Institute of Bioorganic Chemistry, Polish Academy of Sciences
Department of Plant Genomics

DOCTORAL THESIS

mgr inż Anastasiia Satyr

VARIATION OF THE RECEPTOR-LIKE PROTEIN GENE FAMILY IN THE *ARABIDOPSIS THALIANA* GENOME REVEALED BY NANOPORE SEQUENCING AND *DE NOVO* ASSEMBLIES

Supervisor:
Dr hab. Agnieszka Żmieńko, prof. IBCh PAS

Poznań 2025

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor for giving me the opportunity to accomplish my PhD Project. I deeply appreciated being a PhD Student in a beautiful and cozy city of Poznań. I am also grateful for acquiring me curiosity to plant science and for support in both professional and personal growth, support in my scientific ideas, the good ones and the bad ones.

I am sincerely grateful to all my colleagues from European Center for Bioinformatics and Genomics (ECBiG) for welcoming and kind environment and providing help wherever I needed it.

I will never forget the nice time spent in the Institute of Bioinformatics of University of Münster, thanks to my colleagues who taught me bioinformatics tricks and combining science and joy.

This work would not have been possible without constant support from my husband, my parents and my friends.

I dedicate this work to my parents whose stoicism in living in the war inspires me.

FUNDING

This research was supported by three grants: National Science Centre (NCN) Grant SONATA 2017/26/D/NZ2/01079 entitled “Contribution of gene copy number polymorphism to natural variation in response of *Arabidopsis thaliana* ecotypes to biotic stress”, Polish National Agency for Academic Exchange (NAWA) Grant BPN/BDE/2022/1/00021 entitled “The role of transposable elements in shaping structural and transcriptional variation of *Arabidopsis thaliana* genome”, and NCN Grant PRELUDIUM 2023/49/N/NZ2/02102 entitled “The genetic basis underlying immune resistance in selected accessions of the model plant *A*. *thaliana* against the plant pathogen *Pectobacterium carotovorum*”. The principal investigator of the first two grants was dr hab. Agnieszka Żmieńko. The grant holder of the latter grant was the author of this thesis.

The author was supported by PhD Stipend from the Grant SONATA, and IBCh PAS Student Stipend.

TABLE OF CONTENTS

Page

ABSTRACT

Receptor-Like Protein (*RLP*) genes is a gene family encoding proteins associated with plant development and responses to biotic and abiotic stressors. However, among 57 *RLP* genes annotated in the model plant *Arabidopsis thaliana* (Arabidopsis), the biological function of only few has been described in detail. Characterization of natural variation in the genomic regions that overlap *RLP* genes might contribute to finding potential functions of RLP proteins as well as discovering new *RLP* alleles.

The recent emerging of long read sequencing methods remarkably enhanced our ability to detect complex genomic rearrangements and to assemble the entire chromosomes and genomes. Accordingly, I hypothesized that by sequencing and assembling the genomes of eight diverse Arabidopsis accessions, followed by detailed sequence comparison, I would be able to resolve all types of structural variants affecting *RLP* genes. To achieve this aim, I developed the protocol of high-molecular-weight genomic DNA extraction from leaves, sequenced genomic DNA of selected plants by nanopore technology, and generated de novo genome assemblies. Next, I performed extensive quality evaluation of genomic assemblies in three categories: contiguity, completeness, and correctness. The results indicated that the assemblies had high quality across chromosome arms, but were fragmented at repeat-enriched regions, e.g. centromeres and telomeres. I annotated genes and transposable elements (TEs) and examined variation patterns at *RLP* loci, where I found numerous small- and large-scale variants, including deletions, insertions, and genomic rearrangements associated with TEs. For example, the region spanning genes *RLP47-RLP50* genes was found to have an interspersed duplication in one accession, and TE-derived insertion in another accession.

Although *RLPs* are usually not considered as the main source of resistance variation in plants, the findings of this work suggest that many of them display substantial intraspecies diversity. Future extending the description of this diversity to a worldwide Arabidopsis population may serve as the foundation for functional studies and contribute to better understanding of plant immune mechanisms.

STRESZCZENIE

Geny *Receptor-Like Protein* (*RLP*) to rodzina genów kodujących białka, które biorą udział w rozwoju roślin oraz odpowiedzi na czynniki abiotyczne i biotyczne. Jednak spośród 57 genów RLP znalezionych w genomie rośliny modelowej Arabidopsis thaliana (Arabidopsis), tylko w przypadku niektórych dokładnie poznano funkcje biologiczne. Scharakteryzowanie naturalnej zmienności regionów genomu obejmujących geny *RLP* mogłoby przyczynić się do poznania potencjalnych funkcji białek RLP oraz opisania ich nowych alleli.

