I annotated genes with a focus on NLRs. I found variation in the NLR gene repertoire of the 18 lines. Moreover, I annotated NLR genes that were not present in the current reference genome of *A. thaliana*.

The crossing of two inbred parents leads to the generation of F_1 hybrids. Heterosis, a phenomenon where multiple explanations were proposed, describes the superiority of such hybrids compared to their parents. In contrast to this, there are inferior hybrid phenotypes such as hybrid necrosis. Biomass heterosis of F_1 hybrids is widely exploited in agriculture. When analyzing transcriptomes of F_1 hybrids one can compare their gene expression levels to the mean of both parents (Mid-parent value; MPV). Identifying genes that deviate from the MPV can give insights into the molecular basis of heterosis. However, F_1 hybrid genomes are composed of two parental inbred genomes. Thus, having only a single reference genome available impedes hybrid transcriptome analysis. In the second part of my work I utilized full genome information of two *A. thaliana* inbred parents in order to analyze corresponding F_1 hybrid transcriptomes. I established a computational workflow that enabled

the identification of genes with significant deviation from the MPV. Moreover, for these genes I demonstrated that the degree of deviation from the MPV correlates with expression divergence between the two parents.

Together the results of this thesis give an insight into intraspecies genomic and NLR gene variation of *Arabidopsis thaliana* while providing a large dataset for future research projects.

Zusammenfassung

Arabidopsis thaliana verfügt über ein enormes Spektrum an phäno- sowie genotypischer Variation und hat ein kleines und weitgehend homozygotes Genom. Sie war die erste Pflanzenart, für welche ein komplett sequenziertes Referenzgenom verfügbar war. Dieses Referenzgenom hat jedoch eine niedrige Auflösung in hoch repetitiven Bereichen des Genoms. Auch kann ein einzelnes Referenzgenome die genotypische Variation einer Art nicht repräsentieren. Neben der Grundlagenforschung werden die in *A. thaliana* gewonnen Erkenntnisse auch dazu genutzt, Fragestellungen aus dem Bereich der Pflanzenzucht zu untersuchen. Zwei besonders relevante Bereiche sind die Variation von Resistenzgenen, welche für die Abwehr von Pathogenen wichtig sind, sowie Aspekte der Heterosis in F_1 Hybriden.

Häufig kodieren Resistenzgene nukleotid-bindende Rezeptoren mit Leucin-reichen repetitiven Sequenzen (nucleotide-binding leucine-rich repeat receptor; NLR). Diese NLRs ermöglichen es der Pflanze ‚Pathogene wahrzunehmen. Das Repertoire an NLRs einer Spezies zu kennen ist von großer Bedeutung, um resistente Pflanzen gezielt zu züchten. NLR-Gene sind jedoch variabel bezüglich ihrer Sequenz. Zudem sind NLR-Gene oft spezifisch für eine bestimmte Population, weshalb sich das NLR Repertoire verschiedener Individuen stark voneinander unterscheiden kann. Diese Faktoren machen es unwahrscheinlich, dass die Vielfalt an NLR-Genen einer Spezies mit einem einzelnen Referenzgenom zu erfassen ist. Im Rahmen meiner Arbeit habe ich einen Arbeitsablauf etabliert, der es ermöglicht, mit Hilfe moderner Sequenziertechnik lange DNA Sequenzen (long-reads) zu erfassen und daraus 18 qualitativ hochwertige *A. thaliana* Referenzgenome zu erzeugen. Diese neu assemblierten Genome verfügen über eine sehr hohe Auflösung in hoch repetitiven Bereichen. Anhand dieser 18 Genome konnte ich zeigen, dass Unterschiede in der Genomlänge weitestgehend durch unterschiedlich lange hoch-repetitive Centromer Sequenzen erklärbar sind, welche im Referenzgenom, das mit anderer Sequenziertechnologie erstellt worden ist, fehlen. Im weiteren Verlauf habe ich Unterschiede im NLR Repertoire zwischen den 18 Genomen aufgedeckt.

Der Begriff Heterosis beschreibt das Phänomen, dass diese F_1 Hybride ‘besser’ als ihre Eltern sind, was für die Pflanzenzucht relevant ist. Hierbei spielt Genexpression eine große Rolle. Das Genom eines F_1 Hybriden besteht zu je einer Hälfte aus beiden elterlichen Genomen. Genexpressionsanalysen in F_1 Hybriden werden daher erschwert, wenn nur ein elterliches Referenzgenom verfügbar ist. Ich habe hierfür einen Arbeitsablauf etabliert, in dem zwei *A. thaliana* genome miteinander verknüpft werden, um so das Transkriptom von F_1 Hybriden zu analysieren. So konnte ich Gene identifizieren, deren Expression signifikant vom elterlichen Mittelwert abweicht. Außerdem konnte ich für diese Gene zeigen, dass der Grad der Abweichung vom Mittelwert damit korreliert, wie unterschiedlich die beiden Eltern diese Gene jeweils exprimieren.

Die Ergebnisse meiner Arbeit geben Einblick in die genomische Variation von *Arabidopsis thaliana*. Außerdem stellen die hier geschaffenen Datensätze und Arbeitsabläufe eine wertvolle Ressource für zukünftige Forschungsprojekte dar.

1. Introduction

Evolutionary processes such as mutation, selection, recombination and genetic drift shape the architecture of genomes. The term ‘genome architecture’ encompasses characteristics such as karyotype, genome size, gene number, transposable element (TE) density or gene order (Gaut et al. 2007). Despite the fact that the gene number remains relatively stable over millions of years of evolution, plant genomes show a greater plasticity compared to animal genomes. Whereas genomes of various mammalian species differ in size by a factor of five (Gregory, T.R. 2005), plant genome sizes can vary by a factor of 2,000 (Pellicer and Leitch 2020). The individual genomes of sexually reproducing species are usually highly collinear, meaning that their orthologous loci are located on the same chromosome and in the same order (H. Tang et al. 2008). Genome collinearity enables the physical exchange of alleles during meiosis, a process that is crucial not only for generating diversity but also for the purging of deleterious alleles (McDonald, Rice, and Desai 2016). However, full collinearity between two genomes is compromised by the existence of large-scale genomic rearrangements, so-called structural variation. This structural variation is particularly interesting to study in plant immune genes, which are known to vary tremendously both between and within species, and for which processes like illegitimate recombination have been described as molecular mechanisms of evolution (Van de Weyer et al. 2019). Our limited knowledge is mostly attributable to the fact that many species only have a single full-length reference genome available, or their reference genomes are not available as chromosome-level assemblies.

This introduction will give an overview of former technical limitations in the detection of large scale structural variation, the annotation of plant immunity genes, the analysis of F_1 hybrid transcriptomes and the most recent technological innovations that were used for this doctoral research project. Subsequently, I will propose how to use these new technologies in order to overcome former limitations.

1.1 *Arabidopsis thaliana* as a model organism

Arabidopsis thaliana (L.), also known as thale cress, is a self-compatible flowering plant in the *Brassicaceae* (Detlef Weigel and Mott 2009). It is an annual weed that is found all across the northern hemisphere, mostly in temperate regions ranging from the Arctic Circle to North Africa (Sharbel, Haubold, and Mitchell-Olds 2000) and exhibiting great genetic as well as phenotypic diversity (Detlef Weigel and Mott 2009). Recently, it has been recognized that the species has a long history in Africa (Durvasula et al. 2017). The progeny of an *A. thaliana* individual that has been collected from the wild is referred to as an accession.

Because *A. thaliana* outcrossing rates in nature can be around 10% (Bomblies et al. 2010), the genomes of *A. thaliana* accessions can initially be quite heterozygous. However, inbreeding for multiple generations results in a mostly homozygous genome (Mauch-Mani and Slusarenko 1993).

Even though it has no economic value, *A. thaliana* has become one of the most popular model species in plant research (Somerville and Koornneef 2002), contributing to our fundamental understanding of genetics, epigenetics, cell biology, molecular plant development, physiology, metabolism and adaptation (Krämer 2015). The ascent of *A. thaliana* as a model organism can be further attributed to the characteristics that make it a great candidate for basic plant research: *A. thaliana* has a diploid genome with five chromosomes and a total length of only 125 to 150 Megabases (Mb), and less than 30,000 protein-coding genes (Somerville and Koornneef 2002). Another advantage of *A. thaliana* as a model species is its relatively short generation time and its large number of offspring, making it possible to complete a full life cycle in six to eight weeks (Krämer 2015). Moreover, flowering plants emerged only around 100 million years ago, making them evolutionarily recent and relatively closely related. Consequently, fundamental processes such as the regulation of flowering time are conserved between *A. thaliana* and its economically more important relatives such as rice, rapeseed and maize (Kobayashi and Weigel 2007). *Arabidopsis thaliana* is also well-suited for functional studies since specific genotypes can be ordered from seed stock centers, and knock-out mutant lines are widely available (Koornneef and Meinke 2010). *Arabidopsis thaliana* is adapted to many different environments, and accessions can vary dramatically in key traits such as flowering, germination, as well as biotic or abiotic stress tolerance, and pest resistance (D. Weigel 2012).

Given the above advantages, it is not surprising that *A. thaliana* became the first plant species to have its genome sequenced in 2000. The initially published genome of the *A. thaliana* accession Columbia-0 (Col-0) was high-quality, with the 120 Mb sequence represented as chromosome arm scale contigs (Arabidopsis Genome Initiative 2000). The published sequence of Col-0 quickly became the gold standard reference genome for the *A. thaliana* species. This reference genome is referred to as TAIR10. However, centromeres, telomeres, and ribosomal DNA repeats had remained unassembled (Arabidopsis Genome Initiative 2000), because these genomic regions are highly repetitive and thus difficult to assemble. These shortcomings will be described in more detail later during this introduction. In addition, as will also be explained later, it became evident that there is no such thing as ‘the' reference genome of an entire species, as variation is commonplace, including presence/ absence variation of a large fraction of genes (Detlef Weigel and Mott 2009; Hirsch et al. 2014; Y.-H. Li et al. 2014; Danilevicz et al. 2020; Torkamaneh, Lemay, and

Belzile 2021).

1.2 Intraspecific genome variation

Genetic variation between the genomes of individuals can occur on different scales encompassing point mutations on a single nucleotide as well as alterations of whole chromosome arms **(Figure 1)**.

Generally one can differentiate between small variations such as single nucleotide polymorphisms or small insertions/ deletions (indels) and large structural variation. A structural variant (SV) is often defined as a large genomic alteration longer than 50 bp, while smaller variants are commonly called indels polymorphisms. This differentiates SVs thm from other types of variations such as single nucleotide polymorphisms (SNPs) (Mahmoud et al. 2019). Structural variants can be further classified into deletions, duplications, translocations, inversions, and insertions (Sudmant et al. 2015; Sedlazeck et al. 2018; Alkan, Coe, and Eichler 2011). Another structural variant type is copy number variation (CNV) (Carvalho and Lupski 2016). Furthermore, one can differentiate between balanced and imbalanced forms of SVs. Balanced forms of SVs, such as inversions and translocations, do not change the overall size of the genome. In contrast to this, deletions, duplications, and copy number variations are imbalanced SV forms, since they do alter the genome size (Kosugi et al. 2019).

Structural variants are particularly likely to impact phenotypic variation in a species (Yuan et al. 2021). For example, CNVs significantly contribute to traits such as grain size and blast resistance in rice (Xu et al. 2006; Shomura et al. 2008; Deng et al. 2017), resistance to nematodes and increased yield in soybean (Cook et al. 2012; Lu et al. 2017), adaptation to cold environments (Knox et al. 2010), enhanced aluminum tolerance in maize (Maron et al. 2013), and herbicide resistance in *Amaranthus palmeri* (Fernández-Escalada et al. 2017). CNVs have further been involved in the domestication of maize (Hufford et al. 2012) as well as of livestock (Lye and Purugganan 2019).

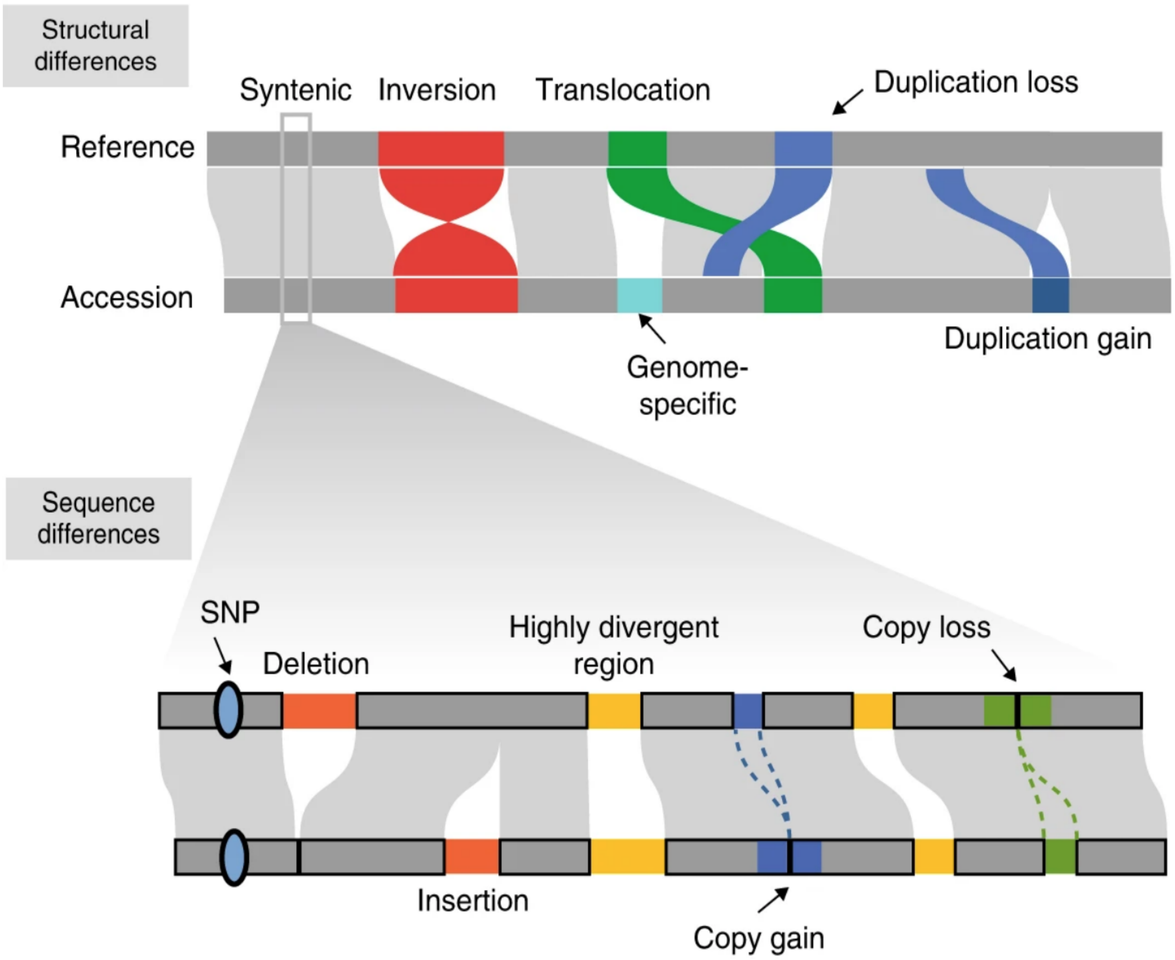

Figure 1: Structural (upper panel) and sequence (lower panel) level variation between two genomes. Inversions are shown in red, translocations in green, and duplications in blue. Moreover, a genome specific insertion is shown in turquoise. Within syntenic regions (shown in gray) there can be variation on the sequence level such as single-nucleotide polymorphisms (SNPs), deletions or insertions (orange), copy gain (blue) and copy loss (green). Highly divergent regions are marked in yellow. Figure source: Jiao and Schneeberger 2020.

1.3 Immunity genes

A class of genes that is particularly prone to presence/ absence variation and other types of SVs are immune genes. Plants rely on a complex immune system to defend themselves against a wide range of enemies (J. L. Dangl and Jones 2001; Ausubel 2005; Chisholm et al. 2006). Plants use two strategies to detect pathogens. The first strategy results in pathogen associated molecular pattern (PAMP) triggered immunity (PTI), while the second strategy results in effector-triggered immunity (ETI) (**Figure 2**) (Jones and Dangl 2006; Chisholm et al. 2006). Both strategies will be introduced in this section.

The PAMP-triggered immunity involves cell surface pattern-recognition receptors (PRRs) that can recognize two different classes of patterns: (i) microbe-associated molecular patterns (MAMPs), and (ii) damage-associated molecular patterns (DAMPs). Once recognized by a PRR on the cell surface, both MAMPS and DAMPs trigger a signal cascade, resulting in cellular changes summarized under the term pattern-triggered immunity (PTI) (Couto and Zipfel 2016). Microbe-associated molecular patterns are slowly evolving structures that are typical for a whole class of microbes (Boller and Felix 2009; Jones and Dangl 2006). A well characterized MAMP is an epitope of bacterial flagellin, which is perceived by the plant receptor FLAGELLIN-SENSING 2 (FLS2). The MAMP recognition in this case depends on the binding of a 22 amino acid stretch of the N-terminus of bacterial flagellin, flg22 (Chinchilla et al. 2006). Orthologs of *FLS2* have been identified in *A. thaliana*, tomato, grapevine, and rice (Chinchilla et al. 2006; Robatzek et al. 2007; Takai et al. 2008; Trdá et al. 2014; Hann and Rathjen 2007). In contrast to MAMPs, the DAMP-triggered immunity does not recognize the pathogen itself. Instead, host-derived molecules that are associated with cell damage as a consequence of pathogen attack are detected. Many different DAMPs have been described, such as cytosolic proteins, nucleotides or cell-wall polysaccharides (Hou et al. 2019). Oligogalacturonide (OG) release from the plant cell wall can be caused by pathogen attack. A prominent example of an *A. thaliana* DAMP receptor is wall associated kinase 1 (WKA1) that triggers immune responses upon detection of OGs (Brutus et al. 2010). Such cell surface localized PRRs can be further grouped into two distinct classes. The receptor kinases (RLK) have an ectodomain for ligand binding, a single pass transmembrane domain, and a cytoplasmic kinase domain. In contrast to that, the receptor-like proteins (RLPs) are lacking the kinase domain while having a short cytoplasmic tail instead (Couto and Zipfel 2016). The most common extracellular domain of plant PRRs are leucine-rich repeats (LRRs) that are involved in PAMP binding and co-receptor association (Shiu and Bleecker 2001). In *A. thaliana* LRR-RLKs and RLP-RLKs constitute a large gene family with 223 and 57 members, respectively (Snoeck et al. 2022).

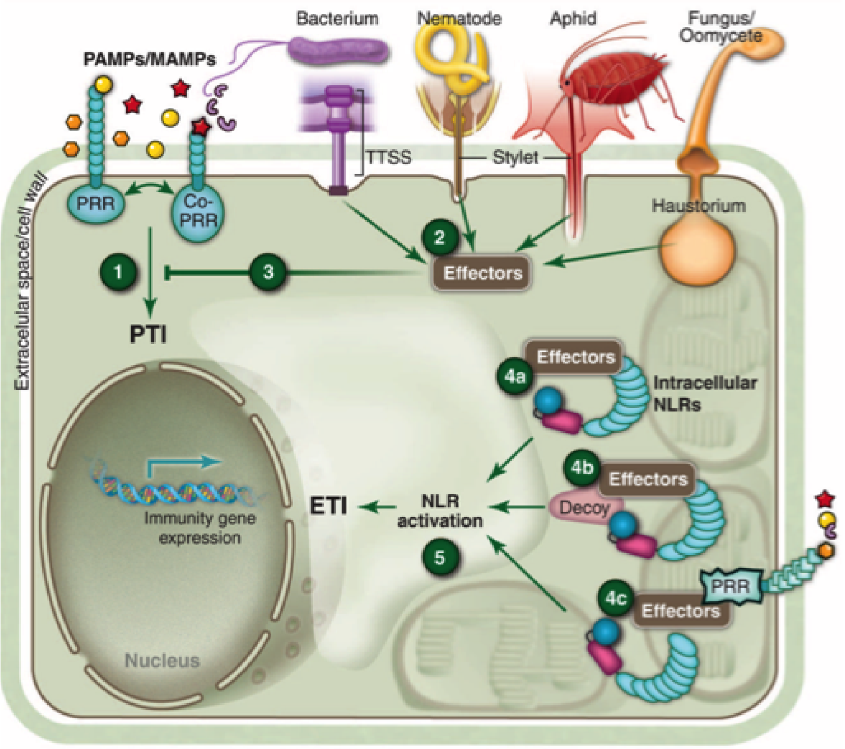

Figure 2: Cartoon showing the plant immune response to pathogen infection. Different pathogens and their PAMPs are color coded. (1) MAMPs and DAMPs are recognized via extracellular PRRs resulting in pattern triggered immunity (PTI). (2) The pathogen releases virulence effectors into the plant cell. (3) Virulence effectors can suppress PTI. (4) Intracellular NLR receptors can perceive effectors in three different ways: (4a) Direct interaction between NLR receptor and ligand, (4b) by mimicking an effector target (decoy strategy), and (4c) by detecting effector-caused alterations of host virulence targets. (5) NLR activation leads to effector triggered immunity (ETI). Figure source: Jeffery L. Dangl, Horvath, and Staskawicz 2013.

However, some pathogens have evolved mechanisms to overcome this first layer of the plant immune system. Such adapted pathogens can evade the PTI by secreting their own effector molecules into the host cell. This is referred to as effector-triggered susceptibility (Jones and Dangl 2006). There are examples of such effectors that allow for successful infection of the plant by either inhibiting or by mimicking eukaryotic cellular functions (Abramovitch, Anderson, and Martin 2006). These effectors are mostly specific to certain pathogens (Boller and Felix 2009). Thus, natural selection has led to a wide range of diversified effector proteins. However, the release of such effectors can be recognized by intracellular host receptors, leading to the so-called effector triggered immunity (ETI). In this case pathogen effectors are recognized by resistance proteins. The majority of these resistance proteins are nucleotide-binding leucine rich repeat (NLR) proteins.

NLRs can detect pathogen effectors by any of the following three strategies: (i) indirectly, by using guardees to detect how the host cell is modified, (ii) directly, through the interaction of the effector and the NLR domains or (iii) directly, via interaction with an integrated domain that mimics an effector target (Kourelis and van der Hoorn 2018). Such detection can trigger resistance responses such as hypersensitive response (HR), leading to rapid localized cell death and inhibiting further pathogen spread (Morel and Dangl 1997). The majority of plant NLRs contain a central nucleotide-binding domain (NB) and either a coiled-coil (CC) domain or a Toll/interleukin-1 receptor (TIR) at the N-terminus (Monteiro and Nishimura 2018). The number of NLR genes greatly varies between different species, ranging from less than a hundred to more than a thousand (Xia et al. 2015; Yue et al. 2012; Baggs et al. 2020). Even within a species the size of the repertoire of NLR genes can vary (Van de Weyer et al. 2019).

Despite their great importance, the study of NLRs remains challenging. One reason is the extraordinarily high diversity in this gene family. The evolutionary forces resulting in such high diversity are explained by Flor's gene-for-gene hypothesis. It states that disease resistance is caused by the interaction of a resistance (R) gene product of the plant with a specific avirulence (Avr) gene product derived from the pathogen (Flor 1971). Thus, pathogens are under constant evolutionary pressure to evolve a diverse set of effectors that do not trigger the interaction with the host R gene product. Therefore the pathogen effector repertoire is highly dynamic. This on the other hand results in high allelic diversity at plant R gene loci (Jeffery L. Dangl and McDowell 2006). Another challenge is presence-absence variation of NLRs that can occur even between closely related individuals (Van de Weyer et al. 2019; Bakker et al. 2008). The high diversity and the extensive presence-absence variation limit the use of a single reference genome for studying NLR genes (Van de Weyer et al. 2019; Christopoulou et al. 2015). Another complication of analyzing NLR genes is the fact that many of them occur in gene clusters with copy number variation (Chae et al. 2014; B. C. Meyers et al. 1998; Leister et al. 1998). However, these gene clusters of tandem NLR repeats are of great interest since they are considered to be hotspots of diversification and generation of novel NLR genes (Michelmore and Meyers 1998; van Wersch and Li 2019; Jiao and Schneeberger 2020). In *A. thaliana* it has been reported that NLR gene clusters can vary in size. Some clusters can span many tens or even hundreds of kilobases (kb) in length, such as the RPP4/RPP5 cluster, others, like the B5 cluster, consist of 11 genes, while the smallest cluster only harbors two NLR genes (Narusaka et al. 2009; Holub 2001; Blake C. Meyers et al. 2003). However, it has been reported that cluster loci with more than two members are less likely to undergo crossing-over events (Rowan et al. 2019). The lack of exchange with the same clusters from different lineages increases the possibility of accumulating mutations. This increases the risk of clusters from different lineages becoming incompatible with each other when combined in a hybrid, resulting in an autoimmune syndrome known as hybrid necrosis (L. Li and Weigel 2021). Despite their great importance for plant breeding and their role in hybrid necrosis, there is only a limited number of studies that assess NLR gene diversity in the context of fully resolved clusters. This is attributable to the short-read DNA sequencing technology that has so far mostly been used for the analysis of NLR genes (Van de Weyer et al. 2019). The limitation of short-reads for assembling repetitive regions such as NLR gene clusters will be explained in the DNA sequencing chapter of this introduction.

1.4 Heterosis in F_1 hybrids

The crossing of two inbred parents leads to the generation of F_1 hybrids. This F_1 generation can show a great variety of non-parental hybrid phenotypes such as hybrid inferiority or hybrid superiority. Both terms describe the deviation of the F_1 hybrids from the parental mean for a given trait. This deviation from the parental mean is also referred to as a non-additive phenotype. Hybrid inferiority can range from mild developmental abnormalities to lethality (Lisa M. Smith, Bomblies, and Weigel 2011; Bomblies et al. 2007; Chae et al. 2014). In *A. thaliana* such F_1 hybrid incompatibilities are observed in 2% of all intraspecific crosses (Bomblies et al. 2007). Many of these incompatibility cases in *A. thaliana* and other species such as rice and tomato involve the interaction of disease resistance (R) gene alleles with other loci in the genome resulting in autoimmunity (Bomblies et al. 2007; Yamamoto et al. 2010; Krüger et al. 2002; Barragan et al. 2021; Chae et al. 2014). This autoimmunity is described by the term hybrid necrosis. Hybrid necrosis can often be explained by the fact that individuals of the same species show a great diversity at R gene loci which makes these loci more likely to cause problems when combined in a hybrid genome (Clark et al. 2007; Jorgensen and Emerson 2008). In contrast to hybrid inferiority, non parental phenotypes can also be superior to their parents. The term heterosis describes the superiority of F_1 hybrid offspring compared to their inbred parents (George H. Shull 1908). It has first been described in the scientific literature by Darwin who noticed, while studying cross-fertilization in plants, that the offspring was often more vigorous (Darwin 1877). Later on George Shull (Shull 1914) introduced the term heterosis for this phenomenon that he observed independently of Edward East in F_1 hybrids of maize. Today heterosis is widely exploited in modern hybrid breeding programs, for example with elite F_1 hybrids in maize having led to a substantial increase in yield over the past century (Duvick 2001; Schnable and Springer 2013). Heterosis is being exploited more and more in other crops important for food and bioenergy production (Longin et al. 2012). Despite the great economic importance of this phenomenon, its molecular basis remains under debate. However, it is widely acknowledged that there is likely not a single explanation and that various genetic mechanisms can be involved in the manifestation of heterosis (Schnable and Springer 2013; Zhai et al. 2013; Birchler et al. 2010). For a given trait, heterosis can either be quantified as the deviation of the progeny from the phenotypic mean of the parental plants (‘mid-parent heterosis’), or, more relevant for agriculture, as the degree of superiority relative to the better parent ('best-parent heterosis'). In either case, heterosis cannot be explained by the addition of the parental trait values (Falconer 1996). At least three non mutually exclusive genetic theories have been proposed to explain such non-additive phenotypes. The dominance theory, first postulated in 1908, argues that dominant alleles from one inbred parent can complement slightly deleterious, recessive alleles from the other parent in the F_1 hybrid (Davenport 1908; Bruce 1910). The overdominance theory, on the other hand, suggests that heterozygosity at an individual locus in the F_1 hybrid gives rise to superiority compared to both homozygous parents. Overdominance was found to be causal for different heterosis phenotypes in *Arabidopsis thaliana* (Rédei 1962), *Solanum lycopersicum* (Vrebalov et al. 2002; Krieger, Lippman, and Zamir 2010), and *Zea mays* (Hollick and Chandler 1998). Lastly, the epistasis model hypothesizes that the interaction between different unlinked parental genes in the hybrid can give rise to hybrid vigor (Williams 1959).

A positive correlation between the genetic divergence and the extent of heterosis has been reported, implying an underlying mechanism that enables hybrids to exploit genetic divergence between the parents to promote heterosis (Chen 2013; Troyer 2006; Birchler et al. 2010). Differences in gene expression between the parents can reflect such genetic divergence (Miller et al. 2015). Gene expression is a phenotype. Thus, the expression of a given gene in the F_1 hybrid can be additive if it follows the mid-parent value or non-additive if it is higher or lower than the parental expression mean (**Figure 3**). Moreover, expression of a given gene in the hybrid can be lower compared to the 'low parent' or higher compared to the 'high parent'. Therefore, the degree of non-additivity can be calculated by comparing the expression level of a gene in the F_1 hybrid vs its average parental expression level (Chen 2013). RNA sequencing (RNA-seq) offers a great opportunity to analyze additive vs non-additive gene expression. Numerous transcriptome studies have shown genome wide expression changes in intraspecific hybrids of rice (He et al. 2010), wheat (Pumphrey et al. 2009), maize (Stupar et al. 2008; Jahnke et al. 2010; Swanson-Wagner et al. 2006; Hu et al. 2016), and *A. thaliana* (Miller et al. 2015; Fujimoto et al. 2012; Meyer et al. 2012). In *A. thaliana* it was possible to link superior hybrid performance in biomass accumulation to transcriptomic mitigation of defense growth tradeoffs (Miller et al. 2015). However, all of the aforementioned RNA-seq studies in F_1 hybrids were carried without having full genome information of both parents. How transcriptome analyses can be performed by the use of telomere-to-telomere genome assemblies for both inbred parents will be introduced in section 4.2.

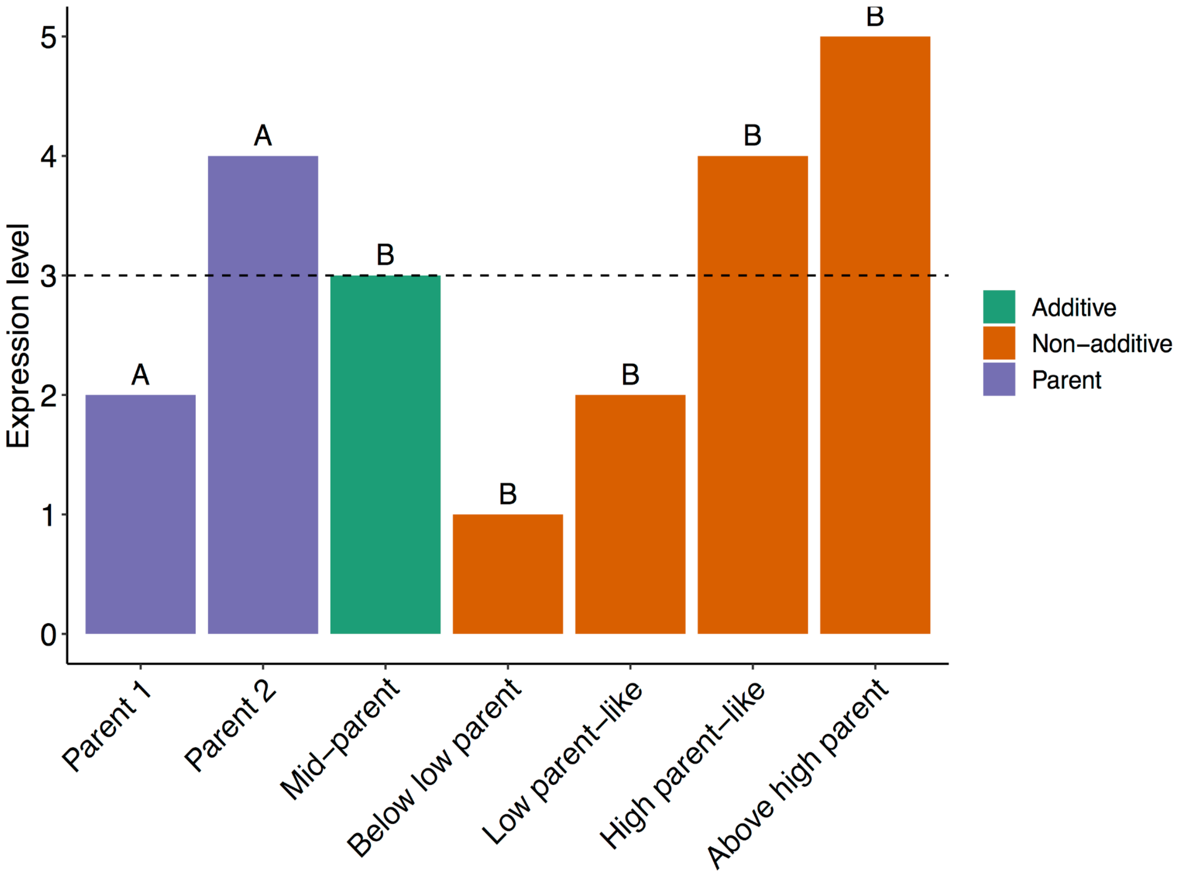

Figure 3: Schematic overview of potential parental (A) and F_1 hybrid (B) expression patterns. A given gene can be additively (orange) or non-additively (green) expressed. Dashed black line represents the mid-parent value (MPV). All potential expression patterns deviating from the MPV are considered to be non-additive.

1.5 DNA sequencing

Genome sequencing is a suite of methods that aim at determining the order of nucleotides in the chromosomes of an individual (Shendure et al. 2017). The technologies used in order to sequence genomes have evolved over time. Therefore, I will give a brief overview of the different approaches as well as their limitations. Moreover, I will describe recent technical advances and how they can be used in order to overcome these (former) limitations.

1.5.1 First and next-generation Sequencing

The three-dimensional structure of the DNA was discovered by Watson and Crick in 1953 (Watson and Crick 1953). However, technologies for actually ‘reading’ a DNA sequence were not available for some time. At the beginning of nucleotide sequencing researchers focused on single-stranded RNA bacteriophages as well as on ribosomal or transfer RNA. Furthermore it was only possible to determine nucleotide composition but not their order (Robert W. Holley et al. 1961). In 1964 Holley and colleagues described a new method for determining the sequence of oligonucleotides. Their approach used an exonuclease that degrades an oligonucleotide in a stepwise manner starting from the 3'-end. The smaller degradation products were separated using chromatography (Robert W. Holley, Madison, and Zamir 1964). One year later, the alanine tRNA of *Saccharomyces cerevisiae* became the first oligonucleotide to be fully sequenced (R. W. Holley et al. 1965).

The first DNA genome, that of PhiX174, has been sequenced in 1977 (Sanger et al. 1977). During the same year, the so-called Sanger method was developed. The method uses electrophoresis and chain-terminating dideoxynucleotides that are randomly incorporated during in vitro DNA replication. Therefore, these and similar technologies are referred to as chain-termination methods. A chain-termination approach uses a single-stranded DNA template, a DNA polymerase, a primer, deoxynucleotide triphosphates (dNTPs), and modified di-deoxynucleotide triphosphates (ddNTPs). These ddNTPs lack a 3'-OH group that causes the DNA polymerase to stop strand elongation after ddNTP incorporation. The target DNA is divided into four separate reactions, each containing all four standard dNTPs (dATP, dCTP, dGTP, or dTTP). Moreover, each reaction contains one of the four chain-terminating ddNTPs (ddATP, ddCTP, ddGTP, or ddTTP). After several cycles of template DNA extension the resulting DNA fragments are denatured and size-separated using gel electrophoresis. Once visualized, the DNA bands can be used to determine the nucleotide order of the DNA template (Sanger, Nicklen, and Coulson 1977). Using radioactively-labeled ddNTPs, this method allowed the sequencing of DNA fragments of different lengths (Sanger, Nicklen, and Coulson 1977). Several improvements such as the replacement of radiolabeling with fluorometric labeling as well as capillary based gel-electrophoresis allowed for an increased automation (W. Ansorge et al. 1986; Luckey et al. 1990). As early as 1979 it was proposed to amplify randomly sampled DNA fragments in bacterial vectors in order to enable the assembly of a genome by finding overlaps between those random sequences (Staden 1979). Because of the random fragmentation of the input DNA, this approach is also referred to as 'shotgun sequencing' (Anderson 1981). However, such approaches were not feasible until fluorescence-based and automated sequencing machines with a higher throughput were developed. These machines enabled the sequencing of approximately 1000 base pairs (bp) per day per reaction (Lloyd M. Smith et al. 1985; L. M. Smith et al. 1986). Commercialized DNA sequencers that allowed sequencing of hundreds of samples in parallel enabled the completion of the first two human reference genomes in 2001 (W. J. Ansorge 2009; Lander et al. 2001; Venter et al. 2001). All of the aforementioned sequencing technologies are referred to as 'first-generation sequencing' since they rely on a chain-termination approach. Application of first-generation sequencing technologies enabled researchers to generate the first *A. thaliana* reference genome in 2000 (Arabidopsis Genome Initiative 2000).

Subsequently, a new approach, not relying on chain-termination and electrophoresis, was developed. Instead, luminescence that gets released upon nucleotide incorporation is monitored during DNA synthesis. First, the single stranded DNA template is immobilized. Now the four dNTPs are sequentially added to and removed from the reaction. Luminescence is released if a nucleotide gets incorporated into the complementary strand. This allows one to determine the sequence of the single stranded template DNA (Nyrén and Lundin 1985; Hyman 1988). This so-called pyrosequencing method was later commercialized by Roche. It marked the first ‘next-generation sequencing’ technology since it is a sequencing by synthesis approach that does not rely on chain-termination technology anymore. Moreover, it allowed parallelizing sequencing reactions and thereby increasing the overall sequence output (Margulies et al. 2005). Later on, new even more massively parallel sequencing technologies such as the Solexa method were developed. In this method the upfront adaptor labeled DNA fragments are ligated to complementary adaptor sequences that are bound to a glass slide (flow cell). Subsequently, a solid phase PCR is performed on the flow cell in order to amplify each DNA fragment in place. Thereby, local clusters of amplified DNA originating from the same molecule are formed that can afterwards be imaged during DNA synthesis. Moreover, the use of highly parallel imaging of the incorporation of fluorescently labeled nucleotides led to an increase in sequencing output (Fedurco et al. 2006; Bentley et al. 2008). This method was commercialized in the form of the Solexa, then Illumina Genome Analyzer that supported sequencing of 1 Gigabase (Gb) of relatively low-error reads with a length of 36 bp (Quail et al. 2012). The number of reactions that could be analyzed on a single flow cell was rapidly increased, and the length of DNA sequencing reads was increased to 150 bp. These innovations led to a rapid decrease in sequencing costs per base (Heather and Chain 2016). Illumina became a leading manufacturer of next-generation sequencing by synthesis instruments. Current Illumina sequencing instruments are capable of producing up to 20 billion reads in less than 45 hours (Pervez et al. 2022). Illumina short-read sequencing quickly replaced alternative methods for population scale analysis of genome diversity, resulting in the ‘1000 Genomes Project’ for humans (1000 Genomes Project Consortium et al. 2015) and the ‘1001 Genomes Project’ for *A. thaliana* (Detlef Weigel and Mott 2009). The efforts in *A. thaliana* were soon followed by efforts in other species, especially maize and rice (Q. Long et al. 2013; Clark et al. 2007; Cao et al. 2011; Ossowski et al. 2008; Hirsch et al. 2016; Schatz et al. 2014).

All of the aforementioned genome sequencing technologies, the first as well as the second-generation, rely on breaking down genomic DNA into fairly small fragments as well as sample amplification prior to the actual sequencing process (Shendure et al. 2017). Moreover, the massively parallel sequencing technologies such as Illumina HiSeq are only capable of producing relatively short-reads with a maximum of 300 bases (Quail et al. 2012). However, these short-reads have a high accuracy. This makes them powerful in the detection of single-nucleotide variants and small indels (1001 Genomes Consortium 2016; 1000 Genomes Project Consortium et al. 2015). On the other hand, certain sequencing applications such as *de novo* genome assembly or the assembly of repetitive regions remain challenging when using short-reads (Ho, Urban, and Mills 2020; Mahmoud et al. 2019; Naish et al. 2021).

1.5.2 The need for longer sequencing reads

In this next section, I want to briefly explain the underlying problem of reconstructing repetitive regions of the genome from short-reads. Ideally a sequencing read would be as long as the chromosome that one wants to assemble. However, as mentioned before, the read length of next-generation sequencing platforms is limited to about 300 bp (Quail et al. 2012). Thus the genome must be assembled from millions of fragments. The assembly from sequencing reads is based on the premise that highly similar sequences originate from the same region of the genome. By finding overlaps, reads are further connected to larger sequence blocks (contigs). Reads from repetitive sequences complicate the assembly, since they may erroneously collapse into a single contig due to their high similarity (Baptista et al. 2018). Therefore, the resulting *de novo* genome assemblies are often fragmented and incomplete when it comes to highly repetitive sequences (Schmid et al. 2018). As mentioned in the chapter ‘1.3 Immunity genes’, it is known that many genes that are particularly important for plants, including NLR genes, occur in repeat clusters. It was reported that half of the NLR genes are organized in clusters (Blake C. Meyers et al. 2003; Van de Weyer et al. 2019). Moreover, large scale genomic rearrangements such as those described in chapter 1.2 are difficult to detect when the sequencing reads are too short to span the structural variant (Sedlazeck et al. 2018). Thus, the usage of short-reads can confound genome assembly as well as read mapping when investigating genome variation.

In order to overcome the aforementioned limitations, new sequencing methods were developed, such as Pacific Bioscience’s (PacBio) single molecule real time (SMRT) or Oxford Nanopore Technologies (ONT) nanopore sequencing, capable of producing long sequencing reads of over 10 kilobases (kb) (Eid et al. 2009; Mikheyev and Tin 2014). This third-generation of sequencing differs from the previously mentioned NGS technologies by no longer relying on the amplification of highly fragmented DNA. Instead, long, single DNA molecules are analyzed in real time (Eid et al. 2009). The disadvantage of analyzing single molecules is much more noise in the measurement, greatly decreasing the accuracy at single nucleotides, from over 99% to under 90%. In the following section, I want to focus on the SMRT sequencing technology developed by PacBio, since that is the sequencing platform that was mostly used for this thesis.

1.5.3 PacBio long-read sequencing

Similar to other sequencing methods, PacBio gathers sequence information while the target molecule is replicated. In the case of PacBio's long-read sequencing, this template molecule is called a SMRTbell (Travers et al. 2010). First, each DNA template is ligated to hairpin adapters (SMRTbell templates) which allow for annealing of the sequencing primer in the next step. Subsequently, a DNA polymerase is bound to the primer. This so-called 'bound complex' is then loaded on a SMRTcell (PacBio Template Preparation and Sequencing Guide; 2014). This SMRTcell consists of small sequencing reactors called zero-mode waveguides (ZMWs). These ZMWs are nanophotonic devices that confine light to a small volume for detection (PacBio Glossary of Terms, v10, April 2019). Such small reaction volumes allow for single-fluorophore detection (Eid et al. 2009). The latest generation of SMRTcells (SMRT Cell 8M) holds up to eight million ZMWs. Each ZMW hold a volume of 10^{-21} liters. Ideally only a single bound complex is captured per ZMW (**Figure 4**). Now, the four different phospholinked dNTPs are added to enable the sequencing-by-synthesis reaction. A fluorescence pulse is generated each time the polymerase retains a nucleotide including its color coded fluorophore in the detection region of the ZMW. This retention time in the detection zone of the ZMW is longer compared to the time that one could expect from diffusion rates. This allows to differentiate between fluorescence caused by the correct incorporation of a nucleotide and background fluorescence from the labeled dNTPs. The light pulse is captured in real time. The fluorophore is released after nucleotide incorporation and diffuses out of the detection zone. The polymerase will now start incorporation of the next nucleotide. The recording of sequential light pulses is then transformed into base calls that produce the continuous long-read (CLR) of the template DNA sequence (Eid et al. 2009). Thus, as with other sequencing-by-synthesis methods, PacBio SMRT sequencing relies on the detection of a fluorescence signal upon nucleotide incorporation. However, in contrast to other methods, it does not use base-linked fluorescent nucleotides, since these are not suited for real-time sequencing. Instead the fluorophore is linked to the terminal phosphate moiety, resulting in the release of the fluorophore via phosphodiester bond formation that is catalyzed by the polymerase upon nucleotide incorporation. Therefore, the synthesized DNA itself does not emit any fluorescence signal (Eid et al. 2009).

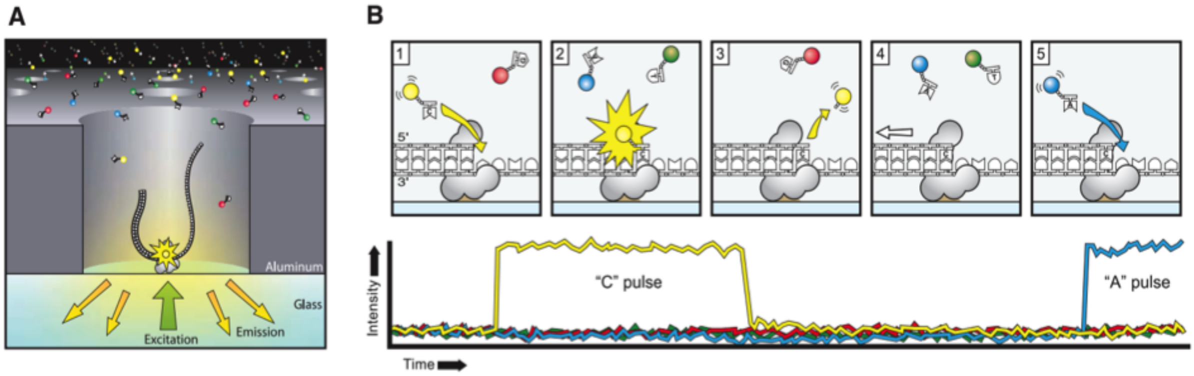

Figure 4: Scheme of single-molecule real-time DNA sequencing using PacBio SMRT sequencing technology. (A) A single DNA molecule, attached to a DNA polymerase, is bound to the bottom of the zero molecule waveguide (ZMW). The ZMW is illuminated using laser light. The small confinement of the ZMW (10^{-21} liter) allows for the detection of individual phospholinked nucleotides while they are incorporated into the DNA strand by the polymerase. (B) The upper panel depicts the incorporation of a phospholinked nucleotide while the lower panel shows the corresponding time trace on the x-axis and fluorescence intensity on the y-axis. First a phospholinked nucleotide binds to the template in the active site of the polymerase. A fluorescence pulse can be detected while the phospholinked nucleotide is retained in the active site of the polymerase. Information about which nucleotide was incorporated is inferred from the color channel. Subsequently, the fluorophore is released from the active site ending the light pulse. The polymerase now shifts to the next position allowing for the next phospholinked nucleotide to bind to the active site. Figure source: Eid et al. 2009.

The great advantage of PacBio SMRT (and Oxford Nanopore) sequencing is the tremendous increase in read length. PacBio CLRs can have a length of up to 100 kb compared to 100 to 300 bases in standard short-read sequencing approaches (Hon et al. 2020). This makes CLRs well suited for *de novo* genome assembly and especially for resolving highly repetitive regions (Chin et al. 2013; Koren et al. 2017). However, long-reads typically have a much higher error rate as compared to short-read sequencing data (Eid et al. 2009). Moreover, it has been shown that continuous long-reads are not accurate enough to assemble centromeres in *A. thaliana* (Rabanal et al. 2022). Several methods for error correction have been developed to offset these limitations. These methods combine either multiple independent long-reads or error-prone long and highly accurate short-reads (Chin et al. 2013; Koren et al. 2017).

1.5.4 PacBio circular consensus sequencing

In 2018 PacBio released the new Sequel II platform that allows for so-called High Fidelity (HiFi) sequencing. This technology improves on the final error rate by reducing effective read length through a process called Circular Consensus Sequencing (CCS). For CCS, a circular DNA template is generated and then sequenced multiple times. The length of the DNA molecule must be a fraction of the maximum read length of the polymerase, so that the same DNA template can be read multiple times. Every instance of the DNA polymerase reading through the entire circular molecule is referred to as a full pass. This generates the so-called subreads (**Figure 5**). Noisy and error-prone, the subreads can then be combined into a single and more accurate CCS read (Travers et al. 2010; Wenger et al. 2019). Since sequencing errors are randomly distributed, multiple passes and subsequent overlaying of the subreads can reduce the total number of errors. Thus, the accuracy of the consensus read increases with an increasing number of full passes (Wenger et al. 2019). The new HiFi sequencing technology now offers the possibility to sequence medium size DNA inserts with a length of approximately 20 kilobases and an accuracy of up to 99.9 % (Rabanal et al. 2022; Wenger et al. 2019). A recent comparison of CLR vs HiFi sequencing has shown that medium sized but highly accurate reads are superior to longer but noisier reads when it comes to genome assembly (Rabanal et al. 2022). Moreover, Hifi sequencing has proven to perform well on highly-repetitive regions such as centromeres, in some cases allowing for telomere-to-telomere *de novo* genome assemblies (Rabanal et al. 2022; Hon et al. 2020; Naish et al. 2021; Wenger et al. 2019). Thus, PacBio HiFi sequencing offers a great possibility to overcome the limitations of short-read sequencing for the analyses of repetitive regions such as NLR gene clusters without compromising on read accuracy as it is the case for continuous long-reads.

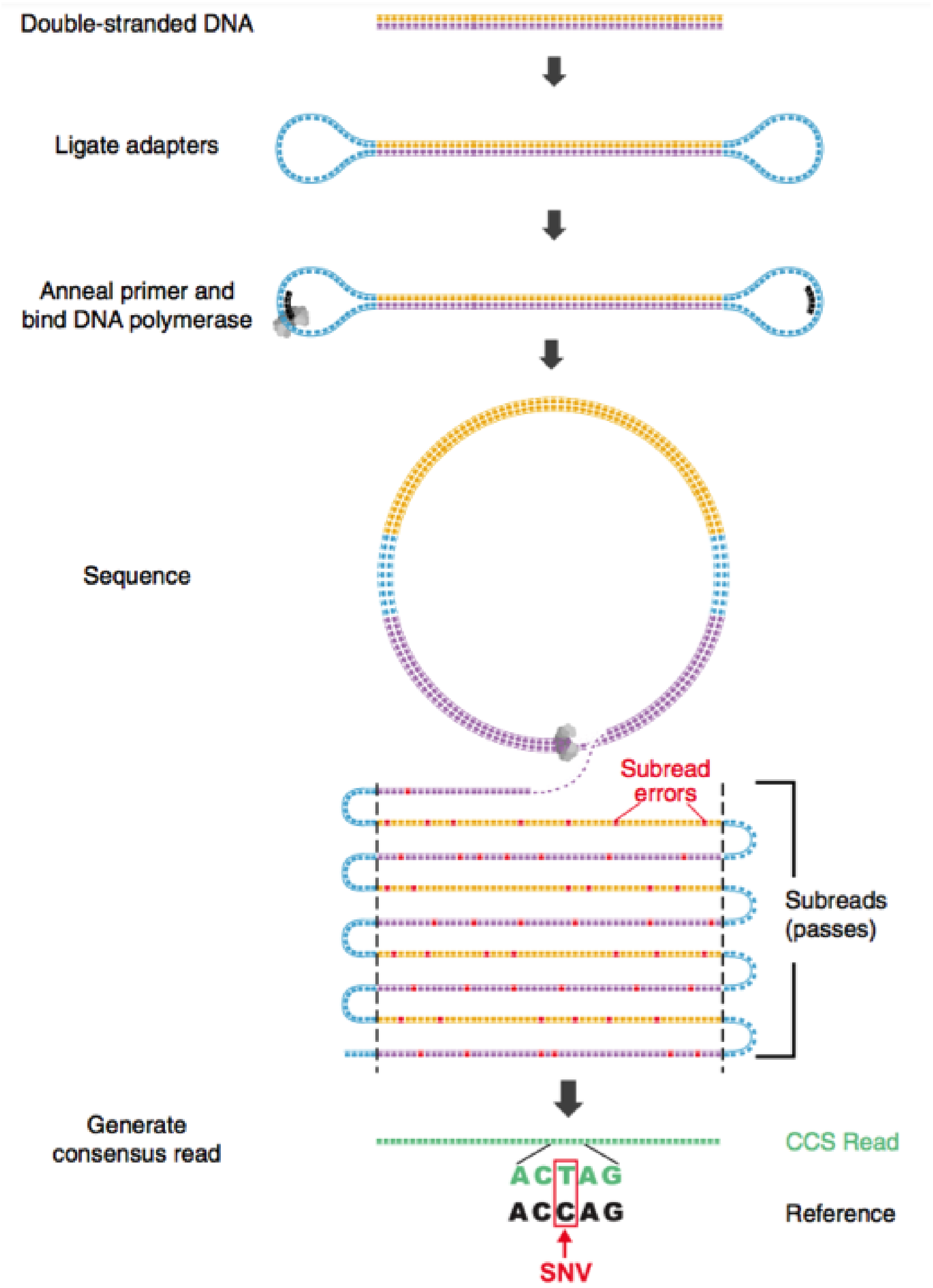

Figure 5: Generating highly accurate long-reads using circular consensus sequencing (CCS). Adaptors are ligated to both ends of the double stranded DNA molecule allowing for the ligation of primers that in turn enable the DNA polymerase to bind. Multiple subreads of the same DNA molecule are generated since the polymerase will pass through the circular structure multiple times. These individual subreads are subsequently overlayed to correct potential sequencing errors before generating a consensus read. Figure source: Wenger et al. 2019.

1.6 Genome assembly

The word genome refers to the combined DNA sequence of all chromosomes of a cell or an individual. The term genome assembly describes both the process and result of reconstructing a genome from DNA sequencing reads. Ideally, a sequencing read would be as long as each chromosome that is the target of the sequencing project. However, as described before, next-generation sequencing technologies are not capable of generating long-reads beyond a few hundred base pairs (Quail et al. 2012). Even more recent long-read sequencing technologies are not yet capable of reading an entire chromosome in one go. The fact that genomes of various species can encompass more than a hundred megabases (e.g. *A. thaliana*) to multiple gigabases (e.g. human) points to a key challenge in genome assembly when using current sequencing methods (Arabidopsis Genome Initiative 2000; Lander et al. 2001): The chromosomes making up a genome cannot be read in one go. Instead, one needs to reconstruct the genome from millions of fragmented DNA sequences by finding overlaps between these reads. Solving this puzzle is the most challenging part of performing a genome assembly (J. R. Miller, Koren, and Sutton 2010; Compeau, Pevzner, and Tesler 2011; Nagarajan and Pop 2013).

The fundamental strategy of genome assemblies can be broken down into two major steps: (1) contig assembly and (2) scaffolding (Paszkiewicz and Studholme 2010). This process is independent of the sequencing strategy and the assembly algorithm that is applied. During the first step, the assembly algorithm tries to find overlaps between the DNA fragments that were sequenced. If reads overlap with each other, a consensus sequence is built and the reads are connected to a contig (El-Metwally et al. 2013). In the second step, the scaffolding, these unconnected contigs are further ordered by comparing them to a known reference and thereby obtaining information on the sequence order. Contigs can also be scaffolded through the use of additional sequencing data spanning the gaps (Sohn and Nam 2018; Simpson and Pop 2015). There are two approaches for performing a whole genome assembly: (1) the *de novo* approach and (2) the comparative approach (El-Metwally et al. 2013). The comparative assembly strategy, also known as reference-based assembly, is relying on the availability of a reference genome from the same species or a closely related one. The reference genome is used as a map to guide the assembly process by aligning the sequencing reads (Pop et al. 2004). In contrast to this, *de novo* genome assembly is performed without the additional guidance of a reference (Martin and Wang 2011).

Different computational tools have been employed to perform genome assemblies. However, most of these tools rely on one of the following algorithms: (a) the overlap-layout-consensus and (b) the de Bruijn graphs dependent algorithms (Paszkiewicz and Studholme 2010). Overlap-layout-consensus-based assemblers try to find overlaps between all possible pairs of reads in a given sequencing dataset. This makes them computationally expensive, since the number of possible combinations increases with the number of sequencing reads (Z. Li et al. 2012). In contrast to this, de Bruijn graph based methods, named after the Dutch mathematician Nicolaas de Bruijn, use a different strategy. First all reads are broken into substrings of length k (kmers). The assembly tool then builds a de Bruijn graph by using these kmers as nodes. An edge (k-1-mer) is assigned to two nodes if the sequence of the two kmers (nodes) overlaps. The algorithm can now reconstruct the initial sequence by finding the way through the graph that traverses each edge once (Compeau, Pevzner, and Tesler 2011).

1.7 Genome annotation

The DNA sequence of a genome alone is of limited use if the encoded information is not deciphered (Mudge and Harrow 2016). The process of decoding the sequence patterns and associating them with features such as transposable elements (TEs), repeat structures or protein-coding genes is referred to as genome annotation (Ejigu and Jung 2020). Repeats and TEs are usually annotated and masked prior to gene annotation (Yandell and Ence 2012). Many different annotation approaches have been developed. The applied methods for both repeat and gene annotation rely on two different types of data: (i) intrinsic information such as *ab initio* predictions from the sequence itself, and (ii) extrinsic information such as transcript or protein alignments (Ejigu and Jung 2020).

1.7.1 Transposable element and repeat detection

The first step in genome annotation usually is the identification and the masking of repeat sequences. In this case, the term repeat summarizes both transposable elements (TEs) and low complexity sequences such as centromeres (Yandell and Ence 2012).

Transposable elements, also known as ‘jumping genes’ are sequences that move from one position in the genome to another. These sequences were initially discovered in maize by Barbara McClintock (McClintock 1950). TEs have the capacity to duplicate or delete genes, to affect gene expression, to delete genes, or to combine genes from different locations into new fusion genes. Thus, transposable elements play an important role in plant genome evolution. TEs are found in almost all eukaryotic genomes that were analyzed so far with prominent examples being maize, rize, *Drosophila*, human, and *Arabidopsis thaliana* (McClintock 1950; International Rice Genome Sequencing Project 2005; Adams et al. 2000; Lander et al. 2001; Arabidopsis Genome Initiative 2000). TEs can be further grouped into two distinct classes based on their replication and insertion strategy (Finnegan 1989). The retrotransposons (Class 1 elements), also known as ‘copy-and-paste’ TEs, are mobilized via a RNA intermediate that is reverse-transcribed into cDNA before re-inserting at a different genomic location (Boeke et al. 1985). Class 2 elements (DNA transposons) are mobilized via a DNA intermediate using a ‘cut-and-paste’ mechanism (Greenblatt and Alexander Brink 1963). Within both classes of TEs there are autonomous and non-autonomous elements. Autonomous elements have open reading frames encoding the proteins that are necessary for transposition. In contrast to that, non-autonomous TEs do not encode transposition proteins. However, they are able to transpose since they carry the cis sequence necessary for transposition. Almost all TEs cause the duplication of short genomic sequences, so called target site duplication (TSD), when integrating into the genome (Feschotte, Jiang, and Wessler 2002). Class 1 TEs can be further grouped into long terminal repeat (LTR) and non-LTR retrotransposons based on their structure and transposition mechanism. As indicated by their name, LTR retrotransposons have a long terminal repeat in direct orientation. Autonomous LTR retrotransposons carry the *gag* gene that encodes a capsid-like protein as well as the *pol* gene encoding a polyprotein that is responsible for protease, reverse transcriptase, RNAse and integrase catalytic activity. In contrast to this, the non-autonomous LTR retrotransposons lack most or all of the coding sequence (Jin and Bennetzen 1989; N. Jiang et al. 2002; Feschotte, Jiang, and Wessler 2002). Non LTR retrotransposons can be further grouped into short interspersed elements (SINEs), which are non-autonomous, and long interspersed elements (LINEs) which are autonomous. Integration of both, SINEs and LINEs, is reached via a process called target-primed reverse transcription (Luan et al. 1993). LINEs and SINEs end with a simple repeat sequence. As mentioned before, Class 2 DNA transposons mobilize via a DNA intermediate in a ‘cut-and-paste’ or in case of helitrons in a ‘peel-and-paste’ manner (Greenblatt and Alexander Brink 1963; Grabundzija et al. 2016). DNA transposons have a terminal inverted repeat in direct orientation. TE subclasses can be further divided into subgroups or superfamilies. These superfamilies can be found across a wide range of different species while having a monophyletic origin. Thus, the two major LTR retrotransposon superfamilies, Ty1/copia and Ty3/gypsy, are found in a wide range of eukaryotic genomes (Malik and Eickbush 2001).

Another fraction of repetitive sequences in plants are centromeric repeats. Centromeres play an important role since they are necessary for cohesion of sister chromatids and for binding of spindle fibers during cell division. Centromeres are organized in large scale tandem repeat arrays (J. Jiang et al. 2003). These repeats have been described to vary in length between 150 bp and 180 bp in most plant species that were studied so far (Oliveira and Torres 2018). However, there are exceptions such as potatoes where repeat monomer length can range from 979 bp to 5.4 kb (Gong et al. 2012). The most abundant satellite repeat in *A. thaliana* is the 178 bp long CEN180 repeat (Martínez-Zapater, Estelle, and Somerville 1986). More recently it was reported that CEN180 can be present with 11,800 to 15,600 copies per chromosome. Thus, these satellite repeats form large tandem arrays (Naish et al. 2021). Such long and repetitive genomic regions often remain unassembled when using short-reads. In *A. thaliana* the estimated genome size is about 135 Mb while the total length of the reference genome TAIR10 is only about 119 Mb (Garcia-Hernandez et al. 2002). This difference between both size estimates is attributable to centromeric regions that remained unassembled in TAIR10 (Arabidopsis Genome Initiative 2000).

Large fractions of eukaryotic genomes can consist of repeat sequences, with transposable elements and other repeats constituting, for example, 80% of the genome of elite maize breeding lines (Springer et al. 2018; S. Sun et al. 2018; Hufford et al. 2021). Similarly, approximately 50% of the human genome consists of repetitive sequences (de Koning et al. 2011; 1000 Genomes Project Consortium et al. 2015). In *A. thaliana* it has been reported that the genome can contain between 10 % (Leutwiler et al. 1984) and 19% repeat sequences (Jiao and Schneeberger 2020). Although many repeats were thought to be non-functional, others have played important roles in generating genetic and phenotypic diversity during evolution (Feschotte and Pritham 2007; Stitzer et al. 2021). Besides studying transposable elements or other repetitive sequences, there is a need to identify such elements since they bias downstream gene annotation (Yandell and Ence 2012). However, the identification and especially correct classification of repetitive elements remains a challenging task (Ou et al. 2019; Flynn et al. 2020). Most tools developed for this task are either *de novo*, homology, or structure-based, although some pipelines also combine different approaches (Han and Wessler 2010; Price, Jones, and Pevzner 2005; Flynn et al. 2020; Ou et al. 2019).

De novo annotation tools aim to identify transposable elements and repeats without prior knowledge about their structure. Many of these tools identify such elements by detecting pairs of similar sequences that are located at different positions in the genome, or by identifying overrepresented sequences. After identification, the overrepresented sequences are clustered into repeat families. However, most of these methods are limited in distinguishing between low complexity repeats and transposable elements. Moreover, low-copy number TEs are often missed (Bergman and Quesneville 2007).

In contrast to *de novo* methods, homology-based approaches rely on extrinsic information about TE protein coding sequences. Thus, these methods are quick and precise in identifying sequences that are similar to already known TEs. In addition, they can be used to accurately identify known repeat motifs such as centromeres (Naish et al. 2021; Rabanal et al. 2022). Moreover, they have better performance in clustering the TEs into families (Ou et al. 2019). However, homology based TE detection can be biased towards identifying known repeats while missing previously unknown or genotype specific ones (Bell et al. 2022).

Structure-based methods (also known as motif based approaches) are relying on prior knowledge about the structure of transposons. Thus, they annotate TEs by identifying combinations of typical TE patterns (Rho and Tang 2009). However, structure-based methods often have a high false discovery rate. Moreover they may fail in identifying TEs with a weak signature (Flutre, Permal, and Quesneville 2012).

All three, *de novo*, structure, and homology-based TE and repeat annotation approaches have specific strengths and weaknesses. Therefore, computational pipelines that combine these methods have been developed (Ou et al. 2019). All annotated transposable elements and repeats are masked prior to gene annotation. Masking converts nucleotides into either lower case letters or into ‘N’. This allows downstream gene annotation tools to recognize and ignore these sequences. Repeat masking is a crucial step, since transposon open reading frames could erroneously be annotated as protein coding genes (Yandell and Ence 2012).

1.7.2 Gene annotation

One can distinguish between *de novo* gene annotation approaches and those methods that use extrinsic data such as transcript or protein alignments. In addition, many *de novo* gene prediction tools such as Augustus or SNAP can additionally utilize extrinsic evidence (Stanke and Waack 2003; Johnson et al. 2008).

De novo gene annotation is also referred to as *ab initio* gene prediction since these tools rely on mathematical models of genic features rather than on extrinsic evidence. Such features can be intron and exon length, promoter sequences, terminator sequences or nucleotide composition (Salzberg et al. 1999; Huang, Chen, and Deng 2016). Different widely used tools predict genes by employing a hidden Markov model (Mahood, Kruse, and Moghe 2020). On one hand, this is an advantage since these approaches can also be used in non-model species with a lack of external information. On the other hand, stand-alone *ab initio* gene prediction has inherent weaknesses. Gene prediction tools are generally not able to annotate alternatively spliced transcripts and untranslated regions. Moreover, these tools still rely on extrinsic data, since the underlying model has assumptions about parameters such as codon frequency or intron and exon length (Yandell and Ence 2012). However, parameters for training a species-specific prediction model in order to annotate a *de novo* genome assembly can be obtained computationally (Simão et al. 2015).

Evidence-based methods can use extrinsic information such as expressed sequence tags (ESTs). ESTs are unedited, short, and randomly selected single pass sequence reads from such a cDNA library. These ESTs give information about transcribed genes in the target tissue or organism (Putney, Herlihy, and Schimmel 1983). Computational tools such as BLAST (Altschul et al. 1990) can be used to identify regions of similarity between features of the query genome and proteins or ESTs (Camacho et al. 2009). Messenger RNA (mRNA) sequences represent the entirety of all expressed genes in a cell. Since RNA cannot be sequenced directly, it first needs to be reversely transcribed into double stranded cDNA. More recent approaches can utilize protein sequences and RNA-seq reads instead of ESTs in order to guide gene annotation. RNA-seq data can be used in two different ways. The reads can either be directly aligned to the query genome or they can be used to first assemble transcripts *de novo.* These are then processed similarly to ESTs. Generally, the use of RNA-seq data improves the annotation of alternatively spliced exons, splice sites and intron-exon boundaries (Yandell and Ence 2012). However, RNA-seq data have the limitation that they rely on the gene to be expressed in the sequenced sample in order to be annotated.

Other evidence-based annotation approaches such as Liftoff (Shumate and Salzberg 2020b) can utilize pre-existing gene annotations from closely related species. In this case, known gene sequences are aligned to the query genome. Exons are then annotated using sequence identity while preserving the transcript and gene structure (Shumate and Salzberg 2020a).

There are tools such as EvidenceModeler that can combine the gene models obtained by the different annotation approaches (Haas et al. 2008). However, even annotations derived from a combination of different sources may still require manual curation (Lewis et al. 2002; Rutherford et al. 2000). Albeit a laborious approach, manual curation is especially useful when annotating complex gene families such as NLRs (Van de Weyer et al. 2019). More recently, deep learning algorithms such as DeepAnnotator have been introduced to the field (Amin et al. 2018). These algorithms can be classified into supervised and unsupervised approaches. Such deep learning algorithms benefit from the increasing amount of sequence data that is generated for all kinds of organisms. Different machine learning methods have been proposed in order to predict sequence features (Yang et al. 2020). There are reports suggesting that supervised machine learning algorithms can outperform traditional HMM based methods when it comes to the prediction of splice or transcript start and stop sites (Sonnenburg et al. 2007; Ben-Hur et al. 2008). However, the availability of high quality training datasets is still a limiting factor when it comes to the application of machine learning algorithms in genome annotation (Mahood, Kruse, and Moghe 2020).

1.8 Limitations in assessing structural and NLR gene diversity

As pointed out in the sections before, there are two major limitations that have so far been hindering the analyses of large-scale structural variants and of NLR gene diversity. I want to give some examples of these limitations before proposing how to use the latest sequencing technology in order to tackle these challenges.

1.8.1 Limitations of evidence from short-read DNA sequencing

Ideally, a sequencing read would be as long as the chromosome that one wants to assemble. However, as mentioned before, the read lengths of next-generation sequencing platforms are limited to about 300 bases (Quail et al. 2012). Therefore, short-reads today are mostly used for the detection of single-nucleotide variants as well as small insertions or deletions by mapping them to known genomes. Even the Sanger sequencing based *A. thaliana* reference genome released in 2000, which is still considered the 'gold standard' today, contains 117 gaps, 29 large mis-assemblies, and misses approximately 25 Mb of repeat sequence (Kawakatsu et al. 2016; Q. Long et al. 2013). The advent of high-throughput next-generation sequencing technologies has led to numerous resequencing efforts trying to assess the genomic diversity of different *A. thaliana accessions* using short-reads (Q. Long et al. 2013; 1001 Genomes Consortium 2016; Clark et al. 2007; Cao et al. 2011; Ossowski et al. 2008). Most of these studies were based on the original *A. thaliana* reference genome and were limited in terms of the detection of large-scale structural variation (Michael et al. 2018). Moreover, multiple studies either performed a reference-guided or a *de novo* genome assembly of *A. thaliana* using short-reads (Schneeberger et al. 2011; Gan et al. 2011). While short-read assemblies are capable of reconstructing genes reasonably well, most of these studies were unable to accurately resolve large-scale structural variations and highly repetitive regions (Michael et al. 2018; B. Wang et al. 2021; Naish et al. 2021; Jiao and Schneeberger 2020; Alkan, Sajjadian, and Eichler 2011). Similar problems were encountered when attempting to *de novo* assemble other species' genomes (such as human or maize) from short-reads (Alkan, Sajjadian, and Eichler 2011; Chaisson et al. 2019; Schnable et al. 2009). Most of these limitations are attributable to the limited read length of short-reads, which is approximately 300 bp (Quail et al. 2012). These reads are too short to span relevant repeat regions and are therefore not suited for assembling centromeres, telomeres, highly repetitive immunity gene clusters, and large structural variants (Rhoads and Au 2015; Hon et al. 2020; Naish et al. 2021). As a consequence, such genomic regions have remained mostly unassembled and therefore understudied (Michael et al. 2018).

1.8.2 Usage of a single reference genome

An increasing number of publications indicates that the full spectrum of genetic diversity within a species cannot be grasped by a single reference genome, also not when re-sequencing hundreds of accessions. For instance, the use of only a single reference genome for an entire species limits the possibility of identifying genetic variants, especially presence-absence variation (PAV), large-scale structural variation (SV), and copy number variation (CNV), simply because one can only detect or miss what is already described in the reference (Scherer et al. 2007; R. Li et al. 2010; Hirsch et al. 2014; Lin et al. 2014; Saxena, Edwards, and Varshney 2014; Yao et al. 2015; Golicz, Batley, and Edwards 2016; Zhou et al. 2017; Wensheng Wang et al. 2018). As a result, it has remained very difficult to describe CNVs in spite of their great importance for the adaptation of wild species as well as agronomically favorable traits (Liu et al. 2020). The diversity of NLR genes makes it unlikely that this gene family can be assessed using a single reference genome (Van de Weyer et al. 2019). Moreover, as mentioned in section 1.4, it is challenging to analyze F_1 hybrid transcriptomes while only having genome information of one parent available.

2. Aims of the dissertation

In my PhD project I aimed to generate multiple reference genomes from different *A. thaliana* accessions. In order to do so, I applied PacBio HiFi sequencing. This highly accurate long-read sequencing technology solves many of the problems in *de novo* genome assembly that are associated with either short-reads or with error-prone continuous long-reads. By doing so, I wanted to overcome (a) the former problems of only a single reference genome being available and of (b) the reference genome being fragmented into highly repetitive regions.

During my doctoral research I strived to characterize differences in genomic architecture of eighteen natural *A. thaliana* accessions from genomes assembled following long-read sequencing. I investigated gene content, transposable element density, large scale structural variation, and NLR genes to describe variation not previously visible from short-read sequencing data. Moreover, I used the power of full genome information to examine non-additive gene expression in F_1 hybrids.

2.1 Eighteen differential *Arabidopsis thaliana* lines

For my main project, the characterization of differences in genomic architecture of eighteen different accessions, I used a set of *A. thaliana* accessions that were previously compiled from different locations within Europe by Dr. Gautam Shirsekar, a postdoc in the lab. He chose those 18 accessions in order to investigate the different responses of the lines when being inoculated with the obligate biotroph oomycete *Hyaloperonospora arabidopsidis*. Thus, one of my key motivations in using exactly these 18 accessions was to provide Dr. Shiresekar's project with a broader data foundation about the genomic architecture as well as the NLR gene diversity within this set. In order to do so, I set out to address the following objectives:

a) Assemble and annotate eighteen diverse accessions from PacBio HiFi long-read sequencing data.
b) Determine the number of genes found in (almost) all accessions.
c) Identify number and prevalence of NLR genes among accessions.

2.2 Differential gene expression in *A. thaliana* F_1 hybrids

In the course of the second project I aimed to answer the following questions by making use of the availability of two full-length genome assemblies:

a) Can parental gene expression predict the magnitude of non-additive gene expression in F_1 hybrids?

b) Does the number of non-additively expressed genes change during plant development?

3. Material & Methods

3.1 Eighteen differential *Arabidopsis thaliana* lines

3.1.1 Accession selection and plant growth

The *A. thaliana* accessions used in the presented work were obtained from the Weigel laboratory at the Max Planck Institute for Biology, Tübingen. Accessions were chosen by Dr. Gautam Shirsekar based on the premise of covering a wide range of genetic diversity. All seeds were surface-sterilized by soaking them in 70% EtOH for 5-10 minutes followed by a wash with 90% EtOH for one minute. Stratification was done by keeping the surface sterilized seeds in 0.1% (w/v) agarose (Sigma Aldrich) at 4 °C for 7 days in darkness. About 5-10 seeds were sown per pot. Plants were kept in a growth chamber in short day conditions (8 h light) under 110-140 µmol m^{-2} s^{-1} light using Philips GreenPower TLED modules (Philips Lighting GmbH, Hamburg, Germany). Ambient temperature was set to 23 °C with a maximum humidity of 65%. Plants were grown under the above described conditions for approximately 30 days. Plant trays were covered in order to keep the plants in darkness for 32 h prior to harvesting the samples. This was done in order to reduce starch content. Plant tissue was harvested and immediately frozen in liquid nitrogen. All samples were stored at -80 °C in order to prevent degradation of the DNA.

3.1.2 High-molecular weight DNA extraction

The high-molecular weight (HMW) DNA extraction approach performed for this study included two steps. First, the plant tissue was lysed in order to extract nuclei while removing organellar DNA. High-molecular-weight DNA was extracted from these nuclei in the second part of the procedure.

Frozen plant material was ground in liquid nitrogen using a pre-cooled mortar and pestle. Finely-ground powder was immediately transferred back into centrifugation tubes and kept on liquid nitrogen until further use. 20 g of frozen plant tissue powder were used per extraction reaction. The powder was transferred into freshly prepared ice-cold Nuclei Isolation Buffer (NIB) and kept on a magnetic stirrer set to low speed for 15 min in a 4°C cold room. The nuclei isolation buffer contains 1 mM Tris (pH 8) (Carl Roth GmbH, Karlsruhe, Germany), 0.1 M KCl (Carl Roth GmbH, Karlsruhe, Germany), 10 mM EDTA (pH 8) (Carl Roth GmbH, Karlsruhe, Germany), 0.5 M sucrose (Carl Roth GmbH, Karlsruhe, Germany), 4 mM spermidine (Sigma-Aldrich, St. Louis, USA), and 1 mM Spermine-4HCl (Sigma-Aldrich, St. Louis, USA). After incubation the nuclei containing solution was filtered through four layers of Miracloth (Merck). Pellets were squeezed in order to allow for maximum recovery of nuclei-containing solution. Subsequently, another 10 mL of NIB containing 20% (v/v) Triton-X-100 (Sigma-Aldrich, St. Louis, USA) were added to the nuclei solution prior to another 15 min of incubation at 4°C on a magnetic stirrer set to low speed. Nuclei solution was distributed into multiple 50 mL centrifuge tubes. Samples were centrifuged at 3250 rpm for 15 min in a tabletop centrifuge set to 4°C. Supernatant was carefully discarded while the nuclei pellets of multiple tubes were combined into a single 50 mL centrifuge tube. Empty tubes were subjected to two serial washes using 17 mL of NIB additionally containing 1% of Triton-X-100. The washing solution was added to the previously collected nuclei. Subsequently, nuclei were again collected by centrifugation with the same conditions as before.

Having isolated intact nuclei, the HMW DNA needed to be released. Therefore, the previously isolated nuclei were lysed in 20 mL of pre-warmed (37°C) G2 lysis buffer (Qiagen, Hilden, Germany). Potential RNA contamination was removed by adding 100 µL (20mg/mL) of RNase A and subsequent incubation for 30 min at 37 °C. Afterwards, 500 µL of proteinase K (20 mg/mL) (AppliChem, Darmstadt, Germany) were added in order to inactivate RNase A (Qiagen, Hilden, Germany) and to degrade proteins. Proteinase K treatment was performed for 3 h at 5 °C while gently inverting the reaction tubes every 30 min. Subsequently, the samples were centrifuged at 8000 rpm for 15 min at 4°C. The DNA-containing supernatant was kept and carefully poured onto a QIAGEN genomic tip100 in order to perform further washing steps. Genomic tip was equilibrated with4 mL QBT buffer (Qiagen, Hilden, Germany) before use. After the DNA containing solution passed through the tip, a total of two washing steps was performed, each using 7.5 mL of QC buffer (Qiagen, Hilden, Germany). High-molecular-weight DNA was eluted using a 5 mL QF buffer (Qiagen, Hilden, Germany) pre-warmed to 50°C. Subsequent DNA precipitation was performed by adding 3.5 mL isopropanol (Carl Roth GmbH, Karlsruhe, Germany). The tube was gently inverted until visible strings of DNA formed that were subsequently transferred to a 1.5 mL DNA low-bind reaction tube containing 300 uL of EB (Qiagen, Hilden, Germany). In order to allow the DNA to fully dissolve, the tubes were incubated without shaking at 4°C for 3-7 days. DNA concentration was measured using the Qubit Fluorometer dsDNA-High-Sensitivity Kit while following the manufacturer's instruction manual.

3.1.3 Library preparation and sequencing

Library preparation and sequencing were done by Dr. Christa Lanz and Theresa Schlegel. Therefore, I will just give an overview about the key steps in this process that differed from the PacBio protocol. Megaruptor 2 (Diagenode s.a., Liege, Belgium) was used with long hydropores in order to shear the HMW DNA to an average size of 15-20 kb. Subsequently, the sheared DNA was concentrated and further purified using AMPure PB beads (Pacific Biosciences, California, USA). Fragment size was checked using the Femto Pulse System (Agilent, Santa Barbara, USA). DNA concentration was measured using Qubit fluorometer (Thermo Fisher Scientific, Waltham, USA). At least 10 µg of the sheared and quality checked DNA was further processed following the PacBio® Procedure & Checklist - Preparing HiFi SMRTbell® Libraries using SMRTbell® Express Template Prep Kit 2.0 (PN101-853-100 Version 01 (September 2019)). Another bead clean up was performed after finishing the library preparation protocol. Subsequently, a size selection for fragments >15 kb was carried out using the BluePippin System (Sage Science, Beverly, USA). The SMRT Link Sample Setup (Pacific Biosciences, California, USA) was used in order to determine exact concentrations of all reagents for each sample individually, based on its size and library concentration. Libraries were loaded onto the instrument via diffusion loading following the manufacturers recommendations and using concentrations ranging from 40 pmol to 55 pmol. A movie time of 30 h with a two hour pre-extension period was used. All sequencing runs were carried out using v2 of the PacBio chemistry bundle.

3.1.4 CCS calling

PacBio fastq read files were filtered for CCS reads with an average of >Q20 read quality, similarly to (Wenger et al. 2019) by using PacBio Circular Consensus Sequencing tool (v6.0.0) with `--minrq 0.99` and `--chunk 10`. Fastq files were split into ten files (chunks) during CCS calling in order to increase speed. The resulting chunked BAM files were merged using pbmerge (v0.23.0) prior to demultiplexing and/ or conversion to FASTA format using bam2fasta (v1.3.0). Both, pbmerge and bam2fasta are supplied with the PacBio SMRTtools.

3.1.5 Evaluation of assembly strategies

Although the results of the assembly tool evaluation led to my decision to use hifiasm (H. Cheng et al. 2021) for the final assemblies in this thesis, I used the Falcon-unzip2 (Chin et al. 2016) assemblies to test the impact of sequencing coverage or contig correction on the *de novo* assembly. This is due to the fact that Falcon-unzip2 was the first tool enabling the assembly of PacBio HiFi read data while hifiasm was released later during the course of the experiments presented in this thesis. Thus, certain tests were only performed using Falcon-unzip2. The manual correction of chimeric contigs was only performed for Falcon-Unzip2 assemblies, since no such errors were observed in the hifiasm contigs. Parameters used to obtain either Falcon-Unzip2 or hifiasm assemblies are listed in the following section.

Assembly tools

I assessed the performance of two different *de novo* genome assembly tools. I assembled all genomes using Falcon-Unzip2 (Chin et al. 2016) and hifiasm (H. Cheng et al. 2021). All tools were used applying the default parameters unless mentioned otherwise. Genomes assemblies with Falcon-Unzip2 (v1.8.1) were performed while applying `genome_size=140000000`, `input_type=preads`, and `overlap_filtering_setting=--min-idt 99.9`. *De novo* assembly of each genome was also performed using hifiasm (v0.15.4-r343) (H. Cheng et al. 2021) with `-l0 (purge level = 0)` and `-f0 (number of bits for bloom filter)`. Contiguity, largest contig, and total assembly length of the primary genome assemblies were compared among the two assembly tools using quast.

Assessing minimum genome coverage

In order to assess the minimum sequencing coverage that allows for *de novo* assembly using Falcon-Unzip2 without compromising on assembly quality, I subsampled the Q20 read files of AT9900, AT9847, AT9830, AT9104 and AT9503 to 150x, 125x, 75x, 50x, and 25x using `seqtk sample`. Genome assembly was performed with Falcon-Unzip2 using the parameters described in the assembly section of this thesis.

Manual correction of chimeric contigs

Primary contigs were aligned to the *A. thaliana* reference genome TAIR10 (Arabidopsis Genome Initiative 2000) using minimap2 (v2.11-r797) (H. Li 2018). Contig alignments were visualized using minidot from the miniasm pipeline (v0.2-r168-dirty) (H. Li 2016) and inspected manually. Contigs that mapped to more than one target chromosome were further inspected by mapping Q20 CCS reads to the respective assembly. Contigs were then checked for regions with low coverage using Integrated Genome Browser (IGV) software (v2.9.4) (Robinson et al. 2011). Contigs were manually broken if only a single read supported a suspicious junction.

The final assembly strategy

The final genome assemblies that were used for this thesis were generated after the above described method evaluation. Hifiasm was used while applying the previously mentioned parameters. Query genome coverage was calculated based on the accumulated length of Q20 filtered subreads. Primary contigs below 100 kb length were removed from the assembly prior to genome scaffolding. Scaffolding was performed by aligning the size-filtered primary contigs of each *de novo* assembly to the Bionano map of AT9852 using ragtag scaffold (v2.0.1) (Alonge et al. 2022) with -q 60 (mapping quality), -f 30,000 (minimum unique alignment length, --remove-small (remove unique alignments shorter than -f 30,000 bp), and -i 0.5 (minimum grouping confidence score). Gaps in the resulting scaffolds were filled by adding stretches of 100 Ns in order to connect contigs.

3.1.6 Completeness assessment

The completeness of each *de novo* assembly was assessed with BUSCO (v4.0.6) (Simão et al. 2015) using -m genome and the odb10_embryophyta database (Zdobnov et al. 2021). Continuity, GC content, overall assembly length and reference coverage of assembled contigs was checked using Quast (v5.0.2) (Gurevich et al. 2013). Structural variation calling among the scaffolded genome assemblies and the *A. thaliana* reference genome TAIR10 was performed using the Synteni and Rearrangement Identifier (SyRi v1.3) (Goel et al. 2019).

3.1.7 Analysis of an outlier sample

During quality assessment I noticed that one assembly differed substantially in almost all quality metrics. Thus, I suspected that the assembly contains genomic reads from more than a single genotype. Thus, kmer counting was performed using jellyfish (v.2.2.8) with count -C, -m 21, and -s 1000000000. The resulting jellyfish output file was converted using jellyfish histo prior to visualization with genomeScope (Ranallo-Benavidez, Jaron, and Schatz 2020).

3.1.8 Contamination removal

Primary contigs of all assemblies were screened for potential contamination with viral or bacterial DNA sequences. Firstly, Q20 CCS reads were mapped against their corresponding *de novo* assembly using minimap2 (v2.17-r941) (H. Li 2018). Resulting SAM files were converted to BAM format and sorted using samtools (v1.9) (H. Li et al. 2009). Secondly, all primary contigs were mapped against a metagenomics database using diamond blastx (v2.0.7) (Buchfink, Xie, and Huson 2014) with --top 10, -b 10, -c 1, --sensitive, and -f 102. Read mappings and diamond results were afterwards combined and visualized using Blobtools (v1.1.1) (Laetsch and Blaxter 2017). Primary contigs that were flagged as ‘Viruses’ or ‘Bacteria’ were removed from any downstream analyses using seqtk (v1.0).

3.1.9 TE and repeat annotation

TEs were identified and annotated using EDTA (v1.9.1) with `--anno 1 (Ou et al. 2019).` Moreover, the `--sensitive 1` setting allowed for the identification of remaining TEs. Additionally, the coding sequences of the *A. thaliana* reference Araport11 (C.-Y. Cheng et al. 2017) were used via the `--cds` option in order to prevent the false annotation and masking of known protein coding genes.

Repetitive elements such as centromeres, telomeres, 5s rDNA, and 45s rDNA were annotated with RepeatMasker (v4.1.0) (Smit, Hubley, and Green 2015) with `-nolow` and `-gff`. A custom repeat library containing centromere and rDNA sequences was specified with `-lib`. Centromeric 178 bp sequence motif as well as telomere and rDNA sequences were obtained from publicly available resources (Maheshwari et al. 2017; Rabanal et al. 2017).

Redundant repeat annotations were removed by cross-referencing the TE annotation obtained from EDTA with the centromere, telomere, and rDNA annotation from RepeatMasker. Transposable element entries from EDTA were removed from the annotation if they exhibited a positional intersection with centromere, rDNA or telomere sequences.

3.1.10 Annotation of protein coding genes

Previously annotated TEs were softmasked using bedtools (v2.26.0) (Quinlan and Hall 2010) `maskfasta -soft` prior to the annotation of protein coding genes. Gene annotation was performed using Augustus (v3.3.3) (Stanke and Waack 2003). Specific retraining parameters for the internal Augustus gene prediction model were obtained by using BUSCO (v4.0.1) (Simão et al. 2015) with the `-m genome` setting. Moreover, Liftoff (v1.5.1) (Shumate and Salzberg 2020b) was used to project the *A. thaliana* TAIR10 reference genome annotation onto each of the here presented *de novo* assemblies with the settings `-exclude partial` and `-copies` in order to generate transmap hints for the subsequent Augustus-based gene annotation. Augustus was used with the following parameters deviating from the default settings: `--softmasking 1`, `--species=BUSCO_retraining`, `--gff3=on`, `--extrinsicCfgFile=Custom_Config`, and `--hintsfile=Liftoff_hints`. Subsequently, sequences in the resulting GFF3 files were translated into amino acid sequences using the Augustus provided script getAnno.pl. Amino acid fasta files were subsetted to only represent the primary isoform of each locus using agat_sp_keep_longest_isoform.pl from the AGAT toolkit (v0.2.3) (Dainat 2020) in order to enable subsequent analysis of homologous genes. Subseqently, homologs of the newly-annotated *A. thaliana* genomes and the publicly available reference annotation TAIR10 were assigned using Orthofinder (v2.2.6) (Emms and Kelly 2019) with `-S diamond`.

3.1.11 Figure generation

All figure panels were generated using R studio (v4.1.0) (RStudio Team 2015), unless stated otherwise. Applied packages included but were not limited to ggplot2 (v3.3.5) (Hadley Wickham 2009), reshape (v4.1.0) (H. Wickham 2007), UpSetR (v1.4.0) (Conway, Lex, and Gehlenborg 2017), and pheatmap (v1.0.12) (Kolde and Others 2012).

3.2 Differential gene expression in F_1 hybrids

3.2.1 Plant material and growth conditions

Arabidopsis thaliana accessions Columbia-0 (Col-0) (CS76778; Ecotype ID 6909) and Landsberg erecta (Ler-0) (CS77020; Ecotype ID 7213) were used as parental lines. F_1 hybrid seeds were obtained by hand pollination. Ler-0 served as the maternal line while Col-0 was used as the paternal line. All seeds were surface sterilized using EtOH and frozen at -80°C overnight in order to kill any insect eggs. Seeds were stratified in darkness at 4°C for five days before sowing. All plants were grown in a randomized incomplete block design with an ambient temperature of 23°C and a 16/8 hour light/dark cycle. Calcined clay was used as a substrate. Nutrients were supplied once a week by watering with nutrient solution as described in Conn et al. 2013. Samples from parents and F_1 hybrids were taken at three different timepoints after germination. Whole seedlings were harvested 3 DAG (days after germination), while roots and shoots were sampled separately at 10 DAG. Finally, at 21 DAG, root, shoot, and flower samples were obtained. All samples were immediately snap frozen in liquid nitrogen and stored at -80°C until further usage.

3.2.2 RNA preparation and sequencing

Total RNA was extracted following the procedure described in Yaffe et al. 2012. RNA integrity was checked using gel electrophoresis. The NEB Next magnetic isolation module (New England Biolabs, Ipswich, USA) was used to enrich for polyA$^+$ mRNA. Sequencing libraries were prepared using NEB Next Ultra II RNA library kit following the manufacturer's protocol. A total of three Illumina HiSeq3000 (Illumina Inc., San Diego, USA) lanes with 18 samples each were sequenced in 150 bp paired-end mode. For each sample I sequenced three biological replicates.

3.2.3 Custom hybrid reference and RNA-seq read mapping

A Ler-0 x Col-0 F_1 hybrid reference was generated by combining full length genome information of TAIR10 (Arabidopsis Genome Initiative 2000) with an inhouse generated reference for Ler-0. Existing TAIR10 annotation was lifted over onto the newly generated Ler-0 genome. RNA sequencing data were used to provide additional evidence for Augustus *de novo* gene prediction (Stanke et al. 2004). Orthologs were assigned in both references using the Comparative Annotation Toolkit software (Fiddes et al. 2017). A combination of Bowtie2 (Langmead and Salzberg 2012) and RSEM (B. Li and Dewey 2011) was applied for read mapping and transcript abundance estimation. Bowtie2 indices for the custom hybrid reference were generated using rsem-prepare-reference with default settings while additionally setting `--gff3` parameters and `--allele-to-gene-map`.

3.2.4 *In silico* hybrids

Mid-parent gene expression levels were obtained by generating *in silico* hybrids. All parental read files were first normalized according to sequencing depth and were subsequently subsampled by randomly drawing 50% of the reads from each file using Seqtk toolkit (https://github.com/lh3/seqtk). Subsampled read files of both parents were combined afterwards in order to obtain in silico hybrid read files containing equal numbers of total reads from both of the inbreds. All possible *in silico* hybrid combinations from the three parental replicates have been generated. This process was done iteratively 5 times by using 5 randomly generated seeds when subsetting the read files. A set of three files per tissue and time point has been randomly selected for further mid-parent value analyses. Principal component analysis of parental samples and *in silico* hybrids were performed to check for intermediate transcriptome composition.

3.2.5 Bioinformatic analyses

All statistical analyses for this chapter were done in Rstudio (version 3.4) (RStudio Team 2015). Expected counts obtained from RSEM were imported into R using the tximport package (Soneson, Love, and Robinson 2015). Significantly differentially expressed genes were identified using the R package DESeq2 (Love, Huber, and Anders 2014) while applying the filtering parameters described in (M. Miller et al. 2015) ($p_{adj} < 0.05$; $\log_2$ fold change > |0.5|). Gene ontology analyses were performed using the agriGO online tool (Z. Du et al. 2010). Hypergeometric tests, relative to the genome background of *A. thaliana*, were applied to identify significantly (FDR adjusted P value <0.05) enriched GO terms. All plots were generated using the R package ggplot2 (Hadley Wickham 2009).

4. Results

The results obtained from the two proposed projects will be presented separately during the course of this section.

4.1 Eighteen differential *A. thaliana* lines

As described in the above method section, Dr. Gautam Shirsekar chose these 18 *A. thaliana* lines in order to investigate their differential response when being inoculated with the obligate biotroph oomycete *Hyaloperonospora arabidopsidis* (*Hpa*). My motivation in using exactly this set of accessions for my doctoral research is to provide Dr. Shirsekar's project with a broad data foundation that enables the analysis of NLR gene diversity in the context of *Hpa* infection and that provides insights into the *A. thaliana* pan-NLRome. Thus, it is crucial to generate high quality genome assemblies and sophisticated genome annotations. In this chapter I will present the results that were obtained from long-read sequencing of different *A. thaliana* accessions. I will start with the origin of the sequenced accessions and with describing the raw sequencing data. Moreover I will show how the assembly method was optimized using only a subset of five out of the eighteen accessions. Afterwards I will present the final results from applying my assembly approach beginning with basic quality control followed by genome assembly, annotation, and the analyses of NLR genes as well as structural variants. Moreover, I will show why two out of the initially chosen *A. thaliana* accessions were not further pursued during the course of this work.

4.1.1 Accession selection

The *A. thaliana* accessions used in the presented study cover a broad geographical range (**Figure 6**) and were selected by Dr. Gautam Shirsekar from the 1001 genomes resource (1001 Genomes Consortium 2016) to best represent the genetic diversity of the species as known at the time. Most accessions (eight out of twenty) were originally collected in Spain. AT9336, an accession from Sweden, is the most northern accession while AT9830 represents the most southern accession. Accessions from Tajikistan (AT6929), Georgia (AT9104), and Romania (AT9744) were analyzed in addition to the ones from central Europe (Appendix: **Table S1**).

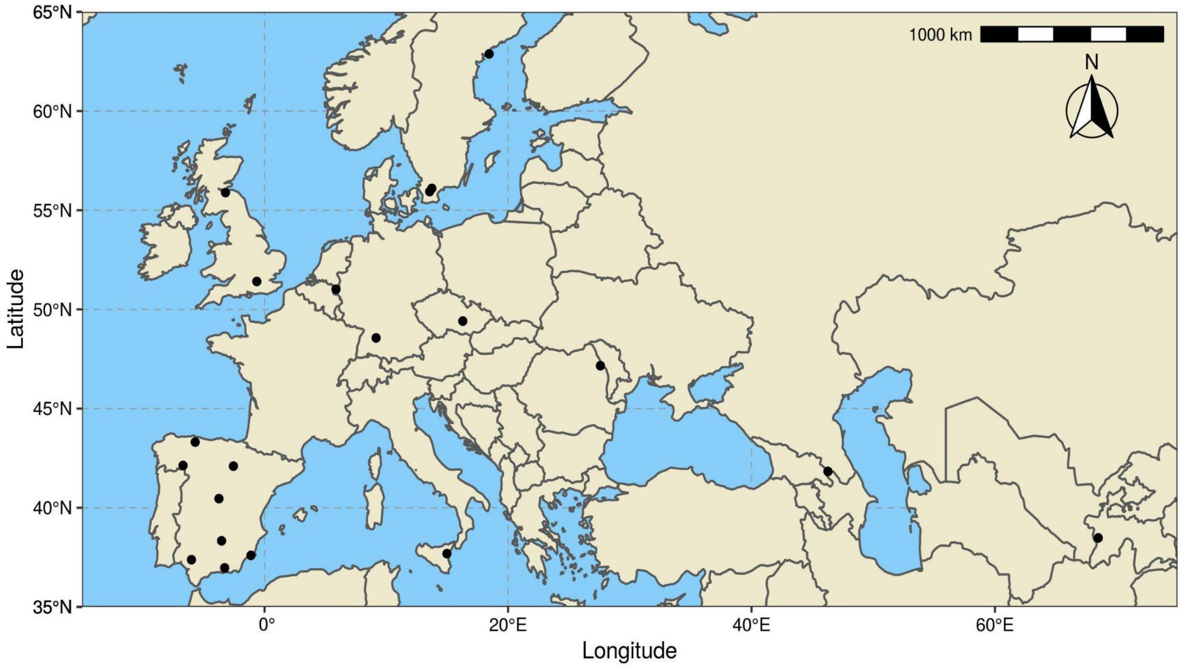

Figure 6: Origins of the 20 differential *A. thaliana* lines. Genotypes (black dots) used for this work cover a broad geographical range within Europe. Most *A. thaliana* lines originated from Spain. The most northern accession was collected in Sweden while the most eastern accession originated from Tajikistan.

4.1.2 Sequencing and CCS calling

Theresa Schlegel performed most of the DNA extractions, and prepared the PacBio sequencing libraries together with Dr. Christa Lanz. Each accession, except for two multiplexed samples, was analyzed on individual SMRT cells, with output of 155 to 555 Gb of raw sequencing data per SMRT cell (**Figure 7**). The average yield per run was 385 Gb. A subread is defined as a single pass of a polymerase on a single strand of an insert with no adaptor sequence within a SMRTbell template (PacBio Glossary of Terms; PN000-710; version 10; April 2019). The subread N50, describing the minimum length of at least half of all subreads, ranged from 15.3 to 21.3 kb, with an average of 17.5 kb. The mean polymerase read length, defined as the mean number of bases produced from a single ZMW, varied from 59.9 to 99.9 kb. On average, mean polymerase read length was 83.1 kb (**Figure 7**). After sequencing it was necessary to perform CCS calling, to generate high quality reads, defined as those that exceed 99% (Q20) accuracy. During CCS calling, circular reads are linearized and the number of adapter occurrences in the linearized molecule are counted, since this number corresponds to the number of full polymerase passes. Reads with ≥5 full polymerase passes meet the quality threshold of Q20 (Wenger et al. 2019). The observation that only a fraction of the initially obtained total sequencing yield is kept for downstream analysis is attributable to the fact that not all CCS reads meet the filtering criteria of Q20 base quality. Moreover, the bases in a read are sequenced several times. Therefore, the total yield is reduced during CCS calling when multiple full passes are combined into a single subread.

Downstream of the CCS calling and filtering for >Q20, I retained between 26.8% (AT9897) and 46.2% (AT9338) of the initial read data (**Figure 7**).

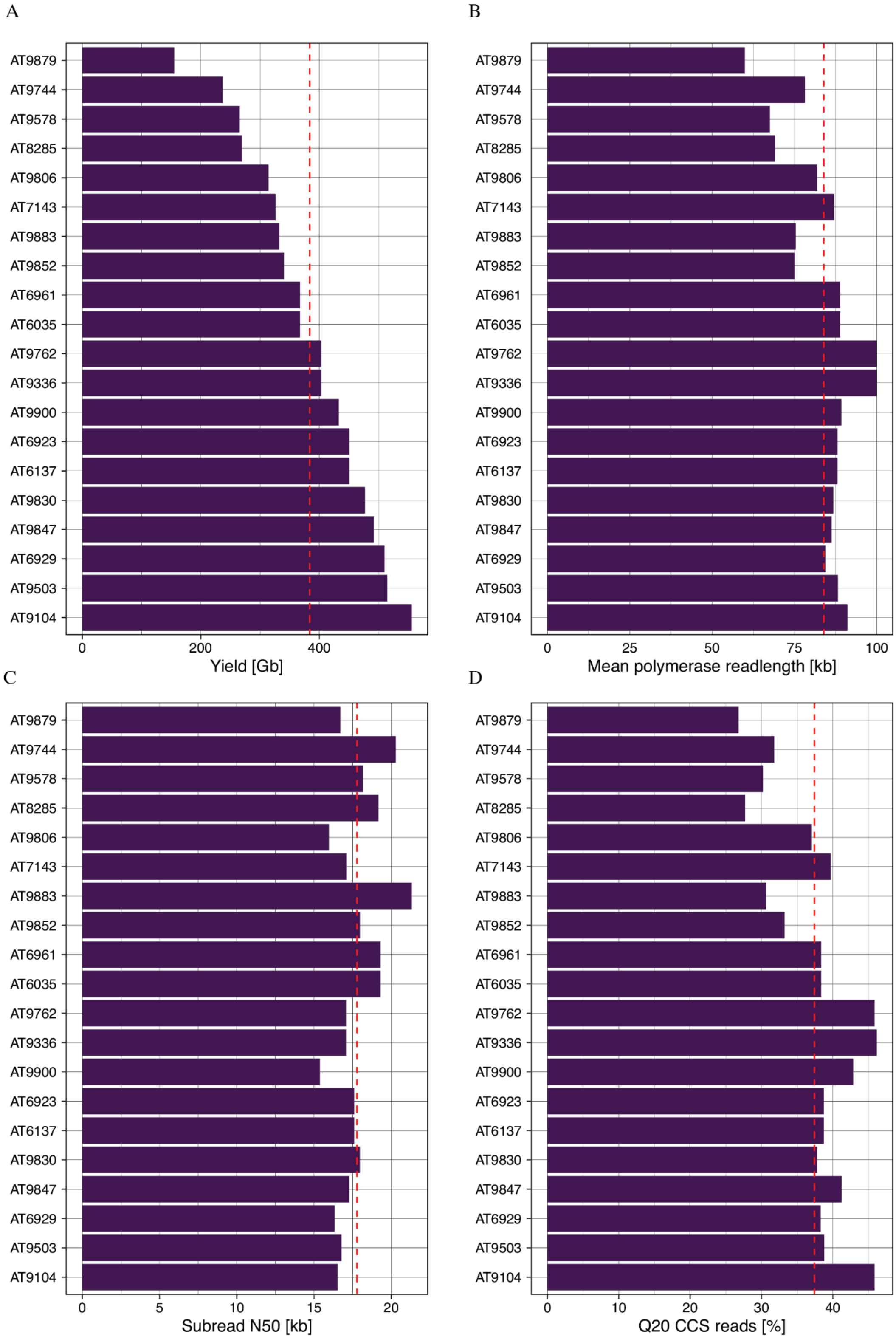

Figure 7: Raw sequencing output statistics of the 20 genomes. Dashed red lines indicate average values. (A) Yield of raw bases in gigabases (Gb). (B) Mean polymerase read length in kilobases (kb). (C) Subread N50 shows that on average 50 % of all bases in the assemblies come from subreads longer than 17.5 kb (dashed red line). (D) Percent of CCS reads with Q >= 20.

Sequencing coverage of the *A. thaliana* genome was calculated based on the reads kept after >Q20 filtering while ignoring the presence of organellar reads that may have a higher coverage and by using the estimated genome size of 135 Mb. The achieved sequencing coverage ranged from 24x (AT6137) to up to 252x (AT9104) (**Figure 8**). Accessions AT6137 and AT6923 were multiplexed on a single SMRT cell, resulting in a lower coverage per genome.

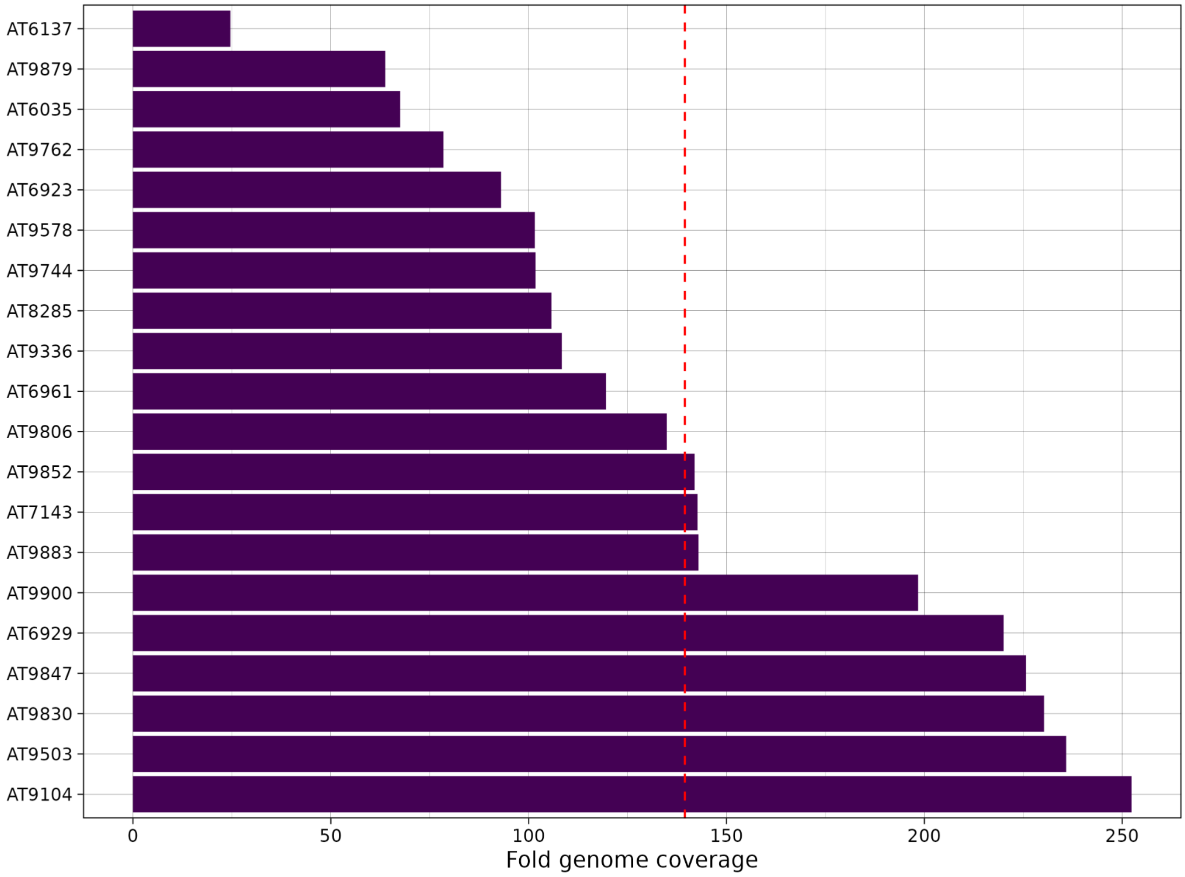

Figure 8: *A. thaliana* genome coverage after Q20 filtering. Fold genome coverage was calculated based on the estimated *A. thaliana* genome size of 135 Mb (Arabidopsis Genome Initiative 2000). The average coverage of 139 fold is indicated by the dashed red line.

The maximum number of full passes for a given circular DNA molecule depends on its length as well as on the maximum read length capability of the polymerase. Thus, at a given maximum polymerase read length, circular DNA molecules with shorter insert sizes have more full DNA polymerase passes, resulting in a higher read accuracy. I observed a moderate negative correlation (*R* = -0.5, *p* = 0.023) between the fraction of > Q20 reads and the subread N50 (**Figure 9**). Further analyses revealed, as expected, a strong positive correlation (*R* = 0.95, *p* = 3.2e-10) between the mean polymerase read length and the fraction of reads that match > Q20 (**Figure 9**).

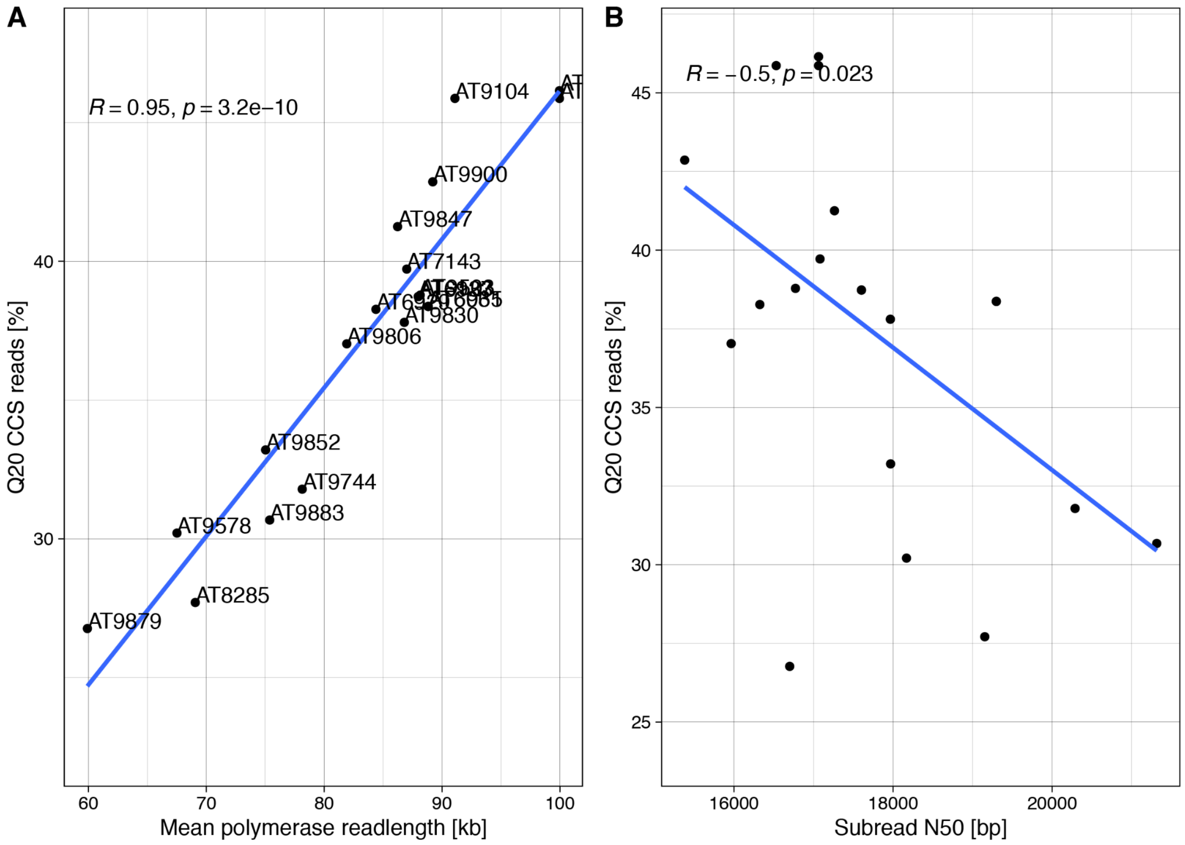

Figure 9: Dependency between the fraction of Q20 CCS reads and two sequencing quality metrics. (A) Longer mean polymerase read length is positively correlated (*R* = 0.95, *p* = 3.2e^{-10}) with greater read fraction passing Q20 filtering. (B) A negative correlation (*R* = -0.5, *p* = 0.023) between subread N50 and percent of reads with >Q20 shows that longer insert size results in fewer reads with >Q20.

4.1.3 Assembly method optimization

To perform long-read sequencing and subsequent assembly in an optimized way, I first tested different parameters and assembly strategies. The main measurement for assessing the quality of *de novo* genome assembly is contiguity. At the beginning I wanted to address the question of how many *A. thaliana* genomes can be sequenced in one run without compromising assembly quality at the end. This matters since each PacBio HiFi run is cost intense. Thus, pooling (multiplexing) several samples into a single sequencing run could reduce future sequencing costs. As mentioned in chapter '4.1.2 Sequencing and CCS calling', there is a strong correlation between the number of full passes and final read accuracy. Therefore, it is tempting to use shorter DNA inserts in order to increase the accuracy of the final CCS reads. However, this is a trade-off between increasing read quality and losing the advantages of long-read sequencing. Moreover, I examined what is the optimum quality threshold for filtering the CCS reads. Filtering the initial raw reads reduces the total number of reads. Thus, I wanted to avoid unnecessary filtering and waste of data. After having assessed these prerequisites, I evaluated the differences in assembly contiguity between state-of-the-art assemblers. Moreover, the impact of an additional correction step after *de novo* assembly was tested.

For the assembly method optimization it is important to note that investigation of multiplexing and optimum DNA insert sizes are meant to enhance the performance of future sequencing runs. In contrast to this, insights from the optimization of quality filtering, the selection of the better assembler as well as the testing of a contig correction tool were directly used here in order to improve the results of this research project. Since many of the following analyses are computationally expensive, I performed method optimization on a subset of five of the eighteen accessions. Only the comparison of two different assembly tools as well as the contig correction were performed on the full set of accessions.

Effect of coverage on assembly contiguity

Although sequencing costs have decreased over the past years, PacBio long-read sequencing remains costly. The latest PacBio sequencing platform, Sequel II, allows multiplexing of up to eight samples in a single run. Thus, reducing the sequencing costs per sample. However, combining multiple samples in the same sequencing run results in lower coverage per sample. Thus, I wanted to investigate the minimum sequencing coverage necessary for obtaining an *A. thaliana de novo* assembly without compromising on contiguity. Five accessions (AT9900, AT9847, AT9830, AT9104 and AT9503) were chosen for this analysis, as all of them have been sequenced at nominal genome coverages exceeding 150x. Q20 filtered CCS reads were subjected to a downsampling approach in which random reads were drawn from the full set in order to reduce the coverage to 150x, 100x, 75x, 70x, 60x, 50x, 40x, 30x, 25x, 20x, and 10x. Subset read files were used to perform *de novo* genome assembly using Falcon-Unzip2 (Chin et al. 2016) as described above. Afterwards, Quast (Gurevich et al. 2013) was used to measure basic quality metrics. The NG50 value describes that contigs of length ≥ NG50 cover 50% of the length of the reference genome. In contrast to that N50 is the length where contigs ≥ N50 produce half of the assembly. Thus, the NG50 value is normalized to the genome size of the target organism. **Figure 10** shows that in three out of five assemblies, the NG50 value did not change when varying the coverage to levels beyond 75x. In the case of AT9900, the NG50 was further improved when using 150x instead of 100x sequencing coverage. However, increasing the sequencing coverage of AT9830 caused the NG50 value to decrease again beyond 100x coverage. The contiguity of all assemblies decreased when genome coverage was below 40x. Collectively, these findings suggest that up to three *A. thaliana* samples can be pooled in a single sequencing run without compromising on assembly contiguity. It is noteworthy that this analysis suffers from the fact that read subsetting is purely based on coverage. Since reads were randomly drawn from the initial read file, some of the differences observed here can be caused by the fact that the read input sets do not only differ in coverage but also in subread N50.

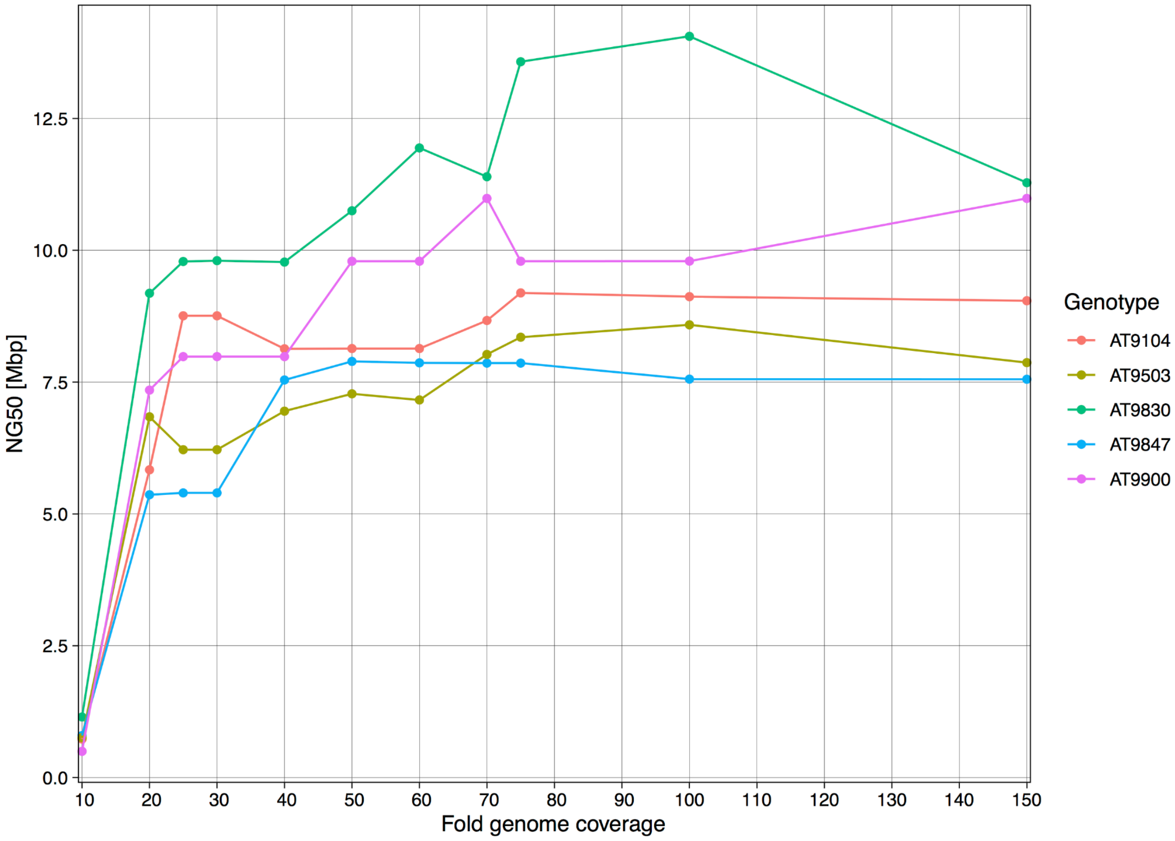

Figure 10: Effect of sequencing coverage on assembly contiguity. Five read sets were subsetted to different fold genome coverages. *De novo* genome assemblies were generated based on the subsetted read files. Subsequent analysis of genome contiguity as measured in NG50 revealed that coverages beyond 75 fold did not improve assembly contiguity any further.

Impact of different input read length distributions on assembly contiguity

Results from the above coverage downsampling showed that sequencing coverages higher than 40x do not necessarily improve assembly contiguity. Therefore, I compared the effect of different input read length distributions. The initial read files of AT9900, AT9104, AT9503, AT9847, and AT9830 were subset to read N25, read N50, and read N75. For example in the case of N75 this means that all reads were sorted by length in a descending manner. Starting from the longest read I then took all reads until 75% of the bases were included in this set. Thus, the average read length of N75 reads is shorter compared to N50 and N25. However, the total number of reads increases from N25 over N50 to N75. Thus, this experiment aims at finding the sweet spot of read length vs sequencing depth. I then performed *de novo* genome assembly and subsequent contig NG50 calculation on each of the datasets. **Figure 11** shows that filtering the input read files for longer reads did not improve the assembly. In all five datasets, the most contiguous assembly was obtained when using the full dataset rather than filtering out shorter reads prior to genome assembly. In the case of AT9830, the overall assembly contiguity did not improve when using the full dataset compared to only using N25, N50 or N75-filtered reads. Instead very similar NG50 values were obtained when using the different read subsets for assembly.

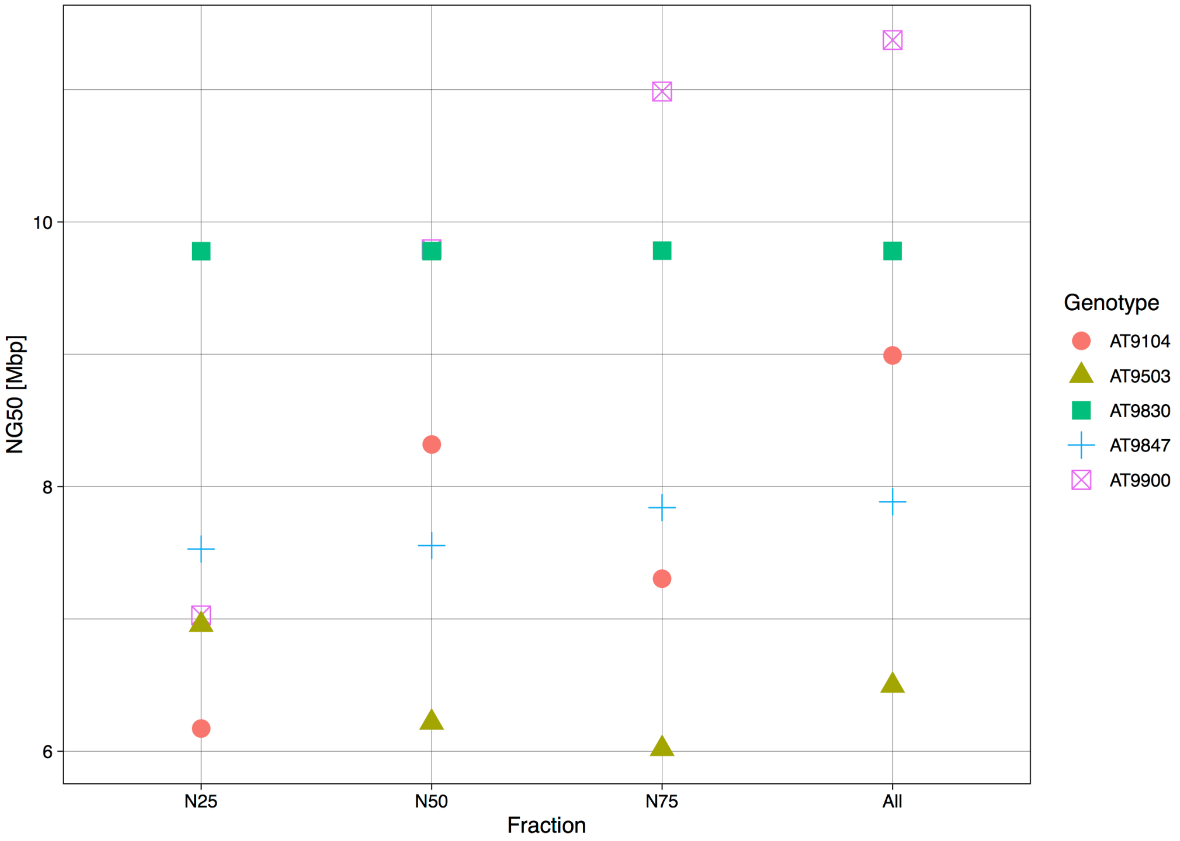

Figure 11: Finding the compromise between read length and sequencing depth. Reads were sorted according to their length. Subsequently, the longest reads representing 25 % (N25), 50 % (N50), and 75 % (N75) of the full set were used for genome assembly. Thus, the average length of reads in the N25 sets is higher compared to the other sets.

Effect of different CCS quality values on assembly statistics

Genome coverage from PacBio HiFi sequencing data not only depends on the amount of obtained sequencing reads, but also on how many CCS reads pass the Q20-filtering criterion. Filtering the initial raw read files for a given Q value represents a trade-off between keeping a larger fraction of erroneous reads vs. only keeping a very small fraction of high quality reads. I investigated how the quality threshold influences the fraction of reads that are kept for downstream assembly and how this affects basic quality metrics of the final genome assembly. Therefore, I performed CCS calling using Q10, Q20, and Q30 as filtering thresholds. **Figure 12** shows the percentage of CCS reads matching a given Q value. In all but one dataset, the fraction of reads that were kept after filtering became smaller with an increasing Q value. The highest fraction of CCS reads was retained when filtering with a Q10 threshold. In the case of AT9900, almost 50% of the initial reads were kept when filtering with

Q10, but this proportion decreased to ~20% when filtering with Q30. These differences were even clearer for AT9847, where an increase from Q10 to Q30 led to a reduction in CCS reads from >45% to < 20%.

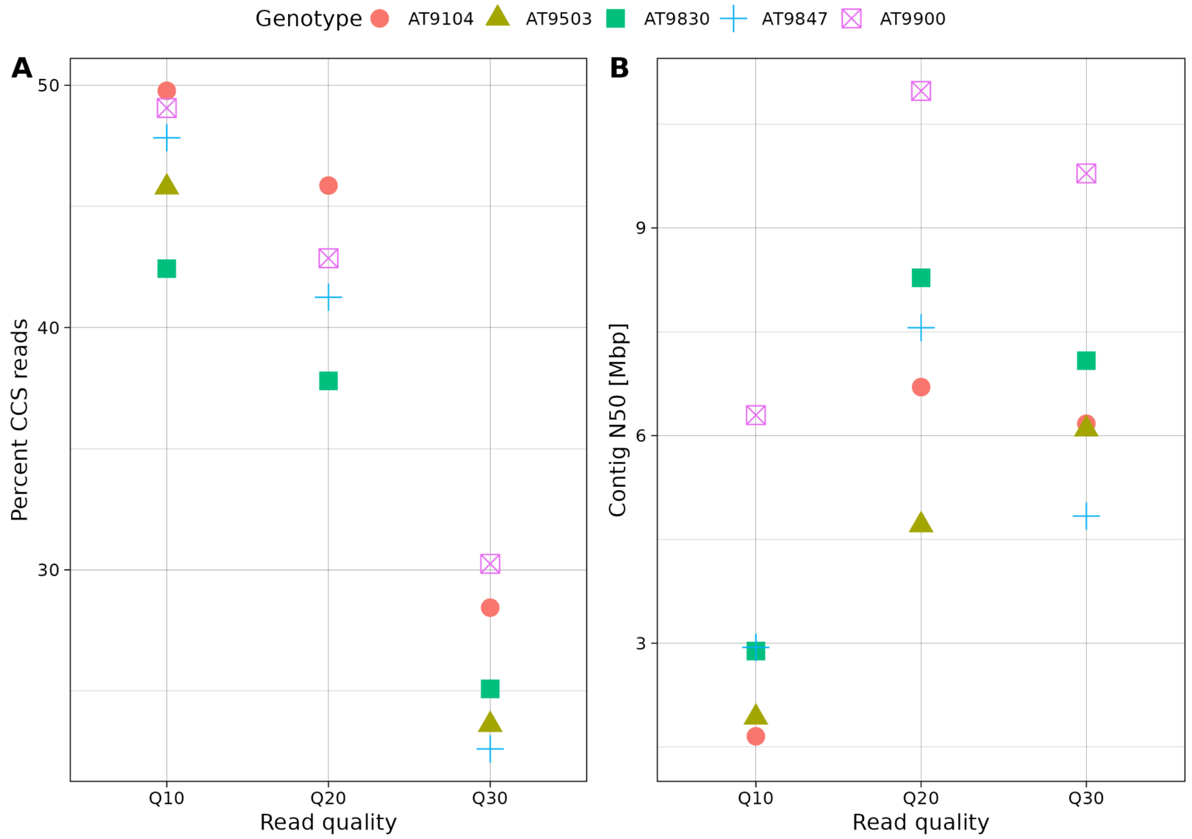

Figure 12: Impact of read quality filtering on fraction of kept reads (A) and on assembly contiguity (B). (A) Using a higher quality threshold for read filtering resulted in a lower percentage of reads that were kept. (B) Negative effects on assembly contiguity when using either reads with Q10 or with Q30. However, in case of reads with Q30 this can be attributable to a lower sequencing coverage.

After CCS calling, all filtered read files were used as input data for *de novo* genome assembly using Falcon-Unzip2 (Chin et al. 2016). In all cases except for AT9503, the highest assembly contiguity was obtained when filtering for Q20 CCS reads (**Figure 12**). Notably, Q30-filtered CCS reads generated more contiguous assemblies compared to Q10-filtered reads, confirming that not only coverage but also read quality impacts the N50 of *de novo* assemblies. Based on these results, all HiFi read sets were filtered for Q20 CCS reads prior to the *de novo* genome assembly.

Manual correction of chimeric contigs in Falcon-Unzip2 assemblies

To identify potential translocations relative to the Col-0 TAIR10 reference genome, or to correct mis-assemblies during the above-described processes, I aligned primary contigs assembled with Falcon-Unzip2 against the *A. thaliana* reference genome TAIR10. In the case of AT8285, a single contig (000000F) partially aligned to the reference chromosomes one and three (**Figure 13**).

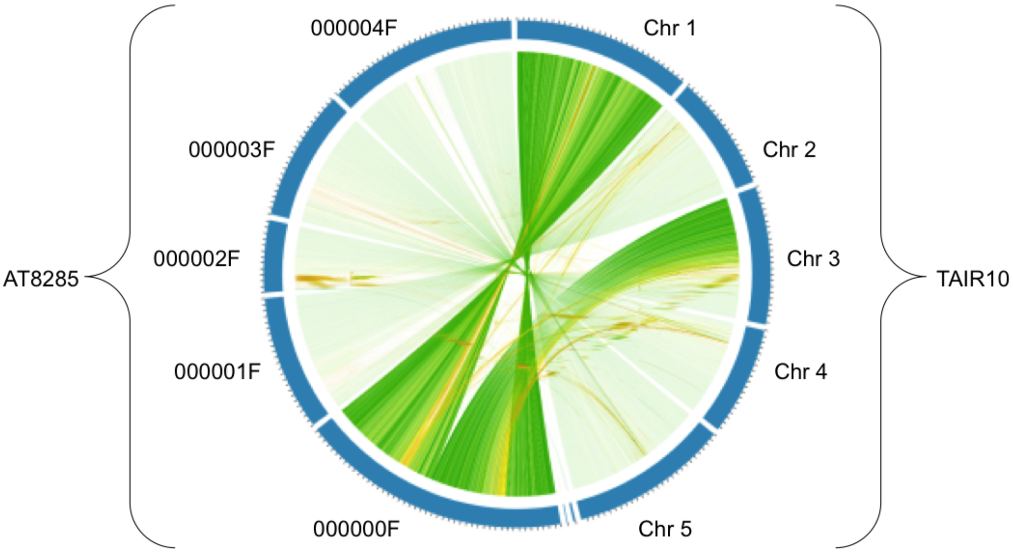

Figure 13: Alignment between the five longest primary Falcon-Unzip2 contigs of AT8285 and the *A. thaliana* reference TAIR10. A single contig (000000F) of AT8285 partially aligns to chromosomes one and three of TAIR10

In order to check this potential translocation, I mapped the Q20-filtered CCS reads against the primary assembly of AT8285. Visualization of the resulting alignments using the Integrated Genome Browser (IGV) revealed that only a single read of 19.492 bp connects both contigs (**Figure 14**). In addition, the read introduced a 124 bp insertion between both contigs. Furthermore, the reads from the 'right side' of the potentially misjoined contigs are soft clipped at the position where the two contigs are joined. The clipped part of those reads maps to TAIR10 chromosome three. Softclipped reads have a mapping quality of zero and are therefore indicated as transparent arrows in **figure 14**. The above findings show that the contig 000000F of the assembly of AT8285 is almost certainly an artifact of the assembly, rather than a chromosomal translocation. Taking into account these results, I decided to manually 'break' and correct this contig at position 14,807,291 bp. Afterwards, the manually corrected contig was scaffolded together with the remaining contigs using RagTag scaffold (Alonge et al. 2022). A similar phenomenon was observed in AT6137 where a single contig (000000F) aligned to chromosomes one and five in the reference genome TAIR10. Subsequent read mapping again revealed that the chimeric contig was generated by connecting two contigs with a single read spanning the junction. Therefore, the corresponding contig 000000F was manually corrected at position 8,934,255 bp.

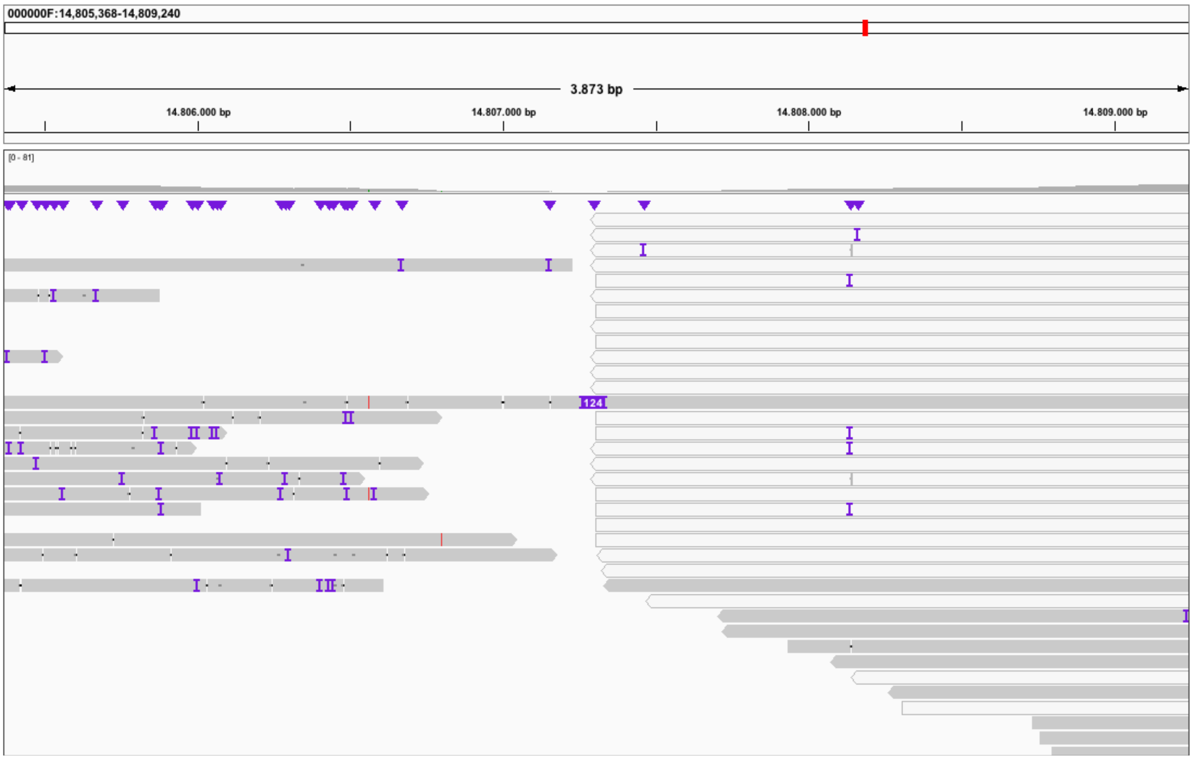

Figure 14: Visualization of Q20 reads mapped to the chimeric contig 000000F. Only a single read spans the junction between the two contigs.

Comparison of assembly tools: Falcon-unzip2 vs Hifiasm

New tools with the ability to assemble genomes from PacBio HiFi read data emerged during the course of this work. Although most of the previously described assembly method optimization was done using Falcon-unzip2 (Chin et al. 2016), which was the only available assembler when the project was initiated, I decided to test Hifiasm, a more recent *de novo* assembler (H. Cheng et al. 2021). Thus, genome assemblies for all eighteen accessions were performed using Falcon-unzip2 and Hifiasm in order to compare their performance on the datasets presented in this study. Subsequently, N50, NG50, and the total number of contigs as a basic quality metric for assembly contiguity were determined using Quast (Gurevich et al. 2013). Additionally, I compared the total length of the resulting genome assemblies between the two assembly tools. Contig N50 in Falcon-unzip2 derived assemblies varied between 4.71 Mb (AT9503) and 15.38 Mb (AT9852), with an average of 9.07 Mb (**Figure 15**). The average contig N50 obtained when using Hifiasm was only slightly lower, at 8.85 Mb. However, Hifiasm performed better in datasets where Falcon-unzip2 generated more fragmented assemblies and vice versa. The contig N50 of AT9879 achieved with Falcon-unzip2 was 5.18 Mb, compared to 11.93 Mb when using Hifiasm. In the case of AT9900, the opposite trend was true. The Hifiasm assembly was more fragmented (6.96 Mb) compared to the Falcon-unzip2 *de novo* assembly (10.98 Mb). Twelve out of eighteen assemblies showed higher contig N50 values when assembled with Falcon-unzip2.