Pojawienie się technik sekwencjonowania długich odczytów znacząco zwiększyło możliwości identyfikacji złożonych rearanżacji genomowych oraz składania kompletnych chromosomów i genomów. Postawiłam hipotezę, że poprzez zsekwencjonowanie i złożenie de novo genomów ośmiu ekotypów Arabidopsis oraz szczegółowym analizom porównawczym, będę mogła poznać wszystkie typy wariantów strukturalnych obejmujących geny *RLP*. W tym celu opracowałam protokół ekstrakcji genomowego DNA o wysokiej masie cząsteczkowej z liści, zsekwencjonowałam DNA genomowe wybranych roślin w technologii nanopore, i złożyłam de novo sekwencje genomów. Przeprowadziłam również dokładną ocenę jakości złożonych genomów pod względem ciągłości, kompletności oraz poprawności sekwencji. Wyniki analiz wykazały, że miały one wysoką jakość w obrębie ramion chromosomów, jednak charakteryzowały się podwyższonym poziomem fragmentacji w regionach powtórzonych, takich jak centromery i telomery. Wykonałam annotację genów i transpozonów oraz przeanalizowałam zakres zmienności loci RLP, gdzie znalazłam liczne polimorfizmy o różnej skali długości, takie jak delecje, insercje, czy rearanżacje genomowe powiązane z działalnością transpozonów. Przykładowo, w klastrze genów *RLP47-RLP50* w jednym ekotypie powstała duplikacja genu *RLP47*, a w innym ekotypie znalazłam insercje transpozonów.

Chociaż rodzina *RLP* nie jest typowo uznawana za główne źródło zróżnicowania odporności roślin na stres, wyniki tej pracy wskazują na istotną zmienność wewnątrzgatunkową wielu tych genów. Rozszerzenie charakterystyki tej zmienności na światową populację Arabidopsis może stworzyć podstawy badań funkcjonalnych, co w przyszłości zaowocuje lepszym zrozumieniem mechanizmów odporności roślin na stres.

LIST OF ABBREVIATIONS

5mC	–	5-methylcytosine
BAC	–	bacterial artificial chromosome
bp	–	base pair
cDNA	–	complementary DNA
CGH	–	comparative genomic hybridization
CNV	–	copy number variation
Col-0	–	Columbia-0
CTAB	–	Cetyltrimethylammonium-bromide
ETI	–	effector-triggered immunity
Gb	–	Gigabase
FoSTeS	–	fork stalling and template switching
HR	–	hypersensitive response
LRR	–	leucine-rich repeat
MAPK	–	mitogen-activated protein kinase
Mb	–	Megabase
MMEJ	–	microhomology-mediated end joining
NAHR	–	non-allelic homologous recombination
NGS	–	next-generation sequencing
NHEJ	–	non-homologous end joining
NLR	–	nucleotide-binding leucine-rich repeat receptor
NOR	–	nucleolus organized region
OLC	–	Overlap-Layout-Consensus
ONT	–	Oxford Nanopore Technologies
ORF	–	open reading frame
PAMP	–	pathogen-associated molecular pattern
PCA	–	principal component analysis
PRR	–	pathogen-recognition receptors
PTI	–	pathogen-triggered immunity
kb	–	kilobase
RLK	–	Receptor-Like Kinase
RLP	–	Receptor-Like Protein
SNP	–	single-nucleotide polymorphism

SV	–	structural variation
T2T	–	telomere to telomere
TAIR	–	The Arabidopsis Information Resource
TRF	–	tandem repeats finder
TE	–	transposable element

1. INTRODUCTION

Disease resistance in plants is maintained by the multi-component regulatory system that detects pathogen-associated molecular patterns (PAMPs), derived from different pathogens such as bacteria, fungi, or viruses, and activates the defense mechanisms against causative agents. To recognize PAMPs directly, plants evolved the variety of immune receptors on their cell surface, classified as pattern-recognition receptors (PRRs). The process initiated by PAMP-PRR interaction is referred to as PAMP-triggered immunity (PTI) and constitutes the first layer of plant defense. To overcome this defense, pathogens secrete virulence protein effectors inside the plant cells. In response, plants enroll intracellular receptors (mostly nucleotide-binding leucine rich repeat-type receptors, NLRs) that interact with these effectors, and activate transcriptional changes of the host genes, leading to pathogen-specific responses, termed effector-triggered immunity (ETI) [Cui 2015]. During the process of plant-pathogen co-evolution, both interacting partners modify the components involved in their mutual recognition, which leads to the sub-species differences in the host plant vulnerability to infection with specific pathogenic strains.

The genes involved in plant resistance undergo intensive natural variation both inter- and intraspecifically, which not only makes it an important area for investigating the genetic mechanisms and selective forces that lead to the formation of new variants, but it may also facilitate searches for new resistance genes of agronomic potential. However, most present methods for variant detection rely on DNA sequencing and alignment of short sequencing reads to the reference genome, the approach which has numerous limitations. Long-read sequencing techniques promise to overcome these limitations, by enabling assembly of structurally complex regions and even entire genomes. This work has been inspired by the recent achievements of nanopore sequencing, which is a rapidly developing sequencing technique. I assumed that performing *de novo* genome assembly of the nanopore sequencing data will enable the discovery of genomic regions exhibiting complex variation, previously inaccessible with short read-based methods, and enable the identification of new gene variants. As a proof-of-concept, this work has been aimed at using this approach to describe the variation patterns within a particular family of genes associated with PAMP recognition, named *Receptor-Like Protein* (*RLP*) genes, in the worldwide representatives of *Arabidopsis thaliana* (hereafter Arabidopsis). The analysis of genetic variants that affect *RLP* genes will contribute to future discovery of their functional consequences and, eventually, to better understanding of plant-pathogen interaction mechanisms in this model plant.