However, comparing the N50 of primary contigs between the two assemblers did not reveal any clear trend (**Figure 15**).

The number of contigs generated by each of the assemblers from the eighteen datasets is shown in **figure 15B**. When using Falcon-unzip2, the number of primary contigs varied between 71 (AT9762) and 1426 (AT9503). The average number of primary contigs in Falcon-unzip2-derived assemblies was 324. The total number of primary contigs in Hifiasm assemblies ranged from 182 (AT9879) to 2410 (AT9900), with an average of 1111 contigs per assembly. In all cases, a Hifiasm-derived assembly had more contigs compared to its corresponding Falcon-unzip2-derived assembly.

The comparison of contig N50 may be biased by differences in the overall assembly length, as well as by the total number of contigs (**Figure 15**). As described above, Hifiasm assemblies invariably had more contigs compared to Falcon-unzip2 assemblies. Therefore, I chose to also compare the NG50 of primary contigs, since this value describes the contiguity of an assembly normalized to a given reference (Gurevich et al. 2013). Hence, the NG50 value is less prone to distortion by the number of primary contigs or by the total assembly length. In this case, the contig NG50 was calculated relative to the *A. thaliana* reference genome TAIR10, and as such it describes the minimum length of contigs that cover half of the reference genome. In the case of Falcon-unzip2, the calculated contig NG50 values ranged from 6.5 Mb (AT9503) to 15.38 Mb (AT9852). The average contig NG50 of Falcon-unzip2 derived genome assemblies was 10.29 Mb. Conversely, the NG50 values in Hifiasm-generated assemblies varied from 7.72 Mb (AT9847) to 19.7 Mb (AT8285), with an average of 11.92 Mb. Based on these results I decided to use Hifiasm for generating the final genome assemblies.

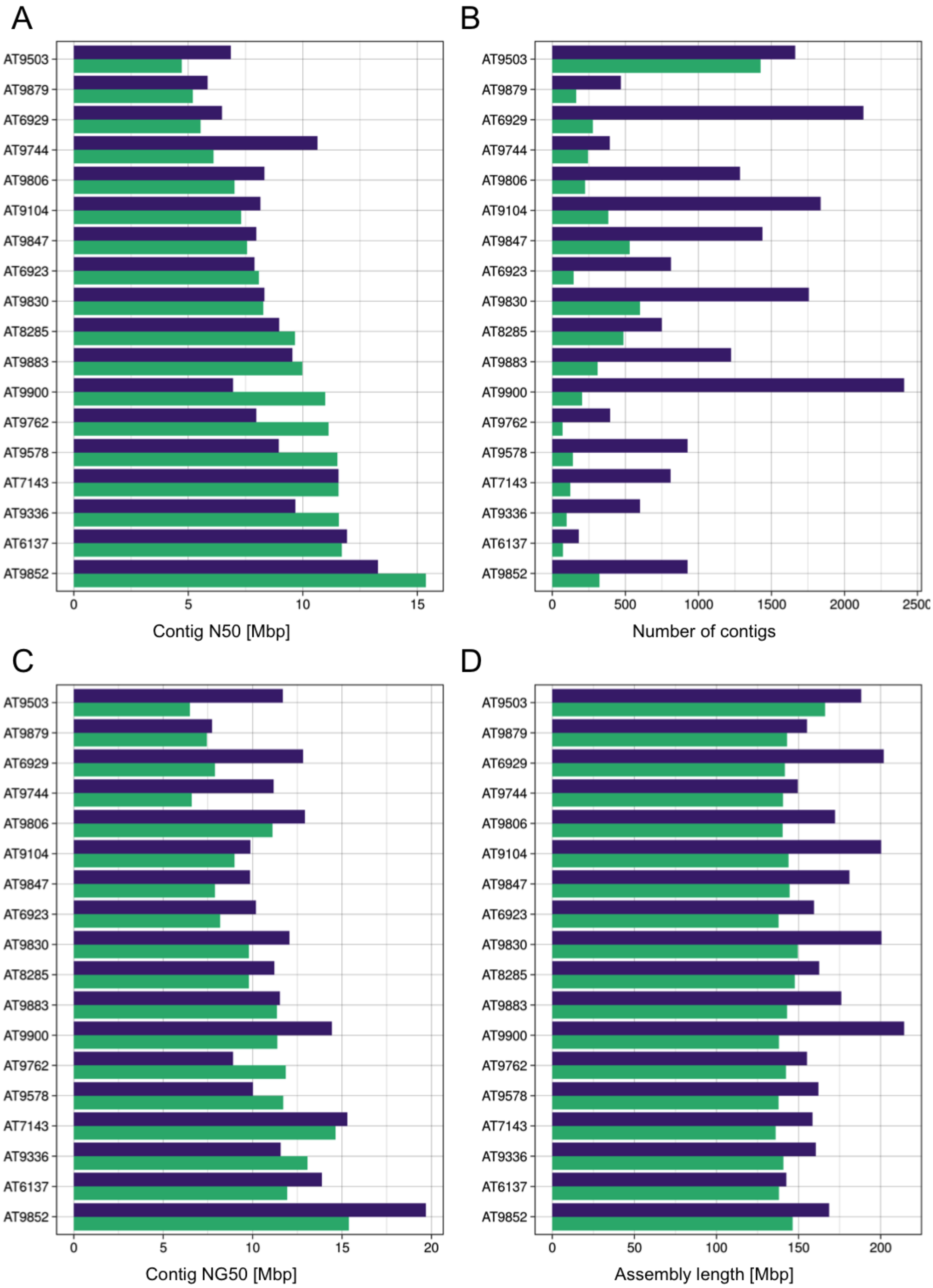

Figure 15: Comparison of *de novo* genome assembly quality metrics between Hifiasm (purple bars) and Falcon-unzip2 (green bars). Comparison of (A) contig N50, (B) number of primary contigs, (C) contig NG50, and (D) total assembly length.

Evaluation of contig correction

Large structural mis-assemblies are often found in draft genome assemblies. Thus, depending on the sequencing method it may be necessary to correct such errors by comparing the *de novo* assembly to a reference genome. However, it is crucial to distinguish potential mis-assemblies from true biological variation between the reference genome and the query assembly (Rhie et al. 2021). Putative mis-assemblies are further evaluated by mapping Q20 CCS reads to the corresponding region. Contigs are 'broken' if no read spans the region of the potential mis-assembly. RagTag Correct (Alonge et al. 2021) was tested in order to identify and confirm such mis-assemblies.

The correction of such errors is a trade-off between overcorrecting the input assembly while making it more similar to the reference, and the carry-over of real mis-assemblies. Thus, it is necessary to evaluate the correction step as well as the impact of using different references for correction. Therefore, an optical map of AT9852 as well as the reference genome TAIR10 were used as references for the correction of all *de novo* genome assemblies. The optical map of AT9852 was generated by the service provider Bionano Genomics (Bionano Genomics, San Diego, CA) following their in house protocols. In short, the Bionano Saphyr instrument was used in combination with the Direct Label and Stain (DLS) kit (Bionano Genomics Catalog 80005) in order to generate chromosome-level scaffolds. Optical mapping works by labeling high molecular weight DNA with a fluorescent dye that binds to a specific sequence motif. Subsequently, the linearized and labeled DNA molecules are passed through a channel with a detector. Thereby it is possible to determine the distance between the fluorescently labeled sequence motifs. Since the sequence motifs that the dye binds to are known one can further use this information in order to align an assembled genome against the optical map and thereby ordering the contigs (Lam et al. 2012; Chan et al. 2018). Such optical data are powerful since they can improve assembly contiguity (Michael et al. 2018; Belser et al. 2018). The number of contig breaks introduced by correcting with either TAIR10 or the Bionano map was counted afterwards. When correcting the draft assembly of AT9852 with Bionano data originating from the same accession, a total of two contig breaks were introduced, but the same assembly was corrected at 13 different positions when using TAIR10 as a reference (**Table 1**). In all but one assembly (AT9830), I observed that the number of corrected contigs differed between the two references. Using TAIR10 for scaffolding introduced more contig breaks in 14 out of 18 assemblies. For further evaluation of the two different reference genomes, I compared, firstly, which contigs were corrected and how often, and secondly, at which positions they were corrected. **Table 1** shows in how many cases both references corrected the same contig and in how many cases the same contig was corrected at the exact same position. Usage of the different references led not only to the correction of a different number of contigs but also to the correction of different sets of contigs (**Table 1**). In all but one assembly (AT9852), the set of corrected contigs partially overlapped between the two references. However, comparing the exact positions that were corrected revealed that using either TAIR10 or the optical map for contig correction introduced breaks, not only at the same positions but also at different positions within the same contig. Taking these findings into consideration, I decided not to use the Ragtag Correct module, and proceeded directly to scaffolding of the contigs.

Table 1: Comparing how many contig breaks were introduced at which contigs and at which position by RagTag correct when using either the Bionano data from AT9852 or TAIR10 as a reference.

Genotype	Bionano	TAIR10	Corrections on same contig	Corrections at same position
AT6137	6	6	2	2
AT6923	9	11	6	1
AT6929	21	25	17	9
AT7143	10	8	4	1
AT8285	11	14	6	8
AT9104	13	15	9	3
AT9336	10	11	4	3
AT9503	17	16	10	6
AT9578	10	11	5	2
AT9744	16	18	9	4
AT9762	9	13	7	3
AT9806	11	15	9	4
AT9830	23	20	14	11
AT9847	12	12	7	4
AT9852	2	8	0	0
AT9879	14	17	9	7
AT9883	11	14	5	4
AT9900	6	14	6	2

4.1.4 Quality control

Sequencing data can be negatively impacted by various artifacts occurring during DNA extraction, library preparation, and sequencing (Trivedi et al. 2014). Therefore, after having optimized the assembly strategy, it is important to perform basic quality control before further analysis. Thus, in this section I am presenting the results obtained from quality control of the sequencing data.

Identification of heterozygous samples

As a selfing species, most inbred *A. thaliana* accessions should have very few, if any, heterozygous sites in the genome. Settings of downstream analysis tools (such as the assembly pipeline) depend on prior assumptions regarding an organism's genome, such as expected genome size or outcrossing vs. inbreeding. Hence, it was necessary to ensure that all accessions chosen for this study were inbred. Moreover, it is of great importance to confirm that the sequenced pool of plants did not comprise more than one genotype.

Kmer counting and frequency analyses were performed in order to assess heterozygosity level and potential contamination with DNA from other accessions. A kmer is a DNA sequence unit of length k. In inbred organisms such as *A. thaliana*, it is expected that the vast majority of kmers is observed at the depth that the sample has been sequenced at. If the sequenced organism was diploid and heterozygous, it is expected that many kmers of lower frequency are observed at exactly half the sequencing coverage.

All Q20 filtered CCS reads were ‘cut’ into 21 nucleotide-long sequences (kmers) before counting and plotting their frequency spectra against their coverage. Kmer frequency spectra for all twenty accessions are shown in appendix **figure S1**. From comparing kmer frequency plots of all twenty accessions, it is evident that AT6961 has an unexpectedly high level of heterozygosity, resulting in two distinct peaks, with the main peak at the sequencing coverage and the second at exactly half the coverage of the main peak (**Figure 16**). This becomes more evident when compared to kmer spectra of the remaining accessions such as the example of AT9104 in **figure 16**. Moreover, the estimated heterozygosity of AT6961 ranges from 0.3% to 0.31% (Appendix: **Table S2**). However, the estimated heterozygosity of all other samples varied between 0.01% and 0.04%, with an average minimum heterozygosity of 0.02% and an average maximum heterozygosity of 0.03%. Thus, AT6961 is on average 10-15 times more heterozygous compared to all other accessions analyzed in this study. Hence, the AT6961 sample was likely not fully inbred, or may even have been contaminated with DNA from a different *A. thaliana* genotype. Therefore, I decided to remove the AT6961 dataset from all downstream analyses.

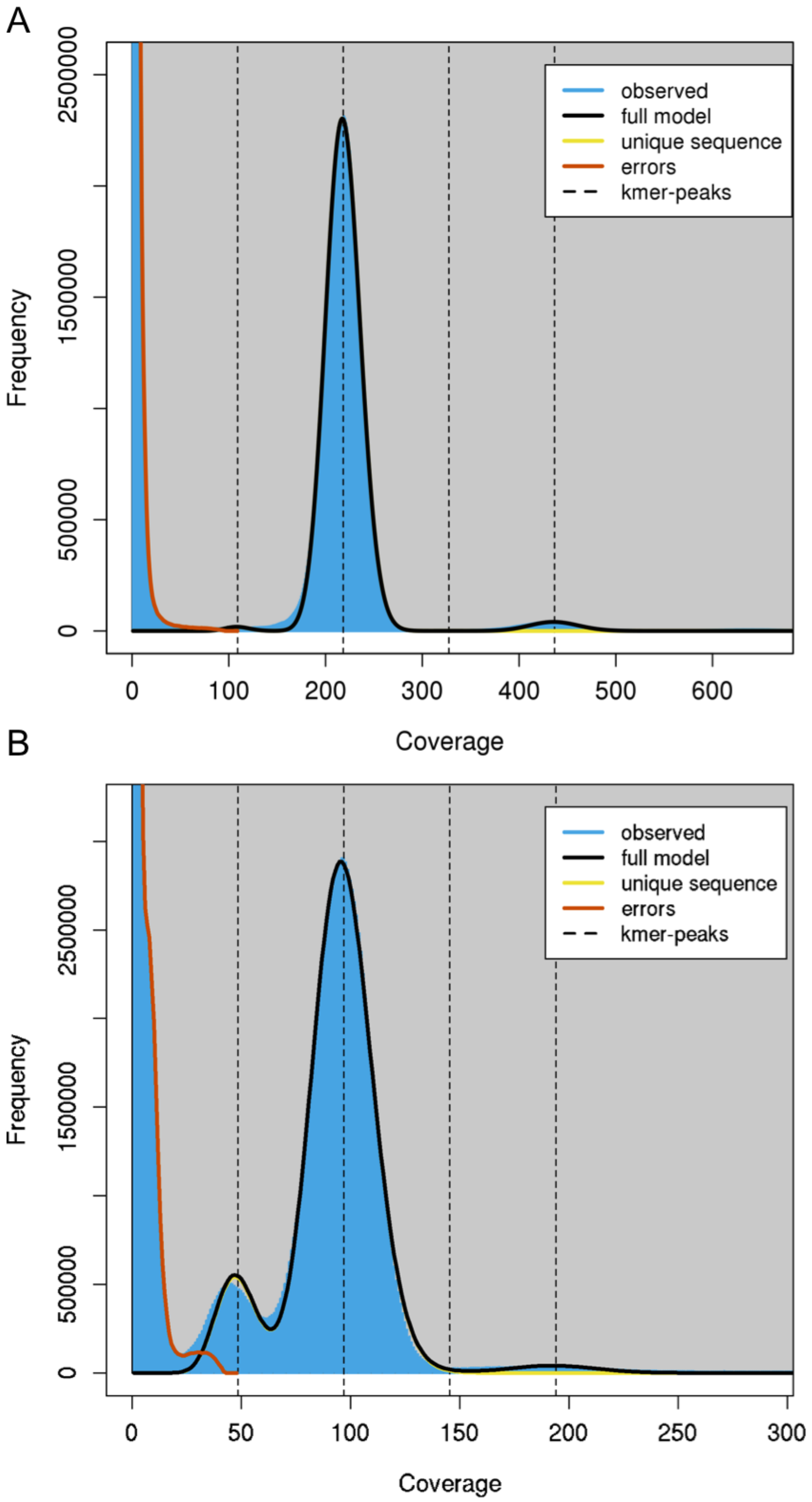

Figure 16: Kmer frequency analyses of AT9104 (A) and AT6961 (B). Due to the homozygous nature of the *A. thaliana* genome it is expected that the vast majority of kmers is observed at the approximate sequencing coverage of the sample. Kmers originating from heterozygous alleles will be observed with half the coverage of the sample. (A) Kmer frequency spectrum of AT9104 showing a single peak with an estimated heterozygosity level of < 0.013 %. (B) Kmer frequency spectrum of AT6961 shows two distinct peaks. The level of heterozygosity was estimated to be 0.304 %.

Removal of bacterial contamination

Based on the optimization steps described in detail above, I proceeded to generate *de novo* genome assemblies for the nineteen accessions that did not appear to be heterozygous or problematic in other ways using Hifiasm (H. Cheng et al. 2021). Primary contigs were first screened for potential bacterial, viral or fungal contamination by aligning them against a metagenomic database and visualizing the results using Blobtools (Laetsch and Blaxter 2017). Moreover, I mapped all Q20 filtered CCS reads against the original, primary contigs to detect contigs with abnormal mapping coverage. Ten out of nineteen assemblies were contaminated with bacterial sequences (**Figure 17**). The total length of contaminated sequences varied from 16.2 kb in AT9852 to 34.78 Mb in AT9104. Most of the bacterial sequences that could be classified originated from *Pseudomonas* as shown for AT9900 in **figure 18**. However, even in this case, the vast majority of input sequences grouped closely together and matched the expected sequencing coverage as well as the GC content. These reads were identified as plant derived (**Figure 18**). Contigs that aligned to bacterial, fungal or viral sequences were removed from each dataset and excluded from all downstream analyses.

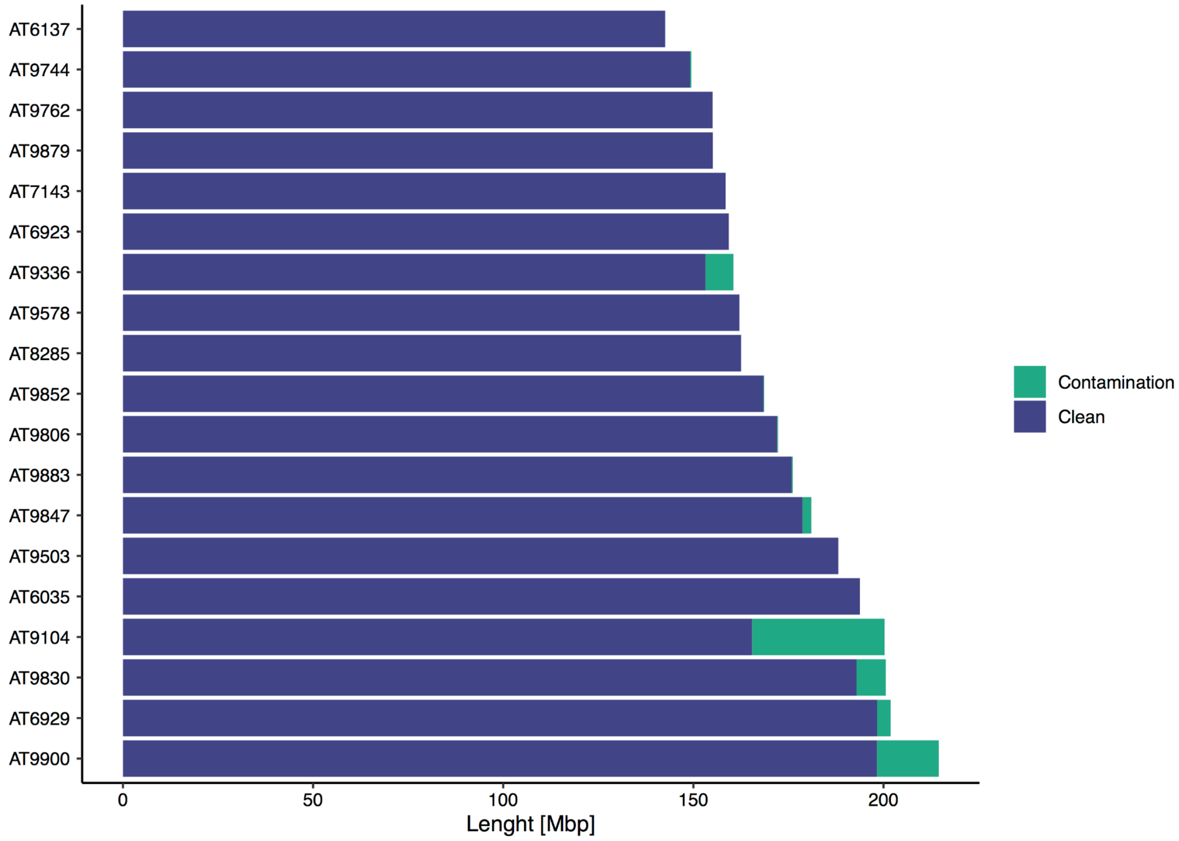

Figure 17: Cumulative length of target (clean) and contaminated sequence per assembly. Ten out of nineteen assemblies were contaminated with non-plant contigs. The contaminated sequence length varied from 16 kb to 34 Mb.

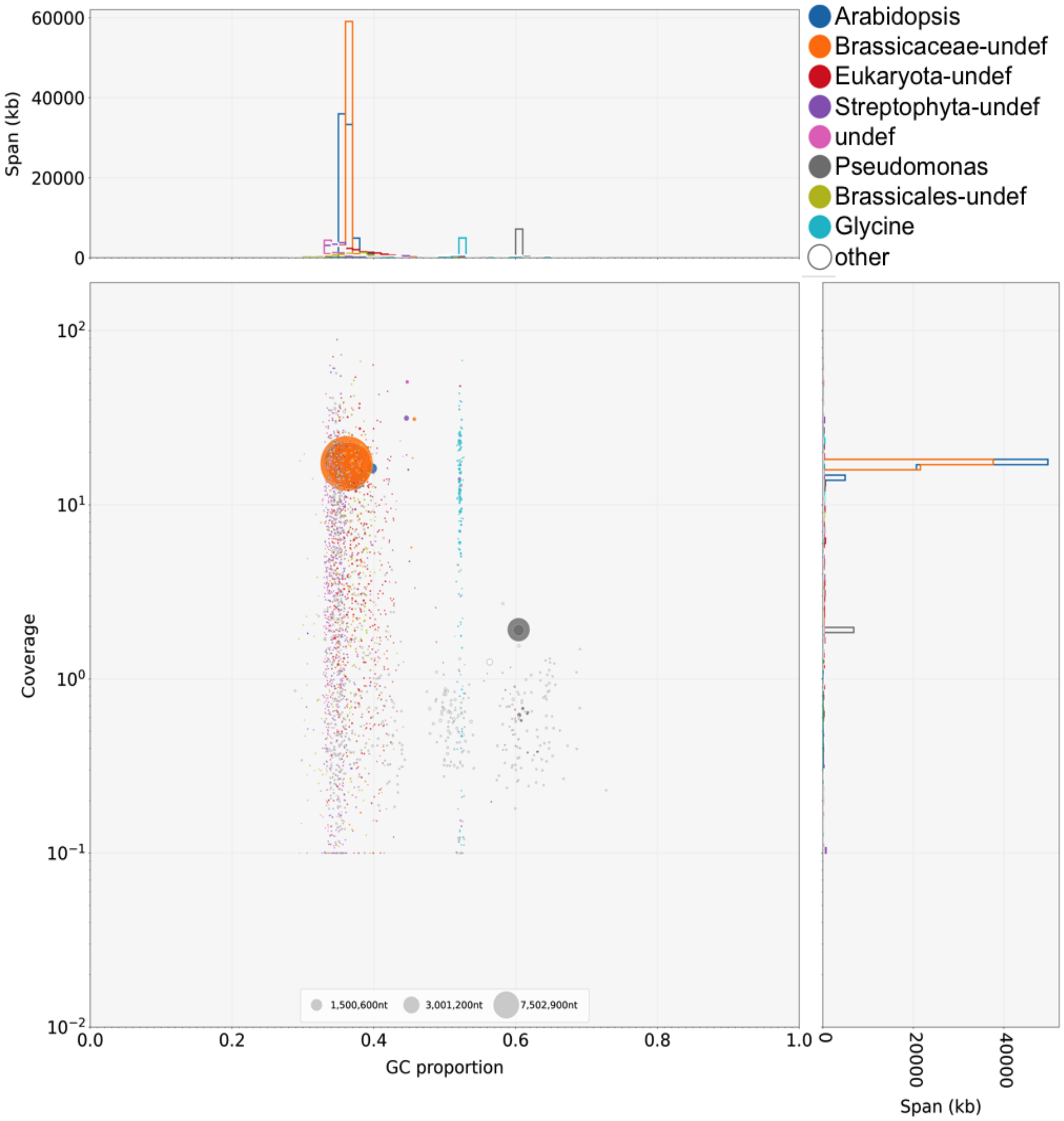

Figure 18: Alignment of primary contigs from AT9900 against a metagenomic database and subsequent separation by GC content and coverage. The vast majority of sequences grouped closely together and could be assigned to 'Arabidopsis' or 'Brassicaceae-undef.' However, 7.85 Mb of the AT9900 primary assembly were classified as 'Pseudomonas'. Moreover, these contaminated contigs showed a very distinct GC proportion and coverage.

4.1.5 Genome assembly

In this chapter I will present the results that were obtained from *de novo* assembly and quality control of the nineteen accessions using the optimized assembly strategy described before. This chapter starts with the assessment of basic quality metrics such as contiguity and completeness of the eighteen assemblies based on Hifiasm. Moreover, I will present results justifying the removal of another accession from the dataset leaving me with a total of eighteen accessions. Subsequently, I will show that in multiple genomes individual chromosomes were assembled telomere-to-telomere and that it was possible to reconstruct highly repetitive centromeric regions. Furthermore, I characterized contigs that remained unplaced after scaffolding the assemblies to chromosome level.

Assembly statistics

In the section 'Assembly completeness' I will show why one more dataset had to be removed from downstream analysis. Even though I removed that dataset after assessing the quality metrics described in this section, I will not show these results for the removed dataset. Median and average quality metrics would be biased if the outlier sample was included. Thus, in this section I am only going to present quality metrics for the eighteen accessions that were of sufficient quality for all downstream analyses. Quality metrics describing the contiguity and length of the assemblies were assessed after contamination and outlier sample removal. The total number of primary contigs varied from 182 (AT6137) to 2410 (AT9900). The median number of primary contigs was 925.5.

Contig N50, describing the minimum length of contigs that represent > 50% of the sequenced bases, was calculated in order to assess assembly contiguity. The lowest contiguity was found for AT9879, with a contig N50 of 5,848,806 bp. In contrast, the most contiguous assembly, AT9852, has a contig N50 of 13,274,449 bp. Four out of the eighteen genomes had N50 values > 10 Mb. The mean contig N50 of all assemblies was 8,847,880 bp.

Subsequently, I aligned all contigs against TAIR10 and compared the median length of contigs that cover 50% of the reference sequence. This is summarized in the NG50 value. Thus, the NG50 represents the N50 value normalized to the reference genome. The most fragmented assembly (AT9879) had an NG50 of 7,730,948 bp, whereas in the most contiguous assembly (AT9852), 50% of the TAIR10 reference sequence are covered by contigs of the size ≥ 19,696,172 bp. Notably, it was possible to assemble fourteen out of eighteen genomes with NG50 values exceeding 10 Mb.

Another metric describing contiguity in assemblies is the LG50 value, which indicates how many contigs with length > NG50 can be found in a given assembly, or how many contigs are needed to represent 50% of the given TAIR10 reference sequence. In the case of AT9852 and AT9900, I observed LG50 values of three, meaning that three contigs represent half of the reference genome. This is consistent with these three assemblies having one chromosome-length telomere-to-telomere contig each. The worst LG50 values were still only six (AT1904 & AT9879), and the average was 4.2.

Moreover, I compared the length of the longest contig between the different *de novo* assemblies. The largest contig with a length of > 30 Mb was found in the assembly of AT9900, but all assemblies had at least one contig with a length of >15 Mb. Twelve of the eighteen genome assemblies had one contig > 19 Mb. The median size of the longest contig was 19.84 Mb. Further details on assembly contiguity will be presented in section ‘Scaffolding’.

The overall length of each assembly was determined after having assessed assembly contiguity, contig numbers, and largest contigs. The longest assembly was the one obtained for AT9900, with a total length of 214,520,093 bp, while the shortest assembly, with a size of 142,588,945 bp, was found in the AT6137 dataset. The mean assembly length was 172.7 Mb.

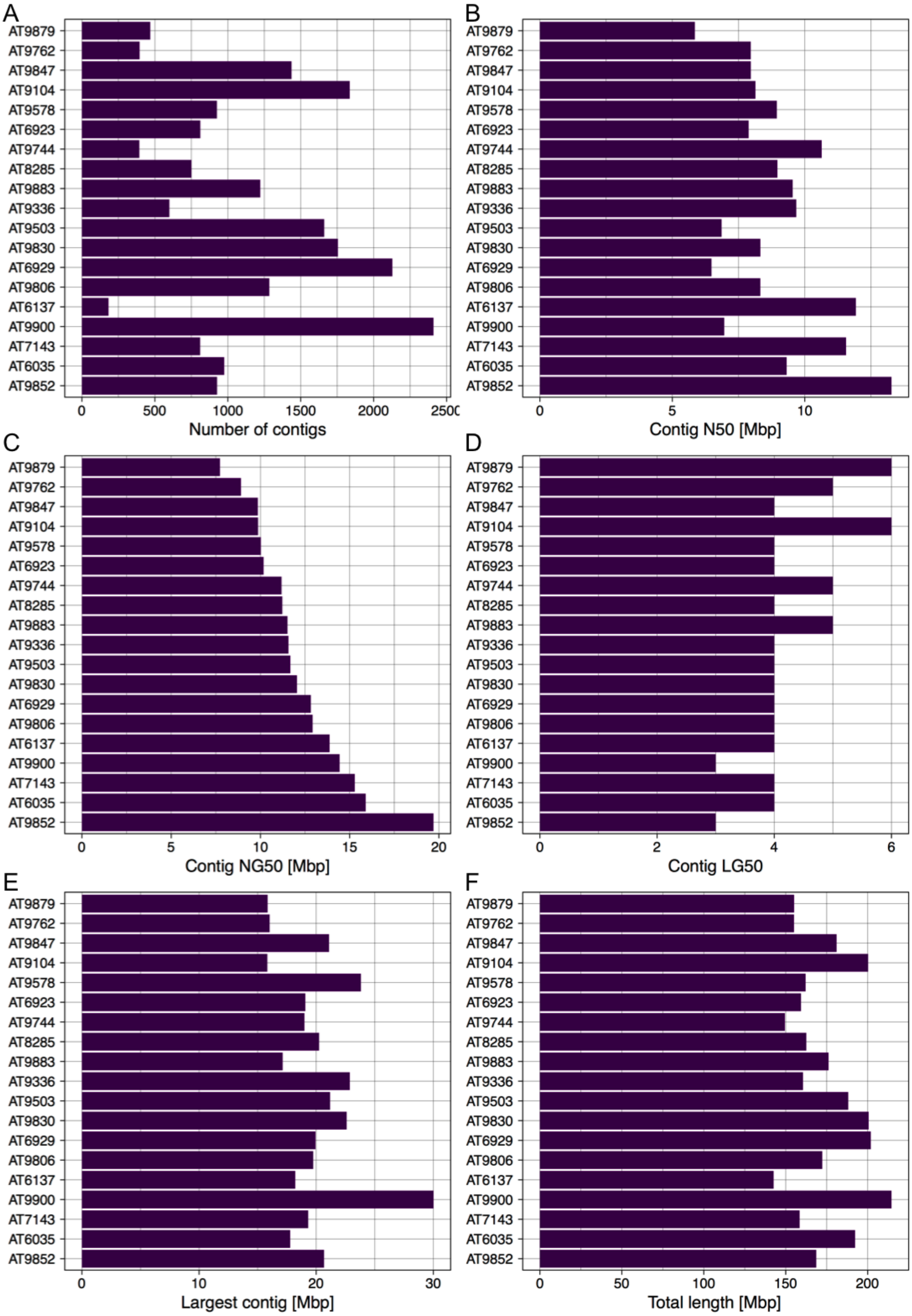

Figure 19: Basic quality metrics of 18 *de novo* genome assemblies. Barplots are showing (A) the total number of contigs, (B) contig N50 (C), contig NG50, (D) contig LG50, (E) length of the largest contig, and (F) the total assembly length.

Assembly completeness

BUSCO (Simão et al. 2015) was used in order to estimate the completeness of each assembly. In this approach, completeness is assessed using an orthologous group of genes that are found as single-copy genes in ~90% of the species in the group. Thus, the presence of these conserved genes can be used as a proxy to infer the overall completeness of the genome assembly (Simão et al. 2015). Assembly completeness is shown in **figure 20**. Completeness levels varied between 99.5%, as observed for the assemblies of AT6929, AT9847, and AT9104, and 99.2%, as observed for the assemblies of AT7143, AT6923, AT9883, and AT9830. The average completeness level was 99.3%. However, the overall completeness is measured as the sum of single-copy and duplicated BUSCO genes. Therefore, I checked how many of the 1614 BUSCO genes are found as single-copy genes and how many are marked as duplicates. On average assemblies had 97.4% single-copy BUSCO genes. However, AT6035 had an exceptionally low-single copy BUSCO completeness level of 76%, with 234% of the 1614 BUSCO genes marked as duplicated. Taking these findings into consideration, I chose to remove the assembly of AT6035 from any downstream analyses. Thus, all values in the rest of this section were calculated without AT6035.

The highest number of complete single-copy BUSCO genes (1595; 98.8%) was found in the assemblies of AT6137, AT6929, and AT9847. The primary contigs of AT9744 and AT9830 showed the lowest number of complete single-copy BUSCO genes (1588; 98.4%). On average, the assemblies showed BUSCO duplication scores of 0.73%. The highest duplication level (0.9%) was now detected in AT9104, AT9762, and AT9744. Duplication rates of 0.6%, as observed for AT6137 and AT7143, were the lowest among the eighteen assemblies. Four assemblies had two fragmented BUSCO genes while the remaining fourteen assemblies had one fragmented BUSCO gene. Of the 1614 BUSCO genes, only 8-10 were not detectable in any of the eighteen assemblies. This means that accessions were rather lacking the same BUSCOs than each missing different BUSCOs. Indeed, six out of these thirteen missing BUSCO genes were absent in all assemblies. Only one BUSCO gene was exclusively missing in one of the assemblies. Thus, all genome assemblies, except for AT6035, were of sufficient completeness for further processing and downstream analysis.

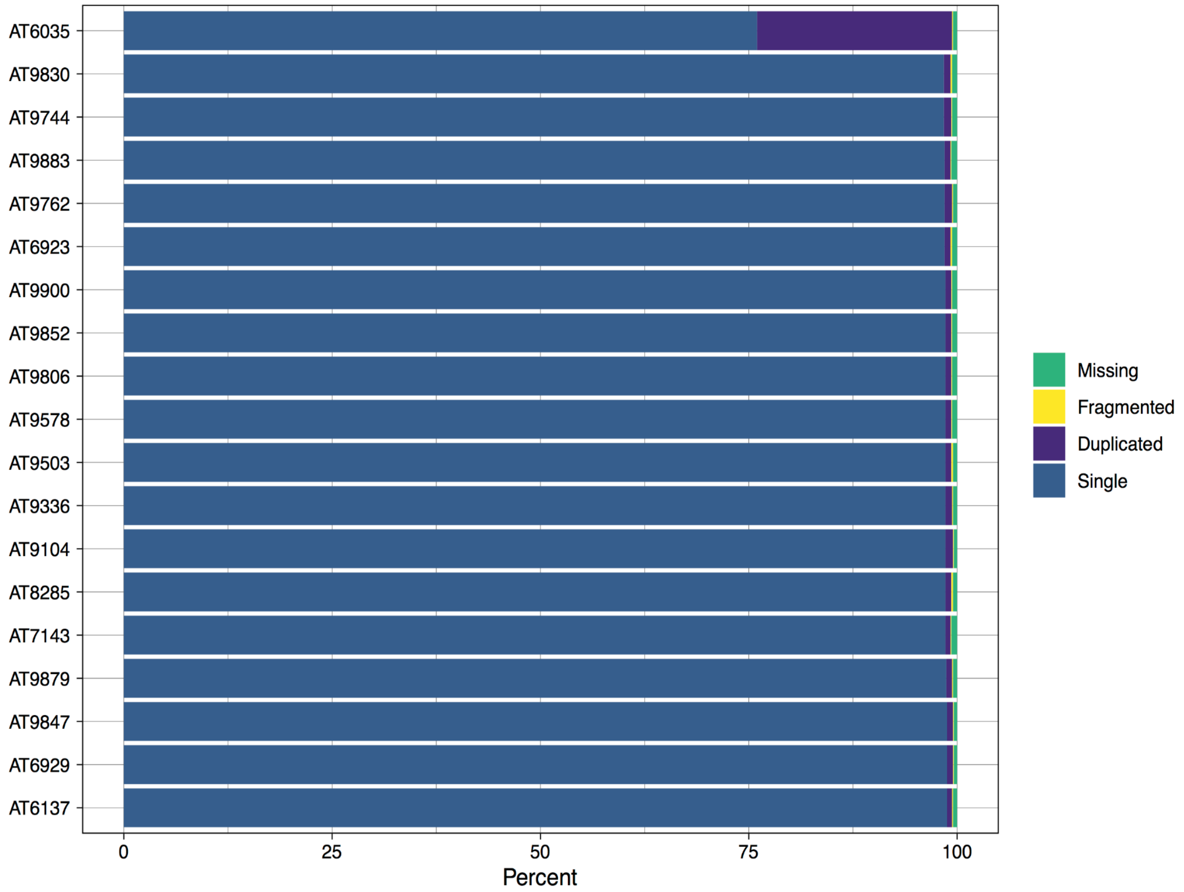

Figure 20: Completeness of 19 *de novo* assemblies before outlier removal. Completeness was inferred by scanning the assemblies for the presence of a set of conserved genes (BUSCOs). AT6035 was removed from downstream analyses due to the high level of duplicated BUSCOs. All other assemblies show completeness levels beyond 99.2 %.

Scaffolding

Contigs of each assembly were further connected into pseudo-chromosomes using RagTag Scaffold (Alonge et al. 2022), with the aforementioned optical map of AT9852 serving as the backbone. Between sixteen (AT7143, AT9830, and AT9852) and 34 (AT9879) contigs per assembly were scaffolded into pseudo-chromosomes (**Table 2**). On average, I could place 21 contigs per assembly, and all but two assemblies could be scaffolded with fewer than 25 contigs. Moreover it was possible to assemble three different chromosomes in three different accessions with just a single contig each. This reflects the high contiguity of the primary assemblies (**Figure 21**).

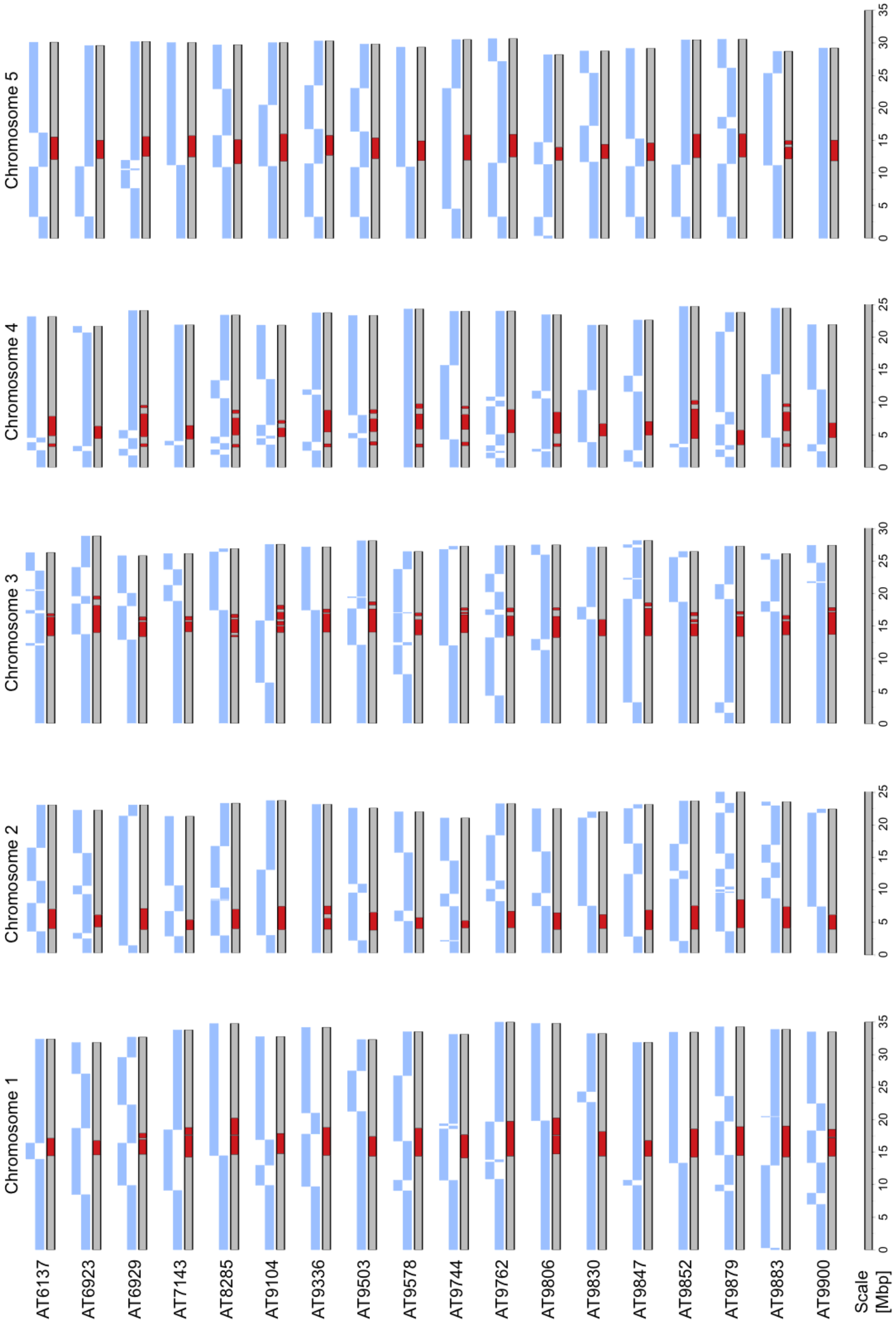

Figure 21: Chromosome-level assemblies of 18 *A. thaliana* genomes. The final scaffold for each chromosome is depicted by the gray bar. Red blocks indicate centromere sequences. Contigs that were placed in order to generate the scaffold are shown as blue bars. Telomere-to-telomere level assemblies were obtained for chromosome 2 of AT9336, chromosome 4 of AT9587, and chromosome 5 of AT9900. Centromeres were often assembled from a single contig while showing variation in size and position among accessions.

Table 2 shows that the N50 value of placed contigs varied from 5.94 Mb (AT9879) to 18.6 Mb (AT9852). In eight of eighteen assemblies, N50 values higher than 10 Mb were observed for placed contigs. The average N50 of placed contigs was 10.71 Mb.

Table 2: Overview the number and N50 of placed contigs per assembly.

Genotype	Number of placed contigs	Placed contig N50 (bp)
AT6137	24	11,929,018
AT6923	22	8,441,029
AT6929	22	12,825,298
AT7143	16	15,287,920
AT8285	23	9,821,434
AT9104	19	9,864,348
AT9336	17	9,714,618
AT9503	21	8,569,625
AT9578	19	9,045,768
AT9744	20	11,178,775
AT9762	30	8,488,419
AT9806	20	11,466,786
AT9830	16	9,782,689
AT9847	24	9,625,350
AT9852	16	18,613,550
AT9879	34	5,951,611
AT9883	21	11,516,797
AT9900	18	10,981,064

Subsequently, I compared the cumulative length of placed sequences, which averaged 135 Mb. However, comparing the length of scaffolded contigs among the assemblies revealed that the cumulative length of placed contigs varied by 8 Mb (**Figure 22**). The shortest scaffold, with 133 Mb, was observed for AT9830, while the longest scaffold (placed sequence) had a length of 141 Mb and was observed in AT9879. In addition, the length of unplaced sequences varied greatly, from 7.8 Mb (AT6137) to over 60 Mb (AT6929 and

AT9900). The median cumulative length of unplaced contigs was 28 Mb. The substantial amount of additional, yet unplaced contigs will be examined later in the thesis.

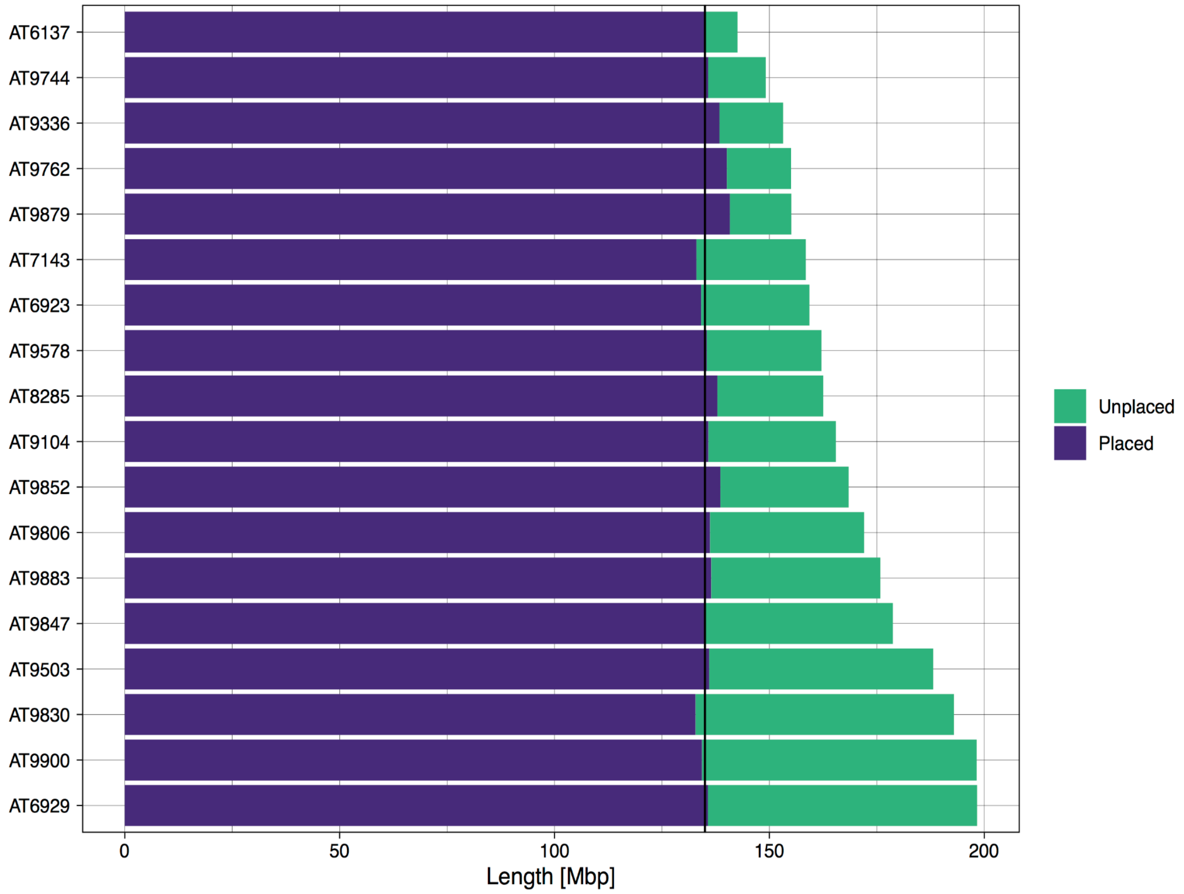

Figure 22: Total assembly length vs length of scaffolded sequences. The cumulative length of scaffolded sequence varied by 8 Mb with an average of 135 Mb (black line). In contrast to this, the length of unplaced sequence was highly variable among the eighteen assemblies.

Next, I assessed the length of all five chromosomes and compared them among the different accessions used in this study (**Table 3**). The overall length of chromosome 1 varied by 3 Mb. The shortest chromosome 1 was found in AT6923 (31.6 Mb), while chromosome 1 of AT9762 was the longest at 34.9 Mb. The average was 33.3 Mb. For chromosome 2, the average length was 22.6 Mb. Its length ranged from 20.8 Mb, in the assembly of AT9744, to 24.8 Mb, in the assembly of AT9879. Chromosome 3 was on average 27.2 Mb long. AT6929 had the shortest chromosome 3 with 25.6 Mb, while the longest chromosome 3 was found in AT6923 (28.6). The overall length of chromosome 4 varied between 21.1 Mb in AT6923 and 24.1 Mb in AT9852, with an average of 22.6 Mb. Chromosome five had an average length of 30.6 Mb. With 29 Mb, the shortest chromosome five was observed in the assembly of AT9806, while the longest chromosome five was found in the AT9762, with 31.5 Mb. All *de novo* assembled and scaffolded chromosomes were longer than in the reference genome TAIR10. However, highly repetitive regions such as centromeres and rDNA repeat regions are mostly unassembled in the current reference genome (Arabidopsis Genome Initiative

2000). Therefore, overall scaffold length, the size differences of chromosomes among accessions as well as between accessions and the reference genome TAIR10 will be further examined in the annotation chapter.

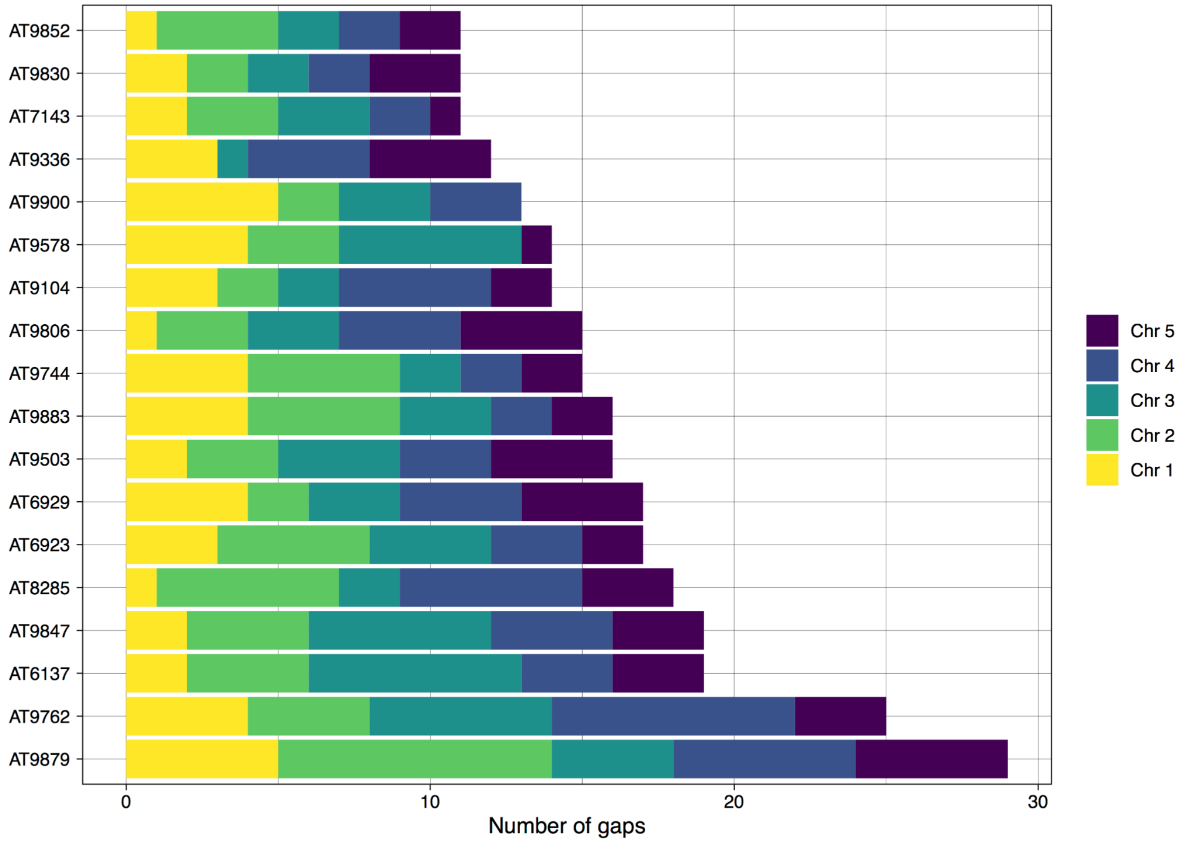

Figure 23: Gaps in scaffolded chromosomes. The gaps per chromosome were counted for each assembly in order to assess contiguity per chromosome after scaffolding. Colored stacks represent the number of gaps per chromosome. AT9336 (chr. 2), AT9578 (chr. 4), and AT9900 (chr. 5) are missing one stack due to one telomere-to-telomere contig each.

Figure 23 shows that the scaffolded assembly of AT9879 had the highest number of gaps (28), while AT9852, AT9830, and AT7143 were the most contiguous (11 gaps). The average number of gaps per assembly was 16. All chromosomes in all assemblies were scaffolded with ≤ 9 gaps (**Figure 23**). Chromosome 2 in AT9336, chromosome 4 in AT9578, and chromosome 5 in AT9900 were assembled from telomere-to-telomere, making further scaffolding unnecessary. Scaffolded chromosome 1 contained between one (AT8285, AT9806, and AT9852) and five gaps (AT9900, and AT9879) with an average of 2.8 gaps. The number of gaps in the scaffolds of chromosome 2 ranged from zero (AT9336) to nine (AT9879) with an average of 3.7 gaps. The mean number of gaps in the scaffolds of chromosome 3 was 3.5, with one being the lowest gap count (AT9336) and seven being the highest gap count (AT6137). The scaffolds of chromosome 4 had between zero (AT9578) and eight (AT9762) gaps. Chromosome 4 scaffolds had 3.1 gaps on average. The lowest average number of gaps, 2.5, was observed in the scaffolds of chromosome 5. The highest number of gaps, five, was again found in the assembly of AT9879, while AT9900 had zero gaps in the scaffold of chromosome 5. In total three chromosomes from three different assemblies were assembled telomere-to-telomere.

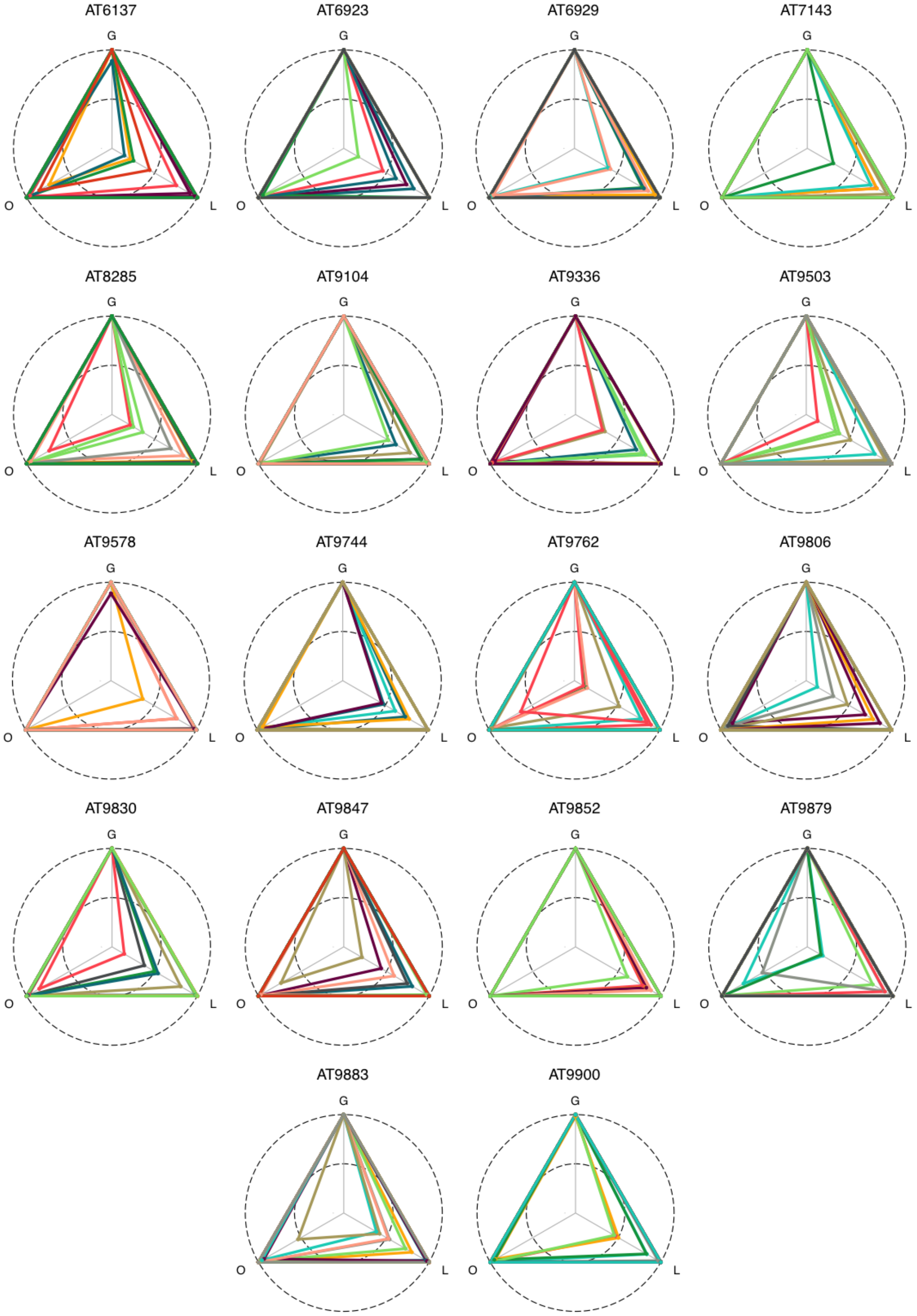

Figure 24: Radar plots of scaffolding confidence scores per assembly. The most outer dashed line indicates a confidence score of 100 %. The next dashed line shows a confidence of 50 % while the center of the circle represents zero. Grouping (G) confidence is the certainty of the given contig to be placed on the correct chromosome. Moreover, location (L) confidence describes the certainty of the contig to be placed at the correct position within the chromosome. Finally the orientation (O) score indicates how likely it is that the contig was placed in the correct orientation. Colored lines connecting grouping, location, and orientation are representing individual contigs.

Grouping confidence (Alonge et al. 2021) describes the likelihood of a given contig to be placed on the correct chromosome. Median contig grouping confidence was > 99% in all assemblies (**Figure 24**). Additionally, the median location confidence, which indicates how likely a contig is to be placed at the correct position on the chromosome, exceeded 99% in all but one assembly (AT7143). All assemblies had median orientation confidence scores above 99%. However, individual contigs of a given assembly may have lower scores (**Figure 24**). The only assembly that did not include such outlier contigs was AT9852, where accession-specific bionano optical map data had guided the scaffolding process.

In summary, I demonstrated that the scaffolds assembled in this study were highly contiguous and that the incorporated contigs were placed with high confidence. Although being longer, all eighteen assemblies met the expectations in terms of chromosome length and overall length as compared to the *A. thaliana* reference genome TAIR10. Thus, all eighteen assemblies are highly suitable for further comparative analyses.

Table 3: Scaffolded chromosome length in base pairs (bp) of the 18 accessions and the current *A. thaliana* reference genome TAIR10.

Genotype	Chromosome 1	Chromosome 2	Chromosome 3	Chromosome 4	Chromosome 5
TAIR10	30,427,671	19,698,289	23,459,830	18,585,056	26,975,502
AT6137	32,271,270	22,775,491	26,194,750	22,667,791	30,926,001
AT6923	31,749,057	21,980,435	28,738,137	21,225,719	30,405,250
AT6929	32,573,738	22,781,885	25,712,594	23,584,199	31,033,573
AT7143	33,668,152	21,027,966	26,044,309	21,422,025	30,876,513
AT8285	34,650,672	23,046,220	26,796,854	22,897,587	30,525,389
AT9104	32,656,548	23,453,113	27,458,818	21,384,833	30,867,744
AT9336	34,071,237	22,870,034	27,060,997	23,254,149	31,130,234
AT9503	32,192,683	22,299,778	28,010,412	22,845,789	30,634,581
AT9578	33,401,629	21,734,620	26,348,032	23,817,584	30,154,809
AT9744	33,001,351	20,772,390	27,175,644	23,477,951	31,338,815
AT9762	34,863,657	22,978,411	27,262,250	23,502,266	31,489,441
AT9806	34,678,793	22,206,001	27,369,393	22,962,254	28,968,011
AT9830	33,113,414	21,730,182	27,041,555	21,387,555	29,547,016
AT9847	31,771,136	22,850,951	28,044,082	22,152,542	29,957,574
AT9852	33,329,944	23,396,451	26,358,531	24,203,097	31,304,789
AT9879	34,155,544	24,779,077	27,173,682	23,300,978	31,396,706
AT9883	33,747,718	23,266,240	26,028,663	23,936,058	29,492,918
AT9900	33,376,575	22,160,010	27,314,136	21,468,527	30,011,449

Collapsed regions

Repetitive or paralogous regions are challenging to assemble. It is possible that the assembler collapses and thereby underestimates the total length of such regions. In case of a perfect assembly one would not expect to call any SNPs when mapping the initial reads back to the *de novo* assembly. However, especially repetitive regions may be collapsed during assembly. Since these repeats may differ from each other by single nucleotides one would expect to call heterozygous SNPs when mapping the initial reads onto a collapsed region (H. Cheng et al. 2021).

Thus, all scaffolded assemblies were checked for collapsed mis-assemblies by first mapping the Q20 filtered HiFi reads to the corresponding assembly. Average genome coverage was calculated based on these mappings. Subsequently, the genome assembly was binned into 5 kb windows. Heterozygous single nucleotide polymorphisms (SNPs) were called. These heterozygous SNPs were further reduced to positions that are supported by n reads, where n was set to 75% of the average sequencing coverage of the given genome. Regions with >= 5 SNPs per 5 kb window are referred to as collapsed mis-assemblies. The cumulative length of all mis-assemblies in a given dataset was estimated based on the average coverage in that 5 kb window compared to the sequencing coverage of the whole assembly. Collapsed regions were detected in all assemblies (**Table 4**). However, the number of mis-assemblies, the number of SNPs per 5 kb window, and the deviation from the average coverage all varied greatly between assemblies. The highest number of collapsed regions (50) was detected in the assemblies of AT6137 and AT9936. The lowest number of such mis-assemblies was found in the scaffolds of AT7143 (2) and AT9883 (4). The average number of collapsed regions per assembly was 24. The estimated cumulative size of mis-assemblies varied between 12.3 kb (AT7143) and 919.7 kb (AT6137). The average cumulative size per assembly was 331 kb. Notably, in the case of AT6137, a total of nine collapsed regions accounted for 578 kb of the total length of all mis-assemblies. Similarly, more than half (504 kb) of the cumulative size of collapsed regions in the assembly of AT9336 was caused by only ten mis-assemblies. The median length of the detected collapsed regions ranged from 4 kb (AT9883) to 23 kb (AT9879). An overlap between collapsed regions and highly repetitive sequences such as centromeres, telomeres, 45S rDNA and 5S rDNA clusters was observed in all assemblies. The collapsed regions intersecting with repetitive sequences accounted for 0.57 Mb in the assembly of AT9336 (**Table 4)**. More than half (330,869 bp) of the cumulative size of the mis-assemblies in the scaffolds of AT9806 intersected with such highly complex regions. All potentially collapsed regions were flagged but kept.

Table 4: Overview about potentially collapsed regions in the 18 genome assemblies. Positions of collapsed regions were intersected with highly repetitive regions such as centromeres, 45S rDNA, and 5S rDNA clusters.

Genotype	Collapsed regions			
	Number	Cumulative length (bp)	Median length (bp)	Number in repeat regions
AT6137	50	919,666	8,677	24
AT6923	29	464,336	7,119	6
AT6929	13	158,150	7,196	3
AT7143	2	12,448	6,224	1
AT8285	10	175,384	13,890.5	1
AT9104	10	80,687	8,093	3
AT9336	50	812,903	6,779.5	33
AT9503	20	219,642	6,503.5	11
AT9578	27	279,487	9,672	20
AT9744	36	529,202	8,906.5	16
AT9762	40	595,757	7,054.5	15
AT9806	19	434,706	8,872	6
AT9830	12	146,015	6,024	3
AT9847	24	204,110	7,188.5	8
AT9852	28	331,044	10,477	14
AT9879	31	773,270	23,283	8
AT9883	4	17,762	4,143.5	4
AT9900	21	310,600	8,141	2

Centromere annotation

Centromeres play a vital role since they are necessary for cohesion of sister chromatids and for binding of spindle fibers during cell division. Centromeres are characterized by large scale tandem repeat arrays (J. Jiang et al. 2003). In many of the so far studied plant species, these satellite repeat monomers are usually between 150 bp and 180 bp long (Oliveira and Torres 2018). In *A. thaliana* the most abundant satellite repeat (*CEN180*) is known to be 178 bp long (Martínez-Zapater, Estelle, and Somerville 1986). Recently it was reported that *CEN180* can be present with 11,800 to 15,600 copies per chromosome. Thus, these satellite repeats form large tandem arrays (Naish et al. 2021). As mentioned in the introduction chapter, these long and repetitive genomic regions often remain unassembled when using short-reads. In *A. thaliana* the estimated genome size is about 135 Mb while the total length of the reference genome TAIR10 is only about 119 Mb. This difference between both size estimates is attributable to highly repetitive centromeric regions that remained unassembled in the current reference genome (Arabidopsis Genome Initiative 2000). In the scaffolding section of this chapter I have shown that the eighteen scaffolded assemblies have a length of approximately 135 Mb (**Figure 22**). To test if the difference in length between the eighteen assemblies and TAIR10 is attributable to assembled centromere sequences, as it has been observed in other long-read assemblies (Naish et al. 2021; Rabanal et al. 2022) I performed annotation of these repeats. Centromeres were annotated by searching for the 178 bp sequence motif using RepeatMasker. Centromeric repeats were detected in the scaffolds of all chromosomes in all eighteen assemblies. It was possible to place all centromeric contigs in ten out of eighteen assemblies (**Figure 25**). Moreover, in multiple assemblies the centromere repeat units were reconstructed without contig breaks (**Figure 21**). The length of unplaced centromeric sequences varied between 7.7 kb (AT9879) and 1.4 Mb (AT6137 and AT9104). The total length of placed centromeric repeats varied by 6.5 Mb among accessions. Thus, it was possible to annotate between 11 Mb (AT9830 and AT7143) and 17 Mb (AT9762 and AT9852) of centromeric sequence. The assemblies contained on average 14.4 Mb of centromeric repeats, with centromere content differing among the five chromosomes. The lowest centromere content on chromosome 1 was found in AT6923 (1.4 Mb). Conversely, chromosome 1 of AT9762 had almost 4.9 Mb of centromeric sequence. The average centromere length of chromosome 1 was 3 Mb. The centromere of chromosome 2 was on average 600 kb shorter as compared to chromosome 1. The shortest chromosome 2 centromere with 1 Mb was observed in the assembly of AT9744. In contrast, the length of the longest chromosome 2 centromere was 4.1 Mb, in the assembly of AT9879. The centromere of chromosome 2 was on average 2.8 Mb long. AT7143 had the shortest centromere (1.5 Mb) when comparing chromosome 3 assemblies, while the longest chromosome 3 centromere was found in the assembly of AT6923 (4.2 Mb). The centromere length of chromosome 4 ranged from 1.8 Mb in AT9104 to 5 Mb in AT9852. The average centromere length of chromosome 4 was 2.8 Mb. A similar average centromere length was observed for chromosome 5 (2.9 Mb). Centromere length of chromosome 5 varied between 1.8 Mb in AT9806 and 3.9 Mb in AT9104. It must be noted that centromere length estimation is less accurate in assemblies which have either contig breaks within the centromere or unplaced centromeric contigs (**Figure 21; Figure 25**).