1.1. Genome variation in plant disease resistance

1.1.1. Structural variation in plant genomes. Structural variation / structural variant (SV) is a term that collectively refers to genetic polymorphisms involving large fragments of the genome. SV is defined as the genomic rearrangements of the sequence of at least 50 base pairs (bp) in length, but frequently range from one kilobase (kb) up to several Megabases (Mb) in size and often cause phenotypic changes [Saxena 2014]. SV occurs between and within species, when comparing the genomes of individual plants. Plant SVs may be balanced, e.g. translocations and inversions, or imbalanced – resulting in gains or losses of the genomic sequence, e.g. insertions, TE-derived insertions, copy number variants (CNVs). A special case is a complex SV event, where different SVs are nested in one genomic location [Ho 2020]. CNV is the variantion type in which a gene or a genomic fragment appears in the genome in a different number of copies among individual genomes. CNVs are represented by deletions and duplications (Figure 1). Presence-absence variation (PAV) is a type of CNV that occurs when the genomic region either persists or is deleted from the genome when compared between the individuals in the studied population.

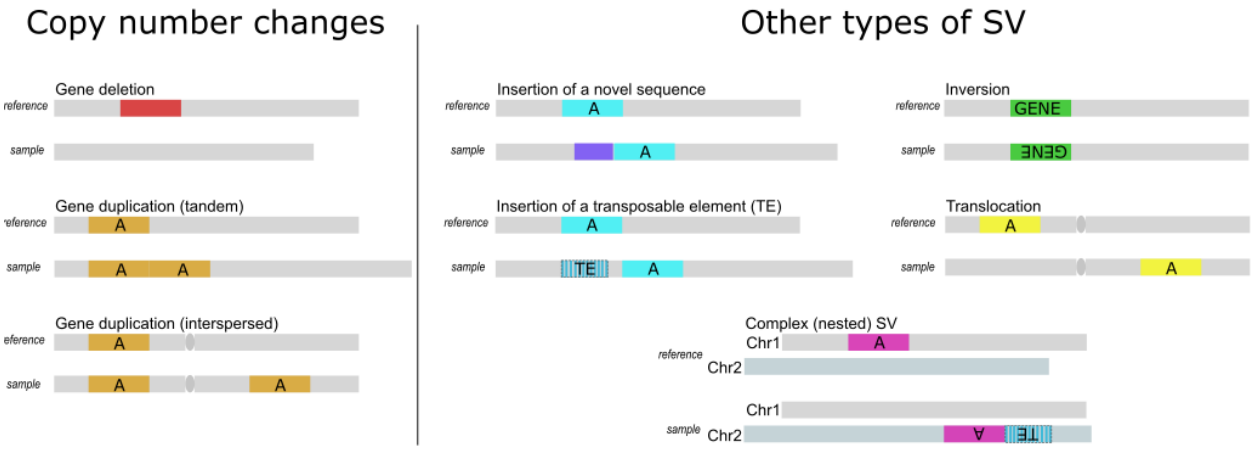


Figure 1. Common types of structural variants in plants (adapted from [Saxena 2014]).

SVs frequently arise in DNA replication phase of the cell cycle and rely on natural DNA recombination and repair mechanisms. The most abundant mechanisms of SV formation include non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS), and microhomology-mediated end joining (MMEJ) [Burssed 2022, McVey 2008]. NAHR operates on long stretches of highly similar sequences (typically about 1 kb in length and at least 97% sequence homology) as the substrates for a rearrangement, when the template of recombination can be derived from the non-allelic

intrachromosomal, intrachromatid, or interchromosomal location. NAHR frequently results in SVs such as tandem duplication, interspersed duplication or deletion [Zmienko 2016]. NHEJ occurs in the process of DNA double-strand break repair and operates on breakpoints without requiring sequence homology, while MMEJ, as the special case of NHEJ, uses sequence microhomology to repair the breaks. NHEJ and MMEJ usually result in deletions, duplications, translocations and inversions. In contrast, FoSTeS often leads to the formation of complex SVs, as during this process the replication forks pause and switch between adjacent templates using complementarity to anneal and resume DNA replication. Several other mechanisms have been described recently [Burssed 2022].

The abundance of intraspecies genetic variants is beneficial for plant fitness, as it creates the basis for the potential of individual plants to adapt to the dynamically changing environment. SV occurs across the whole genome affecting its non-coding regions and altering the protein-coding genes, although with different frequency, due to the stronger selective force on the protein-coding sequence [Lye 2019]. A remarkable example for the variation event has been described in cucumber (*Cucumis sativas*), the important crop [Zhang 2015]. Naturally, male and female flowers are located on the same plant, and after fertilization, the fruit develops from the female ones. A 30-kilobase duplication of the region that includes Female locus, results in the development of the plant that carries only female flowers (a gynoecious plant) and consequently, the sufficient increase of the fruit number. The duplicated region encodes several important proteins: amylocyclopropane-1-carboxylic acid synthase, aminotransferase, and MYB transcription factor. In some accessions that produce gynoecious plants, the copy number of Female locus was increased even 4-fold, suggesting the agronomic importance of this genetic variant [Zhang 2015].