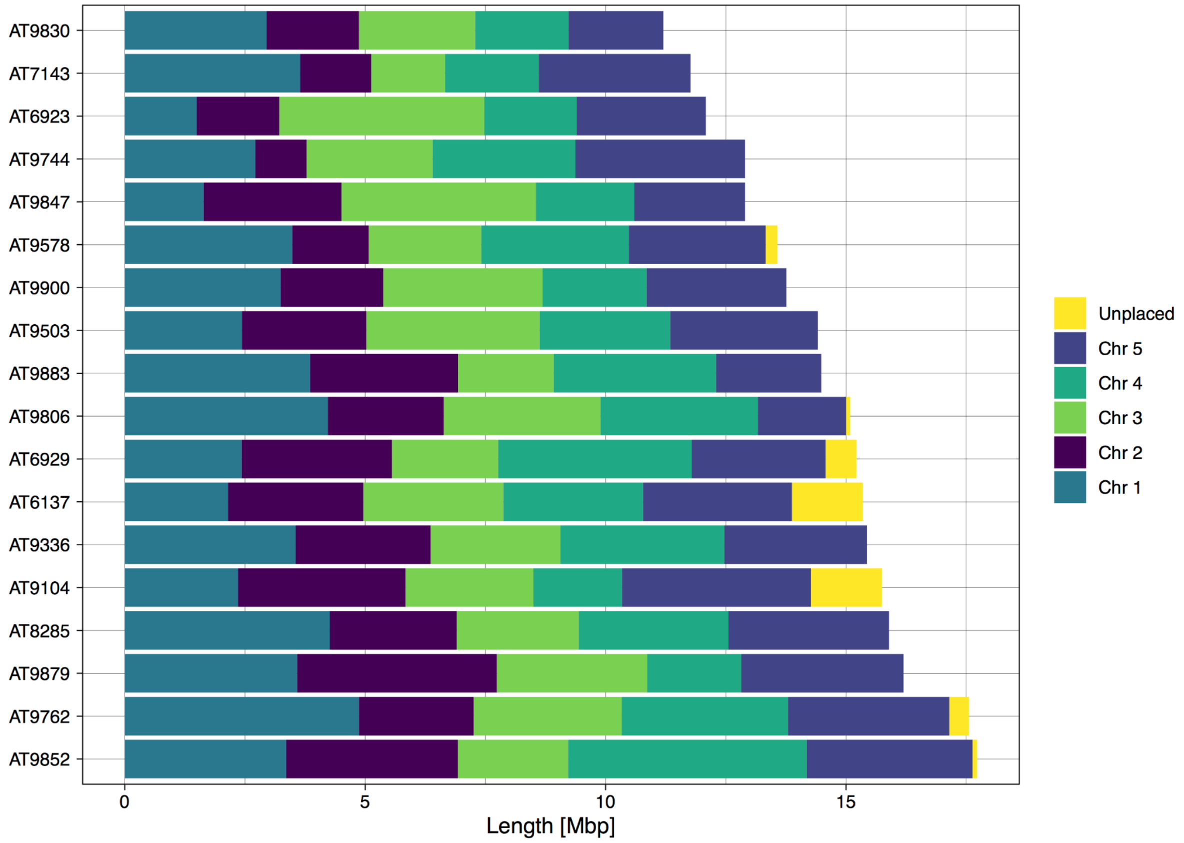

Figure 25: Cumulative length of centromere sequences in the 18 genome assemblies. Different color stacks are indicating the cumulative length of centromere sequence per chromosome. Centromere sequences that remained unplaced after scaffolding are shown in yellow.

Annotation of rDNA repeats

Ribosomal RNAs (rRNAs) are fundamental for maintaining basic cellular functions. The 45S rRNA genes of eukaryotic genomes are arranged in clusters. These clusters are referred to as nucleolus organizer regions (NOR). The 5S rRNA is encoded by a separate gene (E. O. Long and Dawid 1980). A primary transcript of the 45S rRNA gene is generated by transcription via RNA polymerase I. Subsequently, the primary transcript is processed into 18S, 5.8S, and 25S rRNAs. Together with the 5S rRNA they build the catalytic core of ribosomes (Chambon 1975). In *A. thaliana* 5S rDNA is characterized by a 497 bp long repeat unit (Campell et al. 1992).

It is known that the 45S rRNA gene of *A. thaliana* is over 10 kb long. Moreover, the genome contains hundreds of copies of these genes arranged in tandem arrays at the top of chromosome two and four (Copenhaver et al. 1995; Copenhaver and Pikaard 1996). Moreover, it has been demonstrated that variation in genome size between different *A. thaliana* accessions is partially attributable to differences in rRNA gene copy numbers. From short-read NGS data it was estimated that the 45S ribosomal DNA (rDNA) repeats can span up to 40 Mb of the *A. thaliana* genome (Q. Long et al. 2013).

Both 5S rDNA and 45S rDNA repeats were annotated by searching for previously published (Simon et al. 2018; Rabanal et al. 2017) tandem repeats using RepeatMasker.

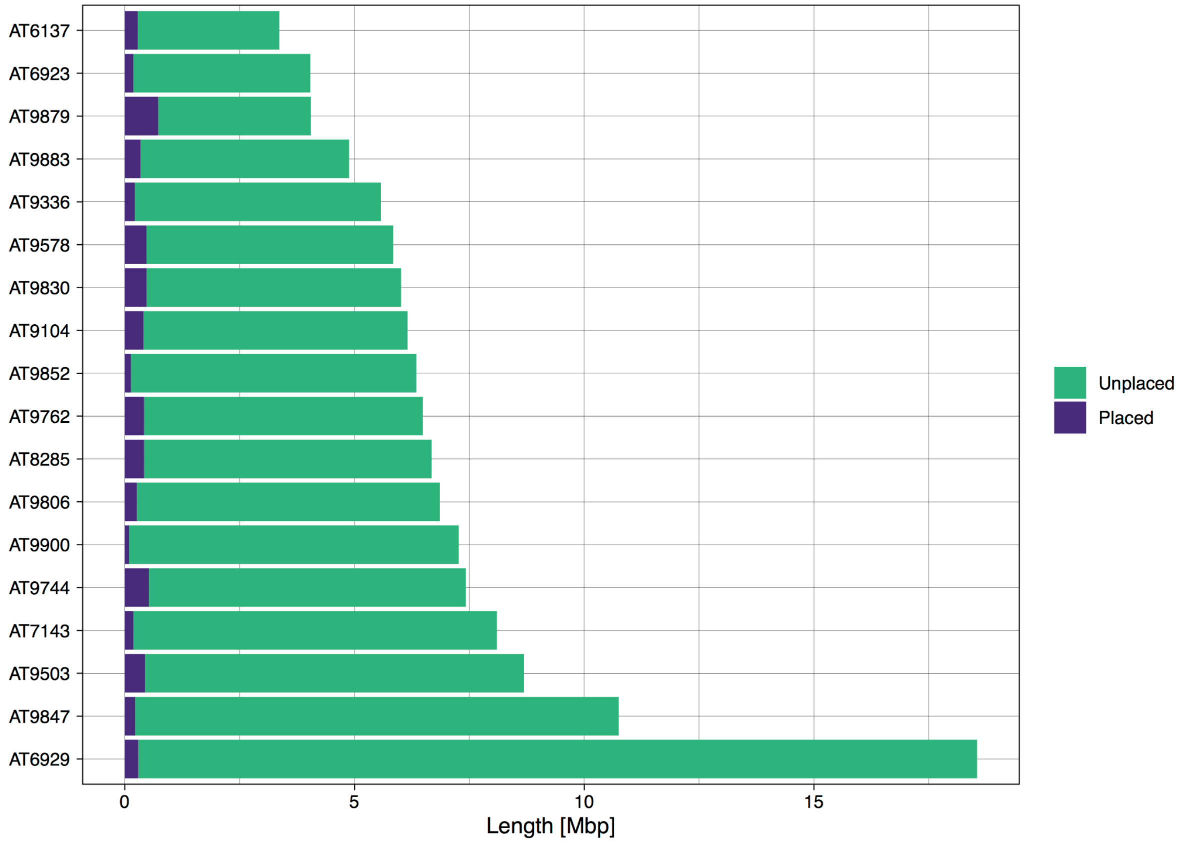

Figure 26: Cumulative length of 45S rDNA repeats per assembly. Purple stacks are showing the 45S rDNA sequences that were annotated on placed contigs whereas green stacks represent 45S rDNA sequences found on unplaced contigs.

I annotated between 3.4 Mb (AT6137) and 18.6 Mb (AT6929) of 45S rDNA sequences (**Figure 26**). The median length of annotated 45S rDNA repeats was 6.3 Mb, but most of the annotated 45S rDNA repeats were located on unplaced contigs. Hence, only between 1.3% (AT9900) and 18% (AT9879) of the total 45S rDNA sequences were annotated on scaffolds. On average, less than 6% of all 45S rDNA clusters were found on scaffolds. Calculating the median length of the annotated 45S rDNA clusters of all accessions per chromosome revealed that most such repeats were, as expected from the literature (Copenhaver and Pikaard 1996) located on chromosome 2 (160 kb) and chromosome 4 (143 kb). The median length of 45S rDNA sequences across accessions was 3.8 kb on chromosome 1, 2.1 kb on chromosome 3, and 1.2 kb on chromosome 5.

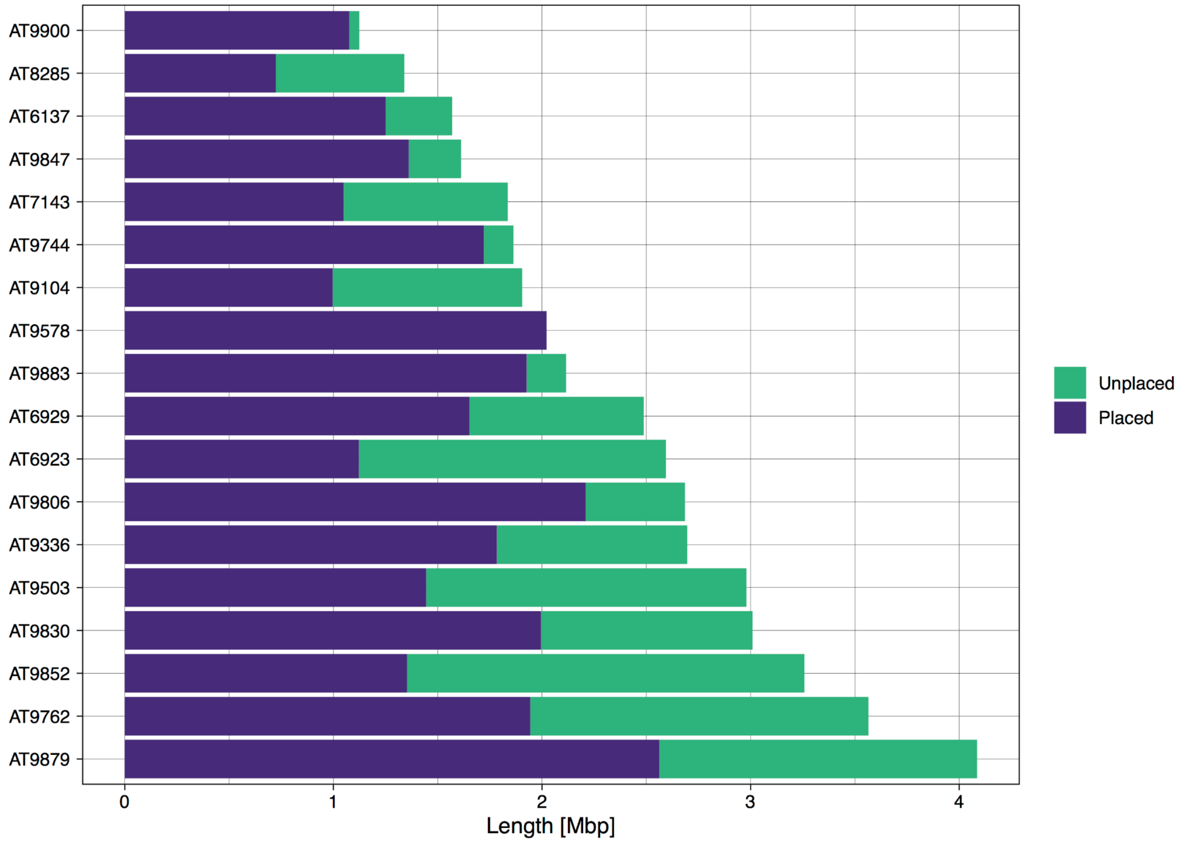

Figure 27: Cumulative length of annotated 5S rDNA repeats. Purple stacks are showing the length of 5S rDNA sequences that were annotated on placed contigs whereas green stacks represent 5S rDNA sequences found on contigs that remained unplaced after scaffolding.

Between 1.1 Mb (AT9900) and 4.1 Mb (AT9879) of 5S rDNA repeats were annotated. The average length of annotated 5S rDNA sequences was 2.3 Mb per assembly. In contrast to the aforementioned 45S rDNA clusters, most of the annotated 5S rDNA sequences were found on scaffolds rather than on unplaced contigs (**Figure 27**). The average length of 5S rDNA clusters found on scaffolds was 1.6 Mb, and on unplaced contigs was 0.9 Mb, although in AT9879 all annotated 5S rDNA clusters were on scaffolds. On average 69% of the annotated 5S rDNA sequences were located on scaffolds.

Chromosome 4 contained between 168 kb (AT7143) and 1.6 Mb (AT9879) of 5S rDNA sequences, with a median length of 598 kb (**Figure 28**). The annotated length on chromosome 5 ranged from 174 kb in the assembly of AT8285 to up to 1.2 Mb in the assembly of AT6929. The median cumulative length of 5S rDNA on chromosome 5 was 743 kb. The other three chromosomes had very small amounts of 5S rDNA sequences. Whether these are assembly or annotation artifacts remains unclear. The total length of 5S rDNA sequences on chromosome 1 varied between 3.3 kb (AT9336) and 5.3 kb (AT9104) (average 4.3 kb). Cumulative 5S rDNA length on chromosome 2 varied between 1.4 kb (AT9852) and

2.7 kb (AT9104) (average 2.3 kb). The cumulative length of 5S rDNA on chromosome 3 varied from 1.1 kb (AT6137) to 451 kb (AT9806) (average 33 kb).

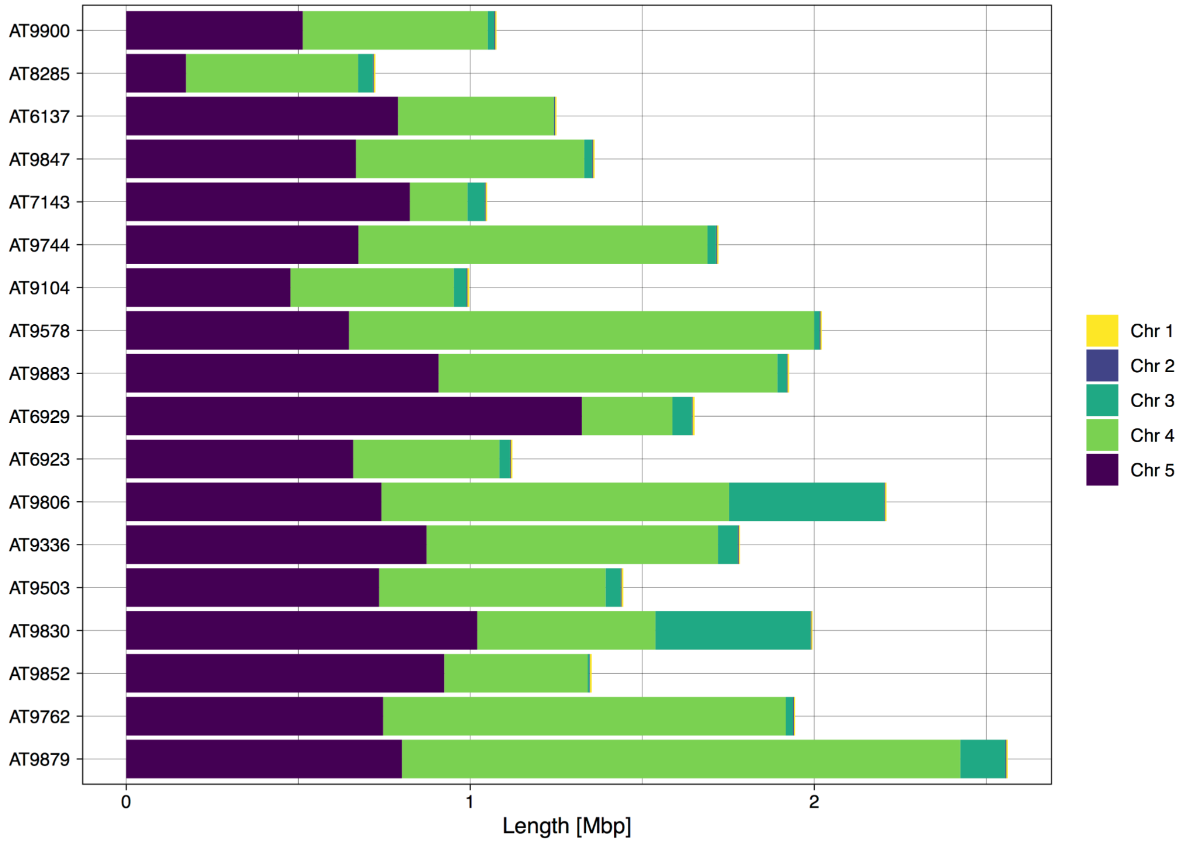

Figure 28: Location and cumulative length of 5S rDNA repeats. The color of the stacks represents the chromosome where the 5S rDNA sequence was annotated.

Differences in chromosome length explained by centromeres

As mentioned before, chromosome length differences across accessions varied from 2.5 Mb for chromosome 5 to 4 Mb for chromosome 2 (**Table 3**). The length of placed centromeric sequences varied by up to 6.5 Mb between genotypes (**Figure 25**). Thus, average chromosome lengths and standard deviations were calculated with and without the centromeric sequences. **Table 5** shows that a substantial part of the variation in chromosome length among the eighteen assemblies can be explained by variation in centromere length. Unsurprisingly, the average chromosome length decreased in all cases when subtracting centromeres. More importantly, the standard deviation among the eighteen assemblies decreased when centromeres were excluded. This indicates that a significant proportion of the variation in length observed before can in fact be explained by differences in centromere length among the eighteen accessions. Moreover, the difference between the mean length of the eighteen assemblies and the chromosome length expected from TAIR10 decreases when centromeres are subtracted. This can be attributed to the fact that centromeres are poorly resolved in the current version of the *A. thaliana* reference genome

Arabidopsis Genome Initiative 2000. These results will be discussed later in the context of recent studies that used similar sequencing technology to assemble centromeres (Rabanal et al. 2022; Naish et al. 2021).

Table 5: Comparison of median assembly length per chromosome with and without centromere sequences. Mean length and standard deviation were calculated based on all 18 assemblies. The *A. thaliana* reference genome TAIR10 is listed for comparison with the *de novo* assemblies.

Chromosome	Including centromeres		Excluding centromeres		Length in TAIR10 (bp)
	Mean length (bp)	Standard deviation (bp)	Mean length (bp)	Standard deviation (bp)	
1	33,292,951	968,386.5	30,168,059	224,354.6	30,427,671
2	22,561,625	945,879.5	20,014,323	309,997.5	19,698,289
3	27,007,380	795,300.9	24,181,964	377,593.2	23,459,830
4	22,749,495	996,942.0	19,910,384	572,969.2	18,585,056
5	30,558,934	734,102.3	27,626,451	393,508.4	26,975,502

Characterization of unplaced contigs

As mentioned previously, all assemblies are markedly longer compared to the estimated genome size of *A. thaliana*. However, most of the additional sequence remained unplaced after scaffolding. The cumulative length of unplaced sequences was highly variable and ranged from 7.8 Mb (AT6137) to up to 64 Mb (AT9900) (**Figure 22**). Proceeding with such a variable length of unplaced sequence would complicate downstream comparison of the eighteen genomes. However, simply removing unplaced contigs runs the risk of losing potentially valuable information. Therefore, I characterized these unplaced contigs further in order to find criteria for filtering them, such as length or sequence composition.

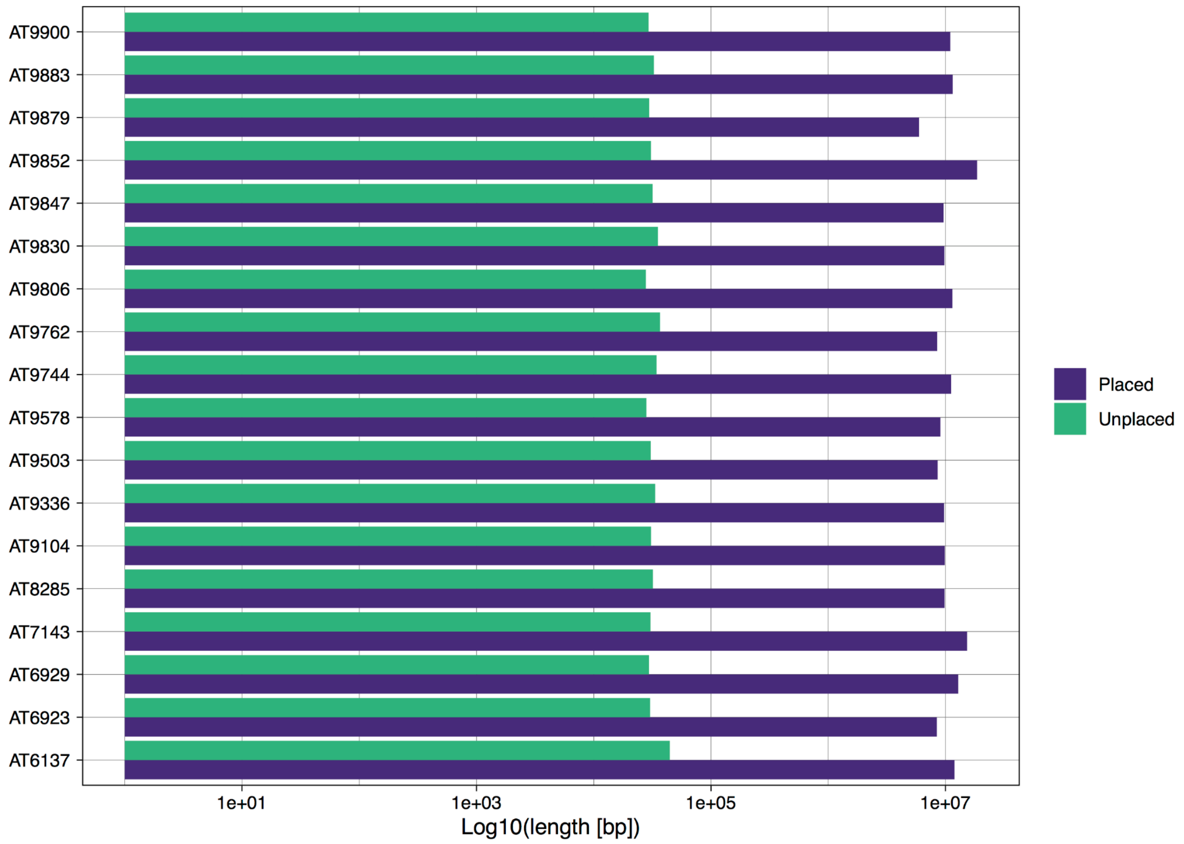

Figure 29: Comparison of N50 values between placed (purple bars) and unplaced (green bars) contigs.

Unplaced contig N50 ranged from 27.8 kb (AT9806) to 44.55 kb (AT6137) (**Figure 29**). Unplaced contigs of all eighteen assemblies were smaller than 1 Mb, with the exception of a single unplaced contig in the assembly of AT6137 (1.43 Mb). Comparison of all assemblies revealed that contig N50 of placed contigs was generally much higher compared to the sequences that remained unplaced after scaffolding (**Figure 29**). The average N50 of placed contigs was 10.6 Mb. Thus, placed contigs were on average 337 times longer compared to unplaced contigs.

Subsequent annotation of unplaced contigs revealed that large fractions of the unplaced contigs are mainly composed of 45S rDNA and chloroplast sequences (**Figure 30**). Between 13% (AT6137) and 39% (AT9900) of the unplaced sequences were chloroplasts, with an average of 28%. The fraction of 45S rDNA sequences varied between 9% (AT9830) and 51% (AT9744). Unplaced contigs contained 24% 45S rDNA sequences on average. In contrast, mitochondrial sequences only accounted for 0.35% (AT6929) to up to 4.5% (AT6137) of the unplaced contigs. Moreover, I found that between 5% (AT6137 and AT6923) and 11% (AT9830 and AT9883) of the unplaced sequences were transposable elements. Between 0% (AT9578) and 11% were 5S rDNA repeats. The fraction of unplaced contigs that could not be classified further (unknown) ranged from 19% (AT9744) to 63% (AT9578).

Notably, in the case of these sequences it was not possible to unambiguously assign them to only one of the aforementioned repeat types. Thus, using different combinations of repeat libraries led to different classification results. Therefore, I classified these sequences as ‘unknown’.

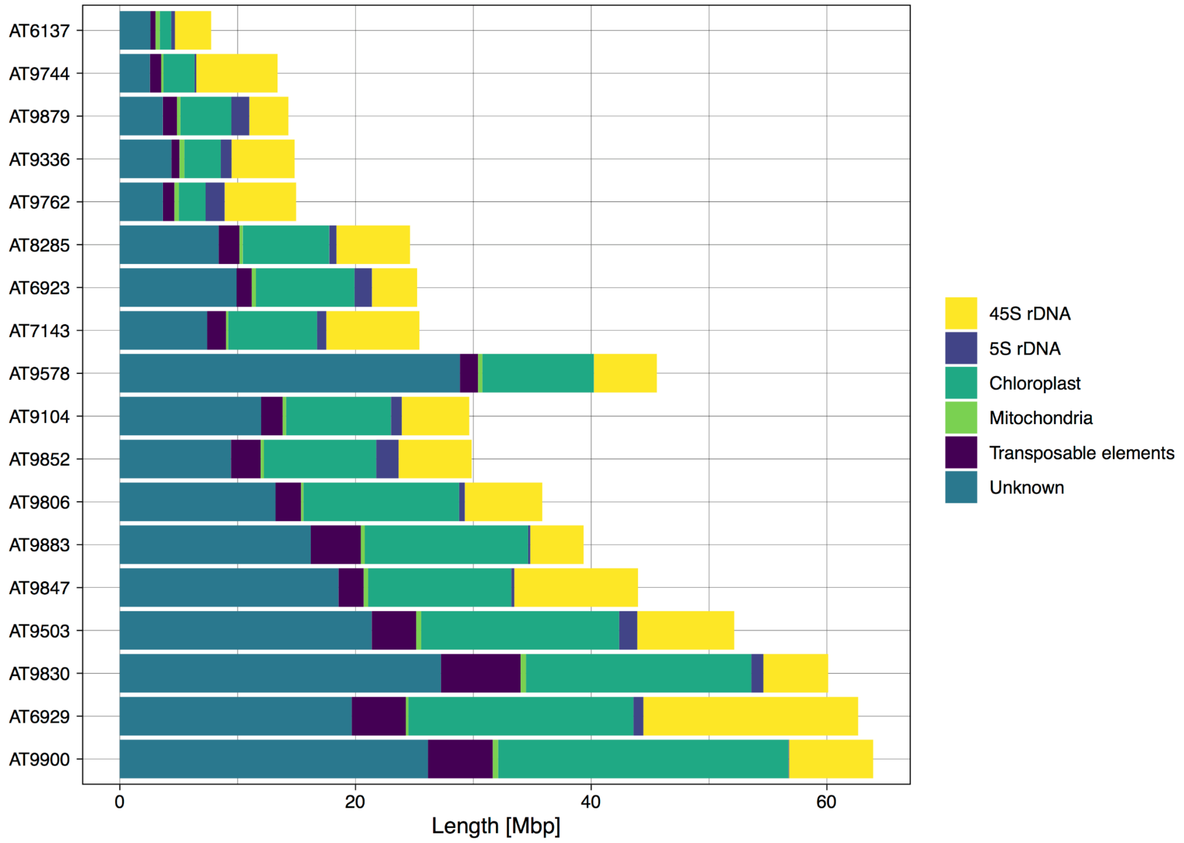

Figure 30: Classification of unplaced sequences. Different color stacks are representing 45S rDNA, 4S rDNA, chloroplast, mitochondria, transposable element, or unknown sequences.

In the next step I investigated if short-reads from previously published experiments can be confidentially mapped to these unplaced contigs. Dr. Kevin Murray, a postdoc in the lab, mapped 135 publicly available short-read datasets onto all assemblies, including unplaced contigs. My subsequent analysis of the average mapping quality of the 135 datasets per contig or scaffold revealed differences not only between unplaced contigs and scaffolds but also among the unplaced contigs (example in **Figure 31**). The majority of mappings on unplaced contigs resulted in a mapping quality of zero meaning that the read mapped to 10 or more positions on the contig (shown in **figure 31** for unplaced contigs of AT9852). This indicates that the majority of unplaced contigs consisted of highly repetitive sequences.

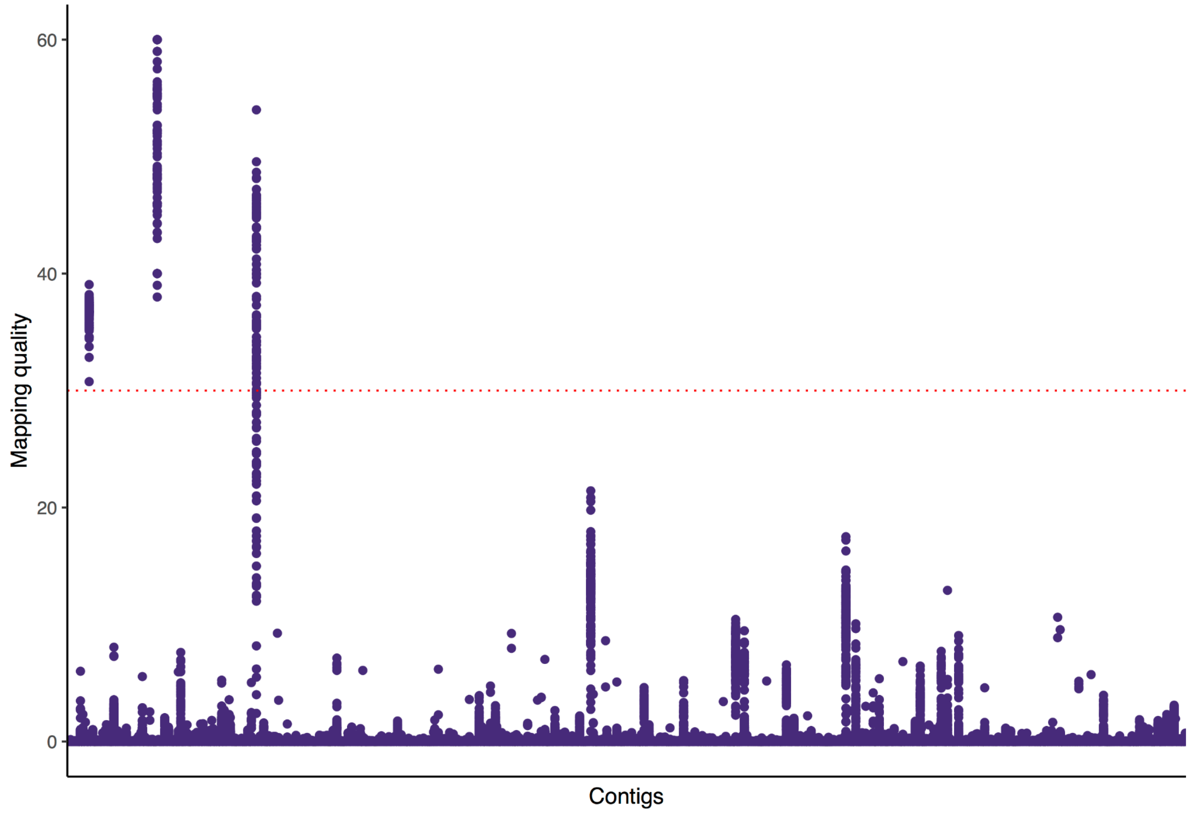

Figure 31: Mapping quality of 135 short-read datasets mapped to unplaced contigs of AT9852. Each dot represents the median mapping quality of one of the 135 readsets. The dashed red line indicates Q30 mapping quality. Contigs are plotted on the x-axis.

In contrast to this, all 135 short-read datasets were aligned to the scaffolded chromosomes with a mapping quality beyond 30 (**Figure 32**).

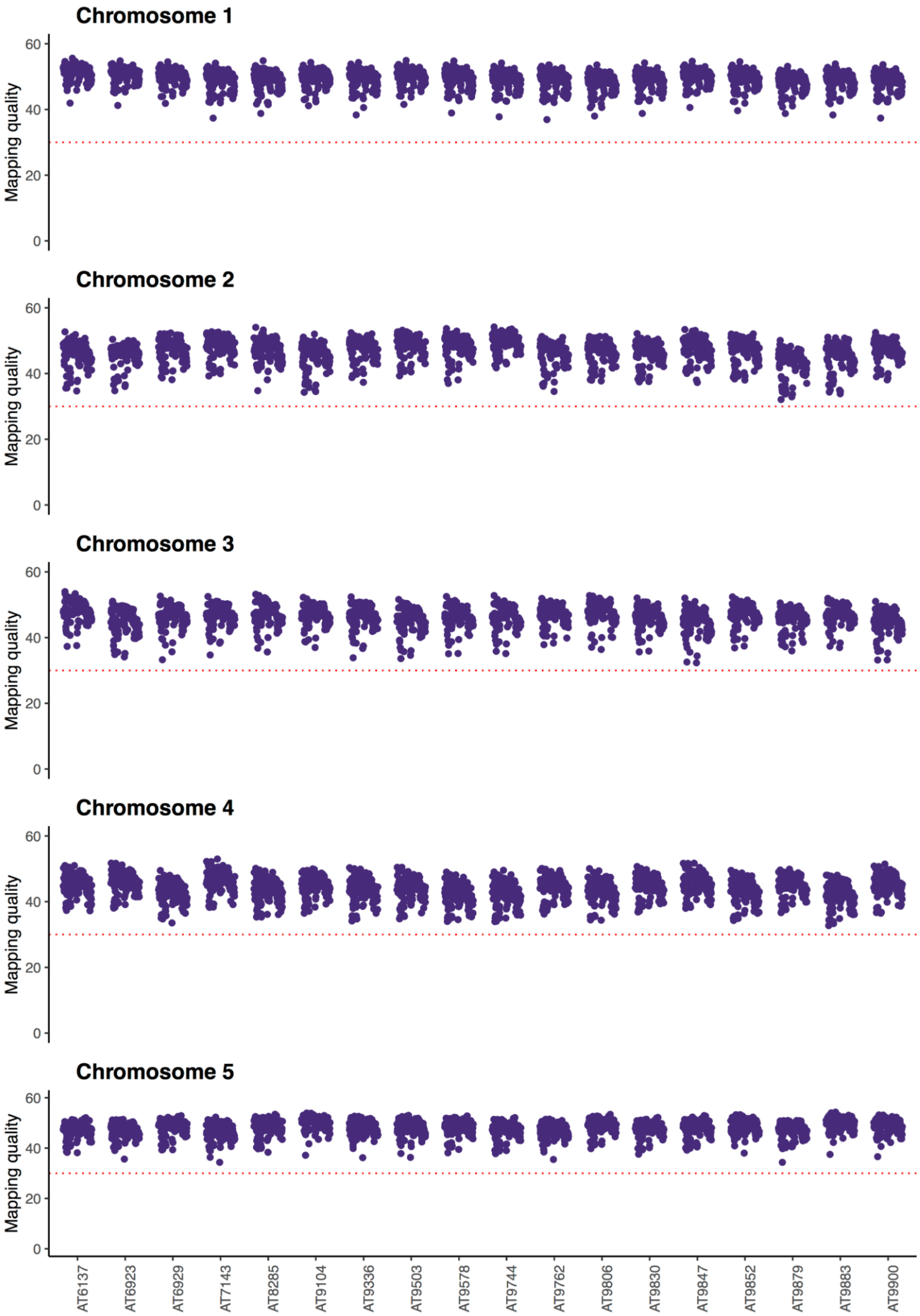

Figure 32: Mapping quality of 135 short-read datasets mapped to the chromosomes of all 18 *de novo* assemblies. Each dot represents the median mapping quality of one of the datasets. The dashed red line indicates the Q30 quality cutoff.

Contig filtering

Taking all these findings from before into account, I decided to filter the unplaced contigs according to the following three criteria: (1) organellar sequence content of less than 20%, (b) a total length above 50 kb, and (c) short-read mapping quality above 30 for at least one of the 135 datasets. When applying these filtering criteria, the number of unplaced contigs decreased drastically (**Table 6**). Before filtering, each assembly had between 163 (AT6137) and 2131 (AT9900) unplaced contigs. The median number of unplaced contigs per assembly was 908. After applying all three filtering criteria, the assemblies had between zero (AT8285, AT9852, and AT9879) and fourteen (AT9336) unplaced contigs, with a median of two unplaced contigs per assembly. The cumulative length of unplaced contigs after filtering ranged from 0 bp (AT8285, AT9852, and AT9879) to 1 Mb (AT9336). The median cumulative length of unplaced contigs after filtering was 28 kb. Thus, the total assembly length of scaffolds, including unplaced contigs, varied between 133 Mb (AT9830) and 140 Mb (AT9879 and AT9762). The median assembly length of the eighteen assemblies was 136 Mb. After applying the above filtering criteria, assembly completeness evaluation revealed that the main change was the reduction of the number of duplicated or fragmented BUSCO genes. Overall, single-copy BUSCO completeness was not significantly affected. Thus, I decided to proceed with the filtered assemblies for further analyses.

Table 6: Assembly statistics before and after filtering of unplaced contigs. Contigs were kept when matching the following three criteria: (1) < 20 % organellar content, (2) total length > 50 kb, and (C) median short-read mapping quality above Q30 for at least one out of 135 short-read datasets.

Genotype	Unplaced contigs			Total assembly length after filtering (bp)
	Number before filtering	Number after filtering	Cumulative length (bp)	
AT6137	163	1	92,208	134,927,511
AT6923	794	2	193,201	134,291,799
AT6929	2,017	6	361,568	136,047,557
AT7143	799	2	219,338	133,258,303
AT8285	732	0	0	137,916,722
AT9104	906	1	409,807	136,230,863
AT9336	422	14	1,043,640	139,430,291
AT9503	1,645	2	424,722	136,407,965
AT9578	911	1	267,603	135,724,277
AT9744	369	1	445,061	136,211,212
AT9762	370	3	372,397	140,468,422
AT9806	1,261	2	409,427	136,593,879
AT9830	1,683	3	224,020	133,043,742
AT9847	1,348	1	56,128	134,832,413
AT9852	909	0	0	138,592,812
AT9879	439	0	0	140,805,987
AT9883	1,196	2	293,805	136,765,402
AT9900	2,131	4	349,617	134,680,314

4.1.6 Structural variation

As described in the introduction, there are various examples of large-scale genomic rearrangements affecting economically relevant traits such as grain size and blast resistance in rice (Xu et al. 2006; Deng et al. 2017). As pointed out, the analysis of large structural variants is often limited by short sequencing read length. Since the genomes assembled here are based on highly accurate long-reads, I chose to use them for the detection of such large-scale genomic rearrangements.

Large scale structural variation compared to TAIR10

Large-scale structural variation (SV) was assessed by first aligning the five pseudo-chromosomes of each assembly to the reference genome TAIR10. Subsequently, SyRi (Goel et al. 2019) was used to measure the lengths of syntenic regions, translocations, duplications, inversions and non-reference sequence blocks. Unplaced contigs were not considered for the SV analysis. All eighteen accessions had a large inversion relative to TAIR10 at the tip of chromosome four as shown for AT9852 in **figure 33**. This inversion has previously been described for other accessions (Jiao and Schneeberger 2020).

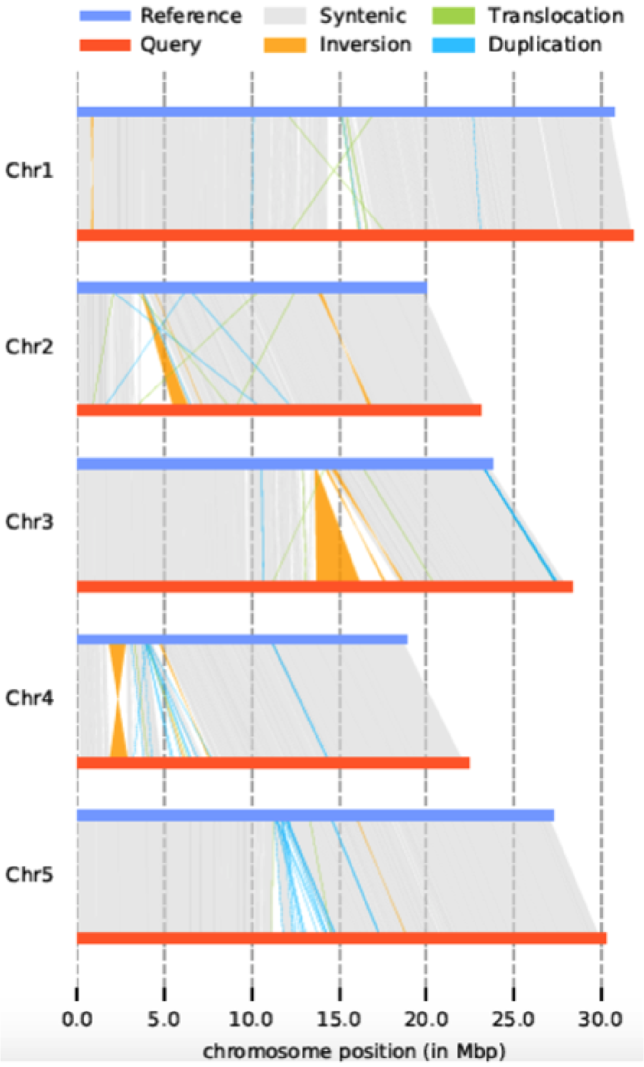

Figure 33: Alignment of AT9852 (query: red) and TAIR10 (reference: blue). Different types of structural variation are color coded in orange (inversion), gray (syntenic), green (translocation), and blue (duplication). SyRi was used for visualization (Goel et al. 2019).

Syntenic regions accounted for the by far largest sequence fraction in all assemblies (**Figure 34**). The total length of sequences that were syntenic compared to TAIR10 varied between 102 Mb in AT9879 and 108 Mb in AT8285, with an average of 105 Mb. Thus, between 85% and 90% of the TAIR10 reference genome were syntenic when compared to the eighteen *de novo* assemblies. Unaligned sequences accounted for the second biggest portion of non-syntenic regions. The accumulated length of unaligned sequences ranged from 18.6 Mb in AT9847 to 28 Mb in AT9879. On average, each assembly contained 21.8 Mb of sequence that could not be aligned to TAIR10. The accumulated length of inverted sequences varied between 3.6 Mb in AT8285 and 9.7 Mb in AT9806. On average, I identified 5.5 Mb of inverted sequences in each of the eighteen *de novo* assemblies. The greatest accumulated length of duplicated sequence was identified in AT9879 with 4.1 Mb. In contrast, AT9578 had only 2.4 Mb of duplicated sequences with respect to TAIR10. The average total length of duplicated sequences was 3.2 Mb. Translocations accounted for the smallest portion of non-syntenic sequences. Their total length varied between 1 Mb in AT6137 and 7.1 Mb in AT9879. All but AT9879 had less than 2 Mb of translocated sequences.

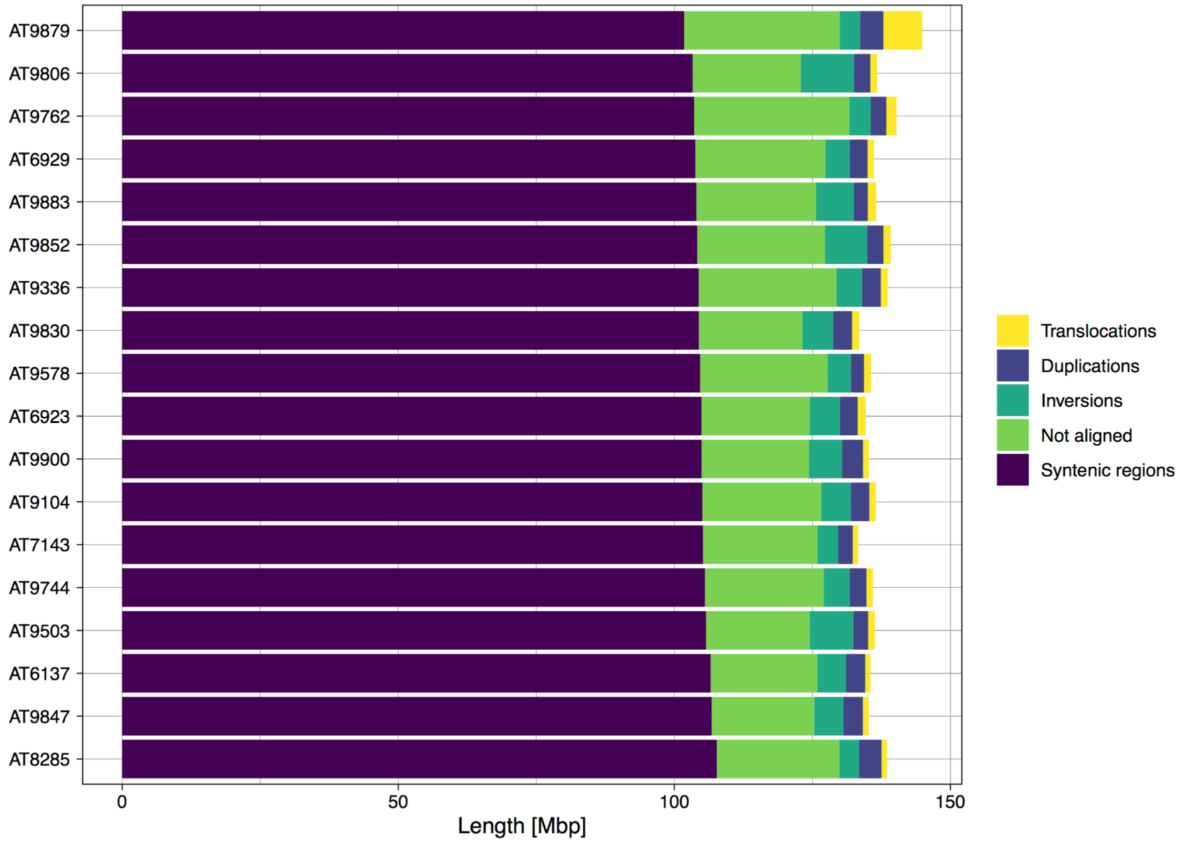

Figure 34: Cumulative length of syntenic (purple) and non-syntenic regions when comparing the 18 assemblies with the *A. thaliana* reference TAIR10. Non-syntenic regions can be either translocated, duplicated, inverted, or unaligned.

Subsequently, I calculated the length of inversions, translocations, and duplications in all eighteen *de novo* assemblies (**Figure 35**). The greatest variation in length was observed for inversions relative to TAIR10. The length of inverted sequence blocks varied between 234 bp and 3.4 Mb. The median length of inversions found in all analyzed assemblies was 5,281 bp. The average length of translocations was 2,321 bp. Translocation length varied between 230 bp and 298,296 bp. However, only three translocations were found to be longer than 80 kb. The length of duplicated sequences ranged from 199 bp to 92,524 bp with a median of 1,658 bp.

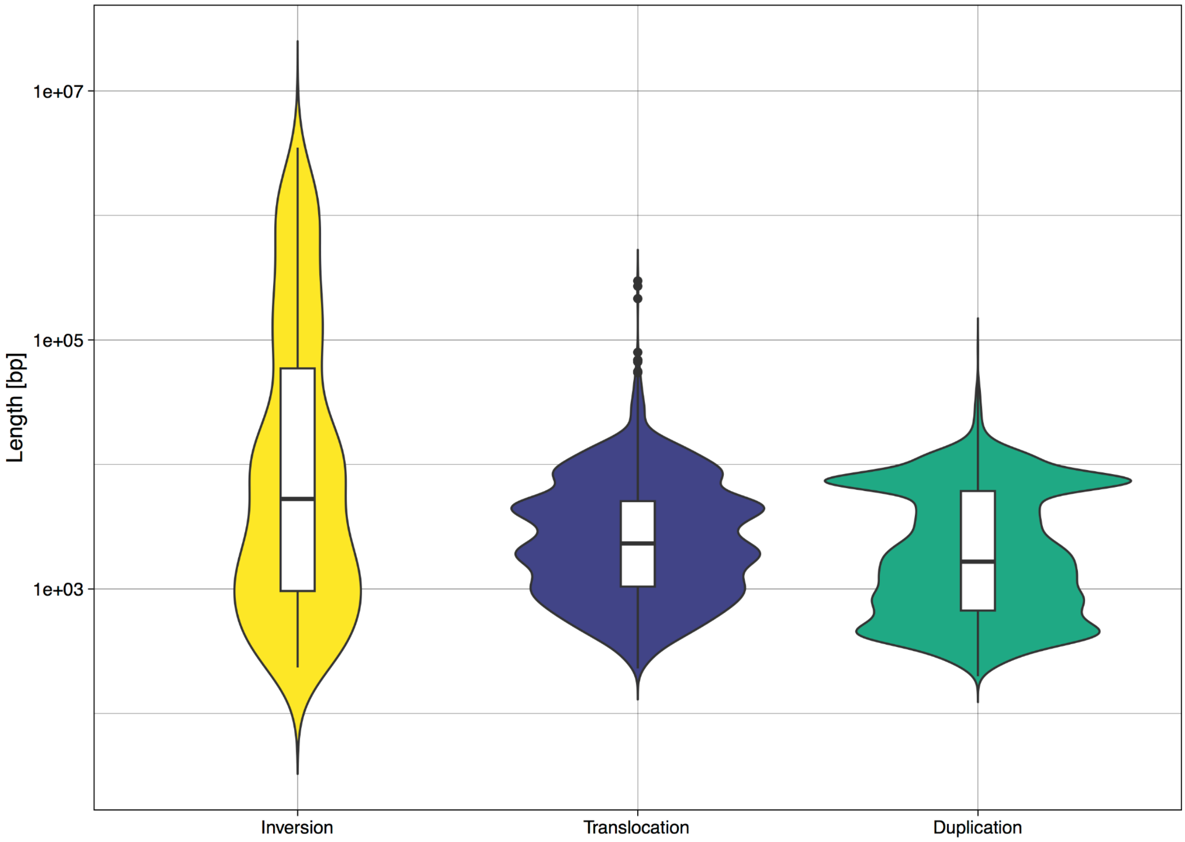

Figure 35: Violin plot showing the length of different structural variants (SV) in all 18 assemblies. Black bars in box plots represent the median value for the given SV type.

Next, I compared how many inversions, translocations, and duplications could be found in each accession (**Table 7**). I assessed the median length of these events in each of the assemblies separately. The highest number of inversions, 39, was detected in AT9744 and AT9762, while AT9900 had the fewest, 27 inversions, and the average was 33. The median length of inverted sequence blocks varied between 1,755 bp in AT9503 and 11,933 bp in AT9806. In contrast, the number of detected translocations ranged from 132 in AT7143 up to 238 in AT9879. On average, 175 translocations were identified per assembly. Their median length varied between 2,004 bp in AT9104 and 2,853 bp in AT9336. Moreover, I found an average of 688 duplication events per assembly. The lowest number, 505, of duplications was found in AT7143, and the highest, 1,100, in AT9336. The median length of duplication events varied between 625 bp in AT9336 and 2,913 bp in AT9830.

Table 7: Frequency and median length of different structural variants. All SVs were identified based on a whole genome alignment of the 18 assemblies against TAIR10.

Genotype	Inversions		Translocations		Duplications	
	Count	Median (bp)	Count	Length (bp)	Count	Median (bp)
AT6137	30	2,842	162	2,095	709	2,208.5
AT6923	35	6,881	203	2,500.5	664	2,224
AT6929	32	10,449	165	2,285	632	2,487
AT7143	36	2,506	132	2,656	505	1,614
AT8285	34	7,336	142	2,120	827	2,109
AT9104	29	3,853.5	183	2,004.5	689	2,082
AT9336	32	10,391	179	2,853.5	1100	625
AT9503	29	1,755	173	2,194.5	595	1,803
AT9578	30	1,855	191	2,215.5	621	1,326
AT9744	39	10,810	156	2,332	604	1,920
AT9762	39	5,112.5	200	2,059	755	1,317.5
AT9806	33	11,933	157	2,476.5	609	1,626
AT9830	36	10,395	192	2,362	603	2,913.5
AT9847	31	4,942	158	2,272	636	2,326
AT9852	33	6,465.5	171	2,380	704	1,688
AT9879	32	5,259	238	2,148	823	1,127
AT9883	33	8,703.5	196	2,287	624	1,462
AT9900	27	8,648.5	158	2,324	704	2,178

Location of unaligned sequence blocks

Figure 34 shows that each assembly contains sequences that cannot be aligned to the *A. thaliana* reference genome TAIR10. The accumulated length of sequence blocks that remained unaligned varied between 18.6 Mb in AT9847 and 28 Mb in AT9879. On average, each assembly contained 21.8 Mb of sequence that could not be aligned to TAIR10. Thus, I analyzed where these unaligned sequence blocks occurred in the *de novo* assembled genomes. To do this, I divided each chromosome of the reference genome into 1-Mb windows with positions counted relative to TAIR10. Subsequently, the number of unaligned sequence blocks per window was counted. Thus, in cases where the query genome had an inserted sequence between two flanking regions that could be aligned to the reference, these inserted and therefore unaligned sequences were counted. **Figure 36** shows that the highest number of unaligned sequence blocks was located in or near the centromeres. This is also visualized for AT9852 in **figure 33**.

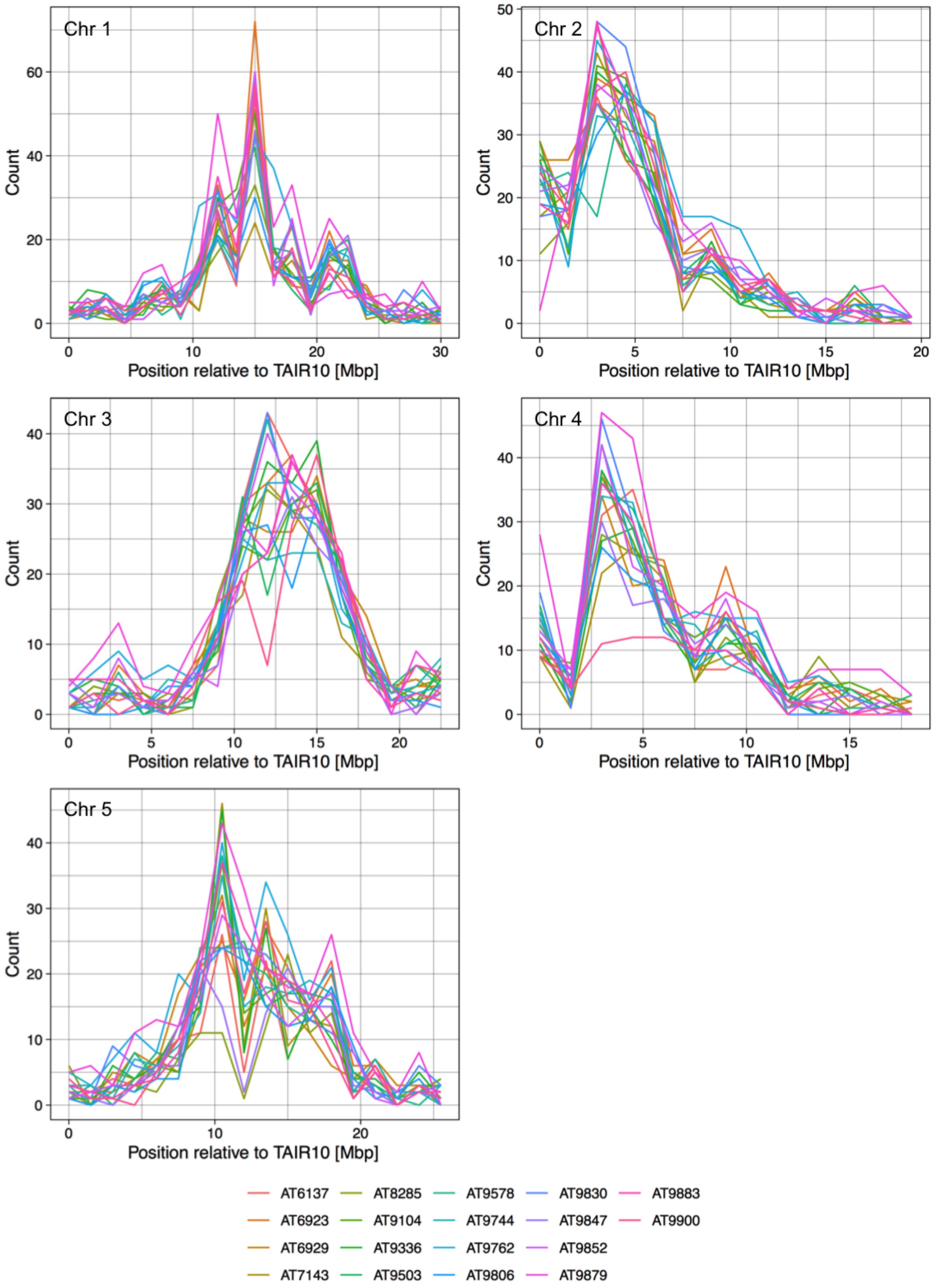

Figure 36: Frequency of unaligned sequence blocks in 1 Mb windows along the chromosomes of the reference TAIR10. Positions are counted relative to TAIR10. Unaligned sequence blocks were identified using SyRi. Colored lines are indicating the number of unaligned sequence blocks per 1 Mb window for a given genome assembly.

4.1.7 Transposable element and gene annotation

Transposable element (TE) annotation was performed prior to the gene annotation in order to avoid the false positive annotation of TEs as genes. TE annotation was done without distinguishing between centromeric and non-centromeric regions.

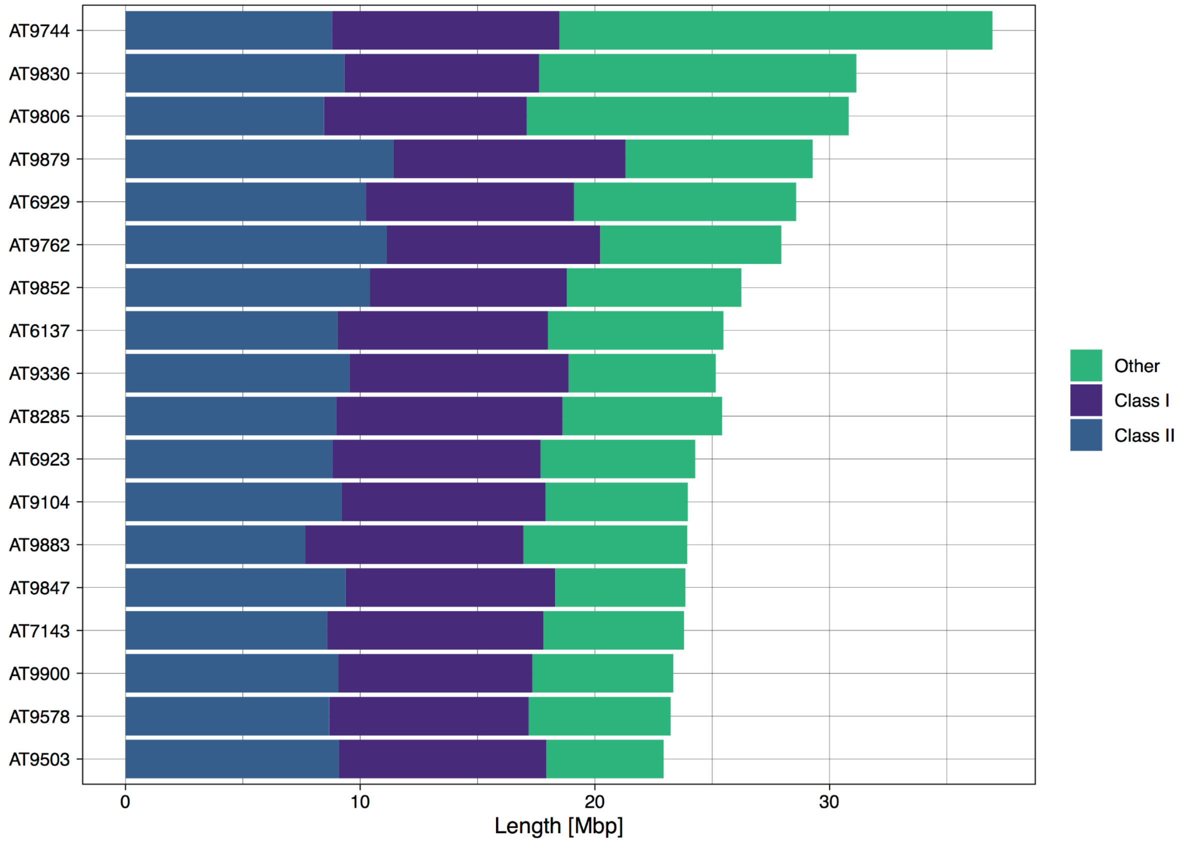

Figure 37: Barplot showing the cumulative length of transposable element sequences. Colored stacks represent cumulative length of class 1 TEs, class 2 TEs, and others. Sequences from the category 'others' were not further classifiable.

Input sequences were categorized into class I retrotransposons and class II DNA transposons using EDTA (Ou et al. 2019). Repeat sequences that could not be assigned to either class were instead labeled as 'other repeats'. The cumulative length of both TE classes and other repeats varied from 23 Mb in the genome of AT9503 to >35 Mb in the assembly of AT9744 (**Figure 37**). Thus, the total TE and repeat content accounted for 15% to 25% of the total genome size. TEs and other repeat sequences on average occupied 25 Mb per genome, which corresponded to an average 18% of the genome size. Class I and class II TEs each accounted for approximately 6.5% of the total genome size. Thus, class I cumulative length ranged from 8.3 Mb in the genome of AT9900 to 9.9 Mb in AT9879. Class II TEs occupied between 7.7 Mb in AT9883 and 11.3 Mb in AT9879 (**Figure 37**). This corresponded to a median cumulative length of 8.9 Mb for each of the TE classes per assembly.

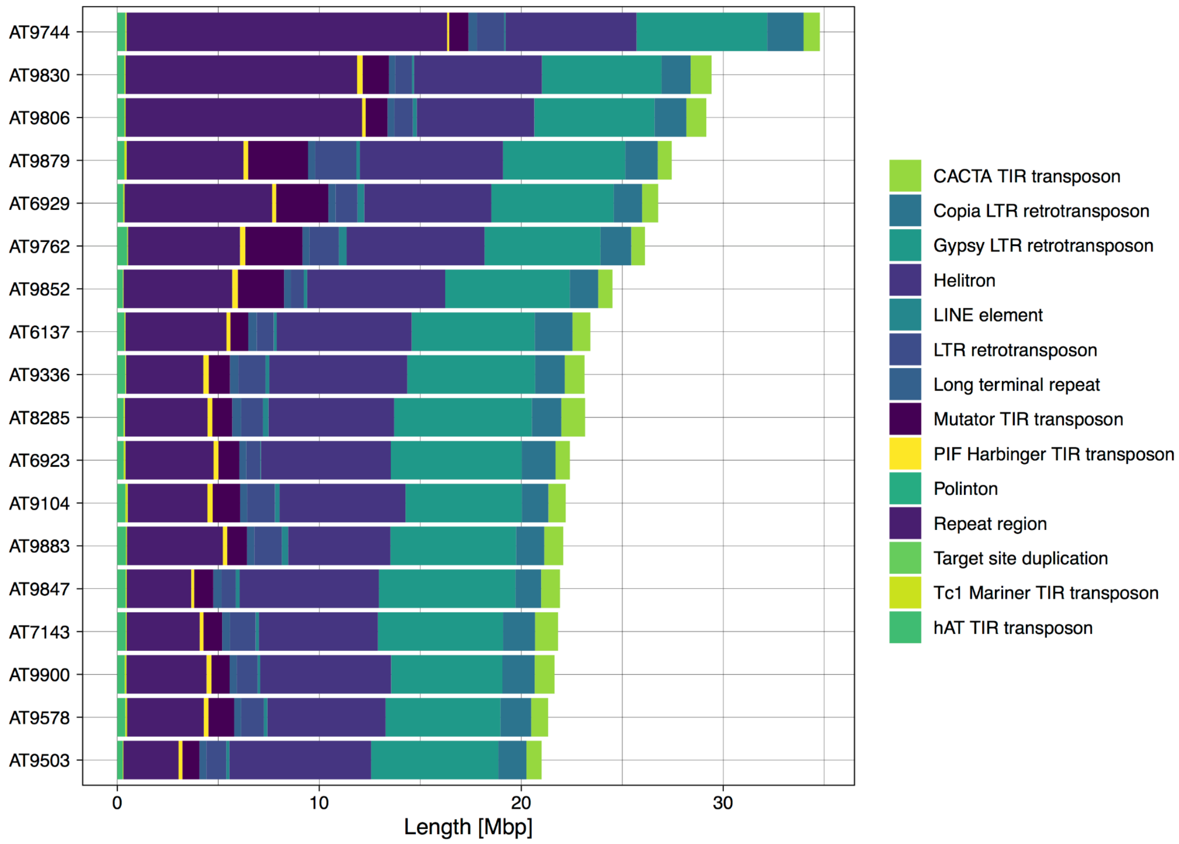

Figure 38: Cumulative length of transposable element families. Sequences from the category ‘repeat region’ were not further classifiable

Class I retrotransposons were further classified, if possible. Of the known class I retrotransposons, Copia LTRs, Gypsy LTRs and LINEs were detected. In addition, not further classifiable LTR retrotransposons were found. The total length of Copia LTR retrotransposons ranged from 1.3 Mb in AT9847 to 1.9 Mb in the genome of AT6137 (**Figure 38**). The median length of Copia LTR retrotransposons was 1.5 Mb. In contrast, the Gypsy LTR retrotransposons accounted for a median length of 6 Mb per assembly. The lowest content of Gypsy LTR retrotransposons was observed in the genome of AT9900 (5.4 Mb), while the highest content of such TEs was found in the assembly of AT8285 (6.7 Mb). Furthermore, I detected LINE retrotransposons that could not be assigned to a superfamily, as well as a small fraction of not further classifiable LTR retrotransposons. The cumulative length of LINE elements varied between 0.03 Mb in the genome of AT6923 and 0.4 Mb in the genome of A9762. LTR retrotransposons that could not be assigned to a superfamily accounted for 0.64 Mb (AT9852) to 1.9 Mb (AT9879) of sequence. The median total length of these not further classifiable retrotransposons was 0.9 Mb. Moreover, the cumulative length of additional LTRs ranged from 0.3 Mb in the genome of AT9830 to up to 0.5 Mb in the genome of AT8285.

Moreover I identified class II DNA transposons, such as Tc1-Mariner Terminal Inverted

Repeat (TIR), hAT TIRs, Mutator TIRs, PIF-Harbinger, CACTA TIR, and Helitrons (**Figure 38**). Tc1 Mariner TIR DNA transposons occupied between 0.03 Mb (AT9852) and 0.1 Mb (AT9879) of the sequence. The median cumulative length of Tc1 Mariner elements was 0.05 Mb per genome. The content of hAT TIR DNA transposons ranged from 0.3 Mb (AT9503) to up to 0.5 Mb (AT9762). In the eighteen genome assemblies, I detected between 0.7 Mb (AT9503) and 3 Mb (AT9879) of Mutator TIR transposon sequences (median 1 Mb). TEs in the PIF-Harbinger superfamily exhibited cumulative lengths between 0.01 Mb (AT9744) and 0.28 Mb (AT8285). The total length of CACTA TIR elements varied between 0.7 Mb in AT9762 and 1.2 Mb in AT8285. The median length of these elements was 0.9 Mb per assembly. By far the most abundant class II DNA transposons detected in the eighteen genome assemblies were Helitrons. Their cumulative length varied between 5.1 Mb in AT9883 and 7.1 Mb in AT9879. The median total length of Helitron elements was 6.5 Mb (**Figure 38**).

In addition to class I and class II TEs, other repeats were detected that could not be assigned to one or the other TE class. The genome fraction taken up by repeats belonging to neither of the two TE classes was more variable compared to the class I and class II TEs. Their cumulative length ranged from 5 Mb in AT9503 (3.6%) to 18 Mb in AT9744 (13.6%). It is noteworthy that ‘other repeats’ accounted for more than 9.5 Mb in three assemblies (AT9830, AT9806, and AT9744) (**Figure 37**). These other repeats were further classified into repeat region, long terminal repeat (LTR), and target site duplication (**Figure 38**). The accession AT8285 exhibited the greatest cumulative length of long terminal repeats at 0.5 Mb, while AT9830 had 0.3 Mb of LTR sequences. Target site duplications that were not already part of an annotated TE were the rarest class of repeat sequences identified in this analysis. They occupied between 1.8 kb in AT9806 and 2.4 kb in AT8285. Variation in the cumulative length of repeat regions among the eighteen accessions turned out to be the main contributor to the overall variation in repeat and TE content. Such repeat-rich regions constituted the most variable fraction of repeats annotated in this section. Their cumulative length varied by 13 Mb. The cumulative length of such repeat sequences varied between 2.6 Mb in AT9503 and 15.9 Mb in AT9744 (**Figure 38**). The median length of repeat-rich regions per genome was 4.6 Mb.

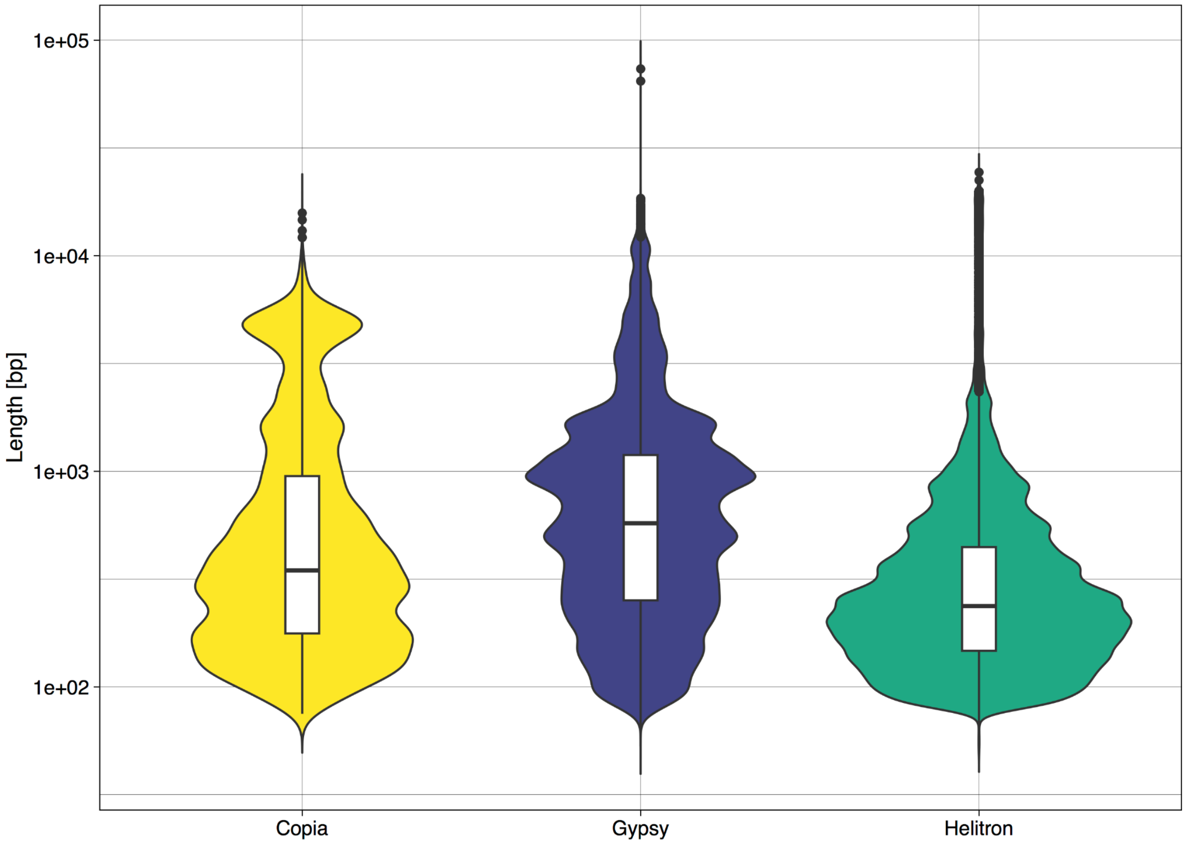

Figure 39: Violin plot showing the length distribution of different TE families in the 18 assemblies. Annotated TEs were not filtered for completeness. Black bars in boxplots represent the median length of the given TE family.

In addition to the total length occupied by a given TE, I also assessed the median length of the three most abundant TE families in all eighteen assemblies (**Figure 39**). TEs were not filtered for being intact since the major goal here was to mask all sequences of TEs prior to the gene annotation. Therefore, the TE length is likely to be underestimated. The comparison of all eighteen genotypes revealed that the median length of Gypsy elements ranged from 507 bp (AT9744) to 668 bp (AT7143) (**Table 8**). Copia elements were shorter compared to Gypsy elements. Their median length varied between 196 bp in AT9883 and 407 bp in AT9847. Helitron elements were even shorter, and their median length ranged from 227 bp in AT9744 to 249 bp in AT9852. Thus, Helitrons varied the least in length among the genotypes in this study. The median length of detected Gypsy elements across all accessions was 547 bp, of Copia elements 347 bp and Helitrons only 237 bp (**Figure 39**).

How the TE annotation results could be improved will be part of the discussion chapter. However, the main purpose at this point was to mask all TE sequences independent of the TE being intact. Thus, all TEs were masked prior to subsequent gene annotation.

Table 8: Median length (in bp) of different transposable element (TE) families. TEs annotated here are not necessarily intact.

Genotype	Median length (bp)		
	Copia	Gypsy	Helitron
AT6137	321	594	240
AT6923	326	642	235
AT6929	387	541	237
AT7143	337	668	231
AT8285	398	592	241
AT9104	368	612	239
AT9336	354	521	237
AT9503	350	650	237
AT9578	342	583	237
AT9744	339	507	227
AT9762	399	517	237
AT9806	353	512	227
AT9830	364	580	235
AT9847	407	573	244
AT9852	368	563	249
AT9879	376	517	243
AT9883	196	574	237
AT9900	357	616	233

4.1.8 Annotation of protein coding genes

The genome sequence of a species alone is of limited use if the encoded information is not deciphered (Mudge and Harrow 2016). Therefore, I performed independent whole genome annotation for all eighteen *de novo* assemblies. The *ab initio* gene predictions were done with accession-specific training parameters for Augustus (Stanke and Waack 2003) gene models as well as using a reference guided transmap (Shumate and Salzberg 2020a) as extrinsic evidence.

Transmap as extrinsic evidence for gene annotation

The transmap was generated by mapping the reference annotation TAIR10, which comprises 28,775 genes, onto each of the eighteen query assemblies. Results are shown in **figure 40**. The number of mapped genes, excluding multi-copies, ranged from 27,784 in AT9852 to up to 28,025 genes in AT9744. Thus, the number of unmapped genes only varied by 241 genes between the accession with the highest and the one with the lowest number of unmapped genes. On average it was possible to map 27,899 genes per assembly. The mapped genes were used as extrinsic evidence in the subsequent gene annotation approach.

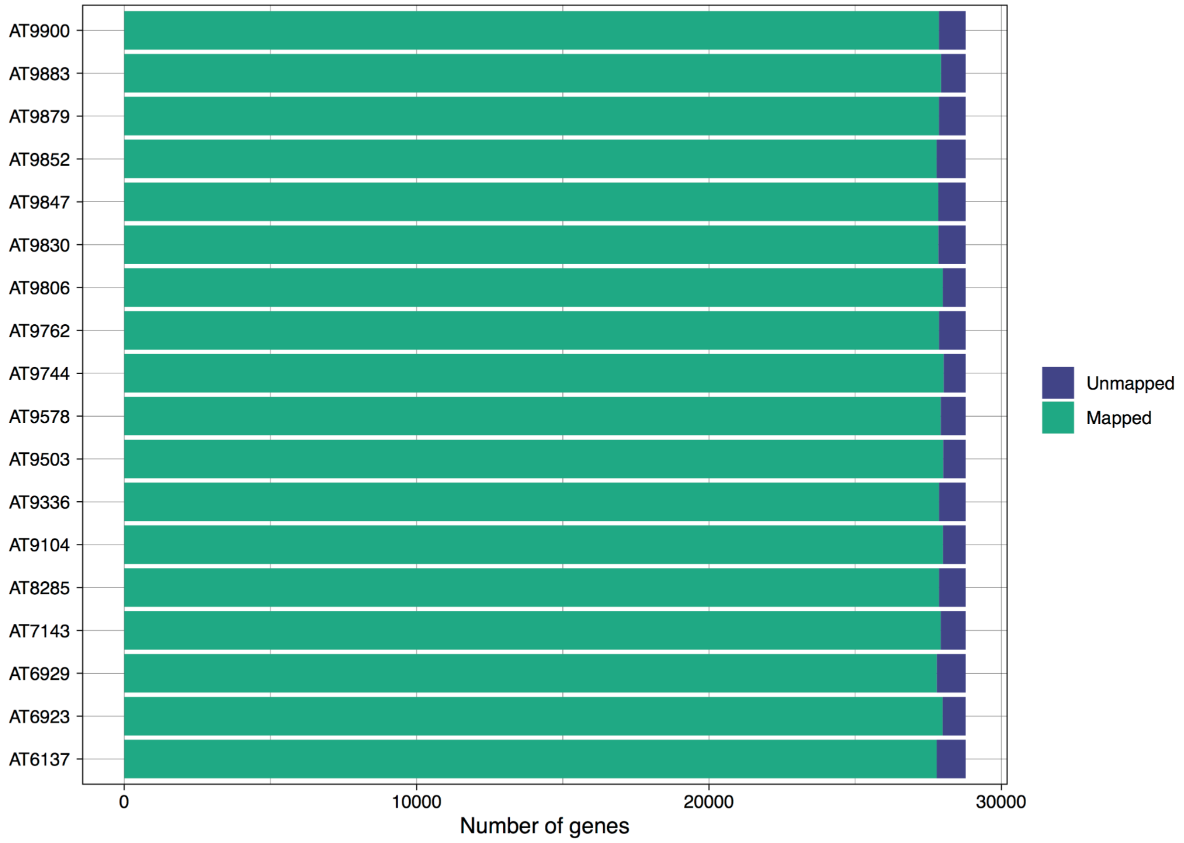

Figure 40: Number of reference genes that were used to generate the transmap evidence for subsequent gene annotation. Hints for annotations were only generated from mapped genes.

Augustus based gene prediction

Subsequent Augustus *ab initio* gene predictions were made using gene models generated for each assembly individually, while taking into account information obtained from the above described annotation liftover. The highest number of genes (32,798) was annotated in AT9336 (**Figure 41**), while the lowest number of genes was found to be 31,642 in AT9852. Thus, the total number of genes only varied by 1,156 between the accession with the lowest and the accession with the highest number of genes. The gene number in all assemblies was higher compared to the number of genes found in the reference genome annotation TAIR10 (Lamesch et al. 2012). However, in none of the assemblies was it possible to annotate as many genes as there are annotated in the more recent reference annotation Araport11 (33,341). This is most likely attributable to the fact that 113 RNA-seq datasets were used to generate 11 tissue specific transcriptome assemblies that were then further processed in order to build Araport11 (C.-Y. Cheng et al. 2017). In contrast to this my annotations were generated without RNA-seq evidence. An average of 32,154 genes was identified per accession. The average number of genes annotated on unplaced contigs was 101 per assembly (**Figure 41**). Thus, 99.7% of all annotated genes were located on one of the five pseudo-chromosomes. On average, most genes were annotated on chromosome one (7,182), followed by chromosome five (7,107) and three (6,632). The average number of genes found on chromosomes four (5,280) and two (5,257) were almost identical.

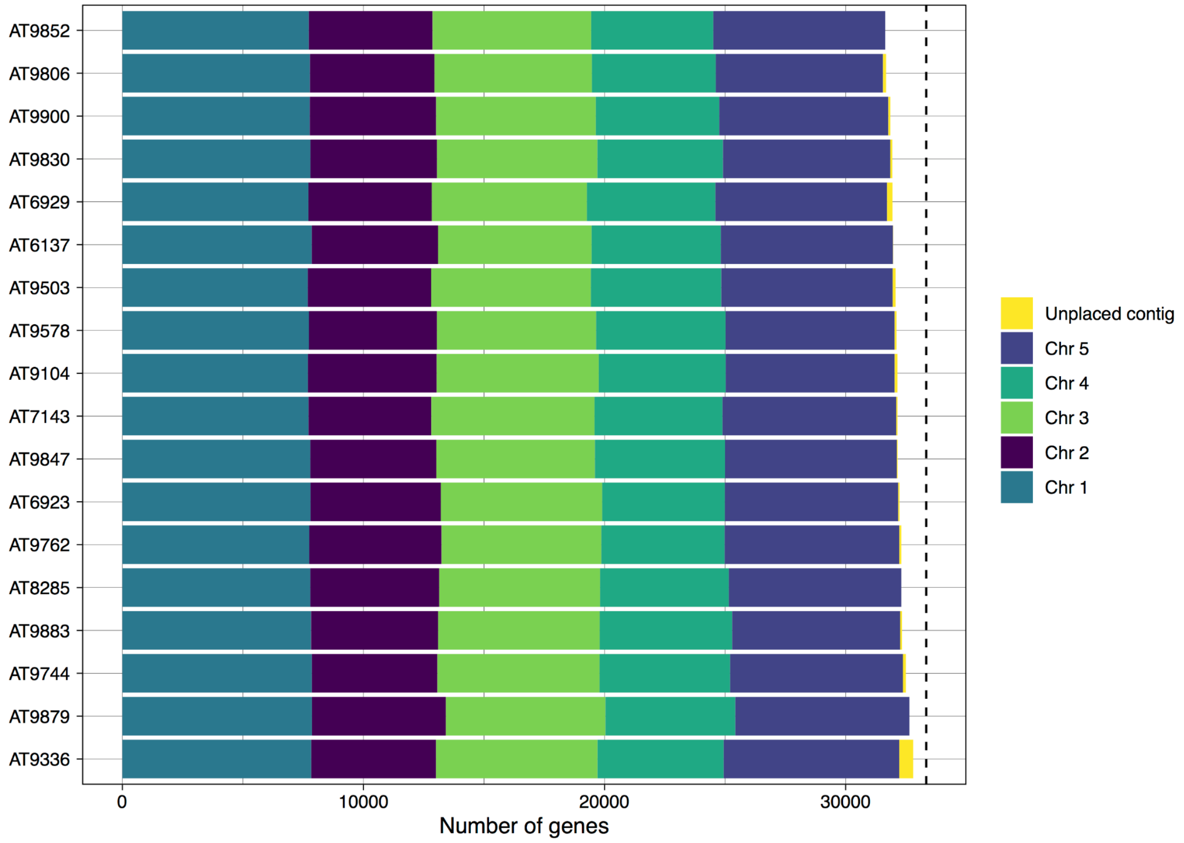

Figure 41: Final number of annotated genes per assembly. Colored stacks are representing chromosomes as well as unplaced contigs. The dashed black line shows the number of genes in the most recent *A. thaliana* reference annotation Araport11.

Chromosome 1 harbored between 7,693 genes in AT9503 and 7,868 genes in the assembly of AT9879. Thus, the accessions with the lowest and highest number only differ by 175 genes. In contrast, I found that the number of genes annotated on chromosome 2 varied between 5,088 in AT7143 and 5,548 in AT9879. The total number of genes detected on chromosome 3 differed by 402. Most genes on chromosome three were found in AT7143 (6,770) while the lowest number was detected in AT6137 (6,368). Chromosome 4 in the assembly of AT9852 had the lowest number of annotated genes, at 5,069. The same chromosome of AT9883 had 425 additional genes. Chromosome 5 harbored between 6,933 genes in AT9806 and 7,282 in AT9336. Thus, chromosome 2 was the most variable in terms of the detected number of genes whereas chromosome 1 showed the lowest variation in gene content among the eighteen accessions. The number of genes that was annotated on unplaced contigs showed even greater variation. Three out of eighteen assemblies did not have unplaced contigs. In the remaining fifteen assemblies, I annotated between nine (AT6137) and 572 genes (AT9336) on unplaced contigs. However, it is noteworthy that only six out of the eighteen assemblies had over 100 genes on unplaced contigs (AT9104, AT9503, AT9744, AT9806, AT6929, and AT9336).

The vast majority of all annotated genes is located on chromosomes rather than unplaced contigs (**Figure 41**). The total number of genes only varied by 1,156 genes between accessions. It was possible to annotate more genes as compared to the reference annotation TAIR10 (Lamesch et al. 2012) while not reaching the number of genes from the most recent *A. thaliana* reference annotation Araport11 (C.-Y. Cheng et al. 2017).

Annotation completeness

Similarly to the assembly completeness, I assessed annotation quality by using conserved protein coding genes. Annotation completeness was assessed by comparing all protein-coding genes to a BUSCO database of 1,614 conserved protein sequences (Seppey, Manni, and Zdobnov 2019). Thus, these BUSCO genes were used as a proxy to estimate the completeness of an annotation, the duplication rate, the rate of fragmented genes as well as the fraction of missing genes. The fraction of single copy BUSCO genes was equal to or above 98.4% in all annotations, with an average of 98.6%. Duplication levels varied between 0.8% in AT9900 and 1.1% in AT9336 and AT9879, respectively. The average duplication rate was 0.95%. The portion of fragmented BUSCO genes ranged from 0.1% (AT9762 and AT9852) to 0.4% (AT9503 and AT9578). On average, I found that 0.2% of the 1,614 BUSCO genes were fragmented. The highest rate of missing BUSCO genes was detected in AT9847 (0.4%), whereas in six accessions only 0.2% of all BUSCO genes were missing (AT8285, AT9336, AT9503, AT9578, AT9744, and AT9900).

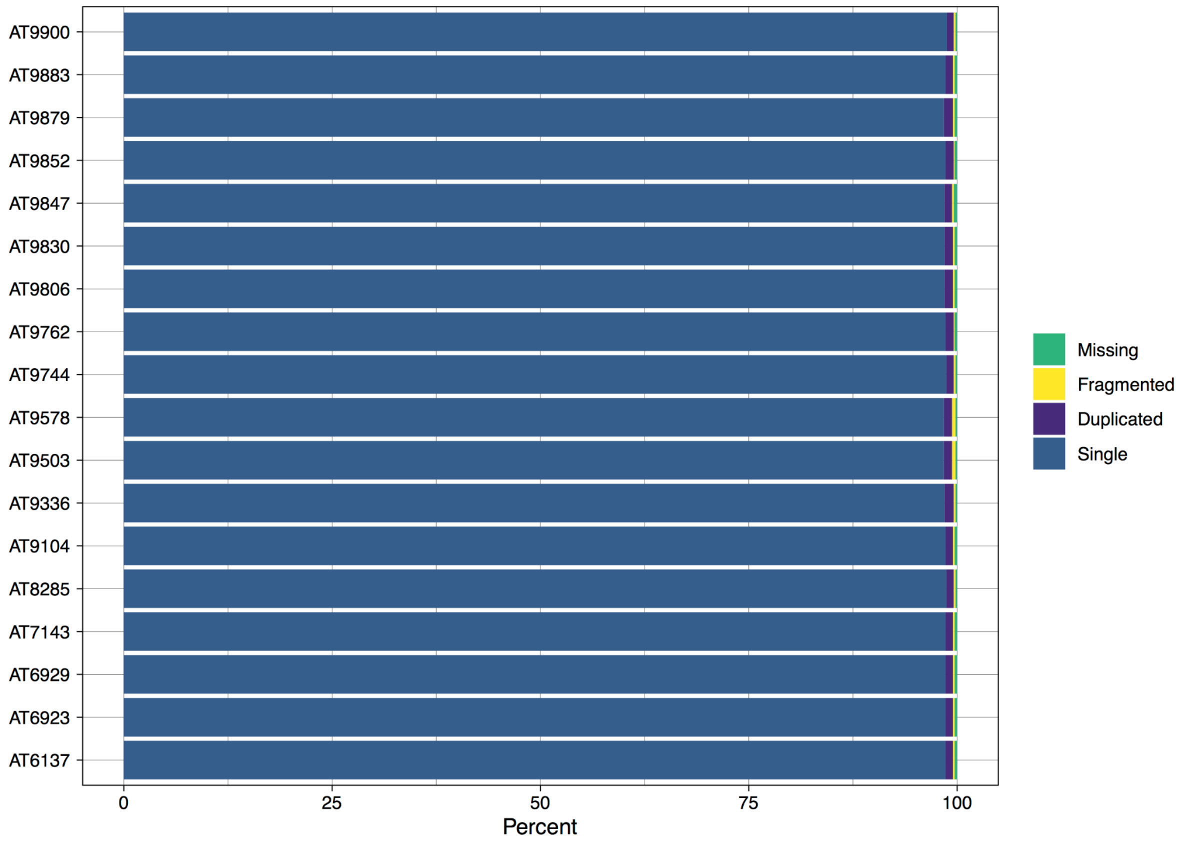

Figure 42: Completeness of protein coding gene annotation. Completeness was inferred by scanning the assemblies for the presence of a set of conserved genes (BUSCOs). All samples exhibit completeness levels beyond 98 %.

Next I wanted to investigate if the same set of BUSCO genes is missing, duplicated or fragmented in multiple accessions (**Table 9**). On average, the eighteen annotations were missing 0.26% of all BUSCO genes. Thus, the total number of different BUSCO genes that were missing in any eighteen annotations was only seven. Four out of these seven missing BUSCO genes were not detected in any of the annotations, while three BUSCOs were only absent in one to three of the annotations. Thus, all annotations showed sufficient completeness for further downstream analyses.

Table 9: Summary of duplicated, fragmented, or missing BUSCO genes. The number of affected accessions shows in how many of the annotations the given BUSCO was found to be duplicated, fragmented, or missing.

BUSCO ID	Status	No. of affected accessions
103418at3193	duplicated	15
103458at3193	duplicated	18
145760at3193	duplicated	12
147386at3193	duplicated	18
161537at3193	duplicated	16
167028at3193	duplicated	9
183573at3193	duplicated	15
189672at3193	duplicated	1
202147at3193	duplicated	18
212908at3193	duplicated	17
217208at3193	duplicated	16
25535at3193	duplicated	17
35121at3193	duplicated	15
40592at3193	duplicated	1
4115at3193	duplicated	10
44817at3193	duplicated	2
49808at3193	duplicated	1
56690at3193	duplicated	1
58451at3193	duplicated	3
59251at3193	duplicated	1
60794at3193	duplicated	4
69847at3193	duplicated	17
7552at3193	duplicated	4
85636at3193	duplicated	18
87297at3193	duplicated	12
92128at3193	duplicated	1
9250at3193	duplicated	14
152036at3193	fragmented	2
159342at3193	fragmented	15
167289at3193	fragmented	3
168457at3193	fragmented	18
174147at3193	fragmented	18
174618at3193	fragmented	1
200083at3193	fragmented	6
28151at3193	fragmented	1
45227at3193	fragmented	1
87393at3193	fragmented	1
111086at3193	missing	18
136033at3193	missing	18
14163at3193	missing	18
180289at3193	missing	3
186149at3193	missing	2
66400at3193	missing	1
79at3193	missing	18

Ortholog assignment

For downstream analysis such as copy number variation of NLR genes, it is relevant to know whether any of the *de novo* assemblies has orthologs in the TAIR10 reference annotation. Therefore, all putative genes were translated into protein sequences *in silico*. The longest predicted protein sequence was chosen for each locus. Orthofinder (Emms and Kelly 2019) was used to assign into orthogroups the genes of all eighteen accessions, as well as those of the reference genome TAIR10 and a previously published *Arabidopsis arenosa* annotation (Barragan et al. 2021).

99.5% of all genes (including TAIR10 and *A. arenosa*) were assigned to a total number of 31,839 orthogroups. The median orthogroup size was 20, while the average orthogroup size was 19.7. This means that genes from 20 genomes (18 assemblies + TAIR10 + *A. arenosa*) were found in at least half of the orthogroups. Moreover, I found 14,750 single copy orthogroups, meaning that each genotype contributed exactly one gene to the respective orthogroup. **Figure 43** shows that the majority of all orthogroups was composed of genes from all twenty input annotations. A total of 17,400 orthogroups contained at least one gene from each of the input annotations. Moreover, 5,006 orthogroups contained genes from nineteen input annotations, while 1,378 orthogroups comprised genes from eighteen accessions. In contrast, 446 orthogroups contained genes from a single input annotation. *A. arenosa* served as an outgroup in the orthofinder analyses. Therefore, the *A. arenosa* annotation was not taken into account for downstream analyses such as private gene counts or NLR copy number variation.

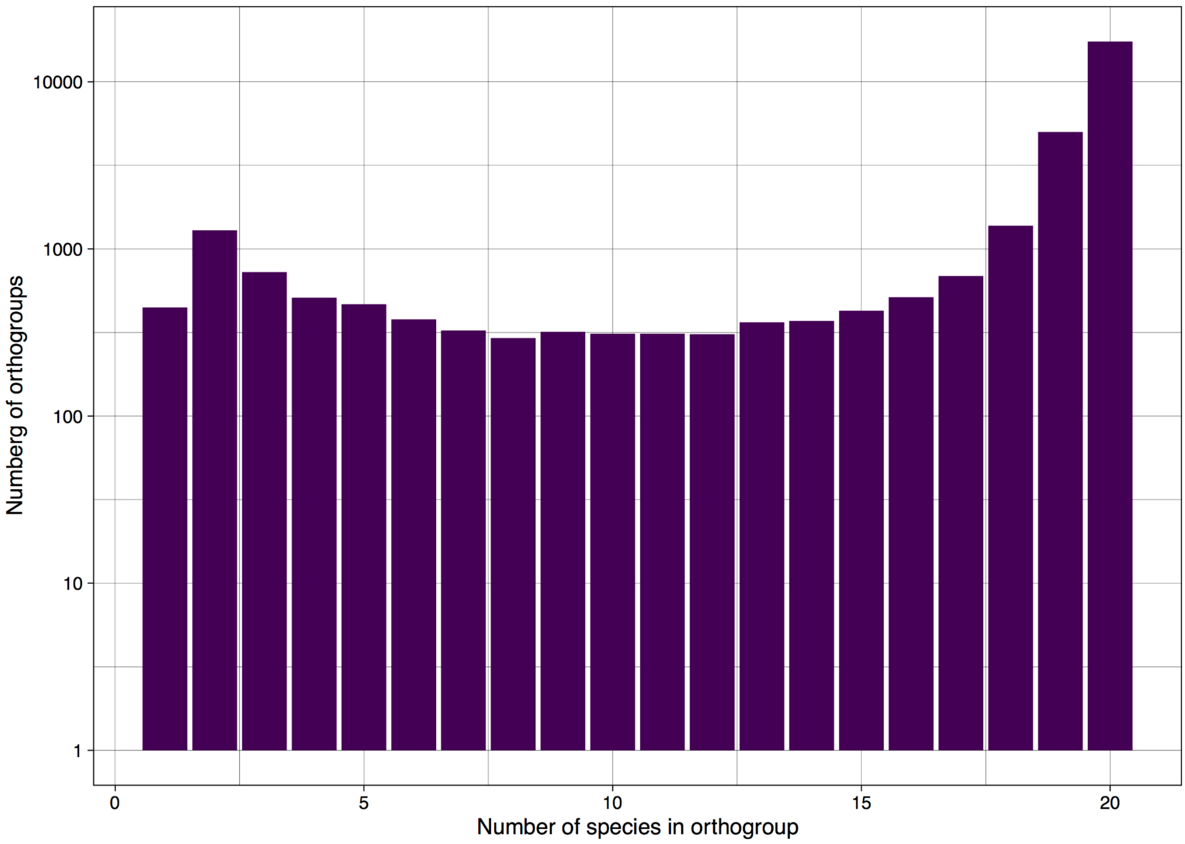

Figure 43: Number of orthogroups that contained genes from a given number of different input genomes (species). *A. arenosa* was used as an outgroup. Moreover, the TAIR10 annotation has been included in addition to the 18 differential lines.

Private genes in the eighteen assemblies

Having assessed how many genes are shared among the different input assemblies, I was able to count the number of genes that can only be found in a single genome. I will be subsequently referring to these genes as private genes.

The number of private genes varied greatly. The lowest number was observed in AT6923 (39), but 310 private genes in AT9336, with a median of 74, and only four accessions having more than 100 private genes (**Figure 44**).

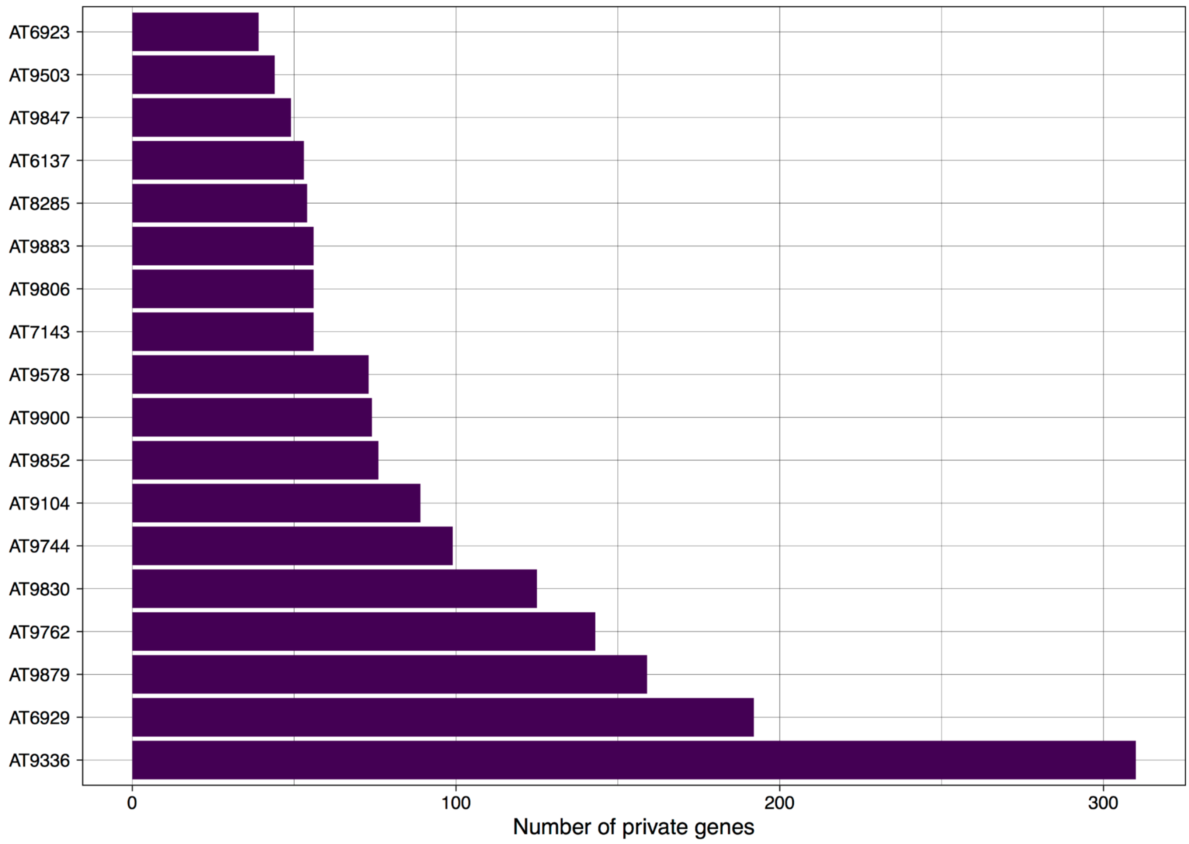

Figure 44: Variation in the number of private genes. Genes were considered as 'private' if no orthologs were found in any of the other input genomes.

Subsequently, I compared the amino acid length of private genes with the length of genes that are shared among all eighteen accessions and the reference annotation TAIR10 (**Figure 45**). This showed that the median length of private genes differed from the length of genes shared among all input annotations. The median length of genes that could be detected in all annotations was 343 amino acids. In contrast, the median amino acid length of private genes was 134. Thus, private genes were generally shorter compared to shared genes. It could be that these genes are either artifacts or novel genes. However, this is going to be the subject of the discussion chapter.

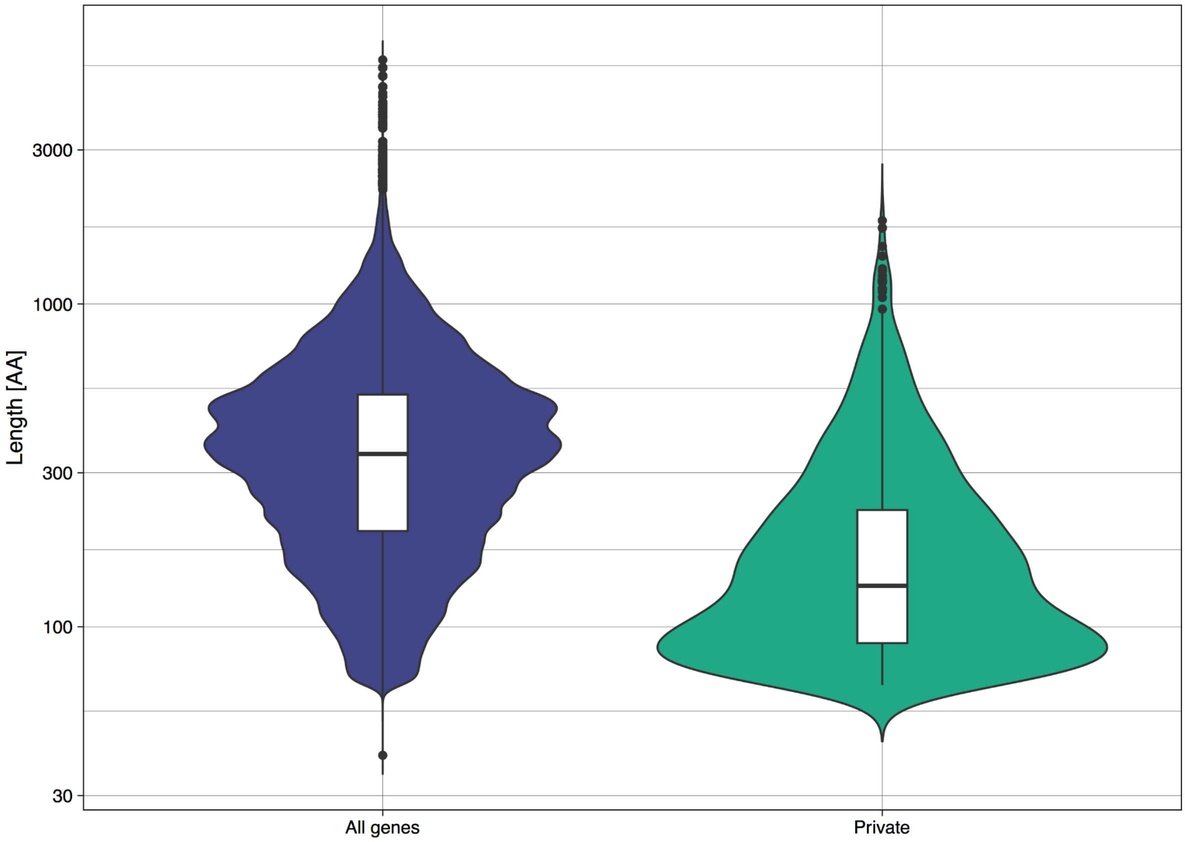

Figure 45: Comparison of the length distribution of all shared genes vs private genes in the 18 annotations. Boxplots are indicating the median value for the given gene type. Genes were considered private if they were only annotated in a single accession while shared genes were found in all 18 genomes.

In the next step, I subjected all genes to a blastp (Altschul et al. 1990) search against a database I generated by using all *A. thaliana* genes known from TAIR10. Subsequently, the distribution of e-values as well as the median e-value were compared between private and shared genes (**Figure 46**). The median e-value of genes shared among all eighteen accessions and TAIR10 was 0 (average 0.0049). In contrast, for genes only found in a single accession, a median e-value of 0.14 (average 1.23) was observed. Thus, the vast majority of private genes did not have a significant hit in the reference annotation TAIR10, and can therefore truly be considered to be private genes.

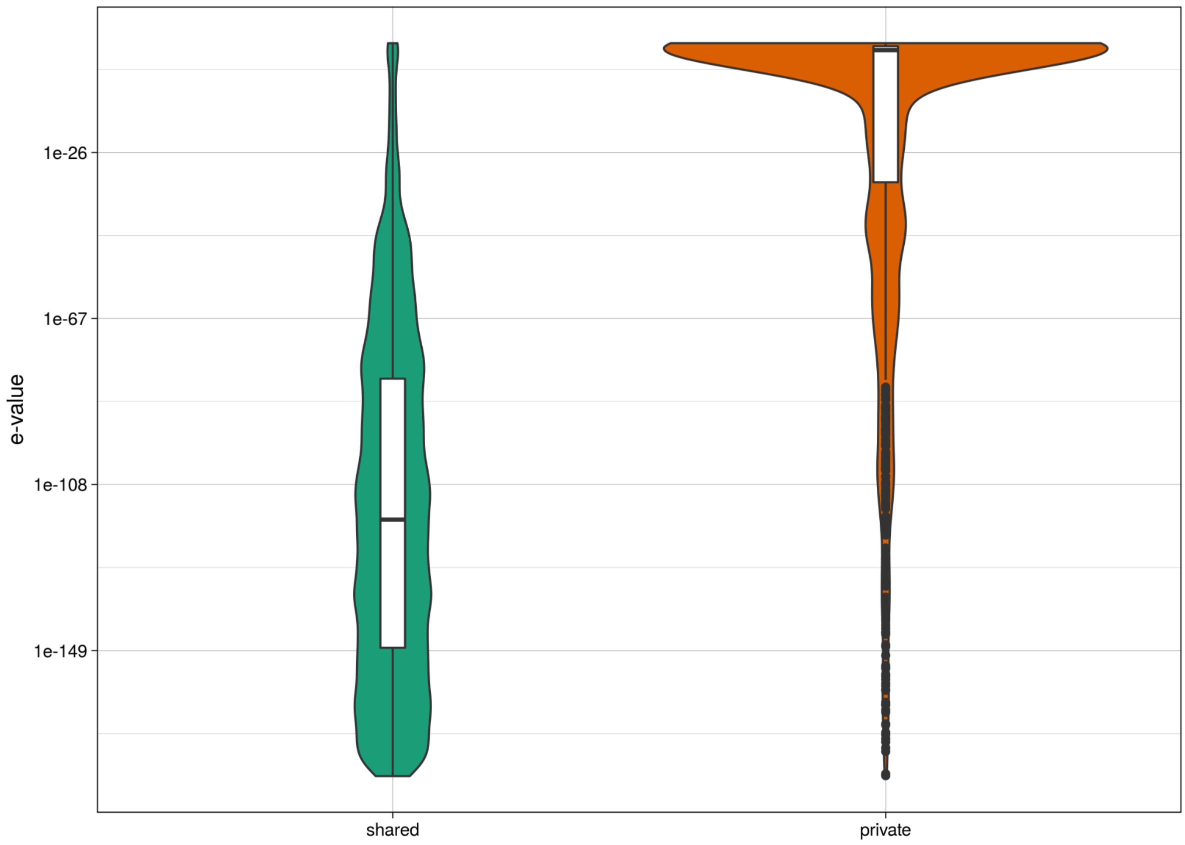

Figure 46: Violin plot showing the distribution of expected (e) values after a blastp analysis of private and shared genes. Boxplots are indicating the median value obtained from the 18 annotations.

Z-scores of orthogroup sizes

Calculating the z-score of a given trait is a simple way to get an overview of which values are ‘typical’ for a dataset and which ones are not. Therefore, z-scores of gene copy numbers were calculated for all orthogroups in order to gain an insight into how variable orthogroups can be among all eighteen accessions (**Figure 47**). An orthogroup in this case is defined as the set of genes from multiple species descended from a single gene in the last common ancestor of that species (Emms and Kelly 2019).

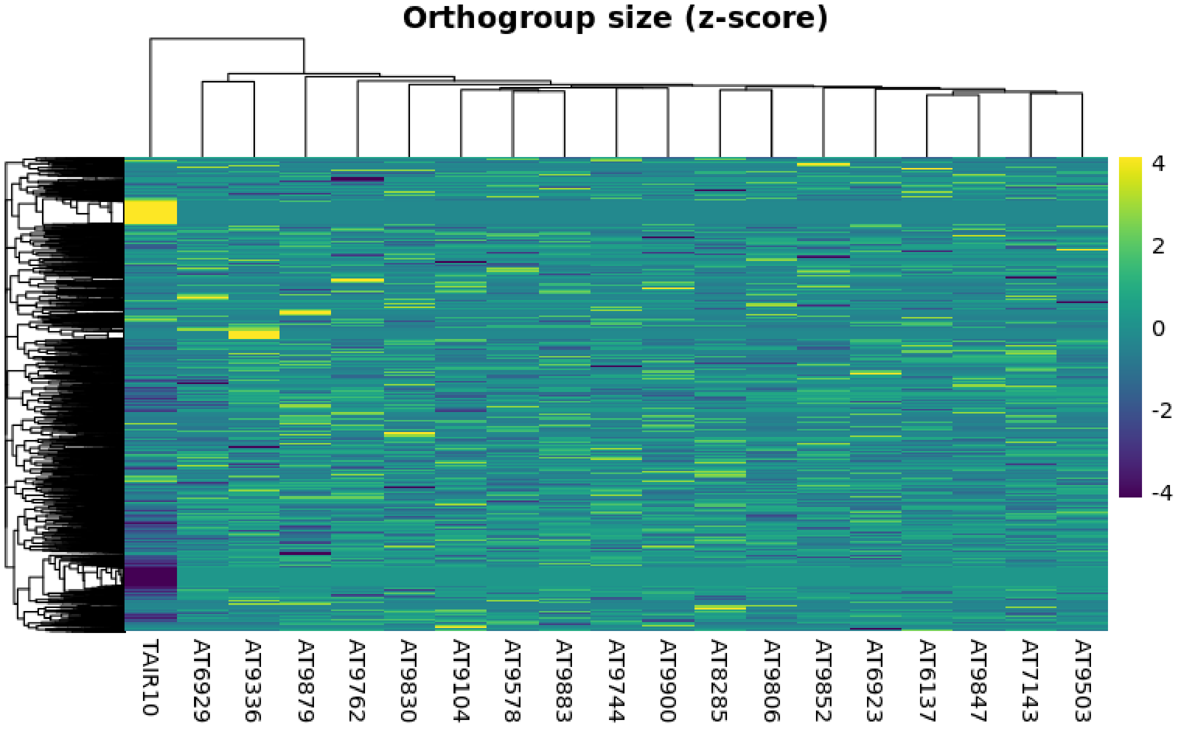

Figure 47: Heatmap showing the z-scores of orthogroups sizes for the 18 accessions and the reference TAIR10. Colors are indicating if an accession has a higher (bright colors) or a lower (dark colors) gene copy number in a given orthogroup.

The z-score analysis revealed that there are orthogroups which are strongly enriched for TAIR10 reference genes (**Figure 47**; bright yellow) while other orthogroups have a lower number of reference genes (dark blue).

Thus, I filtered out all orthogroups where the z-score of TAIR10 was ≥ | ±4 |. In total there were 530 orthogroups where the z-score observed for TAIR10 was greater than or equal to 4, while 451 orthogroups had a TAIR10 z-score below or equal to -4. Subsequently, it was analyzed if genes that are underrepresented in the 18 annotations are associated with specific functions. Therefore, TAIR10 gene names from these filtered orthogroups were used in a gene ontology (GO) analysis (Z. Du et al. 2010) with the TAIR10 reference as a background.

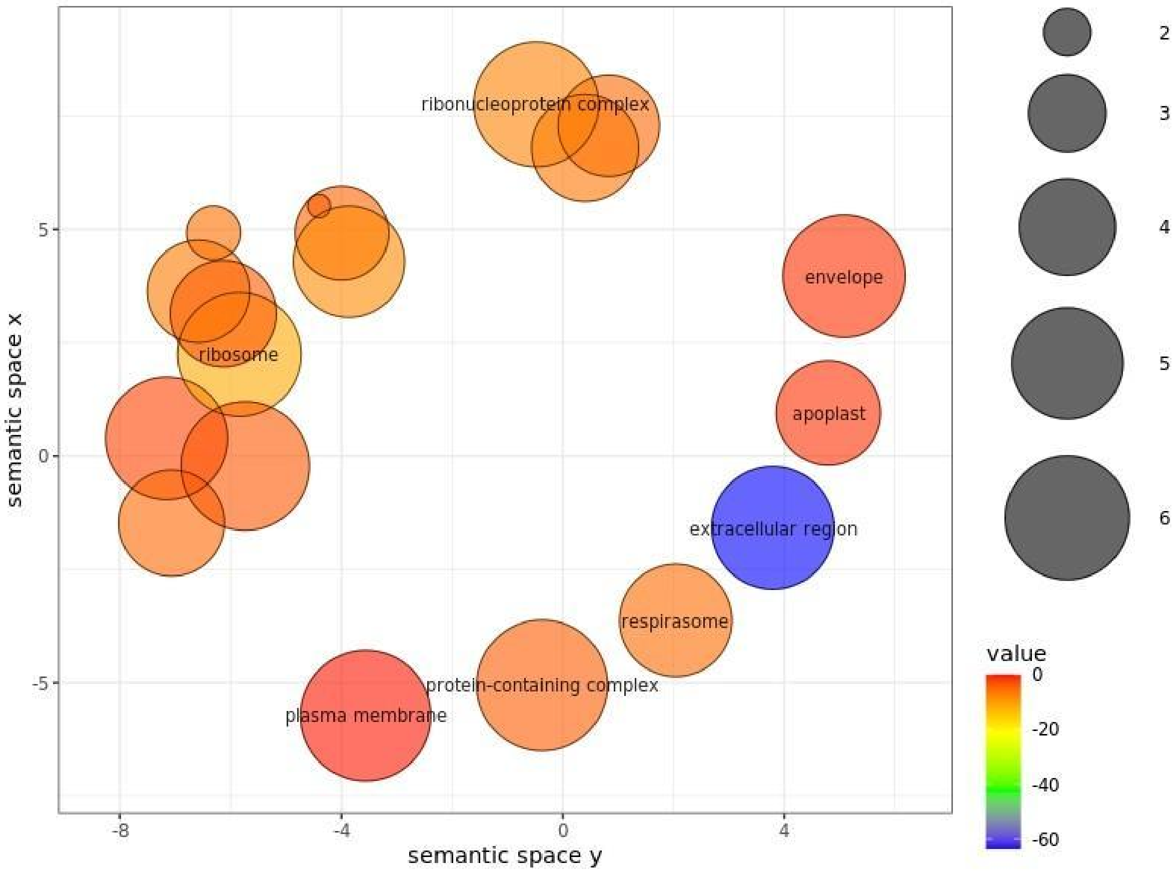

Figure 48: Gene ontology analysis of genes that were underrepresented in the 18 accessions as compared to TAIR10.

GO analysis revealed that orthogroups with TAIR10 z-scores of above or equal to four were significantly enriched for cellular functions associated with ribosomes, respiratory chain, mitochondria, and chloroplasts (**Figure 48**). Next I used only those genes that are overrepresented in TAIR10 compared to the eighteen annotations (z-score > 4). A total of 59 significantly enriched GO terms was identified. Thirteen out of these are associated with 'ribosome' while another thirteen are associated with 'mitochondrium' and 'respiratory chain'. Given the fact that I did not assemble chloroplast or mitochondrial genomes and that I masked rDNA sequences prior to gene annotation, it is likely that these genes are underrepresented in the annotations of the eighteen accessions. Genes that were found in the eighteen accessions but absent from TAIR10 could not be subjected to GO enrichment, since there were no TAIR10 gene identifiers. However, many of these genes were rather short, indicating that they are split genes or annotation artifacts. Further explanation will follow in the discussion section.

Gene copy number variation compared to TAIR10

After orthogroup assignment, I set out to analyze variation in gene copy number between each of the eighteen assemblies and the reference TAIR10 individually.

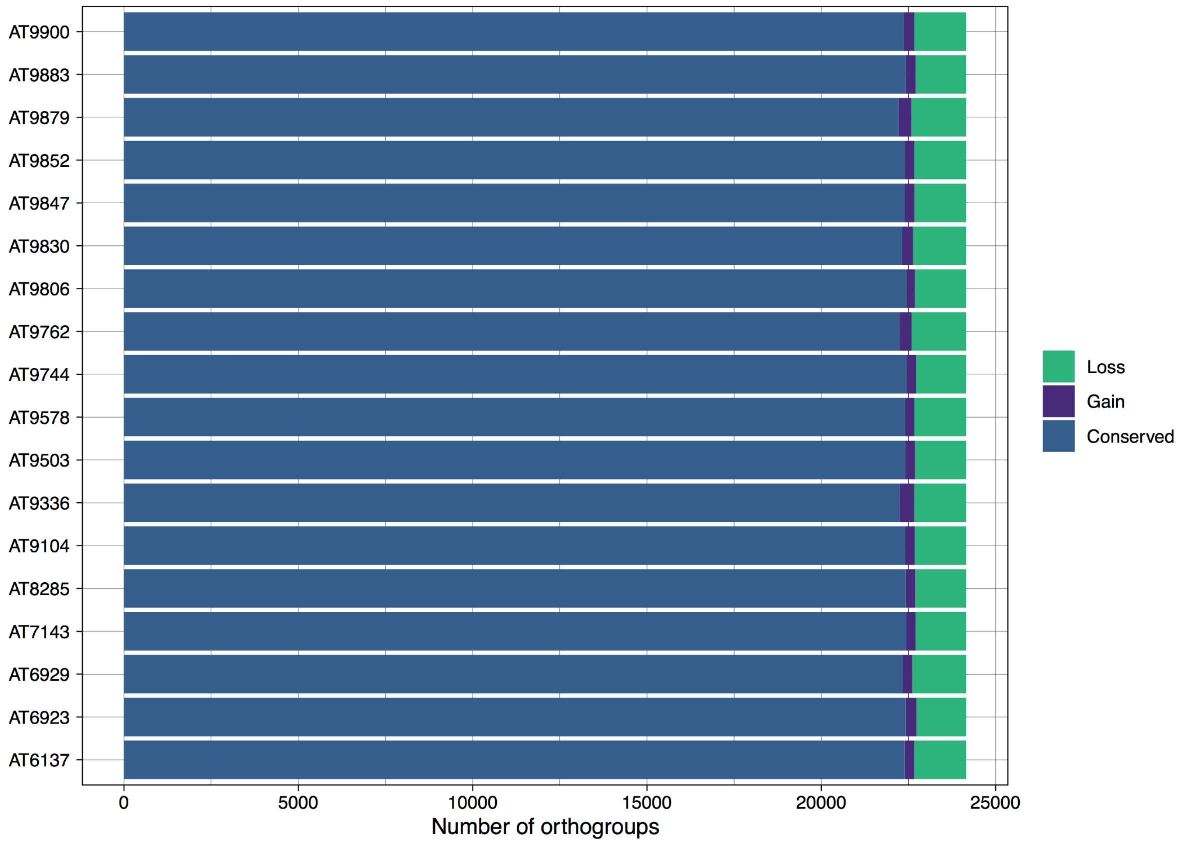

Figure 49: Gene copy number variation between the 18 accessions and TAIR10. Orthogroups (OG) were considered ‘conserved’ when the OG contained equal numbers of genes from both TAIR10 and the given input assembly. Gain means ‘TAIR10 < input assembly’ while loss was considered when ‘TAIR10 > input assembly’.

First, I filtered for orthogroups that contained at least one gene from the reference annotation TAIR10. Subsequently, for each of these orthogroups the number of genes from TAIR10 was compared to the number of genes of a given input assembly assigned to the same orthogroup (**Figure 49**). Orthogroups were then categorized into loss (TAIR10 > input assembly), gain (TAIR10 < input assembly), and conserved (TAIR10 = input assembly). The highest number of 'gain' orthogroups was observed in AT9879 with a total of 1,573 genes. On average, every accession had 1,490 orthogroups with gene copy loss. The number of orthogroups with a gain in gene copy number was generally lower compared to those with gene loss. On average, each accession had 296 orthogroups with a copy number gain. The lowest number of such orthogroups was observed for the assembly of AT9806 (242), while AT9336 had the highest number of orthogroups with a gain in gene copy number (420). However, the vast majority of orthogroups showed a conserved number of genes as compared to TAIR10. The number of these conserved orthogroups varied between 22,213 (AT9879) and 22,442 (AT9744) with an average of 22,371. In summary, most TAIR10 orthologs were present in a conserved copy number, and there were more orthogroups with copy number loss than with copy number gain.

NLR copy number variation in the 18 accessions

After ortholog assignment, I set out to gain insight into copy number variation of the NLR genes known from the TAIR10 reference. Therefore, the size of each orthogroup containing one of the 160 known (Baggs et al. 2020) NLR genes from TAIR10 was compared among all eighteen annotations and TAIR10. 128 out of the 160 known resistance genes were found in at least two assemblies (**Figure 50**). Genes were further categorized according to their domain structure as coiled-coil nucleotide-binding site NLRs (CC-NBS), nucleotide-binding site leucine-rich repeat NLRs (NBS-LRR), coiled-coil nucleotide-binding site leucine-rich repeat NLR (CC-NBS-LRR), toll-interleukin receptor NBS-LRR (TIR-NBS-LRR), and RPW8-NBS-LRR (Van de Weyer et al. 2019; Blake C. Meyers, Morgante, and Michelmore 2002).

Five of the six CC-NBS genes were found as single-copy genes in all eighteen input assemblies (**Figure 50A**). Only AT9762 lacked an ortholog of the CC-NBS NLR AT1G50180 *(CAR1*).

In contrast, the 24 known CC-NBS-LRR NLR genes were more variable in their copy number (**Figure 50B**). *RFL1* (AT1G12210) had a conserved copy number of one in all but three genomes. AT9900, AT9762, and AT9744 had three copies of *RFL1* each. The copy number of AT1G12220 (*RESISTANT TO P. SYRINGAE 5; RPS5*) varied between five (AT6137, AT8285, AT9847, and AT9852) and nine in AT9503. In former studies such presence absence variation has been reported for *RPS5* (Gao, Roux, and Bergelson 2009). The following CC-NBS-LRR NLR genes had a conserved copy number of one in all eighteen genomes, as well as TAIR10: AT1G12280 (*Suppressor of MKK1 MKK2 2; SUMM2*), AT1G51480 (*RESISTANCE SILENCED GENE 1*; *RSG1*), AT1G53350, AT3G14470 (*LEUCINE RICH REPEAT PROTEIN 1, LRR4*), AT3G50950 (*HOPZ-ACTIVATED RESISTANCE 1*; *ZAR1*), and AT5G47250. All other CC-NBS-LRR NLR genes exhibited copy number variation among the eighteen assemblies and/ or TAIR10. AT1G15890 (*L3*) was present in two copies in all but four genomes (TAIR10: 1; AT6929: 1; AT9503: 1; AT9578: 1). AT1G58390 was present in two copies in all but five genomes. TAIR10, AT6929, and AT8285 had three copies while AT7143 and AT9744 had one copy of AT1G58390. Copy numbers of AT1G58602 (*RECOGNITION OF PERONOSPORA PARASITICA 7*; *RPP7*) varied between one and five. Similarly, AT1G59780 copy numbers ranged from one to four with a median of one per assembly. In contrast, AT1G61190 was only detected in three of the eighteen accessions (AT8285, AT9762, and AT9847) with one copy each, and in TAIR10 with one copy as well. AT1G62630 was detected in all genomes. However, its copy number ranged from one to five (AT9744), with a median of three. AT1G63350 was only found in seven genomes (TAIR10, AT6929, AT8285, AT9503, AT9762, AT9806, and AT9883). All of these genomes had exactly one copy of the gene, except for AT9883, which had three copies. Eight of the genomes had two copies of AT3G46730, while all other genomes, including TAIR10, only had one copy of the gene. In the case of AT4G26090 (*RESISTANT TO P. SYRINGAE 2*; *RPS2*) the gene was found to have a conserved copy number of one in all assemblies except for AT9900, which had two copies. Similarly, all genomes had one copy of AT4G27190, except for AT9879 and AT9762, which had two copies each. AT4G27220 was only detected in AT6923, AT7143, AT9104, AT9336, AT9578, AT9744, AT9762, AT9806, AT9847, AT9852, and TAIR10. In these genomes, I annotated exactly one copy of AT4G27220. Similarly, AT5G05400 was only found in ten out of the eighteen accessions. When present, there was only one copy. AT5G35450 was annotated in all eighteen genomes. Its copy number varied between three and five, with a median of four. The *RESISTANCE SILENCED GENE 2* (*RSG2*; AT5G43730) was found in all assemblies in either one or two copies. The copy number of AT5G47260 was observed to be the same among all accessions except for AT9762 where it had two copies instead of one. Similarly, *SUT1* (*SUPPRESSOR OF TOPP4-1*; AT5G63020), had a copy number in all but two assemblies. In AT9578 and AT9744, two copies of AT5G63020 were annotated each.

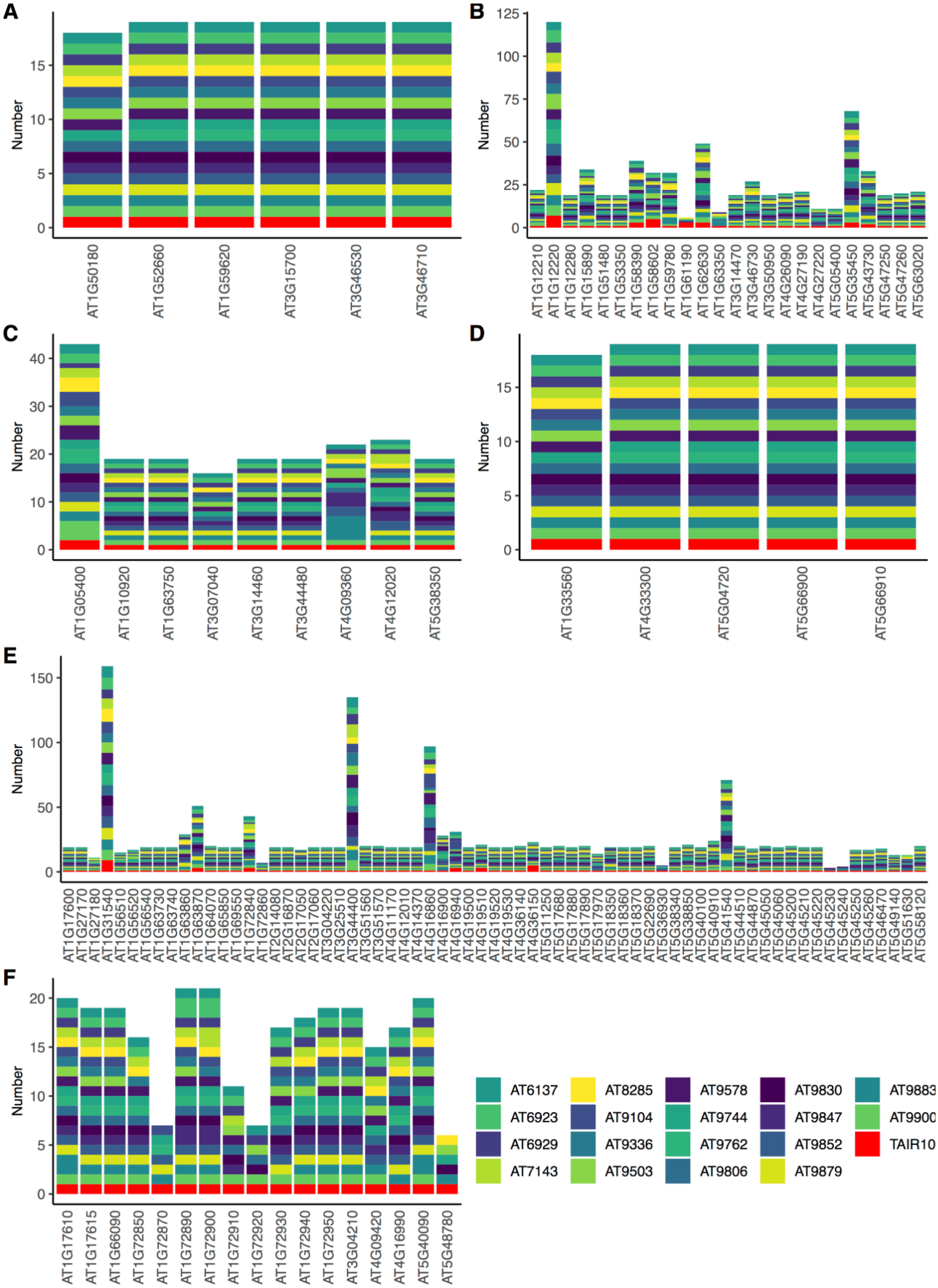

Figure 50: Stacked bar plots showing copy number variation of different NLRs. Copy number of reference (A) CC-NBS NLRs, (B) CC-NBS-LRR NLRs, (C) NBS-LRR NLRs, (D) RPW8-NBS-LRRs, (E) TIR-NBS-LRR NLRs, and (F) TIR-NBS NLRs is plotted on the y-axis. Colored stacks represent the 18 annotations.

All but two (AT3G07040, and AT4G09360) of the known NBS-LRR NLR genes were detected in all eighteen assemblies (**Figure 50C**). For AT1G61190 I annotated one copy each in the assemblies of AT8285, AT9762, and AT9847. In contrast, AT3G07040 (*RESISTANCE TO P. SYRINGAE MACULICOLA 1*; *RPM1*) was detected in fifteen assemblies with a conserved copy number of one per genome. *RPM1* was absent in the annotations of AT7143, AT9762, and AT9830. In the case of *RPM1*, such presence absence variation (PAV) has been described before (Grant et al. 1998). AT4G09360 was found in thirteen assemblies and TAIR10 with a copy number ranging from one to five. Four assemblies had more than one copy. The assemblies of AT9503 and AT9852 had two copies each, while AT9847 had three copies. AT9900 had five copies of AT4G09360. The copy number of AT1G05400 varied between one and four with a median of two. The following NBS-LRR NLR genes had a conserved copy number of one in all eighteen assemblies: AT3G44480 (*RPP1-WsA*), AT1G63750, AT3G14460 (*LEUCINE-RICH REPEAT PROTEIN 1*; *LRRAC1*), and AT5G38350. In contrast, the copy number of AT4G12020 (*WRKY19*) varied between one and two. Four accessions (AT7143, AT9744, AT9847, and AT9852) had two copies of *WRKY19*.

Most RPW8-NBS-LRR NLR genes were found in all eighteen assemblies: AT4G33300 (*ADR1-L1*), AT5G04720 (*ADR1-L2*), AT5G66900 (*N REQUIREMENT GENE 1.1; NRG1.1*), and AT5G66910 (*NRG1.2*). *ADR1* (*ACTIVATED DISEASE RESISTANCE 1*; AT1G33560) was absent in the assembly of AT9762 (**Figure 50D**).

Out of the 67 known TIR-NBS-LRR genes, 46 were found in all eighteen accessions (**Figure 50E**). The following TIR-NBS-LRR genes were detected in all assemblies with a conserved copy number of one: AT1G17600 (*SOC3*), AT1G27170, AT1G56540, AT1G63730, AT1G63740, AT1G65850, AT1G69550, AT2G14080, AT2G16870, AT2G17060, AT3G04220, AT3G25510, AT4G11170 (*RESISTANCE METHYLATED GENE 1*; *RMG1*), AT4G12010 (*DOMINANT SUPPRESSOR OF CAMTA3*; *DSC1*), AT4G14370, AT4G19500, AT4G19510 (*RPP2B*), AT4G19520, AT4G19530, AT4G36140, AT4G36150, AT5G11250 (*BURNOUT1*), AT5G17880 (*CONSTITUTIVE SHADE AVOIDANCE 1*; *CSA1*), AT5G17890 (*CHILLING SENSITIVE 3*; *CHS3*), AT5G18360 (*HOPB-ACTIVATED RESISTANCE1*; *BAR1*), AT5G18370 (*DOMINANT SUPPRESSOR OF CAMTA2*; *DSC2*), AT5G38340, AT5G40100, AT5G44510 (*TARGET OF AVRB OPERATION*; *TAO1*), AT5G45060, AT5G45210, and AT5G45220. In contrast, sixteen out of the 67 TIR-NBS-LRR genes were found in all assemblies but with variable copy numbers. The gene AT1G31540 had between seven and ten copies per assembly, with eight copies on average. AT1G63870 copy numbers ranged from one to three with a median of three. *RLM1* (*RESISTANCE TO LEPTOSPHAERIA MACULANS 1*; AT1G64070) had a copy number of one in all assemblies except for AT7143, which had two copies. The median number of detected copies of AT3G44400 was seven. The lowest number of copies was found in the assembly of AT9979 (three), while four other assemblies had ten copies each (AT7143, AT9336, AT9578, and AT9830). Every assembly had one copy of AT3G51560, except for AT9852 which had two copies. Similarly, AT3G51570 had a conserved copy number of one in all but one assembly. In the genome of AT6929 I annotated two copies of the gene. The number of AT4G16860 (*RECOGNITION OF PERONOSPORA PARASITICA 4*; *RPP4*) copies ranged from one (TAIR10) to eleven (AT9104), with a median of five. In the case of AT9806, two copies of AT5G17680 were detected. In contrast, the other seventeen assemblies had one copy of the gene. Similarly, AT5G22690 had a conserved copy number of one in all assemblies except for AT9830, where two copies were observed. *MIST1* (*MICRORNA-SILENCED TNL1*; AT5G38850) was annotated with one copy per genome in all but two assemblies. In the genomes of AT9744 and AT9900, it was possible to detect two copies of the gene. AT5G40910 was found in all eighteen assemblies. In five assemblies (AT6923, AT6929, AT7143, AT9847, and AT9900) I annotated two copies each, while the remaining assemblies each had one copy of the gene. Copy numbers of AT5G41540 varied between three and five in the eighteen assemblies. TAIR10 has one copy of AT5G41540. The median copy number of AT5G41540 was four per genome. In contrast, *WRK16* (AT5G45050) had a conserved copy number of one in all assemblies except for AT6929, which had two copies. Similarly to that, AT5G45200 had a copy number of one in all but one (AT9104; two copies) of the eighteen assemblies. Likewise, I detected one copy of AT5G58120 (*DANGEROUS MIX 10*; *DM10*) in each of the assemblies with the exception of AT6923, where two copies of the gene were annotated. The TIR-NBS-LRR gene AT1G27180 was found to have one copy each in the assemblies of AT6923, AT8285, AT9336, AT9503, AT9744, AT9847, AT9852, AT9879, AT9883, and AT9900, but was absent in the other eight assemblies. Similarly, AT1G56510 (*ACTIVATED DISEASE RESISTANCE 2*; *ADR2*) was detected with one copy per assembly and was absent in the assemblies of AT8285, AT9104, AT9578, and AT9806. Both AT1G56510 and AT1G56520 were absent in the annotations of AT8285, AT9104, AT9578, and AT9806. In the case of AT1G63860, copy numbers ranged from one to three. However, AT1G63860 was absent in the assemblies of AT9744 and AT9830. *TNL40* (AT1G72840) copy numbers varied between one and four while the gene was absent in AT9104 (**Figure 50E)**. In contrast, *TNL60* (AT1G72860) could only be detected in the assemblies of AT9104, AT9744, AT9762, AT9852, AT9879, and AT9883. It had a conserved copy number of one in these six genomes. AT2G17050 was observed to have one copy in all assemblies except for AT9503 and AT6923, where it was not detectable. Moreover, the assemblies of AT9336, AT9830, and AT9900 lacked the TIR-NBS-LRR gene AT4G16900. In the other assemblies I annotated between one and four copies of the gene. The median copy number of AT4G16900 was one per genome. *SIKIC1* (*SIDEKICK SNC1 1*; AT4G16940) was absent in the assemblies of AT7143, AT8285, AT9806, and AT9830. *SIKIC1* copy numbers ranged from one to six in the assembly of AT9104. However, the median copy number was two. Furthermore, the gene was not detected in the annotations of AT7143, AT8285, AT9806, and AT9830. AT5G17970 had a conserved copy number of one but was absent in the assemblies of AT7143, AT9104, AT9830, AT9879, and AT9900. Copy numbers of the gene AT5G18350 varied between one and two in thirteen assemblies. In contrast, AT5G36930 was only found in four assemblies and TAIR10 (AT9104, AT9762, AT9806, and AT9852) with a copy number of one. *TTR1* (*TOLERANCE TO TOBACCO RINGSPOT VIRUS 1*; AT5G44870) was not detected in AT9900, but had a conserved copy number of one in all other assemblies. Unlike *TTR1*, AT5G45230 was only observed in the genomes of AT9830 and AT9883. Likewise, AT5G45240 was absent in all assemblies except for AT9830 and AT9883. Both AT5G45230 and AT5G45240 had a copy number of one in the aforementioned genomes. In contrast, *RPS4 (RESISTANT TO P. SYRINGAE 4*; AT5G45250) was found in every assembly except for AT9900 with a conserved copy number of one. Likewise, *RPS6* (*RESISTANT TO P. SYRINGAE 6*; AT5G46470) had one copy in each assembly, with the exception of A9879, where it was not detected. Similarly, AT5G45260 (*RESISTANT TO RALSTONIA SOLANACEARUM 1*; *RRS1-R*) exhibited a copy number of one in every assembly except for AT9900 and AT9879 where the gene was absent. AT5G49140 was found in twelve out of eighteen assemblies, where it had one copy per genome. Similarly, AT5G51630 had one copy in twelve assemblies while being absent in the remaining six accessions.