SVs often alter gene expression, by affecting gene-coding or regulatory sequence. For example, the red skin color of the radish roots (*Raphanus sativus*) is achieved by the accumulation of anthocyanins, plant metabolites known for their antioxidant properties. The red-skin phenotype is regulated by the gene *RsMYB1.1*, which encodes a transcription factor. Kim and colleagues [2024] compared *RsMYB1.1* promoter sequence in 34 red-skin and 66 non-red-skin radish accessions and revealed an insertion of a low-complexity sequence enriched in AT-repeats in the latter ones. The insertion negatively regulates the expression of *RsMYB1.1*, resulting in the silencing of anthocyanin biosynthesis pathway genes. The authors also found single-nucleotide polymorphisms (SNPs) in one of the exons of *RsMYB1.1*, indicating high variability of the studied region [Kim 2024]. Another remarkable example is the 17.1-kb duplication in Grain Length on Chromosome 7 (*GL7*) locus in rice (*Oryza sativa*) [Wang 2015]. *GL7* encodes the protein homologous to Arabidopsis *LONGIFOLIA*, associated with longitudinal elongation of cells. The

discovered variant results in the increase of grain length and quality, which was suggested as a beneficial trait for rice breeding strategies.

SV events may be associated with plant resistance to different types of abiotic stresses, for example boron-toxicity tolerance in barley (*Hordeum vulgare*), regulated by *Bot1* gene [Sutton 2007]. The gene encodes the efflux transporter of boron in plant roots and is responsible for regulation of boron uptake from soil. The boron-tolerant barley landrace (Sahara 3771) was shown to have four additional copies of *Bot1*-containing locus when compared to boron-intolerant accession (Clipper). Expectedly, the concentration of *Bot1* mRNA in the roots of Sahara 3771 was higher than in Clipper [Sutton 2007]. Numerous genomic variants associated with other abiotic factors like drought and flood, acidity and salinity changes, temperature and osmotic pressure extremes, have also been reported [Varshney 2024]. The crop tolerance to abiotic stress is an important issue for global food strategies.

Considering the role of SV in the development of important plant traits, associating genome variation with disease resistance seems to be an obvious research focus. Over the years of evolution, plant defense genes underwent sequence variation for diversification, as sequence rearrangements might provide more effective pathogen recognition. The remarkable example for intraspecies polymorphism and resistance was revealed in Arabidopsis [Stahl 1999]. The *Rpm1* gene confers plant resistance to *Pseudomonas* strains, and its product – RPM1 (Resistance to *Pseudomonas syringae* pv. *maculicola*) – is a functional protein that recognizes and binds the pathogen-derived effector AvrRpm1, resulting in the activation of plant resistance mechanisms. To investigate the variation of *Rpm1* gene across 26 natural accessions, the authors cloned the region flanking *Rpm1* and found that in the accessions that were susceptible to *Pseudomonas* infection, the sequence encoding *Rpm1* was absent from the genome, whereas the gene was present in the resistant accessions, demonstrating that *Rpm1* undergoes PAV. The authors also concluded that the polymorphism pattern of functional/null allele was shaped during years of evolution, and the currently existing sequence variants represent the accumulation of sequence rearrangements in the result of the dynamic plant-pathogen interaction.

1.1.2. *Arabidopsis thaliana* – a top model plant for variation studies. Thale cress – or Arabidopsis – was proposed as a model species to study different processes of plant biology due to its universal advantages: short generation time (~4 weeks), self-pollination and the resulting homozygosity, high fertility, small size, and easy handling in laboratory conditions. The broad plant habitat, spanning temperate, subarctic and even tropical climate regions of the world, and small genome size (~140 Mb) are crucial features that make this species also an indispensable model for plant genomic variation research. The reports of genetic variation in Arabidopsis appear

in the context of every living process: plant growth and morphology [Koorneef 2004], organ development timing [Doody 2022], flowering time [Yan 2020, Schmalenbach 2014], rosette branching [González 2020], gene expression levels [Plantegenet 2009], activity of enzymes [Cross 2006], epigenetic patterns [Vaughn 2007], response to abiotic [Clauw 2016] and biotic stresses [Rowe 2008], and others. In addition, diverse environmental conditions, to which natural Arabidopsis accessions adapted, shaped the strong diversity within the species [Weigel 2009].

Sequencing and annotating the genome sequence of the reference Arabidopsis accession Columbia-0 (Col-0) greatly contributed to the discovery of genes associated with phenotypic traits, many of which had been subsequently translated to important crops [Uauy 2025]. The access to the reference genome, which was published in 2000 [AGI 2000], provided the research community with the information on the structure and localization of genes in the genome, and enabled to conduct studies, aimed to understand the fundamentals of plant physiology, molecular biology and evolution. Soon it became clear that the phenomenon of natural variation in this population may serve as a great resource for functional studies, as we can observe different phenotypes that occur in nature and associate them with the traits responsible for the adaptation to different living conditions. The available Arabidopsis germplasm collections include thousands accessions – individual plants that were gathered from different geographic locations [Weigel 2012].

The range and the patterns of variation in Arabidopsis population remain the subject of strong interest, as confirmed by the numerous sequencing projects that were conducted to assess this issue at a global scale, starting with the description of variation among 96 accessions by DNA hybridization methods [Clark 2007], followed by the Arabidopsis Genome Consortium project of resequencing 1135 genomes [1001 Genomes Consortium 2016], which was further complemented by the transcriptome and methylome data for most plants in the collection [Kawakatsu 2016], and other initiatives [Ossowski 2008, Cao 2011]. Most of these data is stored in the numerous publicly available databases, e.g. The Arabidopsis Information Resource (TAIR) (https://www.arabidopsis.org/), Arabidopsis Information Portal (https://www.araport.org/), 1001 Genomes (https://1001genomes.org/). With such a large resource set, Arabidopsis represents an attractive model for variation studies using comparative genomics and multi-omics approaches.