All known TIR-NBS NLR genes were detected in at least five out of eighteen accessions (**Figure 50F**). The following TIR-NBS NLR genes were found to have a conserved copy number of one in all eighteen assemblies: AT1G17615 (*TIR-NBS 2*; *TN2*), AT1G66090 (*TIR-NBS 3*; *TN3*), AT1G72950 (*TIR-NBS 12*; *TN12*), and AT3G04210 (*TIR-NBS 13*; *TN13*). In contrast to the aforementioned genes, I could identify three other TIR-NBS-LRRs with copy number variation in all eighteen assemblies (AT1G17610, AT1G72890, and AT5G40090). In the case of AT1G17610 and TN18 (TIR-NBS 18; AT5G40090), all genomes had one copy, while two copies of each gene were detected in the assembly of AT9883. *TN6* (*TIR-NBS 6*; AT1G72890) had one copy per genome, with the exception of AT6923 and AT9847, where two copies were annotated. *TN4* (*TIR-NBS 4*; AT1G72850) had one copy in fifteen assemblies, but was absent in the genomes of AT6929, AT9104, and AT9503. In contrast, *TN5* (*TIR-NBS 5*; AT1G72870) was only detected in six accessions (AT9104, AT9744, AT9762, AT9852, AT9879, and AT9883). However, *TN7* (*TIR-NBS 7*; AT1G72900) was present in all assemblies except for AT9503. Three accessions (AT6923, AT7143, and AT9847) had two *TN7* copies, while all others had only one copy of the gene. In contrast, *TN8* (*TIR-NBS 8*; AT1G72910) was only detected in ten out of eighteen assemblies, with a conserved copy number of one. *TIR-NBS 9* (*TN9*; AT1G72920) was only detected in the genomes of AT6137, AT6929, AT9503, AT9806, AT9830, and AT9900. In contrast, it was possible to detect *TN10* (*TIR-NBS 10*; AT1G72930) in all but two assemblies (AT8285 and AT9883), with a copy number of one. Similarly, one copy of *TN11* (*TIR-NBS 11*; AT1G72940) was annotated in all genomes except for AT9578. Analyzing *TN15* (*TIR-NBS 15*; AT4G09420) revealed one copy of the gene in all but four genomes (AT9104, AT9762, AT9830, and AT9879). Similarly, one copy of *TN16* (*TIR-NBS 16*; AT4G16990) was detected in all genomes except for AT9503 and AT9900, where the gene was absent. In contrast, *TN20* (*TIR-NBS 20*; AT5G48780) was only detected in the genomes of AT8285, AT9503, AT9744, AT9830, and AT9883. In these assemblies *TN20* had a conserved copy number of one.

In summary, the analysis of NLR orthologs revealed that CC-NBS and RPW8-NBS-LRR genes are less variable in copy number among the eighteen accessions compared to the other analyzed NLR gene families. Further analyses will be required in order to investigate if NLRs from clusters are more likely to exhibit copy number variation.

NLRome liftoff

A special focus of this work was NLR gene diversity. Thus, in addition to the aforementioned transmap-guided *ab initio* whole genome annotation followed by ortholog assignment, resistance genes were annotated additionally using a previously published set of *A. thaliana* NLR genes from 64 accessions (Van de Weyer et al. 2019). Two accessions (AT9762 and AT9879) out of the 18 lines had also been part of the Van de Weyer et al. study. Thus, I benchmarked the NLR liftover approach with these two overlapping accessions. Due to the fact that full genome information was not available in the Van de Weyer et al. study, I expected that the NLR liftover approach presented here should yield at least as many, if not more, annotated NLR genes. Van de Weyer et al. reported a total of 216 NLR genes for AT9762 and a total of 213 genes for AT9879. I annotated 17 additional NLR genes in AT9762, and 9 additional NLR genes in AT9879. All additionally annotated NLR genes were part of the NLR catalog from van de Weyer et al.. However, these genes were either not annotated in the two aforementioned accessions by van de Weyer or they had lower copy numbers. The approach presented here yielded a slightly higher number of NLR genes. Therefore, I used the NLR gene liftover approach for confidently annotating NLR genes.

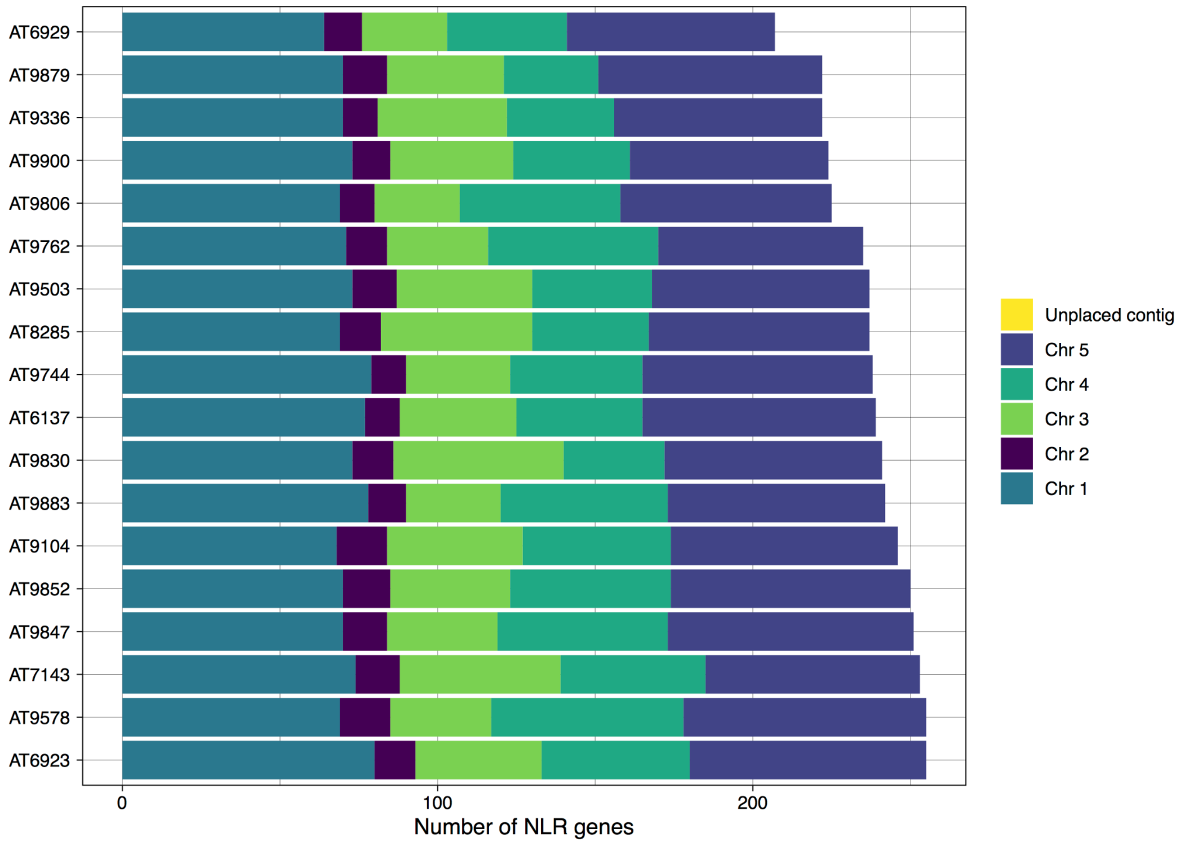

Figure 51: Number of NLR genes from the pan NLRome that were successfully mapped onto the 18 genome assemblies. Colored stacks are representing unplaced contigs and the five chromosomes.

The number of NLR genes annotated by liftover (Shumate and Salzberg 2020a) varied between 207 (AT6929) and 255 (AT6923 and AT9578) as shown in **figure 51**, with an average of 238 NLR genes per genome. Most NLR genes were located on chromosomes one (72) and five (70), while the lowest average was observed on chromosome two (13). These findings are in agreement with previous studies on NLR gene clusters of chromosomes one and five in *A. thaliana* (Chae et al. 2014). Chromosome three had 38 NLR genes on average whereas chromosome four had 44 NLR genes. The number of NLR genes annotated on chromosome one ranged from 64 in the assembly of AT6929 to 80 in AT6923. In contrast, the number of NLR genes annotated on chromosome two only varied between eleven (AT6137, AT9336, AT9744, and AT9806) and sixteen (AT9104 and AT9578). On chromosome three I detected between 27 (AT6929 and AT9806) and 54 NLR genes. Similarly, NLR gene numbers on chromosome four ranged from 30 in the assembly of AT9879 to 61 in AT9578. Between 63 (AT9900) and 78 (AT9847) NLR genes were annotated on chromosome five. Former studies have shown that chromosomes one, three, four, and five each have NLR gene clusters (Chae et al. 2014). Chromosomes with the most NLR genes were also the most variable in terms of the total number of NLRs. Thus, the greatest variation in terms of the number of annotated NLR genes was observed on chromosomes three and four, where the maximum number of NLRs was almost twice as high as compared to the assembly with the lowest number of such genes. It is noteworthy that no NLR genes were annotated on unplaced contigs.

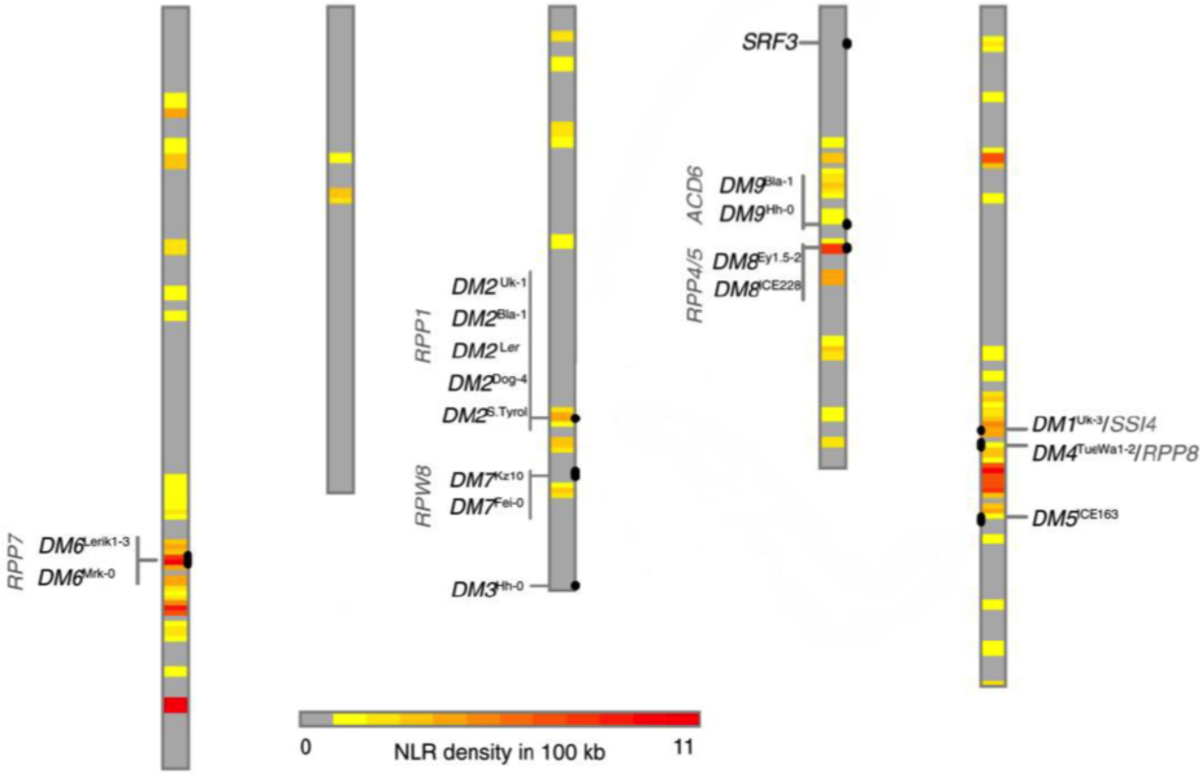

Figure 52: Heatmap of NLR gene densities along chromosomes of the reference Col-0. Darker colors indicate a higher NLR density while gray indicates the absence of NLRs in a given window. Figure adopted from Chae et al. 2014.

As mentioned in the introduction, NLR genes often occur within clusters (**Figure 52**) (Chae et al. 2014; B. C. Meyers et al. 1998; Leister et al. 1998). For further analysis I used a definition of NLR clusters as genes that are less than 200 kb away from each other in the genome (Holub 2001). Following this definition, while calculating the NLR distances, revealed that between 66 % (in AT6929) and 72 % (in AT9762) of the annotated NLRs are located within clusters. This is comparable to Van de Weyer et al. 2019 who reported that 47 % to 71 % of all NLRs can be found in clusters. In order to assess where these clusters are located I separated each chromosome into 100 equally long windows. Subsequently, all NLRs being located in one window (bin) were counted. The counts per bin are visualized for all eighteen assemblies in **figure 53**. When comparing the distribution of all genes vs the distribution of NLR genes it became evident that NLR genes are much more unevenly distributed compared to other genes. This is expected from literature (Chae et al. 2014). This analysis showed that the known NLR clusters on chromosomes one, three, four, and five were recovered by the aforementioned lift over approach. In all assemblies a high density of NLR genes was observed in the region of chromosome 1 where the *RPP7* cluster is located in the reference (**Figure 52 and figure 53**). Based on the heatmaps shown in **figure 53** it became evident that presence absence variation of NLRs is common in all assemblies which is in agreement with my previous findings (**Figure 50**). However, further effort will be required in order to investigate if NLRs located in clusters have a higher likelihood of exhibiting presence absence variation.

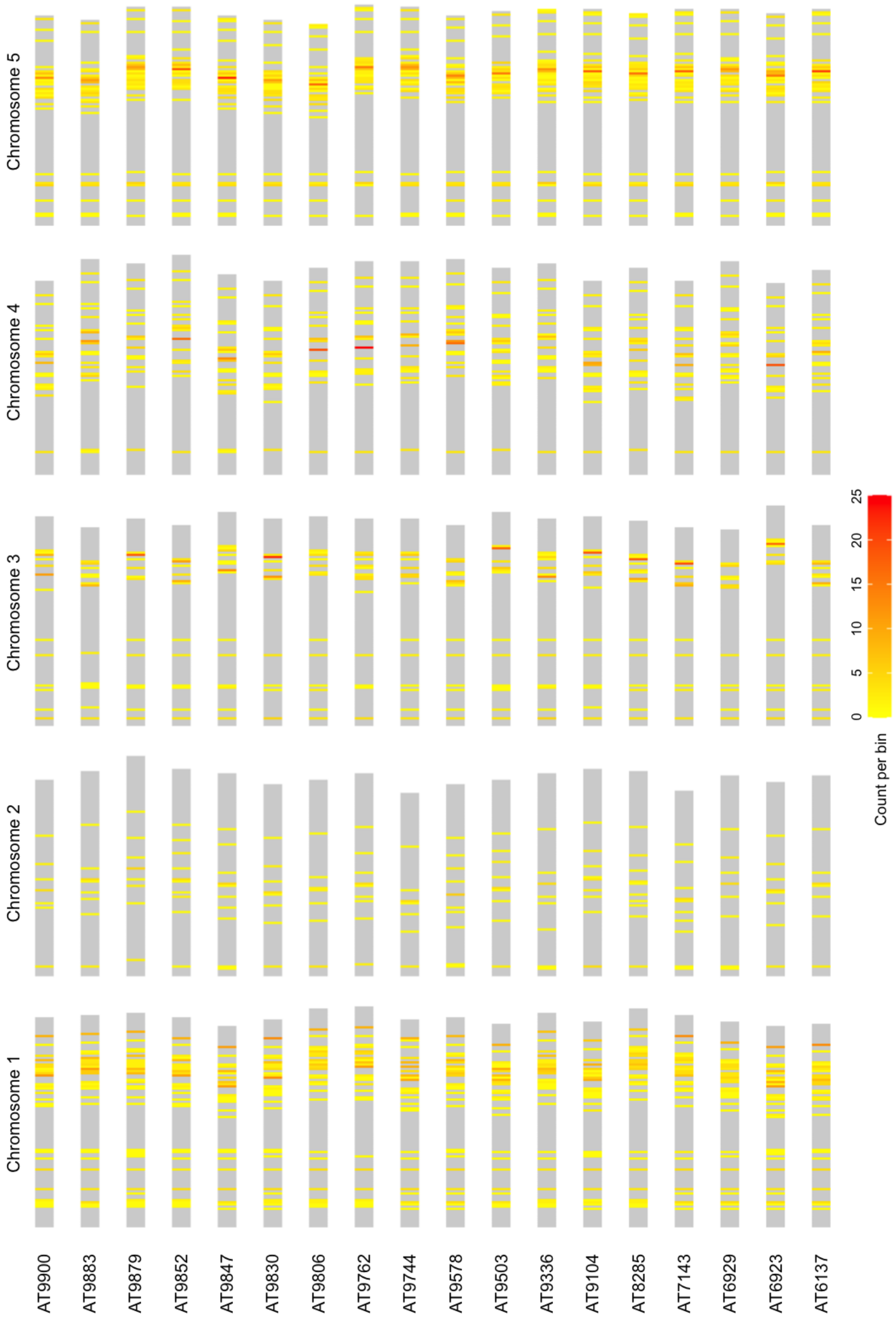

Figure 53: Heatmap of NLR gene densities along chromosomes of the eighteen assemblies. Each chromosome was divided into 100 bins. NLRs from the Van de Weyer 2019 dataset were counted. Brighter colors indicate a lower NLR density while gray regions do not bear any NLRs.

Having assessed the number of annotated NLR genes, I addressed the question of how the NLR types are distributed in the eighteen genomes of this study. The NLR genes were therefore grouped into the four classes TIR-NLR (TNL), CC-NLR (CNL), CC$_R$-NLR (RNL), and NB-and-LRR-only proteins (NL), as described in Van de Weyer et al. 2019. Subsequently, the occurrences of genes from these four classes were counted in each accession (**Figure 54**). In each of the eighteen accessions, TNLs accounted for the vast majority of NLR gene types. The number of annotated TNL genes varied between 130 in the assembly of AT9879 and 162 in AT9578, with an average of 146. The number of detected NL genes ranged from 34 in AT6929 to 58 in the assemblies of AT7143 and AT9847. The average number of annotated NL genes was 47 per genome. The average number of CNLs was 28 per assembly with a minimum of 22 in AT9830 and a maximum of 34 in AT9744. CC$_R$-NLRs represented the rarest class of NLR genes. On average each accession had sixteen such genes. The total number of RNLs varied between nine in AT9806 and 28 in AT9830.

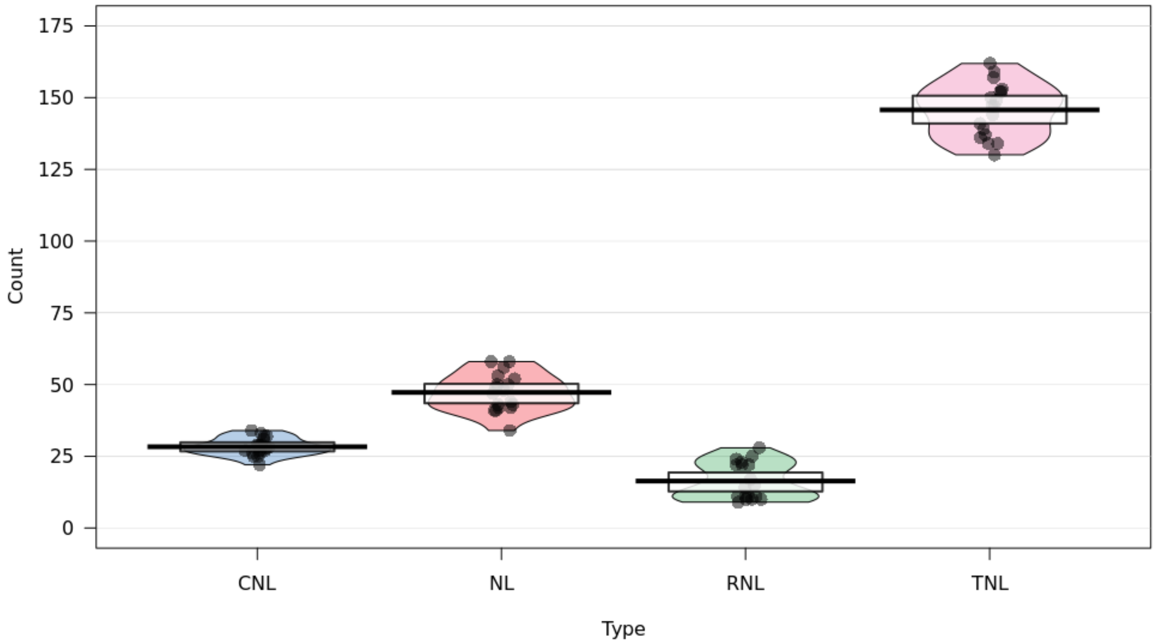

Figure 54: Pirate plot showing the number of NLR genes from a given class that were successfully lifted over from the Van de Weyer 2019 dataset. Each dot represents a single accession.

From the NLR liftover approach it can be summarized that the total number of annotated NLR genes differed by up to 48 between the accession with the highest and lowest number. Moreover, it became evident that most NLR genes were located on chromosomes one, four, and five. Additionally, it can be noted that TNLs represented the most abundant class of NLR genes in the eighteen accessions analyzed in this study and that the majority of NLRs were located in clusters. In conclusion, using the NLRome data from the pan NLRome study (Van de Weyer et al. 2019) it was possible to discover NLR genes in the eighteen accessions that would have been missed when only relying on assignment of orthologous genes with the reference TAIR10. Moreover, I have demonstrated that the usage of a *de novo* whole genome assembly helps to identify more NLR genes as compared to using bait capture based methods.

4.2 Differential gene expression in *A. thaliana* F_1 hybrids

As stated in the introductory section of this thesis I set out to assess if F_1 hybrids of Col-0 and Ler-0 show non-additive gene expression. Thus, RNA sequencing (RNA-seq) was used in order to compare the transcriptomes of Ler-0xCol-0 F_1 hybrids to their inbred parents. Gene expression in hybrids and parents was assessed in different tissues and at different time points. Instead of mapping the RNA-seq reads to a single reference genome I wanted to make use of recently in house generated PacBio long-read assemblies. These high quality genome assemblies of Col-0 and Ler-0 were used in order to generate a custom hybrid reference that contains full genome information of both inbred parents. Subsequently, the RNA sequencing reads were mapped to that custom reference in order to assess differences in gene expression between hybrids and inbred parents.

4.2.1 Custom reference and short-read mapping

The custom hybrid reference was generated using in house generated PacBio long-read genome assemblies from Col-0 and Ler-0. In the following sections and figures I will refer to Col-0 as ‘Parent 1’ while Ler-0 will be referred to as ‘Parent 2’. Both parental genomes were annotated separately using the comparative annotation toolkit (CAT) (Fiddes et al. 2017) with TAIR10 (Lamesch et al. 2012) as the reference. Subsequently, the information from both, the reference as well as the newly annotated genomes were combined by assigning orthologs using Orthofinder (Emms and Kelly 2015). In the parental genome annotations it was possible to recover 96% of all protein coding genes annotated in TAIR10. Moreover, a total of 1545 genes have been *de novo* annotated using AugustusCGP (Stanke and Waack 2003). The ortholog assignment enabled generating a transmap for the hybrid. This map holds the information of which two parental alleles belong to which gene. RNA-seq reads from all genotypes, tissues, and timepoints were mapped against the custom hybrid reference containing full length transcriptome information of Col-0 and Ler-0 with mapping rates ranging from 69% to >90% and an average of 83.43% reads mapped. The number of reads being filtered due to too many alignments was zero in all samples. For comparison I mapped all RNA-seq samples against the public *A. thaliana* reference TAIR10. Comparison between the custom and the public reference did not reveal any significant differences (paired t-test) with regard to the number of aligned (*p* value > 0.6), unaligned (*p* value > 0.6) or filtered reads (*p* value = n/a). Thus, using the custom reference did not improve or worsen the overall number of aligned RNA-seq reads. Subsequently, the fraction of aligned reads was divided into unique, multiple, and uncertain alignments. Between 91 % and 94 % of the reads were uniquely aligned when using the custom reference. Thus, between 5 % to 9 % of the reads aligned to multiple positions. In contrast to this, the fraction of reads aligning to multiple positions ranged from 6 % to 21 % when using TAIR10 as a reference. Thus, the number of uniquely aligned reads was significantly higher (paired t-test; *p* val < 2.18e-13) when using the custom reference. However, in this case multi-mapped reads also include isoform-level multi-mappings. The annotation of multiple isoforms was not done for the custom hybrid reference. Thus, these differences in multiple mappings are most likely explained by the lack of more isoforms in the custom reference. Moreover, RSEM, the tool that was used for downstream processing of the read alignments, is capable of correcting for multiple mappings while estimating the abundance of each transcript (B. Li and Dewey 2011). Having demonstrated that using the custom reference resulted in higher numbers of uniquely aligned reads while not affecting overall mapping rates I decided to perform further downstream analyses using the custom hybrid reference. RSEM was used to estimate the abundance of each transcript from the previously generated alignments. In case of the custom reference the expression level of a given gene is calculated by taking into account the reads that are mapped to the two parental alleles. Estimated transcript abundances were subsequently used to compare the transcriptomes from the different genotypes, tissues, and time points among each other.

4.2.2 Tissue as main driver of variance

A principal component analysis (PCA) was performed on all samples of the presented dataset in order to visualize key differences among the transcriptomes. Transcriptomic differences among tissues can explain much of the overall variance in the dataset (**Figure 55A**), indicated by tight clustering according to samples by tissue of origin. The first principal component, explaining 52% of the variance, separates the four tissue types, with the greatest distance between flowers and roots, and seedlings 3 days after germination (3DAG) in between. PC2 groups roots, seedlings, and shoots closer together while further separating flower samples from the other tissues. In sum, PC1 and PC2 explain 72 % of the overall variance among the transcriptomes of the dataset. Principal component three separates the three genotypes from each other while explaining 8 % of the overall variance (**Figure 55B**). These findings show that transcriptomic differences among tissue types are the main driver of variance in this dataset rather than genotypes or tissue age.

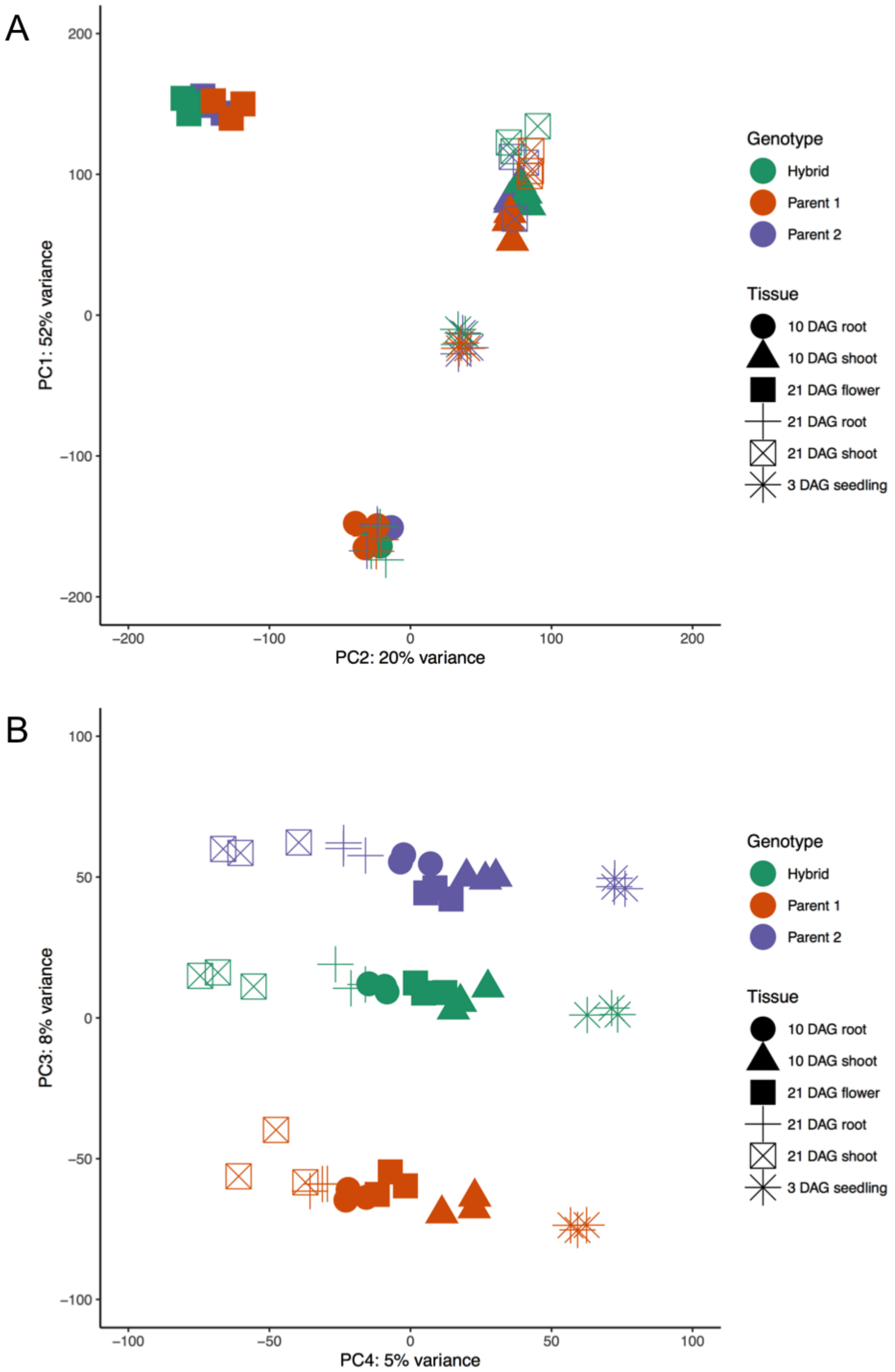

Figure 55: Principal component analyses of the transcriptomes of all samples. Shapes are indicating different tissues and time points while colors are representing the three genotypes. (A) Principal components 1 and 2, (B) principal components 3 and 4.

To explore how parents and hybrids compared to each other, I performed PCA on samples with the same tissue of origin, which separated the three genotypes from each other and placed the hybrid transcriptomes between the two parents. This pattern (**Figure 56**), hybrids being in between the inbred parents was observed for all tissues and timepoints. Differences among the three genotypes explained between 45 % and 63 % of the overall variation within one tissue and at one time point. Moreover, in seedling and flower samples, it was observed that the second principal component grouped the inbred parents and separated them from the F_1 hybrid samples while explaining between 11 % (seedlings) (example in **figure 56**) and 21 % (flowers) of the variation. However, no such clear grouping in the second PC was observed for root and shoot samples.

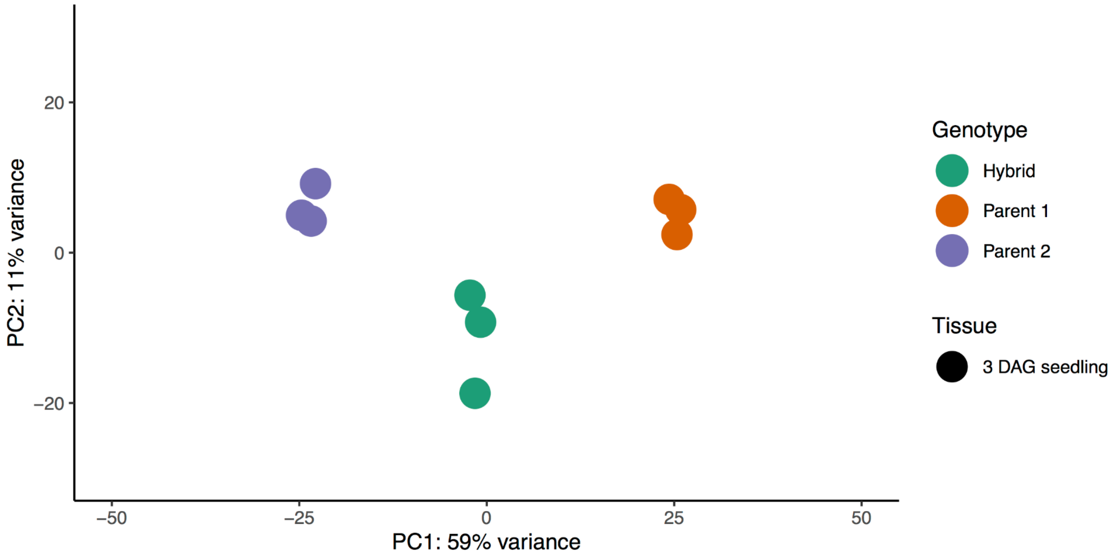

Figure 56: Principal component analysis of the transcriptomes of seedlings 3 days after germination. Colors are representing the three genotypes.

It can be summarized that transcriptomic differences among tissues are greater as compared to differences among genotypes or timepoints. However, when investigating samples from the same tissue, the first principal component always separated the three genotypes from each other with the F_1 hybrid samples clustering in- between the parental transcriptomes. Moreover, in seedling and flower samples, it was observed that the second principal component grouped the inbred parents and separated them from the F_1 hybrid samples.

4.2.3 Non-additive effects in principal component analysis

Non-additive hybrid phenotypes are such phenotypes that deviate from the mean of their inbred parents for a given trait (**Figure 3**). In order to measure such non-additive effects one has to calculate the mean of both parents, the so-called mid-parent value (MPV). The above described results from the PCA of 3 DAG seedlings and 21 DAG flowers grouped the inbred parents while separating them from the F_1 hybrids in the second PC. In order to investigate if this separation of inbred parents and F_1 hybrid offspring indicates non-additive effects, I created *in silico* hybrids. These *in silico* hybrids comprise 50% read data from each parent. Thus, if there were no non-additive effects it would be expected that *in silico* hybrids and F_1 hybrids cluster together in the second principal component.

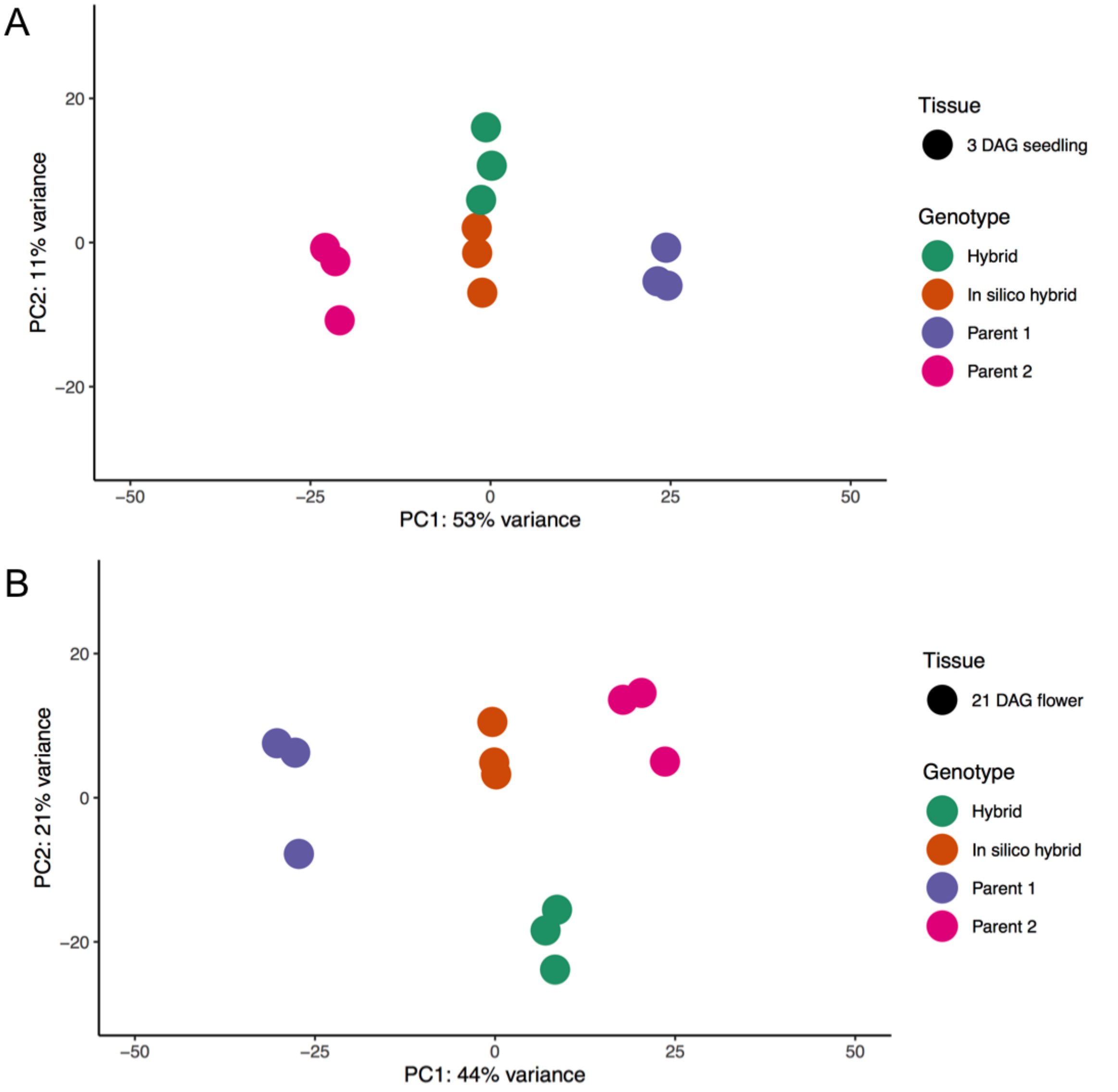

Figure 57: Principal component analyses of the transcriptomes of (A) seedlings 3 and (B) flowers 21 days after germination including *in silico* hybrids. Colors are representing the four genotypes.

Randomly selected *in silico* samples from seedlings 3 DAG formed a cluster with F_1 hybrids samples in the first principal component while clustering with inbred parents in the second principal component (**Figure 57A**). Likewise, the *in silico* hybrids from flower samples clustered with the F_1 hybrids in the first principle component while being grouped with the inbred parents in the second principle component (**Figure 57B**). These findings indicate that the separation of parents and F_1 hybrids in the second principle component is due to non-additive gene expression that cannot be explained by simply mixing parental read data.

Thus, the next section describes how expression of those genes, that separate F_1 hybrids from the parents and the in silico hybrids, was quantified.

4.2.4 Non-additive gene expression in F_1 hybrids

I determined differentially expressed genes (DEGs) among the three different genotypes for each tissue and time point individually using DESeq2 (Love, Huber, and Anders 2014)2014). Mid parent values (MPV) for gene expression of each gene was calculated computationally based on parental values. Hybrid transcriptomes were compared to each of the parents as well as to the MPV. Moreover, differences in gene expression between both inbred parents were examined. A gene was considered to be significantly differentially expressed when showing a log2 fold-change of > 0.5 and p_{adj} < 0.01. The number of genes that showed significant expression changes as compared to the MPV, was counted. This expression pattern is referred to as mid-parent heterosis (MPH). Moreover, the number of genes exceeding the best-parent (best-parent heterosis; BPH) or being below the expression level of the low-parent was examined (low-parent heterosis; LPH). These two expression patterns are summarized as best parent heterosis (BPH). Thus, genes showing BPH are also included in the MPH gene set. However, genes showing MPH do not necessarily also show BPH.

The number of non-additively expressed genes differed among the different tissues and time points is shown in **table 10**. The highest number of genes showing MPH expression patterns was observed in 21 DAG flower samples (1,999) while the lowest number of MPH genes was detected in 3 DAG seedlings (422). In all samples it was observed that the majority of non-additively expressed genes was upregulated compared to the MPV. The portion of genes showing not only MPH but also BPH ranged from 29 % (124) in 3 DAG seedlings to 81 % (1,627) in 21 DAG flowers. Between 15 % (52 in 3 DAG seedling) to 70 % (859 in 21 DAG flowers) of the genes being upregulated compared to MPV also exceed the better parent value.

Table 10: Overview on differentially expressed genes (log2FC > |0.5| && padj < 0.01). Non-additively expressed genes were further separated into mid-parent and best or low parent heterosis.

Sample	Parent 1 (P1) vs Parent 2 (P2)			Non-additively expressed genes					
				Mid parent heterosis			Best parent (BP) & low parent (LP) heterosis		
	P1 > P2	P1 < P2	Sum	F_1 > MPV	F_1 < MPV	Sum	F_1 > BP	F_1 < LP	Sum
3 DAG seedling	1387	1715	3102	347	74	422	52	72	124
10 DAG root	2773	3097	5870	478	27	505	119	25	144
10 DAG shoot	2265	2101	4366	525	238	763	206	211	417
21 DAG root	2110	2497	4607	434	55	489	97	48	145
21 DAG shoot	1865	1924	3789	706	591	1297	350	561	911
21 DAG flower	1506	1866	3372	1210	789	1999	859	768	1627

In contrast to that, it was observed in all but one sample (21 DAG root) that between 89 % to 97 % of the genes being expressed below the MPV were also expressed below the lower parent. In three samples (3 DAG seedling, 10 DAG shoot, and 21 DAG shoot) more BPH genes were expressed below the value of the lower parent while fewer BPH genes were exceeding the better parent. In the remaining two samples the opposite was observed. However, positive as well as negative fitness consequences can result from both, lower and higher expression compared to MPV. In addition, shoot samples from both time points showed a higher number of MPH and BPH genes when compared to the corresponding root samples. Comparing parental transcriptomes with each other led to the detection of 3,102 (3 DAG seedling) to 5,870 (10 DAG root) differentially expressed genes (**Table 10**). Thus, the differences in non-additively expressed genes between the different tissues and time points did not correlate with the number of genes that are differentially expressed between the inbred parents. As mentioned before, most non-additively expressed genes were upregulated as compared to the MPV. In contrast to this, the number of up- and down-regulated genes was more balanced when comparing the parental transcriptomes (**Table 10**; see example **figure 58**). Thus, between 42 % (21 DAG root) and 52 % (10 DAG shoot) of the differentially expressed genes were upregulated in parent 1 when compared to parent 2. If non-additively expressed genes are also differentially expressed between the inbred parents will be examined in a later section of this thesis.

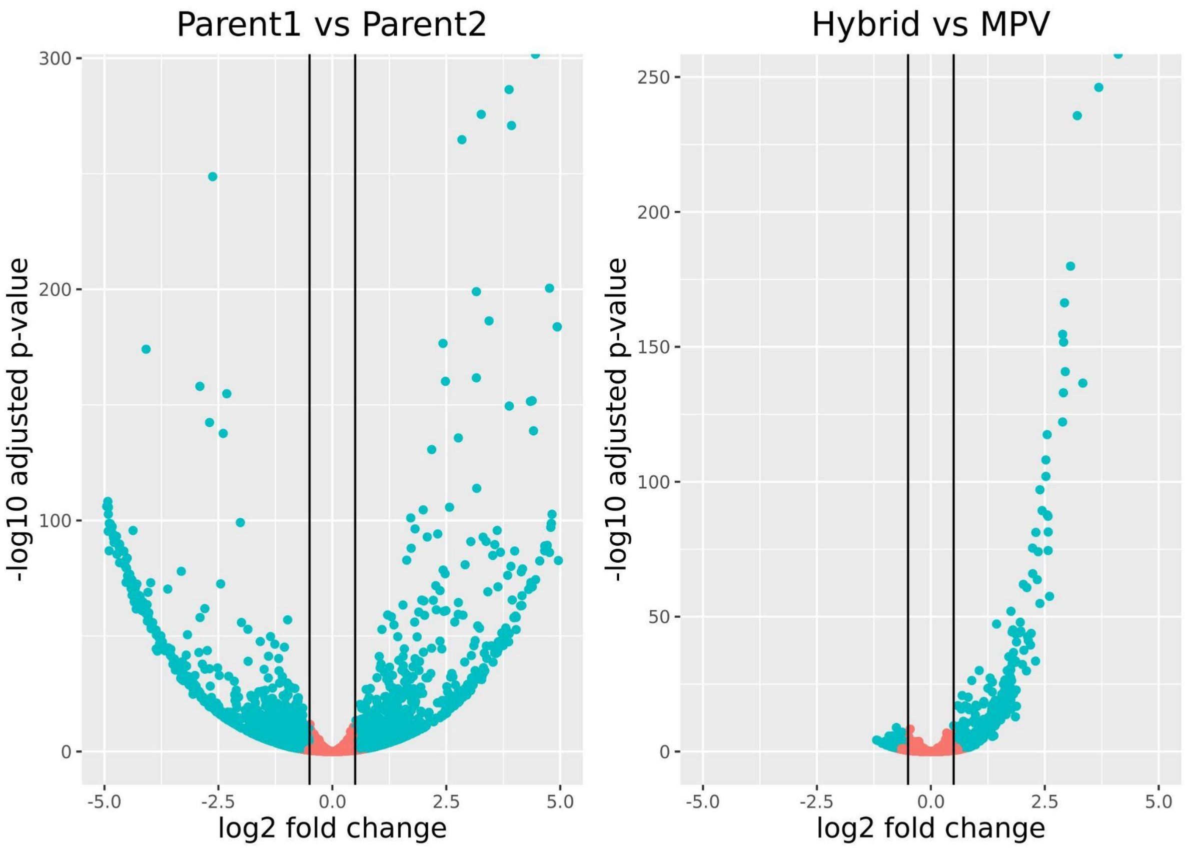

Figure 58: Volcano plot showing gene expression changes in 3 DAG seedlings. Blue dots are representing significantly differentially expressed genes while red dots represent genes with padj > 0.01 && log2FC < |0.5|. Black vertical bars are marking log2FC of 0.5. Left panel: Differentially expressed genes when comparing parent 1 and parent 2. Right panel: Differentially expressed genes when comparing the F_1 hybrid to the mid-parent value.

It can be summarized that the number of non-additively expressed genes increased during plant development (**Table 10**; **Figure 59**). Therefore, I examined if the change in the number of non-additively expressed genes during plant development could be explained by variation in the number of expressed genes in general.

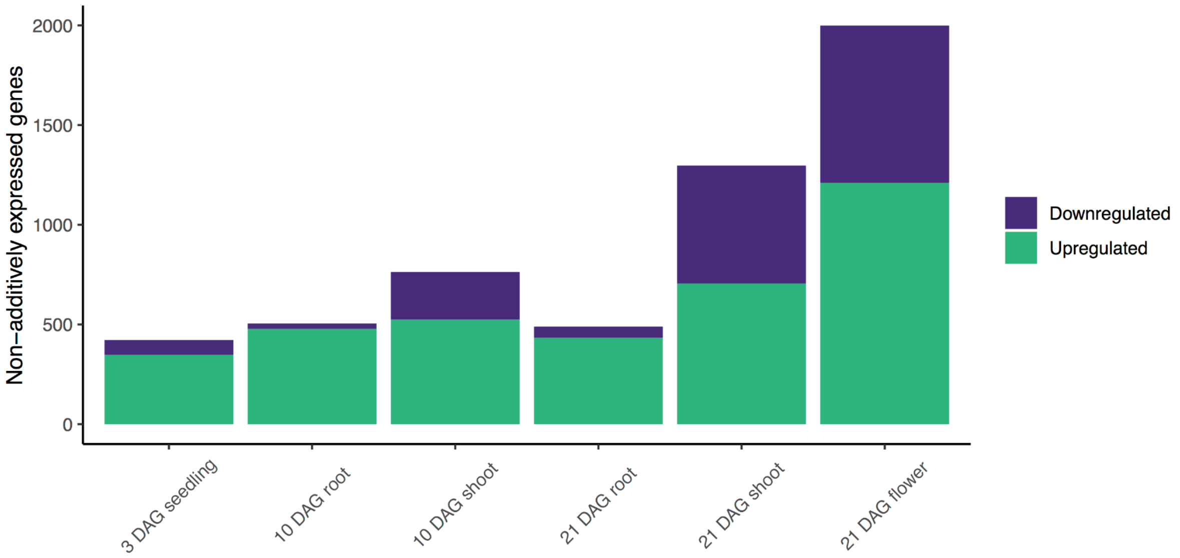

Figure 59: Increase in the number of non-additively expressed genes in F_1 hybrids. Colored stacks are indicating the number of up- and down regulated genes.

The number of genes expressed in each genotype, tissue, and time point was calculated. The total number of expressed genes varied between timepoints and tissues (**Figure 60**). However, the variation in the number of expressed genes did not correlate with the number of non-additively expressed genes.

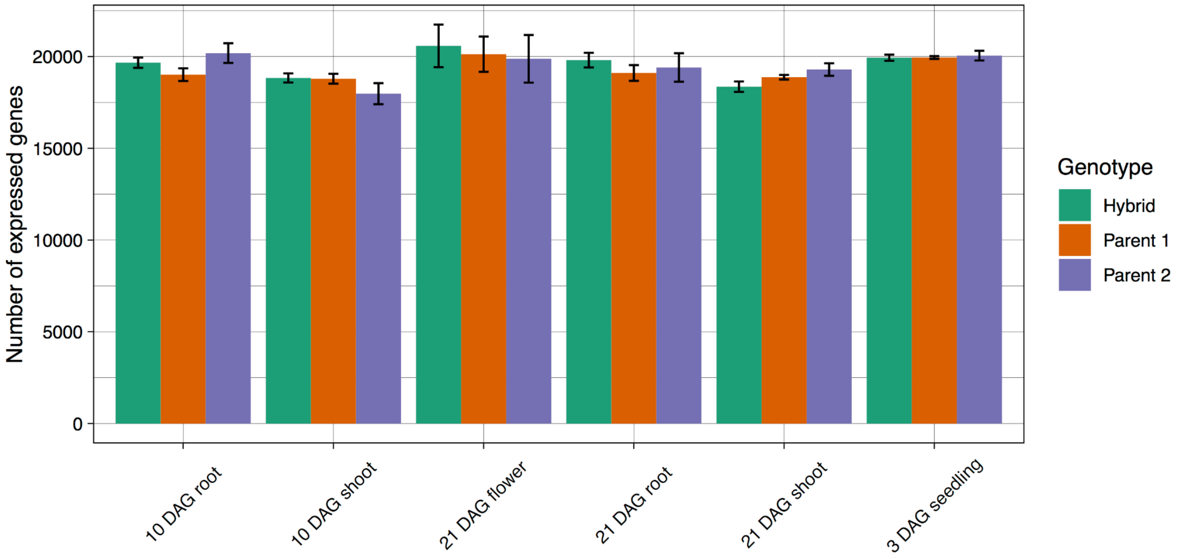

Figure 60: Total number of expressed genes in all samples. Error bars are indicating the standard deviation calculated from the three RNA sequencing replicates.

In summary, the here conducted analyses showed that the number of non-additively expressed genes increased during plant development. Most non-additively expressed genes were up-regulated compared to the mid parent value. Furthermore, in each tissue and at all timepoints, the greatest number of significantly differentially expressed genes was identified when comparing the parental genotypes to each other.

4.2.5 Overlap in non-additively expressed genes

After identifying varying numbers of non-additively expressed genes in different developmental stages and tissues, I tested how many of these MPH genes were shared among samples from different tissues or time points.

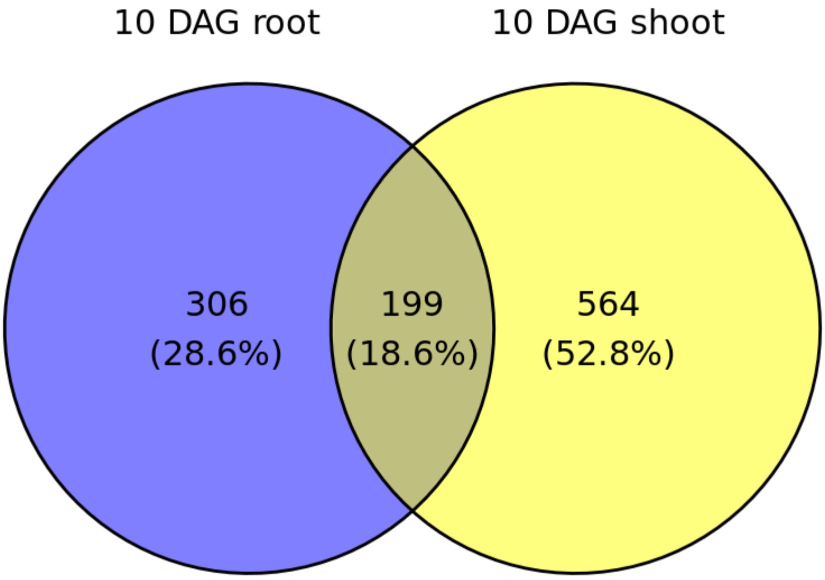

Figure 61: Venn diagram showing the intersection of non-additively expressed genes between root and shoot samples 10 days after germination.

Intersection analysis of root and shoot samples from plants 10 days after germination showed that 18.6 % (199) of the non-additively expressed genes were shared between both tissues (**Figure 61**).

Moreover, when intersecting the gene lists from hybrid root, shoot, and flower samples harvested 21 days after germination, I observed that only 3.3 % of these genes showed non-additive gene expression in all three tissue types (**Figure 62**). Flower samples had 1612 non additively expressed genes that were not shared with any of the other two sets. In contrast, in the root samples only 8.1 % of the non-additively expressed genes were not shared with either shoot or flower samples. Moreover, less than five percent of the MPH genes were shared between root and shoot samples at 21 days after germination.

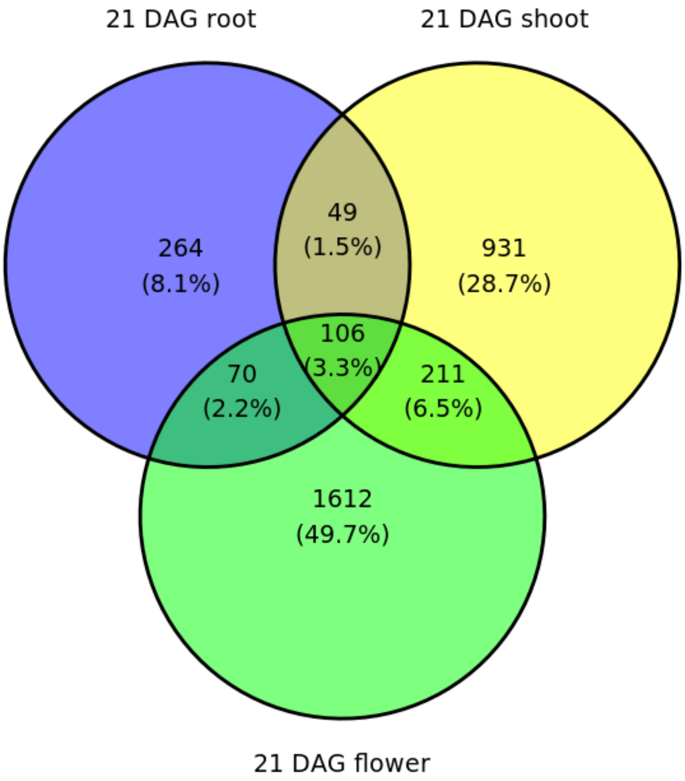

Figure 62: Venn diagram showing the intersection of non-additively expressed genes between root, shoot, and flower samples 21 days after germination.

Next, I intersected the non-additively expressed genes of roots at ten and twenty-one days after germination (**Figure 63**). Here it was observed that 39.8 % (283) of the genes were shared between root samples of both time points.

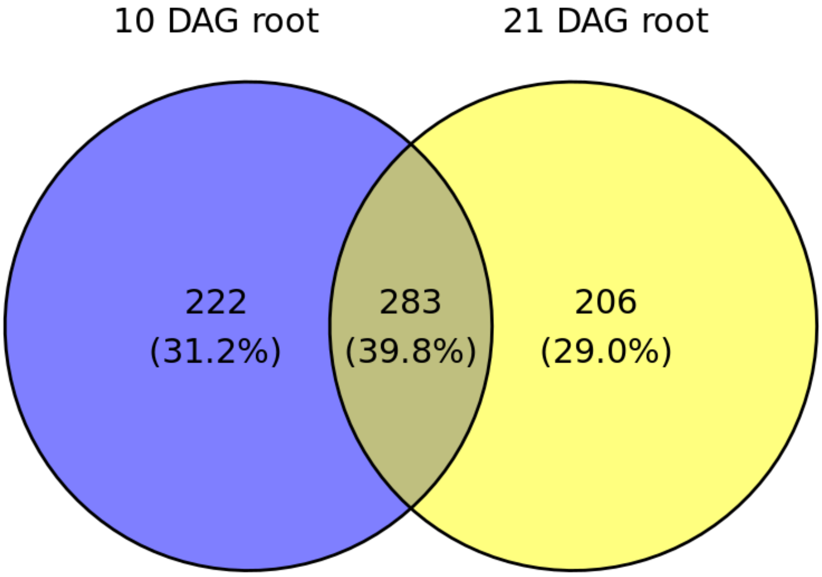

Figure 63: Venn diagram showing the intersection of non-additively expressed genes between root samples.

A similar analysis of shoot samples from ten and twenty-one days after germination showed that shoot samples from both time points only shared 13.3% (242) of the non-additively expressed genes (**Figure 64**).

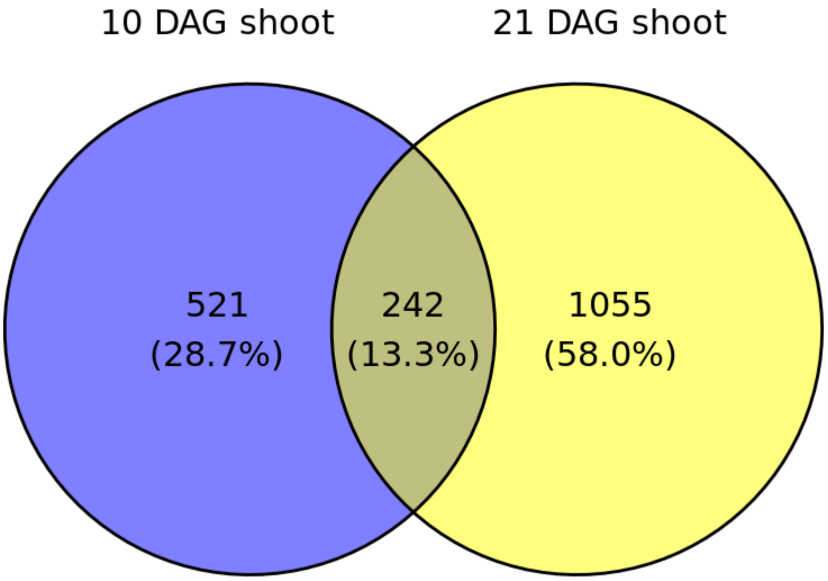

Figure 64: Venn diagram showing the intersection of non-additively expressed genes between shoot samples.

Further intersection analyses revealed that 81 were expressed in a non-additive manner in all analyzed tissues and timepoints. Therefore, it can be summarized that tissues of different time points shared a higher fraction of non-additively expressed genes compared to different tissues from the same time point. This indicates that the genes which are expressed in a non-additive manner are not specific to a certain time point but rather tissue dependent. Moreover, it became evident that there is only a small number of genes exhibiting non-additive expression patterns in all tissues and time points. These genes will be further analyzed in the next section.

4.2.6 Gene ontology analysis of non-additively expressed genes

After having assessed the number of non-additively expressed genes for all tissues and time points I set out to investigate if these MPH genes can be associated with certain biological functions. Therefore, gene ontology (GO) analysis has been applied to the MPH gene lists of all tissues and timepoints. The GO analysis is used in order to identify biological processes, molecular functions, and cellular locations of gene products. The gene ontology consists of two main elements. First there is a collection of biological terms that are put in a hierarchical relationship. Second, there is an annotation that links genes and their corresponding gene products to specific terms. Thus, a GO enrichment analysis allows one to detect if certain biological processes are overrepresented in a given set of genes (Ashburner et al. 2000). GO analysis of MPH genes has been performed using Panther (v17.0) (Mi et al. 2019). The first GO analysis has been performed on each MPH gene list separately. Between 29 (3 DAG seedling and 10 DAG root) and 150 (21 DAG shoot) biological processes were overrepresented when using Fisher’s exact test with $p < 0.05$. The terms 'defense response', 'defense response to other organism', 'response to biotic stimulus', 'response to external biotic stimulus', 'response to external stimulus', 'response to other organism', 'response to stimulus', and 'response to stress' were significantly enriched in the MPH genes of all tissues and time points when performing the GO analysis.

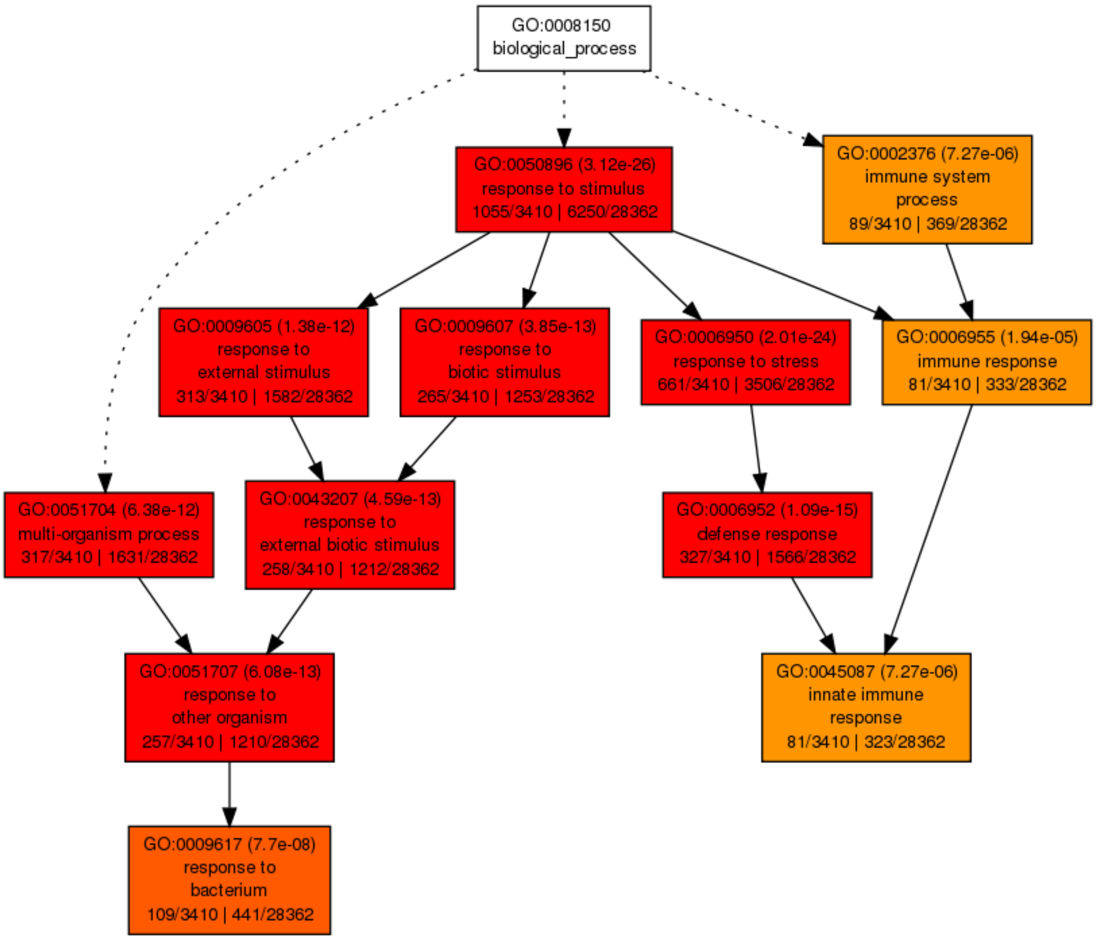

Figure 65: Gene ontology analysis of genes with non-additive expression patterns in all tissues and timepoints.

Subsequently, I performed GO analyses on those 81 genes where the before described intersection analysis showed that they have non-additive expression patterns in all tissues and time points. As expected from the GO analyzes on the individual samples, the categories 'defense response' and 'response to stimulus' were overrepresented. Next GO analysis was performed on the combination of all genes that were expressed at MPH in at least one sample. A subset of the significantly enriched GO terms is visualized using AgriGO (v2) (Z. Du et al. 2010) (**Figure 65**). Again, GO terms related to 'response to stimulus', 'defense response', or 'immune response' were significantly enriched. In summary, I have shown that the number of significantly enriched GO terms differed among the MPH genes of the various samples. Moreover, I showed that a handful of GO terms, mostly related to 'response to stimulus' and ‘defense response’ are enriched in all tissues and time points.

4.2.7 Parental expression divergence correlates with deviation from MPV

It has been described before that heterosis in F_1 hybrids can be correlated with genetic divergence of the corresponding inbred parents (Birchler et al. 2010; Troyer 2006). Therefore, I checked if non-additively expressed genes are also more likely to be differentially expressed between the inbred parents. All samples showed a significant overlap (hypergeometric test for overrepresentation; p <0.01) among genes exhibiting differential expression between parents and genes with non-additive expression levels in hybrids. Between 30% and 80% of the non-additively expressed genes were also differentially expressed when comparing both inbred parents to each other (**Figure 66**). In maize it was also observed that there is a significant overlap between genes being differentially expressed between parents while showing non-additive expression in F_1 hybrids (Paschold et al. 2012).

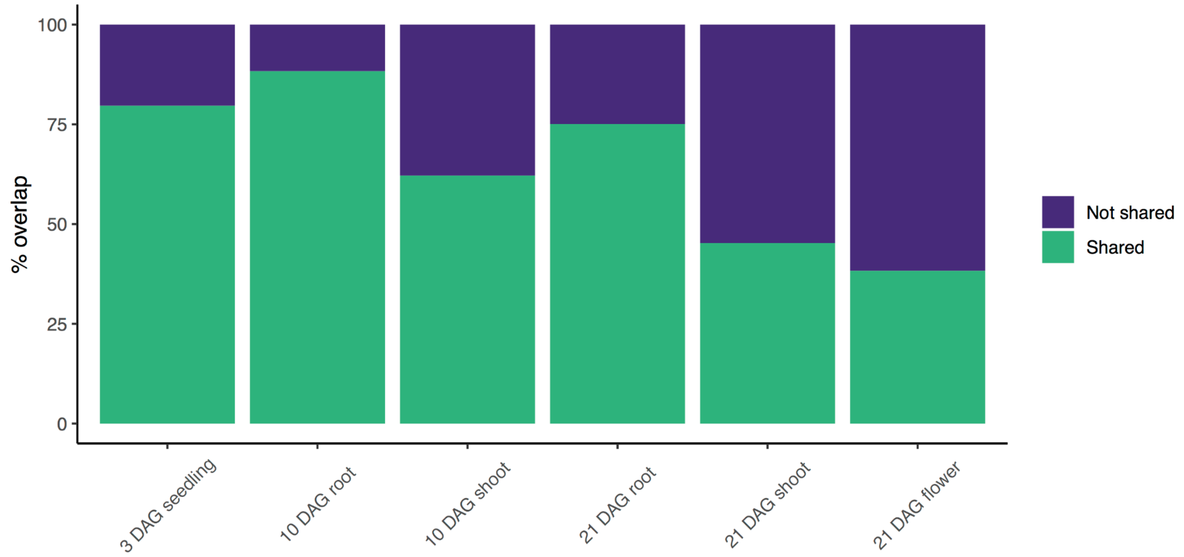

Figure 66: Overlap of non-additively expressed genes with genes that were differentially expressed between the inbred parents.

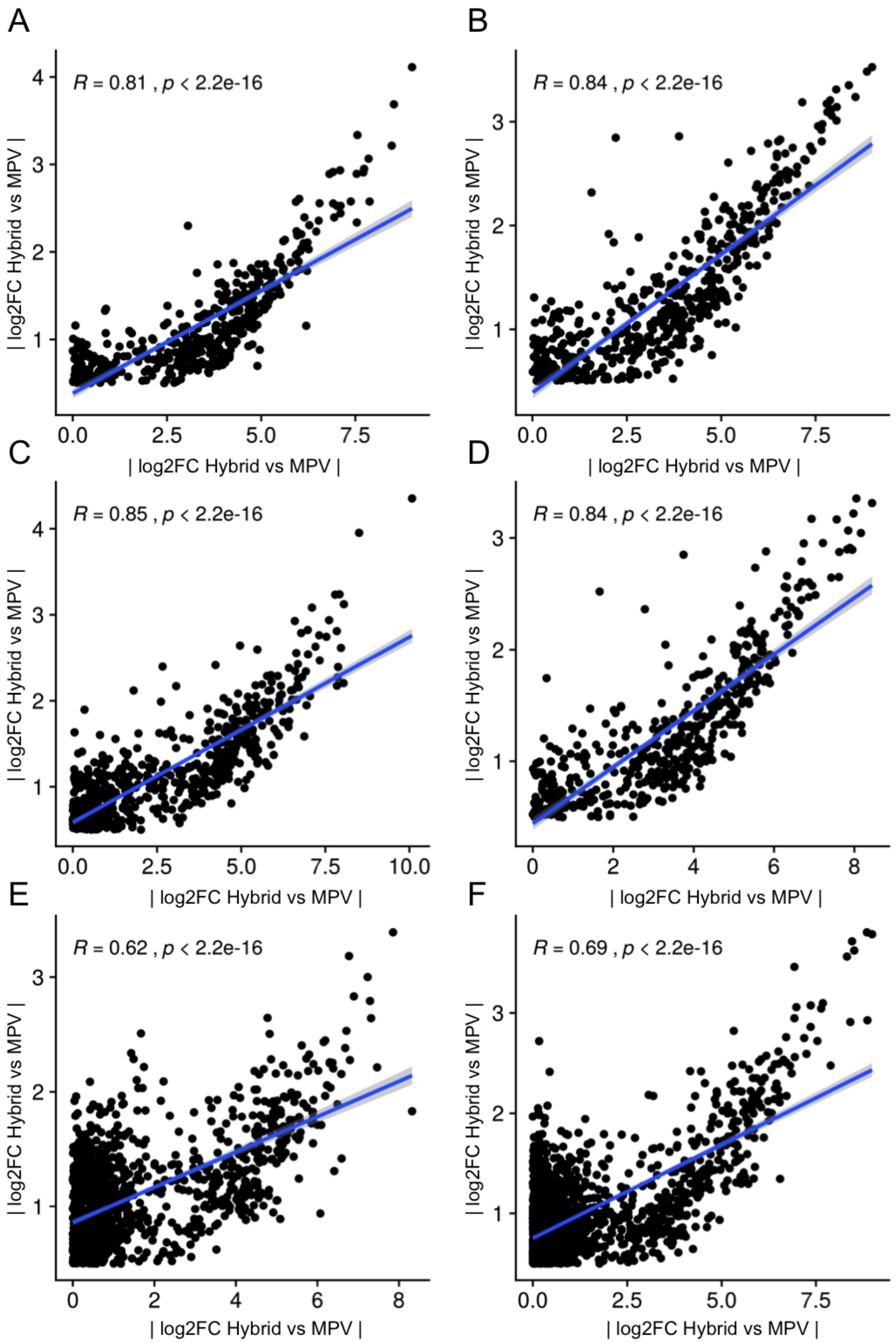

Figure 67: Dot plots showing the correlation between parental expression divergence and deviation from the mid-parent value (MPV) in F_1 hybrids. Blues lines are representing the fitted linear model. (A) 3 DAG seedling, (B) 10 DAG root, (C) 10 DAG shoot, (D) 21 DAG root, (E) 21 DAG shoot, and (F) 21 DAG flower.

After having observed this significant overlap I set out to investigate if the degree of expression divergence between parents correlates with the degree of deviation from the MPV in F_1 hybrids. All genes that were differentially expressed between parents while showing MPH in the F_1 hybrids were analyzed further. For each gene the |log2 fold change| between parents was compared to the |log 2 fold change| of the F_1 hybrid vs the mid-parent value. Pearson's correlation test was performed to determine statistically significant correlations ($P < 0.01$) among absolute values of log2 fold change Parent1 vs Parent2 and the obtained log2 fold change of F_1 hybrids vs MPV. Both variables were positively correlated in all analyzed samples with correlation coefficients (*R*) ranging from 0.62 in shoot transcriptomes 21 DAG and 0.85 in shoot transcriptomes 10 DAG. Thus, the magnitude of expression divergence between parental genotypes was significantly correlated with F_1 hybrid expression deviation from the mid parent value. It has been reported that expression deviation from the MPV in maize hybrids is correlated with expression divergence between parents (Stupar et al. 2008). From this I conclude that the results from this set of *A. thaliana* F_1 hybrids are in agreement with previous findings. If parental expression divergence between parents can be used to predict non-additive gene expression in hybrids will be discussed later in this thesis.

5. Discussion

The goal of my dissertation was to exploit state-of-the-art long-read sequencing technologies to overcome some of the limitations of former studies that relied on short-read data. In chapter 4.1, I showed that PacBio Hifi sequencing allowed for the robust generation of eighteen highly contiguous *de novo* genome assemblies with very similar characteristics. In addition I have explained that two of the initially twenty datasets had to be removed from downstream analyses due to low quality or contamination. Moreover, I demonstrated that these eighteen genomes can be used to study regions of the genome that were inaccessible even in the current gold-standard reference genome of *A. thaliana*. The *de novo* genome assemblies allowed me to confidently assess and compare gene numbers between accessions, with a focus on detection of non-reference NLR genes. In chapter 4.2, I demonstrated how such long-read *de novo* genome assemblies can be used in the analysis of non-additive gene expression of F_1 hybrids from *A. thaliana*. My PhD project will serve as the basis for upcoming projects focusing on the manual annotation and characterization of NLR gene diversity among the eighteen accessions that were the subject of this study. My part of that project has mostly centered on the assembly of high-quality chromosome-scale genomes that had become possible at the start of my PhD, as technological advances became available in-house at the institute. As this was not yet standardized, during the course of my work I undertook several comparisons and some optimization on the way, and the workflow I have described in the results section has now become a standard workflow that others can apply. The biological material I studied is part of a larger project in which genomic diversity at disease resistance loci is particularly interesting, and with the data I generated at hand, in-depth comparisons as well as experimental studies can now take place.

The following chapter discusses the methods and results of this thesis. For the first project of 18 differential lines, issues and decisions related to quality control and the genome assembly workflow are discussed. For the project of differential gene expression in hybrids I will discuss the use of a custom hybrid reference genome for the analyses of non-additive gene expression in hybrids with regards to parental expression divergence.

5.1 Eighteen differential *A. thaliana* lines

5.1.1 Quality control

Performing *de novo* genome assembly requires upfront knowledge of certain properties of the target species' genome such as genome size, ploidy, and heterozygosity (Ranallo-Benavidez, Jaron, and Schatz 2020). *Arabidopsis thaliana* has a diploid genome and a haploid size of approximately 135 Mb (Arabidopsis Genome Initiative 2000). When working with inbred accessions, as it was the case for the work presented here, one can assume that the genome is mostly homozygous (Mauch-Mani and Slusarenko 1993). Moreover, parameters of the assembly tool I used require assumptions on the level of heterozygosity of the genome (H. Cheng et al. 2021). Therefore, it was necessary to ensure that the accessions used for this study were mostly homozygous. Various tools have been developed in order to estimate genome size and heterozygosity directly from sequencing reads, making additional up-front steps unnecessary. One approach relies on counting the frequency of kmers (Chikhi and Medvedev 2013; Melsted and Halldórsson 2014; H. Sun et al. 2017; Ranallo-Benavidez, Jaron, and Schatz 2020). One accession (AT6961) appeared not to be fully homozygous (**Figure 16**). Thus, I decided to remove AT6961 from the dataset.

Non-target contaminants in whole-genome sequencing datasets can impact downstream analyses (Goig et al. 2020). In some cases even published genomes that are deposited in public databases have been found to contain sequences of non-target organisms (Lu and Salzberg 2018; L. Tang 2020). Moreover, it has been described that in the worst-case scenario, such contamination can result in drawing false conclusions about the dataset of interest. One such example is a study on Bdelloid genomes where an unusually high number of genes from non-metazoan species was reported and interpreted as horizontal gene transfer (Debortoli et al. 2016). These results were later questioned as reanalyses of the dataset revealed a contamination with non-target DNA (Wilson, Nowell, and Barraclough 2018). In addition, contaminations in the draft of the human genome have led to false conclusions about potential horizontal gene transfer events (Lander et al. 2001; Willerslev et al. 2002). Another example is a *de novo* assembled cow genome (*Bos taurus*) where further investigation showed that the assembly contained 173 contigs derived from bacterial contamination (Zimin et al. 2009; Merchant, Wood, and Salzberg 2014). Other researchers studying ancient DNA from amber fossil bees found that the analyzed 18S rRNA sequences were actually contaminants (Walden and Robertson 1997). Apart from mostly bacterial contamination, a large study also identified in published genomes 154 assemblies with human DNA contamination (Kryukov and Imanishi 2016). What these examples have in common is that contamination during sample preparation or during the sequencing process lead to false conclusions, highlighting the need for thorough contamination screenings in genome sequencing projects. Therefore, I screened all datasets for potential contaminants. This screening revealed that ten of the generated datasets contained bacterial, mostly *Pseudomonas* spec., sequences. The extent of contamination was highly variable (**Figure 17**). This indicates a non-systematic contamination of the sequenced samples. It has been reported that contaminations with Pseudomonas are ubiquitous in DNA sample preparation (Salter et al. 2014), and in our case likely originated during plant growth, as plants were raised in non-sterile conditions in growth chambers on soil. The contaminant contigs were removed in order to avoid false conclusions during downstream analyses.

5.1.2 Assembly method optimization

The PacBio Sequel II platform allows for multiplexing of different samples on the same SMRT cell. Given the high costs of SMRT cell sequencing, it was important to determine the minimum sequencing coverage that can be used in order to balance assembly quality and sequencing costs. Genome coverage of *de novo* assembly projects in *A. thaliana*, published after I had undertaken similar comparative analyses, are highly variable. For example, Wang and colleagues (B. Wang et al. 2021) achieved a highly contiguous *de novo* assembly with 388x coverage, while Rabanal and colleagues (Rabanal et al. 2022) produced a high quality assembly with 133x coverage. However, Rabanal and colleagues (Rabanal et al. 2022) also showed that the impact of sequencing coverage on assembly quality as measured by contiguity depends on the assembler. For Falcon-Unzip2, the assembler that I used to determine the minimum sequencing coverage, they reported a drop in assembly contiguity below 50x. Moreover they also observed that the contiguity of Falcon-Unzip2 based assemblies is variable at coverages beyond 50x which is very similar to the observations that I made (**Figure 10**). My comprehensive analyses indicate that a total of three *A. thaliana* samples can be pooled on a single SMRT cell without compromising assembly contiguity. However, since the total base yield per SMRT cell differed substantially between different runs, we decided to not multiplex more then two samples in a single run.

It was unsurprising that increasing the quality threshold during CCS calling led to a lower percentage of CCS reads that were kept for genome assembly. Moreover, I could show that filtering for >Q20 reads resulted in a higher assembly contiguity compared to Q10 or Q30 filtered reads. This indicates that assembly contiguity is not only affected by genome coverage but also by read quality. If the assembly contiguity was only affected by coverage it would have been expected that using Q10 filtered reads results in a more contiguous assembly since the total number of available reads is higher compared to Q20 or even Q30. This is in agreement with PacBio company-provided protocols and former research articles that recommend filtering HiFi raw reads with a threshold of Q20 (Wenger et al. 2019; Naish et al. 2021).

Using reads with different lengths as an input for performing *de novo* assemblies had a minor impact on contiguity. This observation is in agreement with the findings reported by Rabanal et al. 2022. Furthermore, they also observed the generation of chimeric contigs when using Falcon-Unzip2, similar to what I have described (**Figure 13 & 14**).

Choosing the appropriate assembly tool can impact the quality of the *de novo* genome assembly (H. Cheng et al. 2021). In general, one can differentiate between assembly tools based on a greedy algorithm and graph based tools (Z. Li et al. 2012). At the moment, there are four commonly used tools for performing *de novo* genome assembly from PacBio HiFi reads: Falcon-Unzip2 (Chin et al. 2016), HiCanu (Nurk et al. 2020), Hifiasm (H. Cheng et al. 2021), and Improved Phased Assembler (IPA) (*Pbipa: Improved Phased Assembler*). Therefore, I compared the performance of Falcon-Unzip2 and Hifiasm, both of which are graph based assemblers (Chin et al. 2016; H. Cheng et al. 2021). I observed that Hifiasm produced more contiguous genome assemblies as measured by the NG50 value (**Figure 15**). Moreover, it outperformed Falcon-Unzip2 when comparing the longest contig of each assembly. Similar findings were reported for the model organisms *A. thaliana* (Rabanal et al. 2022) and *Drosophila melanogaster* (Gavrielatos et al. 2021). In addition, there is an increasing number of studies describing the superiority of Hifiasm for performing *de novo* genome assembly of complex non-model species such as human, *Zea mays*, *Fragaria x ananassa*, *Manihot esculenta*, *Dugesia tigrina*, and *Sminthurides aquaticus* (Gavrielatos et al. 2021; Garg et al. 2021; Qi et al. 2021; Schneider et al. 2021; H. Cheng et al. 2021). Interestingly, Hifiasm assemblies were longer on average and had more contigs compared to Falcon-Unzip2 based *de novo* assemblies (**Figure 15**). These observations are again in agreement with the findings reported by Rabanal and colleagues (Rabanal et al. 2022) and Gavrielatos and colleagues (Gavrielatos et al. 2021). In Rabanal et al. 2022, it was observed that these differences in accumulated contig length between Falcon-Unzip2 and Hifiasm-based assemblies can be attributed to organellar contigs as well as 5S rDNA containing contigs. I made similar observations.

Large structural mis-assemblies at the level of chromosome ‘fusions’ and megabase-scale translocations are often observed in draft genome assemblies (Rhie et al. 2021; Howe et al. 2021; Alonge et al. 2022). However, it is crucial to distinguish potential mis-assemblies from true biological variation, usually by looking at differences between a gold standard reference genome and the query assembly. I therefore tested RagTag 'correct' (Alonge et al. 2021), a contig correction tool that aligns the query contigs to a given reference genome. In order to benchmark the performance of the RagTag correction module, I corrected all primary assemblies using either the published *A. thaliana* genome TAIR10 or the optical map of AT9852 as a reference. The number of contigs that were broken (corrected) was counted afterwards. Since I also assembled AT9852 *de novo* it would be expected that in case of a perfect assembly, no contig would be corrected. However, even when correcting the *de novo* assembly of AT9852 with the optical map of the same accession two break points were introduced. In contrast to this, using TAIR10 for the correction of AT9852 caused more than ten contig breaks. This strongly indicates that the outcome of the RagTag correction step is heavily dependent on the reference genome, and that the contig correction step tends to make the query contigs more similar to the reference. Also in the other assemblies it was observed that the number of introduced breakpoints varied between the two references (**Table 1**). In many cases, contig breaks were introduced at completely different positions, again indicating that overcorrection of the query contigs may occur. I therefore decided to not use an automated contig correction tool.

5.1.3 Genome assembly

Common aims of performing a *de novo* genome assembly are gaining insights into the sequences of all genes and other sequence features as a basis for functional studies as well as generating a ‘reference’ genome that can be compared to other individuals of the species (Rice and Green 2019). Producing a genome assembly is usually a two stage process that starts with the assembly of contigs from the sequencing reads. Subsequently, these contigs are placed onto scaffolds resulting in the generation of pseudo-chromosomes (Hunt et al. 2014).

Choosing an appropriate sequencing technology is important since it affects downstream analyses (Goodwin, McPherson, and McCombie 2016). The use of long-read sequencing technologies is, compared to short-reads, especially well suited for producing contiguous *de novo* genome assemblies (Burgess 2018; Hon et al. 2020), as repetitive sequences can be spanned by individual sequencing reads, and thus resolved. However, long-reads typically suffer from a higher error rate compared to short-reads (Adewale 2020). The higher error rate of stand-alone long-read applications can be compensated by error correction methods or by combining data from long and short-read sequencing platforms (Chin et al. 2013; Koren et al. 2017). The PacBio HiFi sequencing technology used here allows for a compromise since it enables the generation of medium length reads (up to ~25 kb) with a base level precision of >99.8% (Hon et al. 2020; Lang et al. 2020; Wenger et al. 2019). The use of such highly accurate long-reads, as described in several studies, improves the quality of genome assemblies (Nurk et al. 2020; Shumate et al. 2020; Porubsky et al. 2021; Garg et al. 2021; Rabanal et al. 2022). A recent comparison of continuous long-read and circular consensus sequencing (CCS) of the same *A. thaliana* accession has shown that assembly contiguity is improved when using CCS reads (Rabanal et al. 2022). Compared to recent PacBio CLR genome assemblies of *A. thaliana* it was therefore expected to reach a slightly higher contiguity. The assemblies generated during the course of this thesis are meeting these expectations with contig N50 values of up to 13.2 Mb (**Figure 19**) compared to a maximum contig N50 value of 11.2 in the recent CLR assemblies (Jiao and Schneeberger 2020).

Metrics such as N50, cumulative length or NG50 only describe basic assembly properties without assessing the completeness of the assembled sequence (Simão et al. 2015). Therefore, it has been proposed to use a core set of highly conserved eukaryotic genes as a proxy (Parra et al. 2009; Seppey, Manni, and Zdobnov 2019). The completeness of these near-universal single-copy orthologs (BUSCO genes) is then used in order to infer the completeness of the query assembly (Simão et al. 2015). This approach is usually robust for widely studied species. However, a potential disadvantage of these methods is that the newly assembled query genome may contain true copy number variation or even sequence variants that were unknown at the time the core gene set was defined (Rhie et al. 2020). Since *A. thaliana* is a heavily studied model organism it is appropriate to use the new-universal single-copy orthologs as a proxy for completeness. All but one of the *de novo* assembled *A. thaliana* genomes were highly complete. The assembly of AT6035 had an exceptionally high rate of duplicated BUSCO genes compared to all other assemblies of this study. A potential explanation could be a higher level of heterozygosity or a sample contamination. However, kmer analysis of AT6035 did not indicate an increased number of heterozygous sites. In addition, the completeness assessment has been performed after contamination removal. Thus, heterozygosity and sample contamination in this case do not explain the increase in the number of duplicated BUSCO genes. Since the cause of the observed phenomenon remained unclear, I decided to exclude AT6035 from downstream analyses.

As mentioned before, the second step after assembling sequencing reads into contigs is to place those contigs on scaffolds in order to build pseudo-chromosomes. There are different approaches for scaffolding contigs (Rice and Green 2019). Proximity ligation protocols such as Hi-C are one option for scaffolding primary contigs into pseudo-chromosomes (Lieberman-Aiden et al. 2009). Another scaffolding approach is the use of optical maps (Shi et al. 2016; Seo et al. 2016) such as the commercially available Bionano protocol. that makes use of a nicking endonuclease that nicks long DNA fragments. Subsequently, these cutting sites are fluorescently labeled. The labeled DNA strands are then passed through an array of nanochannels that allows to determine the size of the fragment as well as the position of the fluorescent tag. Afterwards, these data are processed to a genome map that can be used for scaffolding contigs (Mak et al. 2016). A third scaffolding approach makes use of synteny between the query contigs and a given reference genome, ideally from the same species or from a closely related one. By aligning the contigs to a reference, information about contig position within a chromosome as well as contig orientation are inferred (Kolmogorov et al. 2014). In contrast to Hi-C or optical maps the synteny based methods do not require additional data. However, the disadvantage of these approaches is that they rely on a reference genome that needs to be assembled on the chromosome level. Furthermore it can be challenging when scaffolding the genomes of species with more structural variation (Rice and Green 2019). The scaffolds obtained here were highly contiguous (**Figure 21**) with fewer gaps compared to the current gold-standard *A. thaliana* reference genome TAIR10 (Berardini et al. 2015). Contigs were placed with high confidence scores (**Figure 24**). It is noteworthy that three contigs were assembled from telomere-to-telomere from the start, making further scaffolding obsolete. With the recent advent of long-read sequencing technologies such as the here applied PacBio Hifi or Oxford Nanopore there is an increasing number of studies reporting telomere-to-telomere assemblies. The higher eukaryotic species where individual chromosomes were assembled gap-free encompass human (Miga et al. 2020), fish (Xue et al. 2021), rice (Song et al. 2021), banana (Belser et al. 2021), and a marine diatom (*Phaeodactylum tricornutum*) (Giguere et al. 2021). Other researchers working with *A. thaliana* also reported telomere-to-telomere assemblies (Naish et al. 2021; Rabanal et al. 2022; B. Wang et al. 2021) when using similar technology. However, in contrast to my work, these studies either used longer insert sizes (Rabanal et al. 2022) or they used a combination of CSS and CLR sequencing (B. Wang et al. 2021). The experiments conducted by Rabanal et al. 2022 showed that insert sizes of PacBio HiFi reads beyond 21 kb increase assembly contiguity. Since the insert lengths used for this thesis are shorter, it was expected to obtain a slightly less contiguous assembly. The scaffolded portion of the *de novo* assemblies matched the estimated *A. thaliana* genome size (**Figure 22**) of ~135 Mb (Somerville and Koornneef 2002). However, all scaffolded genomes exceeded the 119 Mb length of the reference genome TAIR10, which does not contain centromeres (Arabidopsis Genome Initiative 2000). I have demonstrated that most of the variation in total assembly length can be explained by the successful assembly and scaffolding of centromeric repeats that accounted for up to 17 Mb of additional sequence (**Figure 25**; **Table 5**). This is in agreement with recent studies reporting that PacBio Hifi sequencing enables centromere assembly thus resulting in longer overall assemblies (Rabanal et al. 2022; B. Wang et al. 2021; Naish et al. 2021). In contrast to Rabanal et al. 2022 it was not in all cases possible to scaffold all centromere repeats. This could again be an effect of the shorter insert sizes used in my experiments.

A study dealing with the human genome showed that a large number of segmental duplications had been erroneously collapsed (She et al. 2004). In general it has been reported that highly repetitive regions of a genome may be collapsed in the final assembly (Peona et al. 2021; Tørresen et al. 2019; Vollger et al. 2019; Guiglielmoni et al. 2021; Phillippy, Schatz, and Pop 2008; Alkan, Sajjadian, and Eichler 2011). A special purpose of the here generated genome assemblies is to gain insights into NLR genes diversity among the eighteen accessions. However, in other species such as zebrafish it has been observed that some of the NLR genes were collapsed in the initial assembly due to their location in repetitive clusters (Howe et al. 2016). In all assemblies I found potentially collapsed regions. Their median length ranged from 6 to 23 kilobases. Moreover I showed that a large fraction of these collapsed sequences are found in highly repetitive regions such as centromeres, telomeres, and rDNA clusters (**Table 4**). However, compared to another recent study that used PacBio CLR sequencing for *de novo* assembly of six *A. thaliana* accessions I had fewer collapsed regions in the eighteen *de novo* assemblies (Jiao and Schneeberger 2020). All potentially collapsed regions were cataloged so that this information can be taken into account in future analyses.

Assembling highly repetitive regions remains challenging (H. Du and Liang 2019). As mentioned before it was possible to assemble and scaffold complete or nearly complete centromeres in all eighteen assemblies. In addition to the centromeric repeats I annotated 45S and 5S rDNAs. The cumulative length of the annotated 5S rDNA sequences ranged from 1.1 to 4.1 Mb (**Figure 27**). Compared to the assembly of Rabanal et al. 2022, encompassing 1.68 Mb, my assemblies contained longer 5S rDNA sequences (Rabanal et al. 2022). Similarly to the aforementioned study it was possible to scaffold the majority of 5S rDNA contigs.

The contigs containing 45S rDNA repeats accounted for up to 18 Mb. Their cumulative length was observed to be highly variable among the eighteen genomes. This is in agreement with previous reports (Q. Long et al. 2013). Most of these 45S rDNA contigs remained unplaced after scaffolding (**Figure 26**). This is again in agreement with previous findings (Rabanal et al. 2022). The lack of non-repetitive sequences at the edges of these 45S rDNA contigs makes it impossible to correctly anchor them to the scaffolded parts of the assembly (Rabanal et al. 2022). I have shown that HiFi sequencing enables the assembly of highly repetitive rDNA clusters. However, correct scaffolding of these repeats requires additional effort and could be achieved by leveraging different sequencing technologies such as Hi-C in combination with the use of specialized algorithms (Burton et al. 2013; H. Du and

Liang 2019).

As in most genome assembly adventures, several contigs remained unplaced after the scaffolding process. Recent studies (Rabanal et al. 2022) have shown that the choice of the assembly tools has an impact not only on the total assembly length but also on the cumulative length of contigs that cannot be scaffolded with the here used approach. In fact Rabanal et al. reported almost 50 Mb of unplaced contigs which is in agreement with my findings (7.8 to 64 Mb). I demonstrated that the vast majority of unplaced contigs share the following three characteristics: (i) low N50 values (**Figure 29**), (ii) high percentage of organellar or 45S rDNA sequences (**Figure 30**), and (iii) low quality of short-read mappings (**Figure 31**). Mapping of short-read data obtained from the 1001 Genomes project (1001 Genomes Consortium 2016) showed that most of the unplaced contigs are not informative in terms of mapping quality and they were thus removed. However, additional efforts on assembly method optimization could be made in order to avoid the excessive assembly of highly repetitive contigs from the beginning on.

5.1.4 Structural variation

Differences between genomes can range from single nucleotide polymorphisms to large-scale structural rearrangements such as insertions, deletions, inversions, translocations, and duplications. Such genomic differences are typically assessed by comparing a query genome to a reference (Goel et al. 2019). Genomic regions without structural differences are referred to as syntenic or collinear. The comparison between the eighteen genome assemblies and the *A. thaliana* reference genome TAIR10 revealed that the genomes are, as expected, mostly syntenic (**Figure 34**). In fact, the cumulative length of collinear sequences varied between 102 and 105 Mb which is almost identical to the values reported by a previous report that used PacBio CLR sequencing technology to assemble *A. thaliana* genomes (Jiao and Schneeberger 2020). Comparing the median length of different types of structural variants revealed that inversions were longer compared to translocations and duplications. These findings are again in agreement with previous reports using similar computational approaches (Jiao and Schneeberger 2020). However, the amount of sequences that could not be aligned to TAIR10 was higher in my assemblies. I have demonstrated that the vast majority of these unaligned sequence blocks are located around or in the centromeres (**Figure 36**). This is mostly due to the fact that centromeric regions are not well resolved in the current *A. thaliana* reference genome TAIR10 (Naish et al. 2021). Moreover, it has been shown that the continuous long-read sequencing technology is not well suited for disentangling centromeres (Rabanal et al. 2022). Their highly repetitive nature makes these sequences particularly challenging to assemble (Jain et al. 2018; B. Wang et al.

2021; Perumal et al. 2020). However, there is an increasing number of studies demonstrating the ability of accurate long-reads for assembling even such highly repetitive regions (Naish et al. 2021; B. Wang et al. 2021; Song et al. 2021; Rabanal et al. 2022; Miga et al. 2020; Xue et al. 2021; Giguere et al. 2021; Belser et al. 2021). Thus, the fact that I observed more unaligned sequence blocks compared to the study of Jiao and Schneeberger (Jiao and Schneeberger 2020) is likely due to the different sequencing technologies that were used.

5.1.5 Genome annotation

The broad application of Next-Generation-Sequencing technologies has lowered the costs for generating genome assemblies from various species (Mardis 2008). However, the genome sequence alone is of limited use if the encoded information is not deciphered (Mudge and Harrow 2016). The process of decoding the sequence patterns and associating them with protein coding genes is referred to as genome annotation (Ejigu and Jung 2020). However, prior to gene annotation it is necessary to identify and mask transposable elements (Yandell and Ence 2012).

Bursts of transposable element proliferation have contributed to genome size expansion in maize, cotton, rice, and legumes (Mun et al. 2009). TEs account for approximately 85% of the maize and wheat genome (Schnable et al. 2009; Null et al. 2018). Moreover, TEs occupy almost half of the sequence of the human genome (Mills et al. 2007). Different studies have reported that TEs account for approximately 21 Mb (18% to 20%) of the *A. thaliana* genome (Buisine, Quesneville, and Colot 2008; Jiao and Schneeberger 2020). I have shown that the TE content of the eighteen genomes assembled for this thesis varied between 12% and 25% with an average of 18%, although most of this was due to repeats that could not be attributed to known TE classes (**Figure 37**). Another recent study reported that approximately eight percent of their assemblies were made up of class I transposons while class II transposons accounted for almost five percent (Jiao and Schneeberger 2020). Using the before described TE identification pipeline I found that the eighteen genomes on average comprised 6.5% of class I and almost 6.5% of class II TEs. However, each of the assemblies contained between 2.8 Mb and 16 Mb of repeat sequences that could not be classified any further. This could be due to the complex nested structures of TEs that makes them challenging to annotate (Ou et al. 2019). The unexpected short cumulative lengths of the here identified TEs could also be an effect of not only annotating intact but also incomplete TEs. Obtaining a high quality TE annotation including a robust family assignment has been demonstrated to require extensive manual curation and community effort (Ou et al. 2019). However, the main purpose of performing transposable element annotation here was to mask these highly repetitive sequences rather than studying different TE families. Unmasked TE sequences can lead to an overestimation of the number of genes. Thus, the here produced repeat and TE annotations were sufficient to mask these sequences in all eighteen assemblies prior to gene annotation.

Many different annotation approaches have been developed. The applied methods rely on two different types of data: (i) intrinsic information such as *ab initio* predictions from the sequence itself or (ii) extrinsic information such as transcript or protein alignments (Ejigu and Jung 2020). Thus, for this thesis I have made use of both intrinsic *ab initio* predictions and extrinsic gene mapping information. The latter has been obtained by lifting all genes from the *A. thaliana* reference annotation TAIR10 onto the eighteen genome assemblies. This process, the mapping of genes from a previously annotated reference genome, is referred to as 'lift over' (Shumate and Salzberg 2020a). It was possible to lift over 97% of the known reference genes onto the eighteen query assemblies (**Figure 40**). The here obtained information was used as extrinsic evidence in the subsequent genome annotation. *Ab initio* gene predictions were made using a hidden Markov model employed by the annotation tool. Augustus (Stanke and Waack 2003) has been used to make the ab initio predictions and to combine them with the extrinsic lift over data. Thereby, it was possible to annotate approximately 32,100 genes per accession (**Figure 41**). This value is higher compared to the reference annotation TAIR10 (Lamesch et al. 2012) but lower when compared to the more recent, RNA-seq guided reference annotation Araport11 (C.-Y. Cheng et al. 2017). This points out that the result of a genome annotation process can be variable and depending on different factors such as the availability and quality of extrinsic information. However, from studies dealing with the annotation of maize it is known that up to 5% of all predicted gene models can be so called split-gene misannotations (Monnahan et al. 2020). In such cases one gene is incorrectly split and thereby annotated as two distinct genes (Denton et al. 2014). Such cases of mis-annotations have also been observed in the annotations I made. Thus, it is likely that the number of genes is slightly over-estimated. However, resolving such cases either requires extensive manual curation or additional extrinsic information such as transcript mappings (McDonnell, Strasser, and Tsang 2018; Hosmani et al. 2019; Monnahan et al. 2020). Thus, instead of only counting the number of genes, I assessed the completeness of each of the eighteen annotations. The applied procedure is similar to the before described assessment of assembly completeness (Seppey, Manni, and Zdobnov 2019). Here I used a set of known genes in order to infer annotation completeness. The potential of such approaches for assessing completeness has been described in other species as well (Mende et al. 2013; Parra et al. 2009). Thereby, it was possible to demonstrate that all eighteen annotations were highly complete (**Figure 42**). The outcome of performing a whole genome annotation is depending on the quality of the input assembly (Ejigu and Jung 2020). All eighteen assemblies were highly contiguous (**Figure 23**) and showed high completeness levels (**Figure 20**) as assessed by using universal single-copy-orthologs. Thus, it was expected that the quality of the genome assemblies is also reflected by the completeness of the genome annotations.

A key challenge in comparative genomics is determining phylogenetic relationships between genes of different species or in the case of this thesis to assess phylogenetic relationships between genes of different *A. thaliana* accessions (Emms and Kelly 2019). Moreover, the inference of homology relationships between genes allows for the extrapolation of biological knowledge between the eighteen input accessions and the reference TAIR10 (Emms and Kelly 2015). There are different computational methods to infer such relationships (Kristensen et al. 2011; Altenhoff et al. 2016; Nichio, Marchaukoski, and Raittz 2017). Most of these tools rely on the analyses of pairwise sequence similarity scores obtained by BLAST (Altschul et al. 1990) or DIAMOND (Camacho et al. 2009; Buchfink, Xie, and Huson 2014). The here applied approach that uses sequence similarity scores obtained by DIAMOND lead to 95% of all genes being assigned to orthogroups. An orthogroup in this case is defined as the set of genes from multiple species descended from a single gene in the last common ancestor of that species (Emms and Kelly 2019). Moreover, I have shown that the vast majority of annotated genes have a conserved copy number. Similar findings for *A. thaliana* were reported recently (Jiao and Schneeberger 2020). The median protein length of genes that are shared among all eighteen accessions and the reference TAIR10 matched the expected length for *A. thaliana* (Lamesch et al. 2012). Furthermore, the median number of accession specific or private genes in the eighteen assemblies was 74 genes, which is again in agreement with the published data (Jiao and Schneeberger 2020). Moreover, I have shown that the median length of these private genes differs from the length of genes that are found in all input assemblies (**Figure 45**). There are three different potential explanations for the origin of these seemingly private genes. One possibility is that the accession-specific genes are ‘generated’ by the annotation pipeline that has been used for this thesis. As discussed before, split gene mis-annotations can occur. Such mis-annotated genes should have the following characteristics: (i) low sequence similarity with known TAIR10 genes and (ii) short overall length. Indeed, many of the private genes were shorter compared to shared genes. The second potential way of interpreting the accession-specific genes is that these genes are novel. In the literature, there are reports showing a relationship between evolutionary age and gene length (Carvunis et al. 2012; Palmieri, Kosiol, and Schlötterer 2014; Albà and Castresana 2004; Choi and Kim 2006). For example in yeast (*Saccharomyces cerevisiae*), it has been demonstrated that younger genes are significantly shorter compared to evolutionary older ones (Capra, Pollard, and Singh 2010). The third interpretation of these private genes is that they were incorrectly marked as private during the orthofinder analyses. In order to exclude this possibility, I used blast to search for hits in a TAIR10 based database. Most private genes had e-values much higher than the genes that were shared among multiple accessions (**Figure 46**). Thus, an incorrect classification is unlikely. Therefore, it is more likely that these accession-specific genes are either mis-annotated split genes or novel genes. As already discussed before, it is hard to disentangle potentially mis-annotated split genes. However, the use of RNA sequencing could shed light on the question if these genes are really novel or just artifacts, since some of the novel genes should be supported by evidence from transcripts (Durand et al. 2019; Witt et al. 2019; Neme and Tautz 2016; Nagalakshmi et al. 2008; Kapranov and St Laurent 2012). Besides annotating the accession specific genes, I calculated z-scores of orthogroups sizes in order to identify genes that are present in the reference annotation TAIR10 while being absent in the eighteen accessions (**Figure 47**). Hereby, it was possible to show that the list of reference genes which are absent from the *de novo* assemblies was highly enriched for organellar genes (**Figure 48**). This result is unsurprising given that organellar genomes were not assembled as part of my work.

The presented work is embedded in a larger research effort focussing on the diversity of plant immune genes within a species and on how this diversity impacts the interaction of *A. thaliana* and its obligate biotrophic pathogen *Hyaloperonospora arabidopsidis* (Koch and Slusarenko 1990; Mauch-Mani and Slusarenko 1993). The plant immune system relies on immunity receptors that can activate immune signaling upon pathogen contact. These receptors can be grouped into (i) cell-surface proteins that detect microbe associated molecular patterns and (ii) intracellular receptors that recognize pathogen effectors (Jones and Dangl 2006). Many of these intracellular receptors are nucleotide-binding leucine-rich repeat receptors (NLRs) which are encoded by genes with high intraspecific variation (Jones, Vance, and Dangl 2016; Monteiro and Nishimura 2018). This genetic repertoire of NLR genes represents a valuable resource for breeding more disease resistant crops. However, the annotation of such immune genes remains challenging mostly due to their genetic diversity combined with their repetitive structure and the fact that NLR genes often occur in clusters (Zhang 2020; Q. Li, Jiang, and Shao 2021). In *A. thaliana* for example it has been reported that more than half of the known NLR genes are located in clusters (Blake C. Meyers et al. 2003; Van de Weyer et al. 2019). During the course of this thesis I have therefore applied different strategies in order to provide a basis for studying NLR gene diversity among the eighteen accessions. First, I determined the presence of each NLR gene known from the reference TAIR10. Subsequently, an independent lift-over approach, using a set of NLR genes that has been described in a recent *A. thaliana* pan-NLRome project, has been applied (Van de Weyer et al. 2019).

The first approach, assigning NLR orthologs between the eighteen genomes and TAIR10, showed that almost all of the 166 reference NLR genes from the Col-0 accession were detected in at least two out of the eighteen input genomes. Moreover, I have demonstrated that the annotated NLR genes can be highly variable in terms of copy numbers (**Figure 50**). This is in agreement with other studies reporting that many of the *A. thaliana* NLRs exhibit copy number variation (Van de Weyer et al. 2019; Lee and Chae 2020; MacQueen et al. 2019). Massive variation in copy numbers of NLR genes has also been described in other, agriculturally relevant, plant species such as soybean, rice, sorghum or maize (Weidong Wang et al. 2021; Hufford et al. 2021; Schatz et al. 2014; Yu et al. 2011; Read et al. 2020; Zheng et al. 2011). Such gene copies are thought to serve as reservoirs for generating novel resistances by mutation or intergenic recombination (Michelmore and Meyers 1998; Ohno 1970). The more copies of a gene are present the more likely it becomes that beneficial mutations arise since multiple copies constitute a larger target for mutations compared to a single copy (Barragan and Weigel 2021). However, the copy number estimation for the eighteen assemblies could partially be an effect of mis-annotated split genes. In cases where an existing NLR is wrongly split into two distinct genes, both of these genes would be considered to be copies. Thus, further manual curation as it has been done for the *A. thaliana* pan NLRome study will be required for disentangling such cases (Van de Weyer et al. 2019).

For the second, independent NLR annotation approach, the NLR genes annotated by Van de Weyer et al. 2019 were lifted over onto the eighteen assemblies. This approach enabled the detection of up to 255 NLR genes in a single accession (**Figure 51**). Van de Weyer et al. 2019 annotated NLR genes in 64 accessions. Per accession, they detected between 167 to 251 NLR genes. Although being slightly higher, the presented numbers in this thesis are in agreement with the findings of the *A. thaliana* pan NLRome project. The slightly higher number of NLR genes in the eighteen accessions is likely an effect of the different sequencing approaches between the eighteen genome assemblies and the pan NLRome project. The pan NLRome project, although being based on long-read sequencing data, relied on NLR-specific baits that were created to hybridize with more than 730 NLR genes from different *Brassicaceae* (Van de Weyer et al. 2019). Thus, only those NLR genes that show sequence similarity to any of the known NLR genes were annotated. In contrast, NLR genes in the eighteen accessions were annotated on a whole genome level. Thus, it is conceivable that this may lead to the identification of a higher number of NLRs. This hypothesis is indeed supported by the benchmarking that I performed for this thesis. In short, two genomes of accessions that were part of the pan NLRome project were sequenced for this study. When lifting over the pan NLRome genes onto these two genomes, it was observed that the number of NLR genes obtained when using the full genome information was slightly higher compared to the numbers reported by Van de Weyer et al. 2019. This again points out the benefit of having the additional information provided by whole genome assemblies in contrast to the bait approach of Van de Weyer et al. 2019. Counting the occurrences of the different NLR gene families revealed that TNLs represent the most abundant family in the eighteen accessions (**Figure 54**). This observation is again in agreement with the (Van de Weyer et al. 2019).

It can be summarized that using the eighteen *de novo* assemblies for NLR gene annotation allowed for the identification of non-reference NLR genes. Moreover, I demonstrated that the majority of NLR genes are located in clusters. In addition, I showed that copy number variation as well as presence-absence variation are common phenomena when dealing with *A. thaliana* NLR genes. Thus, the here generated dataset provides a robust foundation for studying NLR gene diversity in the eighteen differential lines.

5.2 Differential gene expression in *A. thaliana* F_1 hybrids

5.2.1 Custom F_1 hybrid reference based on long-read assemblies

In order to make use of the availability of full genome information from both parents (Ler-0 and Col-0), I combined the parental genome sequences into one ‘hybrid genome’. Afterwards I assessed mapping rates between the standard approach of short-read mapping to a single reference genome and short-read mapping to the ‘hybrid genome’. Overall mapping rates did only differ slightly between the two references, indicating that the effect of using the 'hybrid genome' is neglectable in this case. A recent study, comparing the genomes of Ler-0 and Col-0 showed that there are only 40 genes specific to Ler-0 and 60 genes specific to Col-0 (Zapata et al. 2016). Thus, the here presented approach might be more powerful when working with more divergent genotypes.

5.2.2 Tissue type explains most of the transcriptomic variation

I demonstrated that transcriptomic differences between tissues are the main driver of variance in the dataset. The observed variation was stronger compared to the effect of different genotypes or timepoints (**Figure 55**). The transcriptomes of flower samples showed the greatest difference when being compared to any other tissue in this dataset. These findings are in agreement with other studies investigating tissue-specific gene expression in plants, birds and mammals (Apelt et al. 2022; Breschi et al. 2016; Laine et al. 2019; Lenz et al. 2016). Another study dealing with the impact of a high sugar diet on gene expression in different drosophila tissues also reported the origin of tissue as the main contributor to variance in the dataset (Ng’oma et al. 2020). Moreover, a study comparing expression patterns of different tissues between mouse and human showed that samples from similar tissues are more similar between the two species compared to samples from different tissues of the same species (Zheng-Bradley et al. 2010). The fact that a strong separation between flower and all other transcriptomes was observed may reflect the heterogeneity of flower tissue in this experiment. Thus, separate RNA-seq of the different flower organs could increase the resolution of the aforementioned analysis. The impact of these findings is increasingly reflected by the growing number of studies that apply single-cell RNA sequencing that aim at compensating for heterogeneity not only among different tissues but also among individual cells of the same tissue (Hwang, Lee, and Bang 2018).

5.2.3 Non-additive gene expression in F_1 hybrids

The crossing of two inbred parents leads to the generation of F_1 hybrids which can exhibit a great variety of non-parental phenotypes. Such non-additive phenotypes deviate from the mean of the parents for a given trait. I focused on identifying genes in a Ler-0xCol-0 hybrid that are expressed in a non-additive manner meaning that their expression levels significantly deviated from the mid-parent value. With this study, I showed that the number of non-additively expressed genes varies among different tissues and time points with flowers showing the highest number of such genes (**Table 10**). However, this could be an effect of flowers containing more different tissues as compared to roots or leaves. Moreover, it was observed that the majority of non-additively expressed genes was upregulated compared to the respective mid-parent value. Furthermore, I demonstrated that the differences in the number of non-additively expressed genes among the different samples cannot be explained by an overall change in the number of expressed genes. Various studies have already compared parental transcriptomes to the transcriptomes of their corresponding F_1 hybrid offspring (Ryo Fujimoto et al. 2018). In maize hybrids, the transcriptomes of roots, embryos, immature ears, and seedlings were analyzed and compared to their corresponding parents (Hu et al. 2016; Jahnke et al. 2010; Paschold et al. 2012; Stupar et al. 2008; Swanson-Wagner et al. 2006). The findings reported in these studies are mostly in agreement with the results that I obtained during my experiments. First, it was reported that in all tissues and at all time points, the number of genes expressed in an additive manner was higher compared to the number of non-additively expressed genes. In samples from all tissues and time points that were analyzed for this thesis, I also found that the majority of genes in F_1 hybrids were exhibiting additive expression levels. Moreover, it was reported in maize and rice that non-additively expressed genes often are specific to a certain tissue or a certain developmental stage (Jahnke et al. 2010; Wei et al. 2009; Meyer et al. 2004). In the case of the here analyzed *A. thaliana* F_1 hybrids it was observed that the number of non-additively expressed genes greatly varies between the different tissues and timepoints (**Figure 59**).

Moreover, I found that only a small number of the non-additively expressed genes were shared among all tissues and timepoints. In addition, it was observed that most of the non-additively expressed genes were upregulated compared to the mid-parent value rather than downregulated. This is also in agreement with the findings that were obtained from a transcriptome analysis of *A. thaliana* Col-0xC24 hybrids (R. Fujimoto et al. 2012). However, the enrichment of non-additively expressed genes with photosynthesis related genes was not observed in the hybrids used for my experiments. Instead, non-additively expressed genes were significantly enriched for terms related to ‘defense response’ (**Figure 65**). These differences could explain the lack of biomass heterosis in the Ler-0xCol-0 hybrids compared to Col-0xC24 hybrids.

5.2.4 Parental expression divergence correlates with deviation from MPV

I reported that the list of genes that were significantly differentially expressed between the parents was enriched for genes where F_1 hybrids showed non-additive expression patterns (**Figure 66**). Subsequent analyses revealed that the degree of expression variation among the inbred parents was positively correlated with the deviation of hybrid expression levels from the mid-parent value (**Figure 67**). Thus, greater deviation from the mid-parent value in the hybrids was associated with greater expression divergence among the parental lines. The overlap of genes that are differentially expressed between the parents while also deviating from the mid parent value in hybrid transcriptomes has also been described in maize (Paschold et al. 2012). Moreover, other studies dealing with maize hybrids have reported that the extent of differential gene expression in maize is positively correlated with genetic distance (Stupar et al. 2008). In addition there are two reports that describe a correlation of parental gene expression and hybrid vigor suggesting that parental transcriptomes could be used in order to predict hybrid performance in maize (Frisch et al. 2010; Thiemann et al. 2010).

6. Conclusion & Outlook

For the first part of my work, ‘18 differential *A. thaliana*’ lines it can be summarized, that the initially proposed research questions were answered during the course of this thesis. I produced 18 highly contiguous *de novo* genome assemblies including transposable element and gene annotations. With further analyses I demonstrated that PacBio HiFi sequencing enables the assembly of highly repetitive regions such as centromeres. Moreover, I used these *de novo* assemblies to assess the number and prevalence of different NLR genes among the 18 accessions. As mentioned before, the work presented here is embedded in a larger research effort focussing on the diversity of plant immune genes within the species and on how this diversity impacts the interaction of *A. thaliana* and its obligate biotrophic pathogen *Hyaloperonospora arabidopsidis.* The genome assemblies that I generated will serve as the basis for upcoming projects focusing on the manual annotation and characterization of NLR gene diversity among the eighteen lines. With the data I generated at hand, in-depth comparisons as well as experimental studies can now take place. The NLR gene annotations as well as the genome assemblies will certainly be a valuable resource for these upcoming projects. The next step for further characterization of NLR gene diversity among the eighteen lines could be to improve the annotations I produced. This could be achieved by generating accession specific long- and short-read RNA sequencing data. Such expression data would allow to refine annotations and to also assess potential isoforms of the NLR genes.

The part of my thesis, dealing with non-additive gene expression in F_1 hybrids of *A. thaliana* confirmed observations that were also made in hybrid offspring of other species such as maize or rice. In addition I demonstrated that in the case of non-additively expressed genes, the extent of parental expression divergence correlated with the deviation from the mid-parent expression value in the F_1 hybrid offspring. Moreover, I proposed a computational approach that allowed me to combine full length genome information from two inbred parents in order to analyze the transcriptomes of F_1 hybrids. Although not being a game changer in terms of overall statistics in my dataset, it would be interesting to test this approach in F_1 hybrids that exhibit more extreme phenotypes such as hybrid necrosis or biomass heterosis.

In summary the main outcomes of this thesis are high quality genome assemblies and annotations of 18 differential accessions, as well as a computational approach that allows us to utilize such resources in order to analyze heterosis in F_1 hybrids. The data resources generated here will certainly aid in future research dealing with genetic variation in *Arabidopsis thaliana*.

7. Appendix

7.1 Additional figures

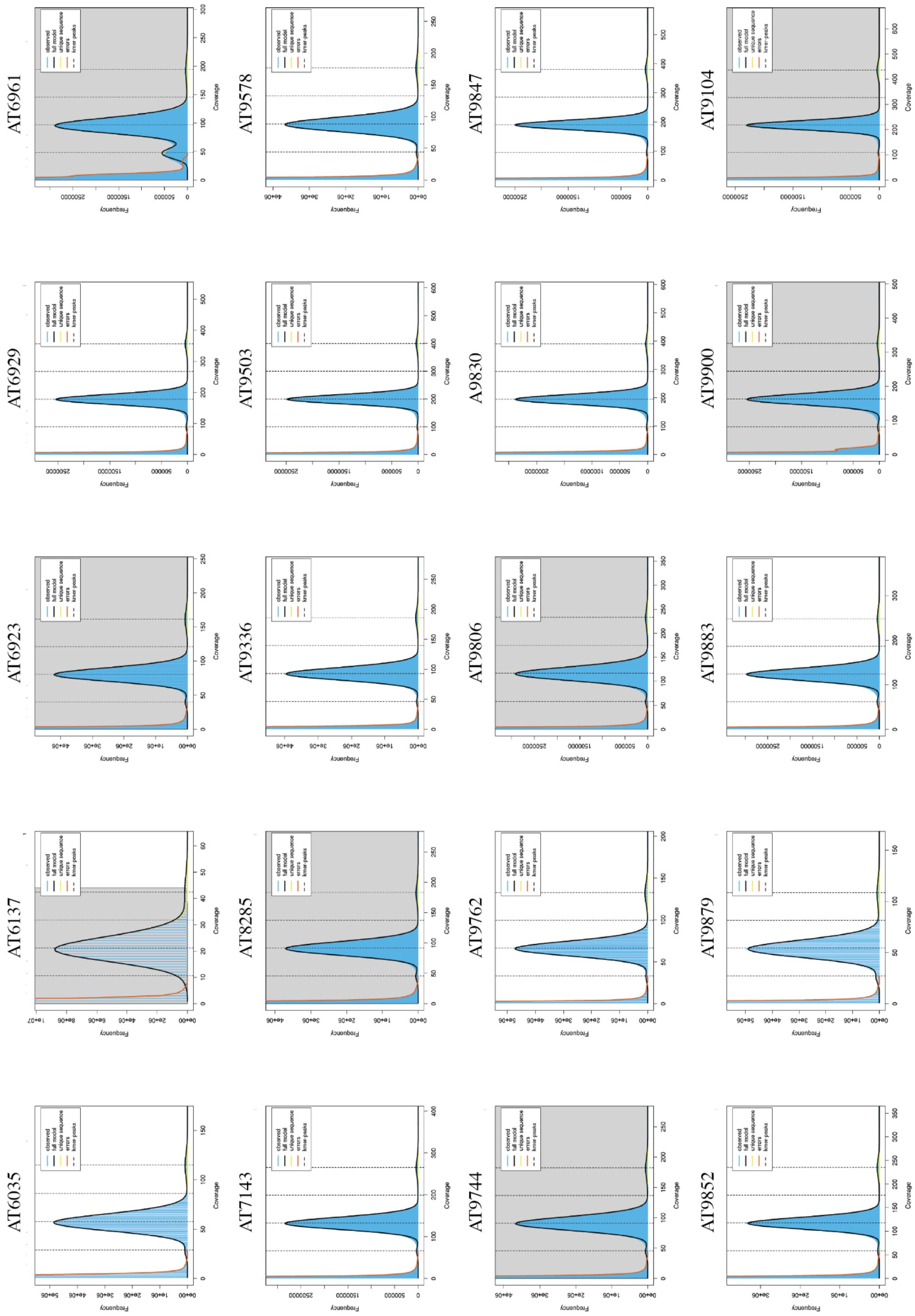

Figure S1: Kmer frequency analyses of the 20 differential lines.

7.2 Additional Tables

Table S1: Origin of the 18 differential lines.

Accession	Country	Long	Lat
AT6035	SWE	13.74	56.1
AT6137	SWE	13.5603	55.9419
AT6923	UK	-0.6383	51.4083
AT6929	TJK	68.49	38.48
AT6961	ESP	-3.53333	38.3333
AT7143	NED	5.86667	51.0167
AT8285	CZE	16.2815	49.4112
AT9104	GEO	46.2831	41.8296
AT9336	SWE	18.4473	62.8794
AT9503	UK	-3.21072	55.8877
AT9578	ESP	-6.7	42.13
AT9744	ROU	27.59	47.16
AT9762	ITA	14.98	37.69
AT9806	GER	9.16	48.56
AT9830	ESP	-3.28	36.97
AT9847	ESP	-5.7	43.31
AT9852	ESP	-3.75	40.46
AT9879	ESP	-1.12	37.6
AT9883	ESP	-2.56	42.1
AT9900	ESP	-6.01	37.38

Table S2: Kmer based estimation of minimum and maximum heterozygosity levels of primary assemblies of the initially 20 differential lines.

Genotype	Min. heterozygosity	Max. heterozygosity
AT6035	0.03%	0.03%
AT6137	0.03%	0.04%
AT6923	0.02%	0.03%
AT6929	0.01%	0.02%
AT6961	0.30%	0.31%
AT7143	0.02%	0.03%
AT8285	0.03%	0.03%
AT9336	0.02%	0.02%
AT9503	0.01%	0.02%
AT9578	0.02%	0.03%
AT9744	0.02%	0.03%
AT9762	0.02%	0.03%
AT9806	0.02%	0.03%
AT9830	0.01%	0.02%
AT9847	0.01%	0.02%
AT9852	0.02%	0.03%
AT9879	0.04%	0.04%
AT9883	0.02%	0.03%
AT9900	0.01%	0.03%

Table S3: Assembly completeness before and after filtering and removal of unplaced contigs.

	Single (%)		Duplicated (%)		Fragmented (%)		Missing (%)	
Genotype	Before filtering	After filtering	Before filtering	After filtering	Before filtering	After filtering	Before filtering	After filtering
AT6137	98.8	98.8	0.4	0.6	0.1	0.1	0.7	0.5
AT6923	98.6	98.5	0.6	0.7	0.1	0.2	0.7	0.6
AT6929	98.8	98.8	0.5	0.7	0.1	0.1	0.6	0.4
AT7143	98.7	98.6	0.4	0.6	0.2	0.1	0.7	0.7
AT8285	98.8	98.6	0.5	0.7	0.1	0.2	0.6	0.5
AT9104	98.7	98.6	0.6	0.9	0.1	0.1	0.6	0.4
AT9336	98.6	98.6	0.6	0.8	0.2	0.1	0.6	0.5
AT9503	98.6	98.6	0.5	0.7	0.2	0.2	0.7	0.5
AT9578	98.7	98.6	0.6	0.7	0.1	0.1	0.6	0.6
AT9744	98.6	98.4	0.7	0.9	0.1	0.1	0.6	0.6
AT9762	98.3	98.5	0.9	0.9	0.1	0.1	0.7	0.5
AT9806	98.6	98.6	0.6	0.7	0.1	0.1	0.7	0.6
AT9830	98.6	98.4	0.6	0.8	0.2	0.2	0.6	0.6
AT9847	98.9	98.8	0.4	0.7	0.1	0.1	0.6	0.4
AT9852	98.6	98.6	0.6	0.7	0.1	0.1	0.7	0.6
AT9879	98.7	98.7	0.5	0.7	0.1	0.1	0.7	0.5
AT9883	98.7	98.5	0.5	0.7	0.1	0.1	0.7	0.7
AT9900	98.8	98.6	0.5	0.7	0.1	0.1	0.6	0.6

8. References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, and 1001 Genomes Consortium. 2016. “1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis Thaliana.” *Cell* 166 (2): 481–91.

Abramovitch, Robert B., Jeffrey C. Anderson, and Gregory B. Martin. 2006. “Bacterial Elicitation and Evasion of Plant Innate Immunity.” *Nature Reviews. Molecular Cell Biology* 7 (8): 601–11.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, et al. 2000. “The Genome Sequence of Drosophila Melanogaster.” *Science* 287 (5461): 2185–95.

Adewale, Boluwatife A. 2020. “Will Long-Read Sequencing Technologies Replace Short-Read Sequencing Technologies in the next 10 Years?” *South African Journal of Laboratory and Clinical Medicine. Suid-Afrikaanse Tydskrif Vir Laboratorium- En Kliniekwerk* 9 (1): 1340.

Albà, M. Mar, and Jose Castresana. 2004. “Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes.” *Molecular Biology and Evolution* 22 (3): 598–606.

Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. “Genome Structural Variation Discovery and Genotyping.” *Nature Reviews. Genetics* 12 (5): 363–76.

Alkan, Can, Saba Sajjadian, and Evan E. Eichler. 2011. “Limitations of next-Generation Genome Sequence Assembly.” *Nature Methods* 8 (1): 61–65.

Alonge, Michael, Ludivine Lebeigle, Melanie Kirsche, Sergey Aganezov, Xingang Wang, Zachary B. Lippman, Michael C. Schatz, and Sebastian Soyk. 2021. “Automated Assembly Scaffolding Elevates a New Tomato System for High-Throughput Genome Editing.” *bioRxiv*. https://doi.org/10.1101/2021.11.18.469135.

Alonge, Michael, Ludivine Lebeigle, Melanie Kirsche, Katie Jenike, Shujun Ou, Sergey Aganezov, Xingang Wang, Zachary B. Lippman, Michael C. Schatz, and Sebastian Soyk. 2022. “Automated Assembly Scaffolding Using RagTag Elevates a New Tomato System for High-Throughput Genome Editing.” *Genome Biology* 23 (1): 258.

Altenhoff, Adrian M., Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A. Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, et al. 2016. “Standardized Benchmarking in the Quest for Orthologs.” *Nature Methods* 13 (5): 425–30.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Amin, Mohammad Ruhul, Alisa Yurovsky, Yingtao Tian, and Steven Skiena. 2018. “DeepAnnotator: Genome Annotation with Deep Learning.” In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 254–59. BCB ’18. New York, NY, USA: Association for Computing Machinery.

Anderson, S. 1981. “Shotgun DNA Sequencing Using Cloned DNase I-Generated

Fragments.” *Nucleic Acids Research* 9 (13): 3015–27.

Ansorge, Wilhelm J. 2009. “Next-Generation DNA Sequencing Techniques.” *New Biotechnology* 25 (4): 195–203.

Ansorge, W., B. S. Sproat, J. Stegemann, and C. Schwager. 1986. “A Non-Radioactive Automated Method for DNA Sequence Determination.” *Journal of Biochemical and Biophysical Methods* 13 (6): 315–23.

Apelt, Federico, Eleni Mavrothalassiti, Saurabh Gupta, Frank Machin, Justyna Jadwiga Olas, Maria Grazia Annunziata, Dana Schindelasch, and Friedrich Kragler. 2022. “Shoot and Root Single Cell Sequencing Reveals Tissue- and Daytime-Specific Transcriptome Profiles.” *Plant Physiology* 188 (2): 861–78.

Arabidopsis Genome Initiative. 2000. “Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana.” *Nature* 408 (6814): 796–815.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29.

Ausubel, Frederick M. 2005. “Are Innate Immune Signaling Pathways in Plants and Animals Conserved?” *Nature Immunology* 6 (10): 973–79.

Baggs, Erin L., J. Grey Monroe, Anil S. Thanki, Ruby O’Grady, Christian Schudoma, Wilfried Haerty, and Ksenia V. Krasileva. 2020. “Convergent Loss of an EDS1/PAD4 Signaling Pathway in Several Plant Lineages Reveals Coevolved Components of Plant Immunity and Drought Response.” *The Plant Cell* 32 (7): 2158–77.

Bakker, Erica G., M. Brian Traw, Christopher Toomajian, Martin Kreitman, and Joy Bergelson. 2008. “Low Levels of Polymorphism in Genes That Control the Activation of Defense Response in Arabidopsis Thaliana.” *Genetics* 178 (4): 2031–43.

Baptista, Rodrigo P., Joao Luis Reis-Cunha, Jeremy D. DeBarry, Egler Chiari, Jessica C. Kissinger, Daniella C. Bartholomeu, and Andrea M. Macedo. 2018. “Assembly of Highly Repetitive Genomes Using Short Reads: The Genome of Discrete Typing Unit III Trypanosoma Cruzi Strain 231.” *Microbial Genomics* 4 (4). https://doi.org/10.1099/mgen.0.000156.

Barragan, A. Cristina, and Detlef Weigel. 2021. “Plant NLR Diversity: The Known Unknowns of Pan-NLRomes.” *The Plant Cell* 33 (4): 814–31.

Barragan, Ana Cristina, Maximilian Collenberg, Jinge Wang, Rachelle R. Q. Lee, Wei Yuan Cher, Fernando A. Rabanal, Haim Ashkenazy, Detlef Weigel, and Eunyoung Chae. 2021. “A Truncated Singleton NLR Causes Hybrid Necrosis in Arabidopsis Thaliana.” *Molecular Biology and Evolution* 38 (2): 557–74.

Bell, Ellen A., Christopher L. Butler, Claudio Oliveira, Sarah Marburger, Levi Yant, and Martin I. Taylor. 2022. “Transposable Element Annotation in Non-Model Species: The Benefits of Species-Specific Repeat Libraries Using Semi-Automated EDTA and DeepTE de Novo Pipelines.” *Molecular Ecology Resources* 22 (2): 823–33.

Belser, Caroline, Franc-Christophe Baurens, Benjamin Noel, Guillaume Martin, Corinne Cruaud, Benjamin Istace, Nabila Yahiaoui, et al. 2021. “Telomere-to-Telomere Gapless Chromosomes of Banana Using Nanopore Sequencing.” *Communications Biology* 4 (1): 1–12.

Belser, Caroline, Benjamin Istace, Erwan Denis, Marion Dubarry, Franc-Christophe Baurens, Cyril Falentin, Mathieu Genete, et al. 2018. “Chromosome-Scale Assemblies of Plant Genomes Using Nanopore Long Reads and Optical Maps.” *Nature Plants* 4 (11): 879–87.

Ben-Hur, Asa, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. 2008. “Support Vector Machines and Kernels for Computational Biology.” *PLoS Computational Biology* 4 (10): e1000173.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry.” *Nature* 456 (7218): 53–59.

Bergman, Casey M., and Hadi Quesneville. 2007. “Discovering and Detecting Transposable Elements in Genome Sequences.” *Briefings in Bioinformatics* 8 (6): 382–92.

Birchler, James A., Hong Yao, Sivanandan Chudalayandi, Daniel Vaiman, and Reiner A. Veitia. 2010. “Heterosis.” *The Plant Cell* 22 (7): 2105–12.

Boeke, J. D., D. J. Garfinkel, C. A. Styles, and G. R. Fink. 1985. “Ty Elements Transpose through an RNA Intermediate.” *Cell* 40 (3): 491–500.

Boller, Thomas, and Georg Felix. 2009. “A Renaissance of Elicitors: Perception of Microbe-Associated Molecular Patterns and Danger Signals by Pattern-Recognition Receptors.” *Annual Review of Plant Biology* 60: 379–406.

Bomblies, Kirsten, Janne Lempe, Petra Epple, Norman Warthmann, Christa Lanz, Jeffery L. Dangl, and Detlef Weigel. 2007. “Autoimmune Response as a Mechanism for a Dobzhansky-Muller-Type Incompatibility Syndrome in Plants.” *PLoS Biology* 5 (9): e236.

Bomblies, Kirsten, Levi Yant, Roosa A. Laitinen, Sang-Tae Kim, Jesse D. Hollister, Norman Warthmann, Joffrey Fitz, and Detlef Weigel. 2010. “Local-Scale Patterns of Genetic Variability, Outcrossing, and Spatial Structure in Natural Stands of Arabidopsis Thaliana.” *PLoS Genetics* 6 (3): e1000890.

Breschi, Alessandra, Sarah Djebali, Jesse Gillis, Dmitri D. Pervouchine, Alex Dobin, Carrie A. Davis, Thomas R. Gingeras, and Roderic Guigó. 2016. “Gene-Specific Patterns of Expression Variation across Organs and Species.” *Genome Biology* 17 (1): 151.

Bruce, A. B. 1910. “THE MENDELIAN THEORY OF HEREDITY AND THE AUGMENTATION OF VIGOR.” *Science* 32 (827): 627–28.

Brutus, Alexandre, Francesca Sicilia, Alberto Macone, Felice Cervone, and Giulia De Lorenzo. 2010. “A Domain Swap Approach Reveals a Role of the Plant Wall-Associated Kinase 1 (WAK1) as a Receptor of Oligogalacturonides.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9452–57.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2014. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Buisine, Nicolas, Hadi Quesneville, and Vincent Colot. 2008. “Improved Detection and Annotation of Transposable Elements in Sequenced Genomes Using Multiple Reference Sequence Sets.” *Genomics* 91 (5): 467–75.

Burgess, Darren J. 2018. “Genomics: Next Regeneration Sequencing for Reference Genomes.” *Nature Reviews. Genetics*.

Burton, Joshua N., Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. 2013. “Chromosome-Scale Scaffolding of de Novo Genome Assemblies

Based on Chromatin Interactions.” *Nature Biotechnology* 31 (12): 1119–25.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (December): 421.

Campell, B. R., Y. Song, T. E. Posch, C. A. Cullis, and C. D. Town. 1992. “Sequence and Organization of 5S Ribosomal RNA-Encoding Genes of Arabidopsis Thaliana.” *Gene* 112 (2): 225–28.

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

Capra, John A., Katherine S. Pollard, and Mona Singh. 2010. “Novel Genes Exhibit Distinct Patterns of Function Acquisition and Network Integration.” *Genome Biology* 11 (12): R127.

Carvalho, Claudia M. B., and James R. Lupski. 2016. “Mechanisms Underlying Structural Variant Formation in Genomic Disorders.” *Nature Reviews. Genetics* 17 (4): 224–38.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, et al. 2012. “Proto-Genes and de Novo Gene Birth.” *Nature* 487 (7407): 370–74.

Chae, Eunyoung, Kirsten Bomblies, Sang-Tae Kim, Darya Karelina, Maricris Zaidem, Stephan Ossowski, Carmen Martín-Pizarro, et al. 2014. “Species-Wide Genetic Incompatibility Analysis Identifies Immune Genes as Hot Spots of Deleterious Epistasis.” *Cell* 159 (6): 1341–51.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. “Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes.” *Nature Communications* 10 (1): 1784.

Chambon, P. 1975. “Eukaryotic Nuclear RNA Polymerases.” *Annual Review of Biochemistry* 44 (1): 613–38.

Chan, Saki, Ernest Lam, Michael Saghbini, Sven Bocklandt, Alex Hastie, Han Cao, Erik Holmlin, and Mark Borodkin. 2018. “Structural Variation Detection and Analysis Using Bionano Optical Mapping.” *Methods in Molecular Biology* 1833: 193–203.

Cheng, Chia-Yi, Vivek Krishnakumar, Agnes P. Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D. Town. 2017. “Araport11: A Complete Reannotation of the Arabidopsis Thaliana Reference Genome.” *The Plant Journal: For Cell and Molecular Biology* 89 (4): 789–804.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. “Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm.” *Nature Methods* 18 (2): 170–75.

Chen, Z. Jeffrey. 2013. “Genomic and Epigenetic Insights into the Molecular Bases of Heterosis.” *Nature Reviews. Genetics* 14 (7): 471–82.