1.1.3. Immune receptors as the key mediators of plant disease resistance. Plant cells are naturally equipped with morphological and physiological defensive systems: rigid cell walls, leaf hairs and external shells, antimicrobial compounds, etc. As opposed to the animal immune system, the plant organism lacks specific cells responsible for immunity, therefore plants enroll all defense mechanisms at once in the place of infection. The “danger” signal is detected through the recognition of the pathogen-derived elicitors (Figure 2). General elicitors, or PAMPs are usually

represented by the components of pathogen cells, for example bacterial flagellin or fungal chitin. The first-contact barrier that recognizes PAMPs, is a system of PRRs on the plant cell surface, and the PAMP-PRR interaction process is referred to as basal immunity or PTI. Other molecules, secreted by pathogens inside host cells, or effectors, enter the cytoplasm of plant cells and activate response mechanism specific to the given pathogen, namely ETI [Jones 2006].

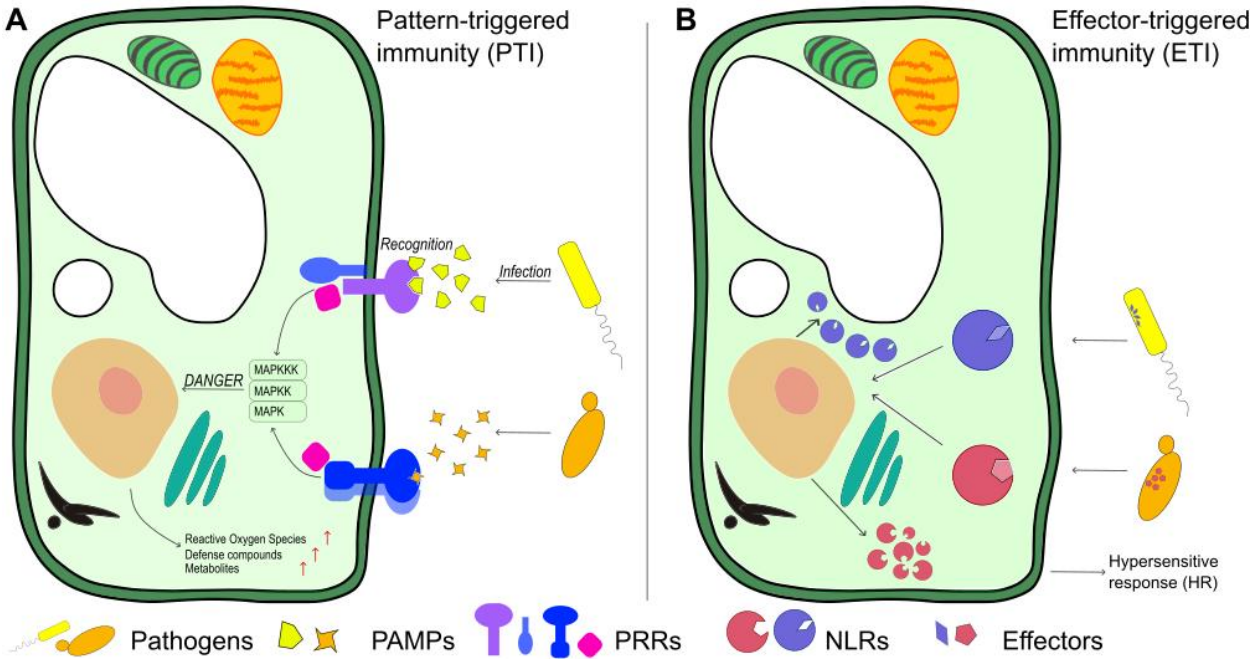


Figure 2. PTI and ETI mechanisms as the components of the plant immune system. A – PAMPs are recognized by PRRs at the plant cell membrane, the "danger" signal is passed inside the cell, leading to activation of defense mechanisms. B – intracellular NLRs bind pathogen effectors and activate pathogen-specific responses.

The mechanism of ETI is predominantly mediated by the proteins that belong to NLR group [Van de Weyer 2019]. The genes encoding ETI-mediating receptors are termed Resistance (R) genes as they evolved to recognize specific pathogen components and trigger more potent defense responses, compared to PTI only. Both PTI and ETI mechanisms are directed towards the reduction of the negative effects of infection on plant organism, and include production of reactive oxygen species, accumulation of calcium ions, activation of mitogen-activated protein kinase (MAPK) cascade, transcriptional reprogramming, and phytohormone biosynthesis [Cui 2015]. In most cases, plant resistance activation leads to hypersensitive response (HR) – the necrosis of plant tissue around the infection site, that prevents spreading the infection across the organism.