Chikhi, Rayan, and Paul Medvedev. 2013. “Informed and Automated K-Mer Size Selection for Genome Assembly.” *Bioinformatics* 30 (1): 31–37.

Chin, Chen-Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. “Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data.” *Nature Methods* 10 (6): 563–69.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. “Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.” *Nature Methods* 13 (12): 1050–54.

Chinchilla, Delphine, Zsuzsa Bauer, Martin Regenass, Thomas Boller, and Georg Felix. 2006. “The Arabidopsis Receptor Kinase FLS2 Binds flg22 and Determines the Specificity of Flagellin Perception.” *The Plant Cell* 18 (2): 465–76.

Chisholm, Stephen T., Gitta Coaker, Brad Day, and Brian J. Staskawicz. 2006. “Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response.” *Cell* 124 (4): 803–14.

Choi, In-Geol, and Sung-Hou Kim. 2006. “Evolution of Protein Structural Classes and Protein Sequence Families.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (38): 14056–61.

Christopoulou, Marilena, Sebastian Reyes-Chin Wo, Alex Kozik, Leah K. McHale, Maria-Jose Truco, Tadeusz Wroblewski, and Richard W. Michelmore. 2015. “Genome-Wide Architecture of Disease Resistance Genes in Lettuce.” *G3* 5 (12): 2655–69.

Clark, Richard M., Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, et al. 2007. “Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis Thaliana.” *Science* 317 (5836): 338–42.

Compeau, Phillip E. C., Pavel A. Pevzner, and Glenn Tesler. 2011. “How to Apply de Bruijn Graphs to Genome Assembly.” *Nature Biotechnology* 29 (11): 987–91.

Conway, Jake R., Alexander Lex, and Nils Gehlenborg. 2017. “UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties.” *Bioinformatics* 33 (18): 2938–40.

Cook, David E., Tong Geon Lee, Xiaoli Guo, Sara Melito, Kai Wang, Adam M. Bayless, Jianping Wang, et al. 2012. “Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean.” *Science* 338 (6111): 1206–9.

Copenhaver, G. P., J. H. Doelling, S. Gens, and C. S. Pikaard. 1995. “Use of RFLPs Larger than 100 Kbp to Map the Position and Internal Organization of the Nucleolus Organizer Region on Chromosome 2 in Arabidopsis Thaliana.” *The Plant Journal: For Cell and Molecular Biology* 7 (2): 273–86.

Copenhaver, G. P., and C. S. Pikaard. 1996. “RFLP and Physical Mapping with an rDNA-Specific Endonuclease Reveals That Nucleolus Organizer Regions of Arabidopsis Thaliana Adjoin the Telomeres on Chromosomes 2 and 4.” *The Plant Journal: For Cell and Molecular Biology* 9 (2): 259–72.

Couto, Daniel, and Cyril Zipfel. 2016. “Regulation of Pattern Recognition Receptor Signalling in Plants.” *Nature Reviews. Immunology* 16 (9): 537–52.

Dainat, Jacques. 2020. *NBISweden/AGAT: AGAT-v0.2.3*. https://doi.org/10.5281/zenodo.3714855.

Dangl, Jeffery L., Diana M. Horvath, and Brian J. Staskawicz. 2013. “Pivoting the Plant Immune System from Dissection to Deployment.” *Science* 341 (6147): 746–51.

Dangl, Jeffery L., and John M. McDowell. 2006. “Two Modes of Pathogen Recognition by Plants.” *Proceedings of the National Academy of Sciences of the United States of America*.

Dangl, J. L., and J. D. Jones. 2001. “Plant Pathogens and Integrated Defence Responses to Infection.” *Nature* 411 (6839): 826–33.

Danilevicz, Monica Furaste, Cassandria Geraldine Tay Fernandez, Jacob Ian Marsh, Philipp Emanuel Bayer, and David Edwards. 2020. “Plant Pangenomics: Approaches, Applications and Advancements.” *Current Opinion in Plant Biology* 54 (April): 18–25.

Darwin, Charles. 1877. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. D. Appleton.

Davenport, C. B. 1908. “DEGENERATION, ALBINISM AND INBREEDING.” *Science* 28 (718): 454–55.

Debortoli, Nicolas, Xiang Li, Isobel Eyres, Diego Fontaneto, Boris Hespeels, Cuong Q. Tang, Jean-François Flot, and Karine Van Doninck. 2016. “Genetic Exchange among Bdelloid Rotifers Is More Likely Due to Horizontal Gene Transfer Than to Meiotic Sex.” *Current Biology: CB* 26 (6): 723–32.

Deng, Yiwen, Keran Zhai, Zhen Xie, Dongyong Yang, Xudong Zhu, Junzhong Liu, Xin Wang, et al. 2017. “Epigenetic Regulation of Antagonistic Receptors Confers Rice Blast Resistance with Yield Balance.” *Science* 355 (6328): 962–65.

Denton, James F., Jose Lugo-Martinez, Abraham E. Tucker, Daniel R. Schrider, Wesley C. Warren, and Matthew W. Hahn. 2014. “Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies.” *PLoS Computational Biology* 10 (12): e1003998.

Du, Huilong, and Chengzhi Liang. 2019. “Assembly of Chromosome-Scale Contigs by Efficiently Resolving Repetitive Sequences with Long Reads.” *Nature Communications* 10 (1): 5360.

Durand, Éléonore, Isabelle Gagnon-Arsenault, Johan Hallin, Isabelle Hatin, Alexandre K. Dubé, Lou Nielly-Thibault, Olivier Namy, and Christian R. Landry. 2019. “Turnover of Ribosome-Associated Transcripts from de Novo ORFs Produces Gene-like Characteristics Available for de Novo Gene Emergence in Wild Yeast Populations.” *Genome Research* 29 (6): 932–43.

Durvasula, Arun, Andrea Fulgione, Rafal M. Gutaker, Selen Irez Alacakaptan, Pádraic J. Flood, Célia Neto, Takashi Tsuchimatsu, et al. 2017. “African Genomes Illuminate the Early History and Transition to Selfing in Arabidopsis Thaliana.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (20): 5213–18.

Duvick, D. N. 2001. “Biotechnology in the 1930s: The Development of Hybrid Maize.” *Nature Reviews. Genetics* 2 (1): 69–74.

Du, Zhou, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. 2010. “agriGO: A GO Analysis Toolkit for the Agricultural Community.” *Nucleic Acids Research* 38 (Web Server issue): W64–70.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. “Real-Time DNA Sequencing from Single Polymerase Molecules.” *Science* 323 (5910): 133–38.

Ejigu, Girum Fitihamlak, and Jaehee Jung. 2020. “Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing.” *Biology* 9 (9).

https://doi.org/10.3390/biology9090295.

El-Metwally, Sara, Taher Hamza, Magdi Zakaria, and Mohamed Helmy. 2013. “Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges.” *PLoS Computational Biology* 9 (12): e1003345.

Emms, David M., and Steven Kelly. 2015. “OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy.” *Genome Biology* 16 (August): 157.

———. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Falconer, D. S. 1996. *Introduction to Quantitative Genetics*. Harlow, England: Prentice Hall.

Fedurco, Milan, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. 2006. “BTA, a Novel Reagent for DNA Attachment on Glass and Efficient Generation of Solid-Phase Amplified DNA Colonies.” *Nucleic Acids Research* 34 (3): e22.

Fernández-Escalada, Manuel, Ainhoa Zulet-González, Miriam Gil-Monreal, Ana Zabalza, Karl Ravet, Todd Gaines, and Mercedes Royuela. 2017. “Effects of EPSPS Copy Number Variation (CNV) and Glyphosate Application on the Aromatic and Branched Chain Amino Acid Synthesis Pathways in Amaranthus Palmeri.” *Frontiers in Plant Science* 8 (November): 1970.

Feschotte, Cédric, Ning Jiang, and Susan R. Wessler. 2002. “Plant Transposable Elements: Where Genetics Meets Genomics.” *Nature Reviews. Genetics* 3 (5): 329–41.

Feschotte, Cédric, and Ellen J. Pritham. 2007. “DNA Transposons and the Evolution of Eukaryotic Genomes.” *Annual Review of Genetics* 41: 331–68.

Fiddes, Ian T., Joel Armstrong, Mark Diekhans, Stefanie Nachtweide, Zev N. Kronenberg, Jason G. Underwood, David Gordon, et al. 2017. “Comparative Annotation Toolkit (CAT) - Simultaneous Clade and Personal Genome Annotation.” *bioRxiv*. https://doi.org/10.1101/231118.

Finnegan, D. J. 1989. “Eukaryotic Transposable Elements and Genome Evolution.” *Trends in Genetics: TIG* 5 (4): 103–7.

Flor, H. H. 1971. “Current Status of the Gene-For-Gene Concept.” *Annual Review of Phytopathology* 9 (1): 275–96.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (17): 9451–57.

Fujimoto, R., J. M. Taylor, S. Shirasawa, W. J. Peacock, and E. S. Dennis. 2012. “Heterosis of Arabidopsis Hybrids between C24 and Col Is Associated with Increased Photosynthesis Capacity.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (18): 7109–14.

Fujimoto, Ryo, Kosuke Uezono, Sonoko Ishikura, Kenji Osabe, W. James Peacock, and Elizabeth S. Dennis. 2018. “Recent Research on the Mechanism of Heterosis Is Important for Crop and Vegetable Breeding Systems.” *Breeding Science* 68 (2): 145–58.

Gan, Xiangchao, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe, et al. 2011. “Multiple Reference Genomes and

Transcriptomes for Arabidopsis Thaliana." *Nature* 477 (7365): 419–23.

Gao, Liping, Fabrice Roux, and Joy Bergelson. 2009. "Quantitative Fitness Effects of Infection in a Gene-for-Gene System." *The New Phytologist* 184 (2): 485–94.

Garcia-Hernandez, Margarita, Tanya Z. Berardini, Guanghong Chen, Debbie Crist, Aisling Doyle, Eva Huala, Emma Knee, et al. 2002. "TAIR: A Resource for Integrated Arabidopsis Data." *Functional & Integrative Genomics* 2 (6): 239–53.

Garg, Shilpa, Arkarachai Fungtammasan, Andrew Carroll, Mike Chou, Anthony Schmitt, Xiang Zhou, Stephen Mac, et al. 2021. "Chromosome-Scale, Haplotype-Resolved Assembly of Human Genomes." *Nature Biotechnology* 39 (3): 309–12.

Gaut, Brandon S., Stephen I. Wright, Carène Rizzon, Jan Dvorak, and Lorinda K. Anderson. 2007. "Recombination: An Underappreciated Factor in the Evolution of Plant Genomes." *Nature Reviews. Genetics* 8 (1): 77–84.

Gavrielatos, Marios, Konstantinos Kyriakidis, Demetrios A. Spandidos, and Ioannis Michalopoulos. 2021. "Benchmarking of next and Third Generation Sequencing Technologies and Their Associated Algorithms for de Novo Genome Assembly." *Molecular Medicine Reports* 23 (4). https://doi.org/10.3892/mmr.2021.11890.

Giguere, Daniel J., Alexander T. Bahcheli, Samuel S. Slattery, Rushali R. Patel, Martin Flatley, Bogumil J. Karas, David R. Edgell, and Gregory B. Gloor. 2021. "Telomere-to-Telomere Genome Assembly of Phaeodactylum Tricornutum." *bioRxiv*. https://doi.org/10.1101/2021.05.04.442596.

Goel, Manish, Hequan Sun, Wen-Biao Jiao, and Korbinian Schneeberger. 2019. "SyRI: Finding Genomic Rearrangements and Local Sequence Differences from Whole-Genome Assemblies." *Genome Biology* 20 (1): 277.

Goig, Galo A., Silvia Blanco, Alberto L. Garcia-Basteiro, and Iñaki Comas. 2020. "Contaminant DNA in Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability." *BMC Biology* 18 (1): 24.

Golicz, Agnieszka A., Jacqueline Batley, and David Edwards. 2016. "Towards Plant Pangenomics." *Plant Biotechnology Journal* 14 (4): 1099–1105.

Gong, Zhiyun, Yufeng Wu, Andrea Koblízková, Giovana A. Torres, Kai Wang, Marina Iovene, Pavel Neumann, et al. 2012. "Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution." *The Plant Cell* 24 (9): 3559–74.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews. Genetics* 17 (6): 333–51.

Grabundzija, Ivana, Simon A. Messing, Jainy Thomas, Rachel L. Cosby, Ilija Bilic, Csaba Miskey, Andreas Gogol-Döring, et al. 2016. "A Helitron Transposon Reconstructed from Bats Reveals a Novel Mechanism of Genome Shuffling in Eukaryotes." *Nature Communications* 7 (March): 10716.

Grant, M. R., J. M. McDowell, A. G. Sharpe, M. de Torres Zabala, D. J. Lydiate, and J. L. Dangl. 1998. "Independent Deletions of a Pathogen-Resistance Gene in *Brassica* and *Arabidopsis*." *Proceedings of the National Academy of Sciences of the United States of America* 95 (26): 15843–48.

Greenblatt, Irwin M., and R. Alexander Brink. 1963. "Transpositions of Modulator in Maize into Divided and Undivided Chromosome Segments." *Nature* 197 (4865): 412–13.

Guiglielmoni, Nadège, Antoine Houtain, Alessandro Derzelle, Karine Van Doninck, and Jean-François Flot. 2021. “Overcoming Uncollapsed Haplotypes in Long-Read Assemblies of Non-Model Organisms.” *BMC Bioinformatics* 22 (1): 303.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.

Haas, Brian J., Steven L. Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E. Allen, Joshua Orvis, Owen White, C. Robin Buell, and Jennifer R. Wortman. 2008. “Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments.” *Genome Biology* 9 (1): R7.

Hann, Dagmar R., and John P. Rathjen. 2007. “Early Events in the Pathogenicity of Pseudomonas Syringae on Nicotiana Benthamiana.” *The Plant Journal: For Cell and Molecular Biology* 49 (4): 607–18.

Han, Yujun, and Susan R. Wessler. 2010. “MITE-Hunter: A Program for Discovering Miniature Inverted-Repeat Transposable Elements from Genomic Sequences.” *Nucleic Acids Research* 38 (22): e199.

Heather, James M., and Benjamin Chain. 2016. “The Sequence of Sequencers: The History of Sequencing DNA.” *Genomics* 107 (1): 1–8.

He, Guangming, Xiaopeng Zhu, Axel A. Elling, Liangbi Chen, Xiangfeng Wang, Lan Guo, Manzhong Liang, et al. 2010. “Global Epigenetic and Transcriptional Trends among Two Rice Subspecies and Their Reciprocal Hybrids.” *The Plant Cell* 22 (1): 17–33.

Hirsch, Candice N., Jillian M. Foerster, James M. Johnson, Rajandeep S. Sekhon, German Muttoni, Brieanne Vaillancourt, Francisco Peñagaricano, et al. 2014. “Insights into the Maize Pan-Genome and Pan-Transcriptome.” *The Plant Cell* 26 (1): 121–35.

Hirsch, Candice N., Cory D. Hirsch, Alex B. Brohammer, Megan J. Bowman, Ilya Soifer, Omer Barad, Doron Shem-Tov, et al. 2016. “Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize.” *The Plant Cell* 28 (11): 2700–2714.

Holley, Robert W., Jean Apgar, Susan H. Merrill, and Paul L. Zubkoff. 1961. “NUCLEOTIDE AND OLIGONUCLEOTIDE COMPOSITIONS OF THE ALANINE-, VALINE-, AND TYROSINE-ACCEPTOR ‘SOLUBLE’ RIBONUCLEIC ACIDS OF YEAST.” *Journal of the American Chemical Society* 83 (23): 4861–62.

Holley, Robert W., James T. Madison, and Ada Zamir. 1964. “A New Method for Sequence Determination of Large Oligonucleotides.” *Biochemical and Biophysical Research Communications* 17 (4): 389–94.

Holley, R. W., J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir. 1965. “STRUCTURE OF A RIBONUCLEIC ACID.” *Science* 147 (3664): 1462–65.

Hollick, J. B., and V. L. Chandler. 1998. “Epigenetic Allelic States of a Maize Transcriptional Regulatory Locus Exhibit Overdominant Gene Action.” *Genetics* 150 (2): 891–97.

Holub, E. B. 2001. “The Arms Race Is Ancient History in Arabidopsis, the Wildflower.” *Nature Reviews. Genetics* 2 (7): 516–27.

Hon, Ting, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, et al. 2020. “Highly Accurate Long-Read HiFi Sequencing Data for Five Complex Genomes.” *Scientific Data* 7 (1): 399.

Hosmani, Prashant S., Teresa Shippy, Sherry Miller, Joshua B. Benoit, Monica Munoz-Torres, Mirella Flores-Gonzalez, Lukas A. Mueller, et al. 2019. “A Quick Guide for Student-Driven Community Genome Annotation.” *PLoS Computational Biology* 15 (4): e1006682.

Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. 2020. “Structural Variation in the Sequencing Era.” *Nature Reviews. Genetics* 21 (3): 171–89.

Howe, Kerstin, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey, and Jonathan Wood. 2021. “Significantly Improving the Quality of Genome Assemblies through Curation.” *GigaScience* 10 (1). https://doi.org/10.1093/gigascience/giaa153.

Huang, Ying, Shi-Yi Chen, and Feilong Deng. 2016. “Well-Characterized Sequence Features of Eukaryote Genomes and Implications for Ab Initio Gene Prediction.” *Computational and Structural Biotechnology Journal* 14 (July): 298–303.

Hufford, Matthew B., Arun S. Seetharam, Margaret R. Woodhouse, Kapeel M. Chougule, Shujun Ou, Jianing Liu, William A. Ricci, et al. 2021. “De Novo Assembly, Annotation, and Comparative Analysis of 26 Diverse Maize Genomes.” *Science* 373 (6555): 655–62.

Hufford, Matthew B., Xun Xu, Joost van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A. Cartwright, Robert J. Elshire, et al. 2012. “Comparative Population Genomics of Maize Domestication and Improvement.” *Nature Genetics* 44 (7): 808–11.

Hunt, Martin, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2014. “A Comprehensive Evaluation of Assembly Scaffolding Tools.” *Genome Biology* 15 (3): R42.

Hu, Xiaojiao, Hongwu Wang, Xizhou Diao, Zhifang Liu, Kun Li, Yujin Wu, Qianjin Liang, Hui Wang, and Changling Huang. 2016. “Transcriptome Profiling and Comparison of Maize Ear Heterosis during the Spikelet and Floret Differentiation Stages.” *BMC Genomics* 17 (1): 959.

Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang. 2018. “Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines.” *Experimental & Molecular Medicine* 50 (8): 1–14.

Hyman, E. D. 1988. “A New Method of Sequencing DNA.” *Analytical Biochemistry* 174 (2): 423–36.

International Rice Genome Sequencing Project. 2005. “The Map-Based Sequence of the Rice Genome.” *Nature* 436 (7052): 793–800.

Jahnke, Stephanie, Barbara Sarholz, Alexander Thiemann, Vera Kühr, José F. Gutiérrez-Marcos, Hartwig H. Geiger, Hans-Peter Piepho, and Stefan Scholten. 2010. “Heterosis in Early Seed Development: A Comparative Study of F1 Embryo and Endosperm Tissues 6 Days after Fertilization.” *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 120 (2): 389–400.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology* 36 (4): 338–45.

Jiang, Jiming, James A. Birchler, Wayne A. Parrott, and R. Kelly Dawe. 2003. “A Molecular View of Plant Centromeres.” *Trends in Plant Science* 8 (12): 570–75.

Jiang, Ning, Zhirong Bao, Svetlana Temnykh, Zhukuan Cheng, Jiming Jiang, Rod A. Wing, Susan R. McCouch, and Susan R. Wessler. 2002. “Dasheng: A Recently Amplified Nonautonomous Long Terminal Repeat Element That Is a Major Component of Pericentromeric Regions in Rice.” *Genetics* 161 (3): 1293–1305.

Jiao, Wen-Biao, and Korbinian Schneeberger. 2020. “Chromosome-Level Assemblies of Multiple Arabidopsis Genomes Reveal Hotspots of Rearrangements with Altered Evolutionary Dynamics.” *Nature Communications* 11 (1): 989.

Jin, Y. K., and J. L. Bennetzen. 1989. “Structure and Coding Properties of Bs1, a Maize Retrovirus-like Transposon.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (16): 6235–39.

Jones, Jonathan D. G., and Jeffery L. Dangl. 2006. “The Plant Immune System.” *Nature* 444 (7117): 323–29.

Jones, Jonathan D. G., Russell E. Vance, and Jeffery L. Dangl. 2016. “Intracellular Innate Immune Surveillance Devices in Plants and Animals.” *Science* 354 (6316). https://doi.org/10.1126/science.aaf6395.

Jorgensen, T. H., and B. C. Emerson. 2008. “Functional Variation in a Disease Resistance Gene in Populations of Arabidopsis Thaliana.” *Molecular Ecology* 17 (22): 4912–23.

Kapranov, Philipp, and Georges St Laurent. 2012. “Dark Matter RNA: Existence, Function, and Controversy.” *Frontiers in Genetics* 3 (April): 60.

Kawakatsu, Taiji, Shao-Shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J. Schmitz, Mark A. Urich, Rosa Castanon, et al. 2016. “Epigenomic Diversity in a Global Collection of Arabidopsis Thaliana Accessions.” *Cell* 166 (2): 492–505.

Knox, Andrea K., Taniya Dhillon, Hongmei Cheng, Alessandro Tondelli, Nicola Pecchioni, and Eric J. Stockinger. 2010. “CBF Gene Copy Number Variation at Frost Resistance-2 Is Associated with Levels of Freezing Tolerance in Temperate-Climate Cereals.” *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 121 (1): 21–35.

Kobayashi, Yasushi, and Detlef Weigel. 2007. “Move on Up, It’s Time for Change--Mobile Signals Controlling Photoperiod-Dependent Flowering.” *Genes & Development* 21 (19): 2371–84.

Koch, Eckhard, and Alan Slusarenko. 1990. “Arabidopsis Is Susceptible to Infection by a Downy Mildew Fungus.” *The Plant Cell* 2 (5): 437–45.

Kolmogorov, Mikhail, Brian Raney, Benedict Paten, and Son Pham. 2014. “Ragout-a Reference-Assisted Assembly Tool for Bacterial Genomes.” *Bioinformatics* 30 (12): i302–9.

Koning, A. P. Jason de, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, and David D. Pollock. 2011. “Repetitive Elements May Comprise over Two-Thirds of the Human Genome.” *PLoS Genetics* 7 (12): e1002384.

Koornneef, Maarten, and David Meinke. 2010. “The Development of Arabidopsis as a Model Plant.” *The Plant Journal: For Cell and Molecular Biology* 61 (6): 909–21.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36.

Kourelis, Jiorgos, and Renier A. L. van der Hoorn. 2018. “Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function.” *The Plant Cell* 30 (2): 285–99.

Krämer, Ute. 2015. “Planting Molecular Functions in an Ecological Context with Arabidopsis Thaliana.” *eLife* 4 (March). https://doi.org/10.7554/eLife.06100.

Krieger, Uri, Zachary B. Lippman, and Dani Zamir. 2010. “The Flowering Gene SINGLE FLOWER TRUSS Drives Heterosis for Yield in Tomato.” *Nature Genetics* 42 (5): 459–63.

Kristensen, David M., Yuri I. Wolf, Arcady R. Mushegian, and Eugene V. Koonin. 2011. “Computational Methods for Gene Orthology Inference.” *Briefings in Bioinformatics* 12 (5): 379–91.

Krüger, Julia, Colwyn M. Thomas, Catherine Golstein, Mark S. Dixon, Matthew Smoker, Saijun Tang, Lonneke Mulder, and Jonathan D. G. Jones. 2002. “A Tomato Cysteine Protease Required for Cf-2-Dependent Disease Resistance and Suppression of Autonecrosis.” *Science* 296 (5568): 744–47.

Kryukov, Kirill, and Tadashi Imanishi. 2016. “Human Contamination in Public Genome Assemblies.” *PloS One* 11 (9): e0162424.

Laetsch, Dominik R., and Mark L. Blaxter. 2017. “BlobTools: Interrogation of Genome Assemblies.” *F1000Research* 6 (1287): 1287.

Laine, Veronika N., Irene Verhagen, A. Christa Mateman, Agata Pijl, Tony D. Williams, Phillip Gienapp, Kees van Oers, and Marcel E. Visser. 2019. “Exploration of Tissue-Specific Gene Expression Patterns Underlying Timing of Breeding in Contrasting Temperature Environments in a Song Bird.” *BMC Genomics* 20 (1): 693.

Lam, Ernest T., Alex Hastie, Chin Lin, Dean Ehrlich, Somes K. Das, Michael D. Austin, Paru Deshpande, et al. 2012. “Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly.” *Nature Biotechnology* 30 (8): 771–76.

Lamesch, Philippe, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, et al. 2012. “The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools.” *Nucleic Acids Research* 40 (Database issue): D1202–10.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Lang, Dandan, Shilai Zhang, Pingping Ren, Fan Liang, Zongyi Sun, Guanliang Meng, Yuntao Tan, et al. 2020. “Comparison of the Two up-to-Date Sequencing Technologies for Genome Assembly: HiFi Reads of Pacific Biosciences Sequel II System and Ultralong Reads of Oxford Nanopore.” *GigaScience* 9 (12). https://doi.org/10.1093/gigascience/giaa123.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Lee, Rachelle R. Q., and Eunyoung Chae. 2020. “Variation Patterns of NLR Clusters in Arabidopsis Thaliana Genomes.” *Plant Communications* 1 (4): 100089.

Legrand, Sylvain, Thibault Caron, Florian Maumus, Sol Schvartzman, Leandro Quadrana, Eléonore Durand, Sophie Gallina, et al. 2019. “Differential Retention of Transposable

Element-Derived Sequences in Outcrossing Arabidopsis Genomes." *Mobile DNA* 10 (July): 30.

Leister, D., J. Kurth, D. A. Laurie, M. Yano, T. Sasaki, K. Devos, A. Graner, and P. Schulze-Lefert. 1998. "Rapid Reorganization of Resistance Gene Homologues in Cereal Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 95 (1): 370–75.

Lenz, Michael, Franz-Josef Müller, Martin Zenke, and Andreas Schuppert. 2016. "Principal Components Analysis and the Reported Low Intrinsic Dimensionality of Gene Expression Microarray Data." *Scientific Reports* 6 (June): 25696.

Lewis, S. E., S. M. J. Searle, N. Harris, M. Gibson, V. Lyer, J. Richter, C. Wiel, et al. 2002. "Apollo: A Sequence Annotation Editor." *Genome Biology* 3 (12): RESEARCH0082.

Li, Bo, and Colin N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics* 12 (August): 323.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.

Li, Heng. 2016. "Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences." *Bioinformatics* 32 (14): 2103–10.

———. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Li, Lei, and Detlef Weigel. 2021. "One Hundred Years of Hybrid Necrosis: Hybrid Autoimmunity as a Window into the Mechanisms and Evolution of Plant-Pathogen Interactions." *Annual Review of Phytopathology*, May. https://doi.org/10.1146/annurev-phyto-020620-114826.

Lin, K., N. Zhang, E. I. Severing, H. Nijveen, F. Cheng, R. G. F. Visser, X. Wang, D. de Ridder, and G. Bonnema. 2014. "Beyond Genomic Variation - Comparison and Functional Annotation of Three Brassica Rapa Genomes: A Turnip, a Rapid Cycling and a Chinese Cabbage." *BMC Genomics* 15 (1). https://doi.org/10.1186/1471-2164-15-250.

Li, Qian, Xing-Mei Jiang, and Zhu-Qing Shao. 2021. "Genome-Wide Analysis of NLR Disease Resistance Genes in an Updated Reference Genome of Barley." *Frontiers in Genetics* 12 (May): 694682.

Li, Ruiqiang, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, et al. 2010. "Building the Sequence Map of the Human Pan-Genome." *Nature Biotechnology* 28 (1): 57–63.

Li, Ying-Hui, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-Guo Jin, Zhouhao Zhang, Yong Guo, et al. 2014. "De Novo Assembly of Soybean Wild Relatives for Pan-Genome Analysis of Diversity and Agronomic Traits." *Nature Biotechnology* 32 (10): 1045–52.

Li, Zhenyu, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, et al. 2012. "Comparison of the Two Major Classes of Assembly Algorithms: Overlap-Layout-Consensus and de-Bruijn-Graph." *Briefings in Functional Genomics* 11

(1): 25–37.

Long, E. O., and I. B. Dawid. 1980. “Repeated Genes in Eukaryotes.” *Annual Review of Biochemistry* 49: 727–64.

Longin, Carl Friedrich Horst, Jonathan Mühleisen, Hans Peter Maurer, Hongliang Zhang, Manje Gowda, and Jochen Christoph Reif. 2012. “Hybrid Breeding in Autogamous Cereals.” *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 125 (6): 1087–96.

Long, Quan, Fernando A. Rabanal, Dazhe Meng, Christian D. Huber, Ashley Farlow, Alexander Platzer, Qingrun Zhang, et al. 2013. “Massive Genomic Variation and Strong Selection in Arabidopsis Thaliana Lines from Sweden.” *Nature Genetics* 45 (8): 884–90.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Luckey, J. A., H. Drossman, A. J. Kostichka, D. A. Mead, J. D’Cunha, T. B. Norris, and L. M. Smith. 1990. “High Speed DNA Sequencing by Capillary Electrophoresis.” *Nucleic Acids Research* 18 (15): 4417–21.

Lu, Jennifer, and Steven L. Salzberg. 2018. “Removing Contaminants from Databases of Draft Genomes.” *PLoS Computational Biology* 14 (6): e1006277.

Lye, Zoe N., and Michael D. Purugganan. 2019. “Copy Number Variation in Domestication.” *Trends in Plant Science* 24 (4): 352–65.

MacQueen, Alice, Dacheng Tian, Wenhan Chang, Eric Holub, Martin Kreitman, and Joy Bergelson. 2019. “Population Genetics of the Highly Polymorphic RPP8 Gene Family.” *Genes* 10 (9). https://doi.org/10.3390/genes10090691.

Maheshwari, Shamoni, Takayoshi Ishii, C. Titus Brown, Andreas Houben, and Luca Comai. 2017. “Centromere Location in Arabidopsis Is Unaltered by Extreme Divergence in CENH3 Protein Sequence.” *Genome Research* 27 (3): 471–78.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. “Structural Variant Calling: The Long and the Short of It.” *Genome Biology* 20 (1): 246.

Mahood, Elizabeth H., Lars H. Kruse, and Gaurav D. Moghe. 2020. “Machine Learning: A Powerful Tool for Gene Function Prediction in Plants.” *Applications in Plant Sciences* 8 (7): e11376.

Mak, Angel C. Y., Yvonne Y. Y. Lai, Ernest T. Lam, Tsz-Piu Kwok, Alden K. Y. Leung, Annie Poon, Yulia Mostovoy, et al. 2016. “Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays.” *Genetics* 202 (1): 351–62.

Malik, H. S., and T. H. Eickbush. 2001. “Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses.” *Genome Research* 11 (7): 1187–97.

Mardis, Elaine R. 2008. “The Impact of next-Generation Sequencing Technology on Genetics.” *Trends in Genetics: TIG* 24 (3): 133–41.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, et al. 2005. “Genome Sequencing in Microfabricated High-Density Picolitre Reactors.” *Nature* 437 (7057): 376–80.

Martínez-Zapater, J., M. Estelle, and C. Somerville. 1986. “A Highly Repeated DNA Sequence in Arabidopsis Thaliana.” *Molecular & General Genetics: MGG*. https://doi.org/10.1007/BF00331018.

Martin, Jeffrey A., and Zhong Wang. 2011. “Next-Generation Transcriptome Assembly.” *Nature Reviews. Genetics* 12 (10): 671–82.

Mauch-Mani, B., and A. J. Slusarenko. 1993. “Arabidopsis as a Model Host for Studying Plant-Pathogen Interactions.” *Trends in Microbiology* 1 (7): 265–70.

McClintock, Barbara. 1950. “The Origin and Behavior of Mutable Loci in Maize.” *Proceedings of the National Academy of Sciences* 36 (6): 344–55.

McDonald, Michael J., Daniel P. Rice, and Michael M. Desai. 2016. “Sex Speeds Adaptation by Altering the Dynamics of Molecular Evolution.” *Nature* 531 (7593): 233–36.

McDonnell, Erin, Kimchi Strasser, and Adrian Tsang. 2018. “Manual Gene Curation and Functional Annotation.” *Methods in Molecular Biology* 1775: 185–208.

Melsted, Páll, and Bjarni V. Halldórsson. 2014. “KmerStream: Streaming Algorithms for K -Mer Abundance Estimation.” *Bioinformatics* 30 (24): 3541–47.

Mende, Daniel R., Shinichi Sunagawa, Georg Zeller, and Peer Bork. 2013. “Accurate and Universal Delineation of Prokaryotic Species.” *Nature Methods* 10 (9): 881–84.

Merchant, Samier, Derrick E. Wood, and Steven L. Salzberg. 2014. “Unexpected Cross-Species Contamination in Genome Sequencing Projects.” *PeerJ* 2 (November): e675.

Meyer, Rhonda C., Ottó Törjék, Martina Becher, and Thomas Altmann. 2004. “Heterosis of Biomass Production in Arabidopsis. Establishment during Early Development.” *Plant Physiology* 134 (4): 1813–23.

Meyer, Rhonda C., Hanna Witucka-Wall, Martina Becher, Anna Blacha, Anastassia Boudichevskaia, Peter Dörmann, Oliver Fiehn, et al. 2012. “Heterosis Manifestation during Early Arabidopsis Seedling Development Is Characterized by Intermediate Gene Expression and Enhanced Metabolic Activity in the Hybrids.” *The Plant Journal: For Cell and Molecular Biology* 71 (4): 669–83.

Meyers, B. C., D. B. Chin, K. A. Shen, S. Sivaramakrishnan, D. O. Lavelle, Z. Zhang, and R. W. Michelmore. 1998. “The Major Resistance Gene Cluster in Lettuce Is Highly Duplicated and Spans Several Megabases.” *The Plant Cell* 10 (11): 1817–32.

Meyers, Blake C., Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W. Michelmore. 2003. “Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis.” *The Plant Cell* 15 (4): 809–34.

Meyers, Blake C., Michele Morgante, and Richard W. Michelmore. 2002. “TIR-X and TIR-NBS Proteins: Two New Families Related to Disease Resistance TIR-NBS-LRR Proteins Encoded in Arabidopsis and Other Plant Genomes.” *The Plant Journal: For Cell and Molecular Biology* 32 (1): 77–92.

Michael, Todd P., Florian Jupe, Felix Bemm, S. Timothy Motley, Justin P. Sandoval, Christa Lanz, Olivier Loudet, Detlef Weigel, and Joseph R. Ecker. 2018. “High Contiguity Arabidopsis Thaliana Genome Assembly with a Single Nanopore Flow Cell.” *Nature Communications* 9 (1): 541.

Michelmore, R. W., and B. C. Meyers. 1998. “Clusters of Resistance Genes in Plants Evolve by Divergent Selection and a Birth-and-Death Process.” *Genome Research* 8 (11): 1113–30.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *Nature* 585 (7823): 79–84.

Mikheyev, Alexander S., and Mandy M. Y. Tin. 2014. “A First Look at the Oxford Nanopore MinION Sequencer.” *Molecular Ecology Resources* 14 (6): 1097–1102.

Miller, Jason R., Sergey Koren, and Granger Sutton. 2010. “Assembly Algorithms for next-Generation Sequencing Data.” *Genomics* 95 (6): 315–27.

Miller, Marisa, Qingxin Song, Xiaoli Shi, Thomas E. Juenger, and Z. Jeffrey Chen. 2015. “Natural Variation in Timing of Stress-Responsive Gene Expression Predicts Heterosis in Intraspecific Hybrids of Arabidopsis.” *Nature Communications* 6 (July): 7453.

Mills, Ryan E., E. Andrew Bennett, Rebecca C. Iskow, and Scott E. Devine. 2007. “Which Transposable Elements Are Active in the Human Genome?” *Trends in Genetics: TIG* 23 (4): 183–91.

Monnahan, Patrick J., Jean-Michel Michno, Christine O’Connor, Alex B. Brohammer, Nathan M. Springer, Suzanne E. McGaugh, and Candice N. Hirsch. 2020. “Using Multiple Reference Genomes to Identify and Resolve Annotation Inconsistencies.” *BMC Genomics* 21 (1): 281.

Monteiro, Freddy, and Marc T. Nishimura. 2018. “Structural, Functional, and Genomic Diversity of Plant NLR Proteins: An Evolved Resource for Rational Engineering of Plant Immunity.” *Annual Review of Phytopathology* 56 (August): 243–67.

Morel, J. B., and J. L. Dangl. 1997. “The Hypersensitive Response and the Induction of Cell Death in Plants.” *Cell Death and Differentiation* 4 (8): 671–83.

Mudge, Jonathan M., and Jennifer Harrow. 2016. “The State of Play in Higher Eukaryote Gene Annotation.” *Nature Reviews. Genetics* 17 (12): 758–72.

Mun, Jeong-Hwan, Soo-Jin Kwon, Tae-Jin Yang, Young-Joo Seol, Mina Jin, Jin-A Kim, Myung-Ho Lim, et al. 2009. “Genome-Wide Comparative Analysis of the Brassica Rapa Gene Space Reveals Genome Shrinkage and Differential Loss of Duplicated Genes after Whole Genome Triplication.” *Genome Biology* 10 (10): R111.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.” *Science* 320 (5881): 1344–49.

Nagarajan, Niranjan, and Mihai Pop. 2013. “Sequence Assembly Demystified.” *Nature Reviews. Genetics* 14 (3): 157–67.

Naish, Matthew, Michael Alonge, Piotr Wlodzimierz, Andrew J. Tock, Bradley W. Abramson, Anna Schmücker, Terezie Mandáková, et al. 2021. “The Genetic and Epigenetic Landscape of the *Arabidopsis* Centromeres.” *Science* 374 (6569): eabi7489.

Narusaka, Mari, Ken Shirasu, Yoshiteru Noutoshi, Yasuyuki Kubo, Tomonori Shiraishi, Masaki Iwabuchi, and Yoshihiro Narusaka. 2009. “RRS1 and RPS4 Provide a Dual Resistance-Gene System against Fungal and Bacterial Pathogens.” *The Plant Journal: For Cell and Molecular Biology* 60 (2): 218–26.

Neme, Rafik, and Diethard Tautz. 2016. “Fast Turnover of Genome Transcription across

Evolutionary Time Exposes Entire Non-Coding DNA to de Novo Gene Emergence.” *eLife* 5 (February): e09977.

Ng’oma, E., P. A. Williams-Simon, A. Rahman, and E. G. King. 2020. “Diverse Biological Processes Coordinate the Transcriptional Response to Nutritional Changes in a Drosophila Melanogaster Multiparent Population.” *BMC Genomics* 21 (1): 84.

Nichio, Bruno T. L., Jeroniza Nunes Marchaukoski, and Roberto Tadeu Raittz. 2017. “New Tools in Orthology Analysis: A Brief Review of Promising Perspectives.” *Frontiers in Genetics* 8 (October): 165.

Null, Null, Rudi Appels, Kellye Eversole, Nils Stein, Catherine Feuillet, Beat Keller, Jane Rogers, et al. 2018. “Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome.” *Science* 361 (6403): eaar7191.

Nurk, Sergey, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. 2020. “HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads.” *Genome Research*, August. https://doi.org/10.1101/gr.263566.120.

Nyrén, P., and A. Lundin. 1985. “Enzymatic Method for Continuous Monitoring of Inorganic Pyrophosphate Synthesis.” *Analytical Biochemistry* 151 (2): 504–9.

Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Springer, Berlin, Heidelberg.

Oliveira, Ludmila Cristina, and Giovana Augusta Torres. 2018. “Plant Centromeres: Genetics, Epigenetics and Evolution.” *Molecular Biology Reports* 45 (5): 1491–97.

Ossowski, Stephan, Korbinian Schneeberger, Richard M. Clark, Christa Lanz, Norman Warthmann, and Detlef Weigel. 2008. “Sequencing of Natural Strains of Arabidopsis Thaliana with Short Reads.” *Genome Research* 18 (12): 2024–33.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 20 (1): 275.

Palmieri, Nicola, Carolin Kosiol, and Christian Schlötterer. 2014. “The Life Cycle of Drosophila Orphan Genes.” *eLife* 3 (February): e01311.

Parra, Genis, Keith Bradnam, Zemin Ning, Thomas Keane, and Ian Korf. 2009. “Assessing the Gene Space in Draft Genomes.” *Nucleic Acids Research* 37 (1): 289–97.

Paschold, Anja, Yi Jia, Caroline Marcon, Steve Lund, Nick B. Larson, Cheng-Ting Yeh, Stephan Ossowski, et al. 2012. “Complementation Contributes to Transcriptome Complexity in Maize (Zea Mays L.) Hybrids Relative to Their Inbred Parents.” *Genome Research* 22 (12): 2445–54.

Paszkiewicz, Konrad, and David J. Studholme. 2010. “De Novo Assembly of Short Sequence Reads.” *Briefings in Bioinformatics* 11 (5): 457–72.

Pbipa: Improved Phased Assembler. n.d. Github. Accessed February 25, 2022. https://github.com/PacificBiosciences/pbipa.

Pellicer, Jaume, and Ilia J. Leitch. 2020. “The Plant DNA C-Values Database (release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies.” *The New Phytologist* 226 (2): 301–5.

Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, et al. 2021. "Identifying the Causes and Consequences of Assembly Gaps Using a Multiplatform Genome Assembly of a Bird-of-Paradise." *Molecular Ecology Resources* 21 (1): 263–86.

Perumal, Sampath, Chu Shin Koh, Lingling Jin, Miles Buchwaldt, Erin E. Higgins, Chunfang Zheng, David Sankoff, et al. 2020. "A High-Contiguity Brassica Nigra Genome Localizes Active Centromeres and Defines the Ancestral Brassica Genome." *Nature Plants* 6 (8): 929–41.

Pervez, Muhammad Tariq, Mirza Jawad Ul Hasnain, Syed Hassan Abbas, Mahmoud F. Moustafa, Naeem Aslam, and Syed Shah Muhammad Shah. 2022. "A Comprehensive Review of Performance of Next-Generation Sequencing Platforms." *BioMed Research International* 2022 (September): 3457806.

Phillippy, Adam M., Michael C. Schatz, and Mihai Pop. 2008. "Genome Assembly Forensics: Finding the Elusive Mis-Assembly." *Genome Biology* 9 (3): R55.

Pop, Mihai, Adam Phillippy, Arthur L. Delcher, and Steven L. Salzberg. 2004. "Comparative Genome Assembly." *Briefings in Bioinformatics* 5 (3): 237–48.

Porubsky, David, Peter Ebert, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Pierre Marijon, Jana Ebler, et al. 2021. "Fully Phased Human Genome Assembly without Parental Data Using Single-Cell Strand Sequencing and Long Reads." *Nature Biotechnology* 39 (3): 302–8.

Pumphrey, Michael, Jianfa Bai, Debbie Laudencia-Chingcuanco, Olin Anderson, and Bikram S. Gill. 2009. "Nonadditive Expression of Homoeologous Genes Is Established upon Polyploidization in Hexaploid Wheat." *Genetics* 181 (3): 1147–57.

Putney, S. D., W. C. Herlihy, and P. Schimmel. 1983. "A New Troponin T and cDNA Clones for 13 Different Muscle Proteins, Found by Shotgun Sequencing." *Nature* 302 (5910): 718–21.

Qi, Weihong, Yi-Wen Lim, Andrea Patrignani, Pascal Schläpfer, Anna Bratus-Neuenschwander, Simon Grüter, Christelle Chanez, et al. 2021. "The Haplotype-Resolved Chromosome Pairs and Transcriptome of a Heterozygous Diploid African Cassava Cultivar." *bioRxiv*. https://doi.org/10.1101/2021.11.16.468774.

Quail, Michael A., Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." *BMC Genomics* 13 (July): 341.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

Rabanal, Fernando A., Maike Graeff, Christa Lanz, Katrin Fritschi, Victor Llaca, Michelle Lang, Pablo Carbonell-Bejerano, Ian Henderson, and Detlef Weigel. 2022. "Pushing the Limits of HiFi Assemblies Reveals Centromere Diversity between Two Arabidopsis Thaliana Genomes." *bioRxiv*. https://doi.org/10.1101/2022.02.15.480579.

Ranallo-Benavidez, T. Rhyker, Kamil S. Jaron, and Michael C. Schatz. 2020. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11 (1): 1432.

Read, Andrew C., Matthew J. Moscou, Aleksey V. Zimin, Geo Pertea, Rachel S. Meyer, Michael D. Purugganan, Jan E. Leach, Lindsay R. Triplett, Steven L. Salzberg, and Adam J. Bogdanove. 2020. “Genome Assembly and Characterization of a Complex zfBED-NLR Gene-Containing Disease Resistance Locus in Carolina Gold Select Rice with Nanopore Sequencing.” *PLoS Genetics* 16 (1): e1008571.

Rédei, György P. 1962. “Single Locus Heterosis.” *Zeitschrift Für Vererbungslehre* 93 (1): 164–70.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. “Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies.” *Genome Biology* 21 (1): 245.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89.

Rho, Mina, and Haixu Tang. 2009. “MGEScan-Non-LTR: Computational Identification and Classification of Autonomous Non-LTR Retrotransposons in Eukaryotic Genomes.” *Nucleic Acids Research* 37 (21): e143.

Rice, Edward S., and Richard E. Green. 2019. “New Approaches for Genome Assembly and Scaffolding.” *Annual Review of Animal Biosciences* 7 (February): 17–40.

Robatzek, Silke, Pascal Bittel, Delphine Chinchilla, Petra Köchner, Georg Felix, Shin-Han Shiu, and Thomas Boller. 2007. “Molecular Identification and Characterization of the Tomato Flagellin Receptor LeFLS2, an Orthologue of Arabidopsis FLS2 Exhibiting Characteristically Different Perception Specificities.” *Plant Molecular Biology* 64 (5): 539–47.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.

Rowan, Beth A., Darren Heavens, Tatiana R. Feuerborn, Andrew J. Tock, Ian R. Henderson, and Detlef Weigel. 2019. “An Ultra High-Density Arabidopsis Thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features.” *Genetics* 213 (3): 771–87.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. 2000. “Artemis: Sequence Visualization and Annotation.” *Bioinformatics* 16 (10): 944–45.

Salter, Susannah J., Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nick Loman, and Alan W. Walker. 2014. “Reagent Contamination Can Critically Impact Sequence-Based Microbiome Analyses.” *bioRxiv*. https://doi.org/10.1101/007187.

Salzberg, S. L., M. Pertea, A. L. Delcher, M. J. Gardner, and H. Tettelin. 1999. “Interpolated Markov Models for Eukaryotic Gene Finding.” *Genomics* 59 (1): 24–31.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. 1977. “Nucleotide Sequence of Bacteriophage Phi X174 DNA.” *Nature* 265 (5596): 687–95.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating

Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.

Saxena, Rachit K., David Edwards, and Rajeev K. Varshney. 2014. "Structural Variations in Plant Genomes." *Briefings in Functional Genomics* 13 (4): 296–307.

Schatz, Michael C., Lyza G. Maron, Joshua C. Stein, Alejandro Hernandez Wences, James Gurtowski, Eric Biggers, Hayan Lee, et al. 2014. "Whole Genome de Novo Assemblies of Three Divergent Strains of Rice, Oryza Sativa, Document Novel Gene Space of Aus and Indica." *Genome Biology* 15 (11): 506.

Scherer, Stephen W., Charles Lee, Ewan Birney, David M. Altshuler, Evan E. Eichler, Nigel P. Carter, Matthew E. Hurles, and Lars Feuk. 2007. "Challenges and Standards in Integrating Surveys of Structural Variation." *Nature Genetics* 39 (7 Suppl): S7–15.

Schmid, Michael, Daniel Frei, Andrea Patrignani, Ralph Schlapbach, Jürg E. Frey, Mitja N. P. Remus-Emsermann, and Christian H. Ahrens. 2018. "Pushing the Limits of de Novo Genome Assembly for Complex Prokaryotic Genomes Harboring Very Long, near Identical Repeats." *Nucleic Acids Research* 46 (17): 8953–65.

Schnable, Patrick S., and Nathan M. Springer. 2013. "Progress toward Understanding Heterosis in Crop Plants." *Annual Review of Plant Biology* 64 (February): 71–88.

Schnable, Patrick S., Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, et al. 2009. "The B73 Maize Genome: Complexity, Diversity, and Dynamics." *Science* 326 (5956): 1112–15.

Schneeberger, Korbinian, Stephan Ossowski, Felix Ott, Juliane D. Klein, Xi Wang, Christa Lanz, Lisa M. Smith, et al. 2011. "Reference-Guided Assembly of Four Diverse Arabidopsis Thaliana Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 108 (25): 10249–54.

Schneider, Clément, Christian Woehle, Carola Greve, Cyrille A. D'Haese, Magnus Wolf, Michael Hiller, Axel Janke, Miklós Bálint, and Bruno Huettel. 2021. "Two High-Quality de Novo Genomes from Single Ethanol-Preserved Specimens of Tiny Metazoans (Collembola)." *GigaScience* 10 (5). https://doi.org/10.1093/gigascience/giab035.

Sedlazeck, Fritz J., Hayan Lee, Charlotte A. Darby, and Michael C. Schatz. 2018. "Piercing the Dark Matter: Bioinformatics of Long-Range Sequencing and Mapping." *Nature Reviews. Genetics* 19 (6): 329–46.

Seo, Jeong-Sun, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, et al. 2016. "De Novo Assembly and Phasing of a Korean Human Genome." *Nature* 538 (7624): 243–47.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. "BUSCO: Assessing Genome Assembly and Annotation Completeness." *Methods in Molecular Biology* 1962: 227–45.

Sharbel, T. F., B. Haubold, and T. Mitchell-Olds. 2000. "Genetic Isolation by Distance in Arabidopsis Thaliana: Biogeography and Postglacial Colonization of Europe." *Molecular Ecology* 9 (12): 2109–18.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. "DNA Sequencing at 40: Past, Present and Future." *Nature* 550 (7676): 345–53.

She, Xinwei, Zhaoshi Jiang, Royden A. Clark, Ge Liu, Ze Cheng, Eray Tuzun, Deanna M.

Church, Granger Sutton, Aaron L. Halpern, and Evan E. Eichler. 2004. “Shotgun Sequence Assembly and Recent Segmental Duplications within the Human Genome.” *Nature* 431 (7011): 927–30.

Shi, Lingling, Yunfei Guo, Chengliang Dong, John Huddleston, Hui Yang, Xiaolu Han, Aisi Fu, et al. 2016. “Long-Read Sequencing and de Novo Assembly of a Chinese Genome.” *Nature Communications* 7 (1): 1–10.

Shiu, S. H., and A. B. Bleecker. 2001. “Receptor-like Kinases from *Arabidopsis* Form a Monophyletic Gene Family Related to Animal Receptor Kinases.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10763–68.

Shomura, Ayahiko, Takeshi Izawa, Kaworu Ebana, Takeshi Ebitani, Hiromi Kanegae, Saeko Konishi, and Masahiro Yano. 2008. “Deletion in a Gene Associated with Grain Size Increased Yields during Rice Domestication.” *Nature Genetics* 40 (8): 1023–28.

Shull, George Harrison. 1914. “Duplicate Genes for Capsule-Form in Bursa Bursa-Pastoris.” *Zeitschrift Fur Induktive Abstammungs- Und Vererbungslehre* 12 (1): 97–149.

Shumate, Alaina, and Steven L. Salzberg. 2020a. “Liftoff: An Accurate Gene Annotation Mapping Tool.” *bioRxiv*. https://doi.org/10.1101/2020.06.24.169680.

———. 2020b. “Liftoff: Accurate Mapping of Gene Annotations.” *Bioinformatics* , December. https://doi.org/10.1093/bioinformatics/btaa1016.

Shumate, Alaina, Aleksey V. Zimin, Rachel M. Sherman, Daniela Puiu, Justin M. Wagner, Nathan D. Olson, Mihaela Pertea, Marc L. Salit, Justin M. Zook, and Steven L. Salzberg. 2020. “Assembly and Annotation of an Ashkenazi Human Reference Genome.” *Genome Biology* 21 (1): 129.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12.

Simpson, Jared T., and Mihai Pop. 2015. “The Theory and Practice of Genome Sequence Assembly.” *Annual Review of Genomics and Human Genetics* 16 (April): 153–72.

Smit, A. F. A., R. Hubley, and P. Green. 2015. “RepeatMasker Open-4.0. 2013--2015.”

Smith, Lisa M., Kirsten Bomblies, and Detlef Weigel. 2011. “Complex Evolutionary Events at a Tandem Cluster of Arabidopsis Thaliana Genes Resulting in a Single-Locus Genetic Incompatibility.” *PLoS Genetics* 7 (7): e1002164.

Smith, Lloyd M., Steven Fung, Michael W. Hunkapiller, Tim J. Hunkapiller, and Leroy E. Hood. 1985. “The Synthesis of Oligonucleotides Containing an Aliphatic Amino Group at the 5′ Terminus: Synthesis of Fluorescent DNA Primers for Use in DNA Sequence Analysis.” *Nucleic Acids Research* 13 (7): 2399–2412.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. 1986. “Fluorescence Detection in Automated DNA Sequence Analysis.” *Nature* 321 (6071): 674–79.

Sohn, Jang-Il, and Jin-Wu Nam. 2018. “The Present and Future of de Novo Whole-Genome Assembly.” *Briefings in Bioinformatics* 19 (1): 23–40.

Somerville, Chris, and Maarten Koornneef. 2002. “A Fortunate Choice: The History of Arabidopsis as a Model Plant.” *Nature Reviews. Genetics* 3 (11): 883–89.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. “Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences.” *F1000Research* 4 (December): 1521.

Song, Jia-Ming, Wen-Zhao Xie, Shuo Wang, Yi-Xiong Guo, Dal-Hoe Koo, Dave Kudrna, Chenbo Gong, et al. 2021. “Two Gap-Free Reference Genomes and a Global View of the Centromere Architecture in Rice.” *Molecular Plant* 14 (10): 1757–67.

Sonnenburg, Sören, Gabriele Schweikert, Petra Philips, Jonas Behr, and Gunnar Rätsch. 2007. “Accurate Splice Site Prediction Using Support Vector Machines.” *BMC Bioinformatics* 8 Suppl 10 (Suppl 10): S7.

Springer, Nathan M., Sarah N. Anderson, Carson M. Andorf, Kevin R. Ahern, Fang Bai, Omer Barad, W. Brad Barbazuk, et al. 2018. “The Maize W22 Genome Provides a Foundation for Functional Genomics and Transposon Biology.” *Nature Genetics* 50 (9): 1282–88.

Staden, R. 1979. “A Strategy of DNA Sequencing Employing Computer Programs.” *Nucleic Acids Research* 6 (7): 2601–10.

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. 2004. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.” *Nucleic Acids Research* 32 (Web Server issue): W309–12.

Stanke, Mario, and Stephan Waack. 2003. “Gene Prediction with a Hidden Markov Model and a New Intron Submodel.” *Bioinformatics* 19 Suppl 2 (October): ii215–25.

Stitzer, Michelle C., Sarah N. Anderson, Nathan M. Springer, and Jeffrey Ross-Ibarra. 2021. “The Genomic Ecosystem of Transposable Elements in Maize.” *PLoS Genetics* 17 (10): e1009768.

Stupar, Robert M., Jack M. Gardiner, Aaron G. Oldre, William J. Haun, Vicki L. Chandler, and Nathan M. Springer. 2008. “Gene Expression Analyses in Maize Inbreds and Hybrids with Varying Levels of Heterosis.” *BMC Plant Biology* 8 (April): 33.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.

Sun, Hequan, Jia Ding, Mathieu Piednoël, and Korbinian Schneeberger. 2017. “findGSE: Estimating Genome Size Variation within Human and Arabidopsis Using K-Mer Frequencies.” *Bioinformatics* 34 (4): 550–57.

Sun, Silong, Yingsi Zhou, Jian Chen, Junpeng Shi, Haiming Zhao, Hainan Zhao, Weibin Song, et al. 2018. “Extensive Intraspecific Gene Order and Gene Structural Variations between Mo17 and Other Maize Genomes.” *Nature Genetics* 50 (9): 1289–95.

Swanson-Wagner, Ruth A., Yi Jia, Rhonda DeCook, Lisa A. Borsuk, Dan Nettleton, and Patrick S. Schnable. 2006. “All Possible Modes of Gene Action Are Observed in a Global Comparison of Gene Expression in a Maize F1 Hybrid and Its Inbred Parents.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (18): 6805–10.

Takai, Ryota, Akira Isogai, Seiji Takayama, and Fang-Sik Che. 2008. “Analysis of Flagellin Perception Mediated by flg22 Receptor OsFLS2 in Rice.” *Molecular Plant-Microbe Interactions: MPMI* 21 (12): 1635–42.

Tang, Haibao, John E. Bowers, Xiyin Wang, Ray Ming, Maqsudul Alam, and Andrew H.

Paterson. 2008. “Synteny and Collinearity in Plant Genomes.” *Science* 320 (5875): 486–88.

Tang, Lin. 2020. “Contamination in Sequence Databases.” *Nature Methods*.

Torkamaneh, Davoud, Marc-André Lemay, and François Belzile. 2021. “The Pan-Genome of the Cultivated Soybean (PanSoy) Reveals an Extraordinarily Conserved Gene Content.” *Plant Biotechnology Journal* 19 (9): 1852–62.

Tørresen, Ole K., Bastiaan Star, Pablo Mier, Miguel A. Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, et al. 2019. “Tandem Repeats Lead to Sequence Assembly Errors and Impose Multi-Level Challenges for Genome and Protein Databases.” *Nucleic Acids Research* 47 (21): 10994–6.

Travers, Kevin J., Chen-Shan Chin, David R. Rank, John S. Eid, and Stephen W. Turner. 2010. “A Flexible and Efficient Template Format for Circular Consensus Sequencing and SNP Detection.” *Nucleic Acids Research* 38 (15): e159.

Trdá, Lucie, Olivier Fernandez, Freddy Boutrot, Marie-Claire Héloir, Jani Kelloniemi, Xavier Daire, Marielle Adrian, et al. 2014. “The Grapevine Flagellin Receptor Vv FLS 2 Differentially Recognizes Flagellin-Derived Epitopes from the Endophytic Growth-Promoting Bacterium Burkholderia Phytofirmans and Plant Pathogenic Bacteria.” *The New Phytologist* 201 (4): 1371–84.

Trivedi, Urmi H., Timothée Cézard, Stephen Bridgett, Anna Montazam, Jenna Nichols, Mark Blaxter, and Karim Gharbi. 2014. “Quality Control of next-Generation Sequencing Data without a Reference.” *Frontiers in Genetics* 5 (May): 111.

Troyer, A. Forrest. 2006. “Adaptedness and Heterosis in Corn and Mule Hybrids.” *Crop Science* 46 (2): 528–43.

Van de Weyer, Anna-Lena, Freddy Monteiro, Oliver J. Furzer, Marc T. Nishimura, Volkan Cevik, Kamil Witek, Jonathan D. G. Jones, Jeffery L. Dangl, Detlef Weigel, and Felix Bemm. 2019. “A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis Thaliana.” *Cell* 178 (5): 1260–72.e14.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507): 1304–51.

Vollger, Mitchell R., Philip C. Dishuck, Melanie Sorensen, Annemarie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J. P. Chaisson, and Evan E. Eichler. 2019. “Long-Read Sequence and Assembly of Segmental Duplications.” *Nature Methods* 16 (1): 88–94.

Vrebalov, Julia, Diane Ruezinsky, Veeraragavan Padmanabhan, Ruth White, Diana Medrano, Rachel Drake, Wolfgang Schuch, and Jim Giovannoni. 2002. “A MADS-Box Gene Necessary for Fruit Ripening at the Tomato Ripening-Inhibitor (rin) Locus.” *Science* 296 (5566): 343–46.

Walden, K. K., and H. M. Robertson. 1997. “Ancient DNA from Amber Fossil Bees?” *Molecular Biology and Evolution* 14 (10): 1075–77.

Wang, Bo, Xiaofei Yang, Yanyan Jia, Yu Xu, Peng Jia, Ningxin Dang, Songbo Wang, et al. 2021. “High-Quality Arabidopsis Thaliana Genome Assembly with Nanopore and HiFi Long Reads.” *Genomics, Proteomics & Bioinformatics*, September. https://doi.org/10.1016/j.gpb.2021.08.003.

Wang, Weidong, Liyang Chen, Kevin Fengler, Joy Bolar, Victor Llaca, Xutong Wang, Chancelor B. Clark, et al. 2021. “A Giant NLR Gene Confers Broad-Spectrum Resistance to Phytophthora Sojae in Soybean.” *Nature Communications* 12 (1): 6263.

Wang, Wensheng, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, et al. 2018. “Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice.” *Nature* 557 (7703): 43–49.

Watson, J. D., and F. H. Crick. 1953. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38.

Wei, Gang, Yong Tao, Guozhen Liu, Chen Chen, Renyuan Luo, Hongai Xia, Qiang Gan, et al. 2009. “A Transcriptomic Analysis of Superhybrid Rice LYP9 and Its Parents.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (19): 7695–7701.

Weigel, D. 2012. “Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics.” *Plant Physiology* 158 (1): 2–22.

Weigel, Detlef, and Richard Mott. 2009. “The 1001 Genomes Project for Arabidopsis Thaliana.” *Genome Biology* 10 (5): 107.

Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. “Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome.” *Nature Biotechnology* 37 (10): 1155–62.

Wersch, Solveig van, and Xin Li. 2019. “Stronger When Together: Clustering of Plant NLR Disease Resistance Genes.” *Trends in Plant Science* 24 (8): 688–99.

Wickham, H. 2007. “Reshaping Data with the Reshape Package.” *Journal of Statistical Software*. https://www.jstatsoft.org/article/view/v021i12.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer Publishing Company, Incorporated.

Willerslev, Eske, Tobias Mourier, Anders J. Hansen, Bent Christensen, Ian Barnes, and Steven L. Salzberg. 2002. “Contamination in the Draft of the Human Genome Masquerades as Lateral Gene Transfer.” *DNA Sequence: The Journal of DNA Sequencing and Mapping* 13 (2): 75–76.

Wilson, Christopher G., Reuben W. Nowell, and Timothy G. Barraclough. 2018. “Cross-Contamination Explains ‘Inter and Intraspecific Horizontal Genetic Transfers’ between Asexual Bdelloid Rotifers.” *Current Biology: CB*.

Witt, Evan, Sigi Benjamin, Nicolas Svetec, and Li Zhao. 2019. “Testis Single-Cell RNA-Seq Reveals the Dynamics of de Novo Gene Transcription and Germline Mutational Bias in Drosophila.” *eLife* 8 (August): e47138.

Xia, Rui, Jing Xu, Siwaret Arikit, and Blake C. Meyers. 2015. “Extensive Families of miRNAs and PHAS Loci in Norway Spruce Demonstrate the Origins of Complex phasiRNA Networks in Seed Plants.” *Molecular Biology and Evolution* 32 (11): 2905–18.

Xue, Lingzhan, Yu Gao, Meiying Wu, Tian Tian, Haiping Fan, Yongji Huang, Zhen Huang, Dapeng Li, and Luohao Xu. 2021. “Telomere-to-Telomere Assembly of a Fish Y Chromosome Reveals the Origin of a Young Sex Chromosome Pair.” *Genome Biology* 22 (1): 203.

Xu, Kenong, Xia Xu, Takeshi Fukao, Patrick Canlas, Reycel Maghirang-Rodriguez, Sigrid

Heuer, Abdelbagi M. Ismail, Julia Bailey-Serres, Pamela C. Ronald, and David J. Mackill. 2006. “Sub1A Is an Ethylene-Response-Factor-like Gene That Confers Submergence Tolerance to Rice.” *Nature* 442 (7103): 705–8.

Yaffe, Hila, Kobi Buxdorf, Illil Shapira, Shachaf Ein-Gedi, Michal Moyal-Ben Zvi, Eyal Fridman, Menachem Moshelion, and Maggie Levy. 2012. “LogSpin: A Simple, Economical and Fast Method for RNA Isolation from Infected or Healthy Plants and Other Eukaryotic Tissues.” *BMC Research Notes* 5 (January): 45.

Yamamoto, Eiji, Tomonori Takashi, Yoichi Morinaka, Shaoyang Lin, Jianzhong Wu, Takashi Matsumoto, Hidemi Kitano, Makoto Matsuoka, and Motoyuki Ashikari. 2010. “Gain of Deleterious Function Causes an Autoimmune Response and Bateson-Dobzhansky-Muller Incompatibility in Rice.” *Molecular Genetics and Genomics: MGG* 283 (4): 305–15.

Yandell, Mark, and Daniel Ence. 2012. “A Beginner’s Guide to Eukaryotic Genome Annotation.” *Nature Reviews. Genetics* 13 (5): 329–42.

Yang, Aimin, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang. 2020. “Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA.” *Frontiers in Bioengineering and Biotechnology* 8 (September): 1032.

Yao, W., G. Li, H. Zhao, G. Wang, X. Lian, and W. Xie. 2015. “Exploring the Rice Dispensable Genome Using a Metagenome-like Assembly Strategy.” *Genome Biology* 16 (1). https://doi.org/10.1186/s13059-015-0757-3.

Yuan, Yuxuan, Philipp E. Bayer, Jacqueline Batley, and David Edwards. 2021. “Current Status of Structural Variation Studies in Plants.” *Plant Biotechnology Journal* 19 (11): 2153–63.

Yue, Jia-Xing, Blake C. Meyers, Jian-Qun Chen, Dacheng Tian, and Sihai Yang. 2012. “Tracing the Origin and Evolutionary History of Plant Nucleotide-Binding Site--Leucine-Rich Repeat (NBS-LRR) Genes.” *The New Phytologist* 193 (4): 1049–63.

Yu, Ping, Caihong Wang, Qun Xu, Yue Feng, Xiaoping Yuan, Hanyong Yu, Yiping Wang, Shengxiang Tang, and Xinghua Wei. 2011. “Detection of Copy Number Variations in Rice Using Array-Based Comparative Genomic Hybridization.” *BMC Genomics* 12 (July): 372.

Zapata, Luis, Jia Ding, Eva-Maria Willing, Benjamin Hartwig, Daniela Bezdan, Wen-Biao Jiao, Vipul Patel, et al. 2016. “Chromosome-Level Assembly of Arabidopsis Thaliana Ler Reveals the Extent of Translocation and Inversion Polymorphisms.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (28): E4052–60.

Zdobnov, Evgeny M., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley, and Evgenia V. Kriventseva. 2021. “OrthoDB in 2020: Evolutionary and Functional Annotations of Orthologs.” *Nucleic Acids Research* 49 (D1): D389–93.

Zhai, Rongrong, Yue Feng, Huimin Wang, Xiaodeng Zhan, Xihong Shen, Weiming Wu, Yingxin Zhang, et al. 2013. “Transcriptome Analysis of Rice Root Heterosis by RNA-Seq.” *BMC Genomics* 14 (January): 19.

Zhang, Wei. 2020. “NLR-Annotator: A Tool for De Novo Annotation of Intracellular Immune Receptor Repertoire.” *Plant Physiology*.

Zheng, Lei-Ying, Xiao-Sen Guo, Bing He, Lian-Jun Sun, Yao Peng, Shan-Shan Dong, Teng-Fei Liu, et al. 2011. “Genome-Wide Patterns of Genetic Variation in Sweet and

Grain Sorghum (Sorghum Bicolor).” *Genome Biology* 12 (11): R114.

Zhou, P., K. A. T. Silverstein, T. Ramaraj, J. Guhlin, R. Denny, J. Liu, A. D. Farmer, et al. 2017. “Exploring Structural Variation and Gene Family Architecture with De Novo Assemblies of 15 Medicago Genomes.” *BMC Genomics* 18 (1). https://doi.org/10.1186/s12864-017-3654-1.

Zimin, Aleksey V., Arthur L. Delcher, Liliana Florea, David R. Kelley, Michael C. Schatz, Daniela Puiu, Finnian Hanrahan, et al. 2009. “A Whole-Genome Assembly of the Domestic Cow, Bos Taurus.” *Genome Biology* 10 (4): R42.