The functional protein domain that is common for many plant immune receptors and involved in pathogen recognition and protein-protein interaction, is leucine-rich repeat (LRR) [Cho 2023]. A typical LRR is a chain of about 23-25 amino acid residues organized in LxxLxxLxx

motif, in which x means any amino acid, folded in β-sheets [Padmanabhan 2009]. This structure creates much space for variation, as the conserved part is only a pattern of hydrophobic leucine residues. The genes encoding LRRs have been found to have many alternative transcripts, indicating the evidence for the extensive variation of LRR at the genome and transcript levels [Tan 2007]. LRRs were found to play a role in diverse processes associated with protein-protein interactions [Cho 2023], and in plants this domain is mostly associated with immune function. In PRRs, which are cell-surface receptors, LRR domain (if present) is located extracellularly.

PRRs function as guards of basal plant immunity, recognizing numerous PAMPs anchoring to the cell membrane. PRRs are represented by the two main groups, receptor-like kinases (RLKs) and RLPs. The three main functional domains of RLK proteins are: extracellular domain for ligand perception, transmembrane domain, and intracellular domain with kinase activity. RLPs have similar domain composition except they lack the kinase domain, they have a short cytoplasmatic tail instead. Therefore, to pass the signal of danger inside the cell, RLPs form complexes with the proteins from the RLK family. The two main proteins that participate in complex formation are BAK1 (BRI1-ASSOCIATED RECEPTOR KINASE 1), which functions as the co-receptor, and SOBIR1 (SUPRESSOR OF BIR1 - 1), as the adapter protein for downstream signal transduction [Jamieson 2018]. The formation of RLP-BAK1-SOBIR1 complex leads to the activation of regulatory signaling pathways, such as calcium-dependent protein kinases and MAPKs, initiating gene expression changes and subsequent defense responses. Additional proteins, which function along with both RLPs and RLKs, are cytoplasmic kinases: they possess only a kinase domain and are aimed to strengthen the danger signal in the cell (Figure 3) [Hailemariam 2024]. It should be added that not only RLPs, but all types of immune receptors may function in complexes with co-receptors and helper proteins.

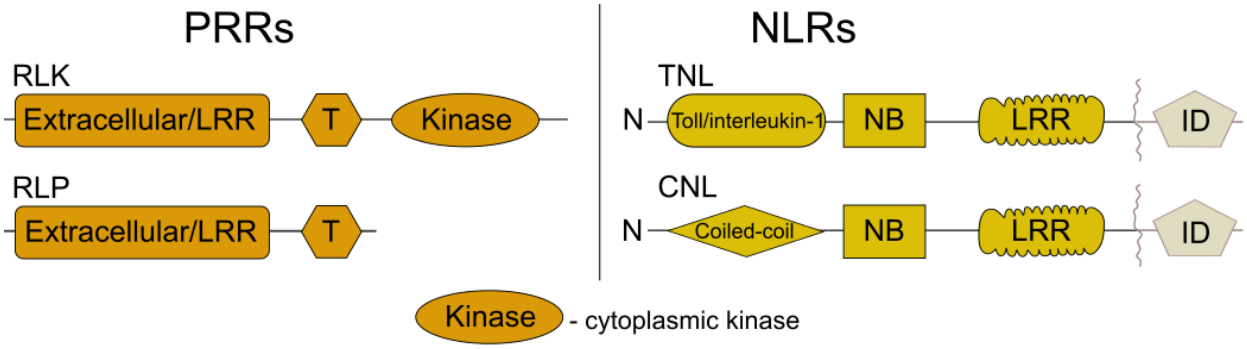


Figure 3. The main functional domains of plant immune receptors. PRRs – pattern-recognition receptors, NLRs – nucleotide-binding leucine-rich repeat receptors, RLK – receptor-like kinase, RLP – receptor-like protein, T – transmembrane domain, TNL – Toll/Interleukin-1 domain-including NLR, CC – coiled-coil domain-including NLR, NB – nucleotide-binding domain, ID – integrated domain (an optional domain to strengthen the attraction of effectors).

In the Arabidopsis genome, RLK superfamily consists of more than 400 gene members, further classified into more than 20 classes. Of total, about 200 RLK proteins contain LRR domains (LRR-RLK); therefore, they can be associated with plant immunity [Cho 2023]. Non-LRR RLKs play an important role in other processes such as sensing the environmental conditions by the plant cell and communication between adjacent cells. RLPs are encoded by 57 genes and all of them contain extracellular LRR domains. RLPs have been found to play a role in the recognition of PAMPs, as well as in plant growth and development. Intracellular NLRs are classified by their N-terminal domain, which can be either Toll/interleukin-1 (TNL) or Coiled-coil (CNL) structure. The two remaining functional domains in NLRs are a central nucleotide-binding domain and an LRR-containing domain. Among 165 NLR-encoding genes in the Arabidopsis genome, the majority (106 genes) belong to the TNL class [Shao 2016]. Some NLRs can involve short integrated domains (IDs) that help to decoy a target effector (Figure 3) [Van de Weyer 2019].

In the process of plant-pathogen co-evolution, both membrane-located PRRs and intracellular NLRs undergo different types of variation. The genes related to pathogen recognition and response belong to the most diverse gene families in plants [Baumgazten 2003]. Interestingly, NLRs, which are effector-specific, are believed to be highly variable, while PRRs are thought to be more conserved because they recognize PAMPs, which also share conservation between different pathogens.

The variation repertoire of *NLR* genes in Arabidopsis has been comprehensively described recently [Van de Weyer 2019]. The authors aimed to characterize the full extent of *NLR* variation patterns using the high-quality sequence datasets for 64 distinct accessions. The authors found that more than a half of *NLR* genes localize within gene clusters in the genome, often in a so-called “head-to-head” orientation, indicating their expression and functioning in pairs. Functional analysis showed that the variation in *NLR* genes was represented by changes in composition of their functional domains, suggesting that the variation events mainly occurred within individual genes. The authors identified 97 different domain compositions, only 22 of which were present in the reference genome. These results showed that in NLR family, the gene sequence variation mainly shapes the final conformation of the receptor protein, and its interaction with the pathogen effector.

1.1.4. Diversity and functions of Arabidopsis *RLP* genes. Stuctural composition of Arabidopsis RLP proteins involves an extracellular LRR-containing domain, a transmembrane domain and a short intracellular cytoplasmic tail. They share sequence homology with the toll-like receptors of animal immune system [Sameer 2021]. The fist functionally described RLPs were

reported to regulate plant developmental processes: RLP17, referred to as TOO MANY MOUTHS (TMM/RLP17), is involved in the regulation of stomatal development on plant leaves; and CLAVATA2 (CLV2/RLP10) plays its role in meristem maintenance throughout the plant life cycle [Wang 2008]. However, knowledge about the functions and diversity of the *RLP* gene family remains very limited. Of the 57 annotated *RLP* genes in the reference Arabidopsis genome (including *RLP8* and *RLP49* which are currently annotated as pseudogenes), only 12 have been functionally characterized so far (Table 1). From the discovered functions of RLP proteins, we can observe that these receptors participate in the development of plant organs (aforementioned CLV2 and TMM), plant resistance to biotic (RLP1, RLP3, RLP32) and abiotic (RLP48) stresses. Such a broad range of functions suggests that the *RLP* gene family has dynamically evolved therefore may be strongly variable, but this variation has been understudied, as indicated by few reports on this matter. In one such study, the authors analyzed publicly available data from variation catalogues of more than 1000 Arabidopsis accessions (based on short read sequencing), aiming to systematize the current knowledge about this group of genes [Steidele 2021]. They subdivided *RLP* family members into two potentially functional categories of genes: developmental and defense-associated. Based on phylogenomic analysis, the authors confirmed the defense-associations of *RLPs* with known functions (for example *RLP23*, *RLP30*, *RLP32*) and added more than 30 *RLPs* to the category of potentially defense-related genes. The authors did not reveal differences in variation patterns between the two categories, however they reported overall higher diversity of the defense-related *RLPs*, when compared to the development-related ones.

The two remarkable studies present evidence for variation of individual *RLPs* and its effect for shaping natural diversity of Arabidopsis resistance. The first study reports the deletion of gene *RLP1* in the genome of the Sha accession, which does not recognize the bacterial PAMP eMax, derived from Xanthomonads [Jehle 2013]. The other accessions, including Col-0, reacted to pathogen exposure. The authors concluded that the mechanism of resistance in Sha lays in the lack of this specific PRR. The second investigation concerns variation in Arabidopsis susceptibility to *Fusarium oxysporum*-caused wilt in two accessions – Col-0 and Ty-0 [Shen 2013]. The susceptibility was conferred by the variable region that encodes *RLP2-RLP3* gene cluster – in the susceptible accession Ty-0, *RLP3* gene, which is referred to as *RFO2* (*RESISTANCE TO FUSARIUM OXYSPORUM 2*) in the mentioned study, was deleted and the sequence encoding RLP2 was altered. However, the response to *Fusarium* was shown to be a multigenic trait, because some pathogen-exposed accessions engaged other proteins, such as RESISTANCE TO FUSARIUM OXYSPORUM1 and Tyrosine-Sulphated Peptide Receptor 1, both belonging to

RLK group. In the case of observed *RLPs*, the variation events altered the entire genes, by changing gene location, genomic context or copy number.

Table 1. The known functions of Arabidopsis *RLPs*.

Gene ID	Symbols	Function	Reference
At1g07390	*RLP1, ReMAX*	Receptor of eMax, a PAMP from bacteria of *Xanthomonas spp.*	[Jehle 2013]
At1g17250	*RLP3, RFO2*	Confers resistance to the fungus *Fusarium oxysporum*	[Shen 2013]
At1g28340	*RLP4*	Participates in the cell wall growth control	[Elliott 2024]
At1g65380	*RLP10, CLV2 (CLAVATA2)*	Receptor for the regulation of shoot meristem growth and development	[Jeong 1999, Bleckmann 2010]
At1g80080	*RLP17, TMM*	Regulates cell differentiation in the development of stomata	[Geisler 2000]
At2g32680	*RLP23*	Recognizes the PAMP nlp20	[Albert 2015]
At3g05360	*RLP30*	Receptor of Sclerotinia culture filtrate elicitor (SCFE1) from the fungus *Sclerotinia sclerotiorum*	[Zhang 2013]
At3g05650	*RLP32*	Receptor of translation initiation factor (IF1) from the bacterium *Pseudomonas syringae*	[Fan 2022]
At3g25020	*RLP42, RBPG1*	Recognizes the conserved motif of fungal polygalacturonases	[Zhang 2014]
At3g49750	*RLP44*	Positive regulator of the brassinosteroid signaling pathway	[Wolf 2014]
At4g13880	*RLP48*	Involved in root hair density regulation in response to low phosphate concentration	[Stetter 2015]
At4g18760	*RLP51, SNC2*	Down-regulator of the expression of *NPR1* (*NONEXPRESSOR OF PATHOGENESIS-RELATED GENES 1*)	[Zhang 2010]

1.1.5. TEs contribution to genome variation. TEs are genomic elements capable of moving and replicating across the genome, which in plants frequently occupy substantial genomic space. Based on the mechanism of transposition, TEs are classified into two main classes: retrotransposons or RNA transposons (Class I), that move by replicating their sequence via RNA intermediate, using reverse transcription, and DNA transposons (Class II) that change their location by getting cut from one genomic locus and inserted into another one [Wicker 2007]. TEs of both classes may encode enzymes necessary for their transposition – these are autonomous TEs,

however most TEs use the transposition machinery of other TEs (non-autonomous) [Liu 2022]. Transposing activity of TEs facilitates the formation of TE-derived duplications, deletions, and insertions, e.g. by increasing the amount of repetitive content and the frequency of nonallelic recombination events, therefore they must be considered in the genome-wide SV studies.

TEs may affect functioning of genes in numerous ways, by inserting in genic location and disrupting open reading frames (ORFs) or into regulatory regions (like enhancers), leading to altered transcription splicing, or creation of novel regulatory elements. These effects can lead to genetic variation, contribute to evolution, and in some cases, drive rapid adaptation or cause disease. One of the beneficial examples is a retrotransposon-mediated duplication of the *SUN* locus in tomato (*Solanum lycopersicum*), which results in fruit elongation: a 27-kb duplication was identified in the landrace Sun1642, which has elongated fruits, while the round-fruited landrace LA1589 carried only a single copy of the *SUN* locus [Xiao 2008]. Another example is active DNA transposon mPing in rice that inserts in gene promoter sequences. The mPing-derived insertion was shown to increase the expression of more than 700 target genes in rice genome, and its copy number was different between rice accessions, ranging from 50 up to 1000 [Naito 2009]. Throughout evolution, some TEs undergo domestication by losing their transposition activity and acquiring a gene function, as was shown in Arabidopsis transposase-derived protein ALP1 (ANTAGONIST OF LIKE HETEROCHROMATIN PROTEIN 1), that functions as a transcription factor during the modulation of plant light sensing [Liang 2015].

The entire set of mobile genetic elements in the Arabidopsis genome makes at least one fifth of its genome content. Arabidopsis TEs are further categorized into more than 320 families, of which the most abundant are *Gypsy* and *Copia* TEs from Class I, and Helitron and Tandem Inverted Repeat (TIR) -containing elements from Class II [Joly-Lopez 2014, Quesneville 2020]. TEs are thought to be among the primary drivers for intraspecies differences in genome size. Remarkably, a comprehensive survey of TE representation and activity across 211 Arabidopsis accessions revealed the majority of TEs to be shared by a few accessions or even accession-specific [Quadrana 2016]. This study reported recent activity for 9 TE families, while most TE sequences were evolutionary silenced by DNA methylation. The relatively recent TE activity in Arabidopsis was reported for retrotransposons LTR/COPIA2 and LTR/COPIA78, and differences in their copy numbers among the accessions were correlated with the annual temperature range [Quadrana 2016]. Because of the potential to alter gene function, the activity of TEs is tightly repressed by epigenetic modification of chromatin in their localization, including direct methylation of cytosine to 5-methylcytosine (5mC) in TE sequence or methylation of histones that fold TE-overlapping regions [Liu 2022].

1.2. Methods for SV identification

Since the publication describing the duplication of *BAR* gene in *Drosophila melanogaster*, which was identified by microscopy by Calvin Bridges in 1936 [Bridges 1936], the number of reports of SV events has been growing intensively. In the early days, individual genomic variants inaccessible by microscopy were identified with molecular biology methods, mostly based on PCR. Developed by Sanger and colleagues in late 70th of the previous century [1977], the first DNA sequencing method, namely chain-termination sequencing, enabled to sequence the genome of a human and other model organisms including Arabidopsis, however due to the limitation of the method's performance the sequencing of a human genome took 13 years (https://www.genome.gov/human-genome-project/timeline). The possibility of working with reference genomes led to the development of methods to analyze variation at a genome scale. The timeline of the most widely used approaches for the genome-wide identification of SVs complemented by the most remarkable surveys of Arabidopsis genome variation is described in this chapter.

1.2.1. Hybridization-based methods. DNA can be annealed to the probe complementary sequence, the method termed as hybridization [Carter 2007]. In the comparative genome hybridization (CGH), the DNA from the two genomes – the reference and the sample – is labeled with fluorescent dyes, then hybridized to a set of probes fixed to a microarray (Figure 4A). The DNA amount is detected by the comparison of fluorescence intensity signals between the samples. The method allows for the detection of DNA gains and losses; therefore, it was developed to detect CNVs. Initially the probes were designed using bacterial artificial chromosomes (BACs), as demonstrated in the remarkable study in which 55 human genomes were compared [Iafrate 2004]. The authors reported 255 loci with CNV and described more than 20 variants abundant within the sample set. This study played an important role in the development of aCGH (array-based CGH method), which subsequently stimulated the development of other types of probes, such as complementary DNA (cDNA) or oligonucleotide probes that provided much better resolution and introduced SNP microarrays [Alkan 2011].

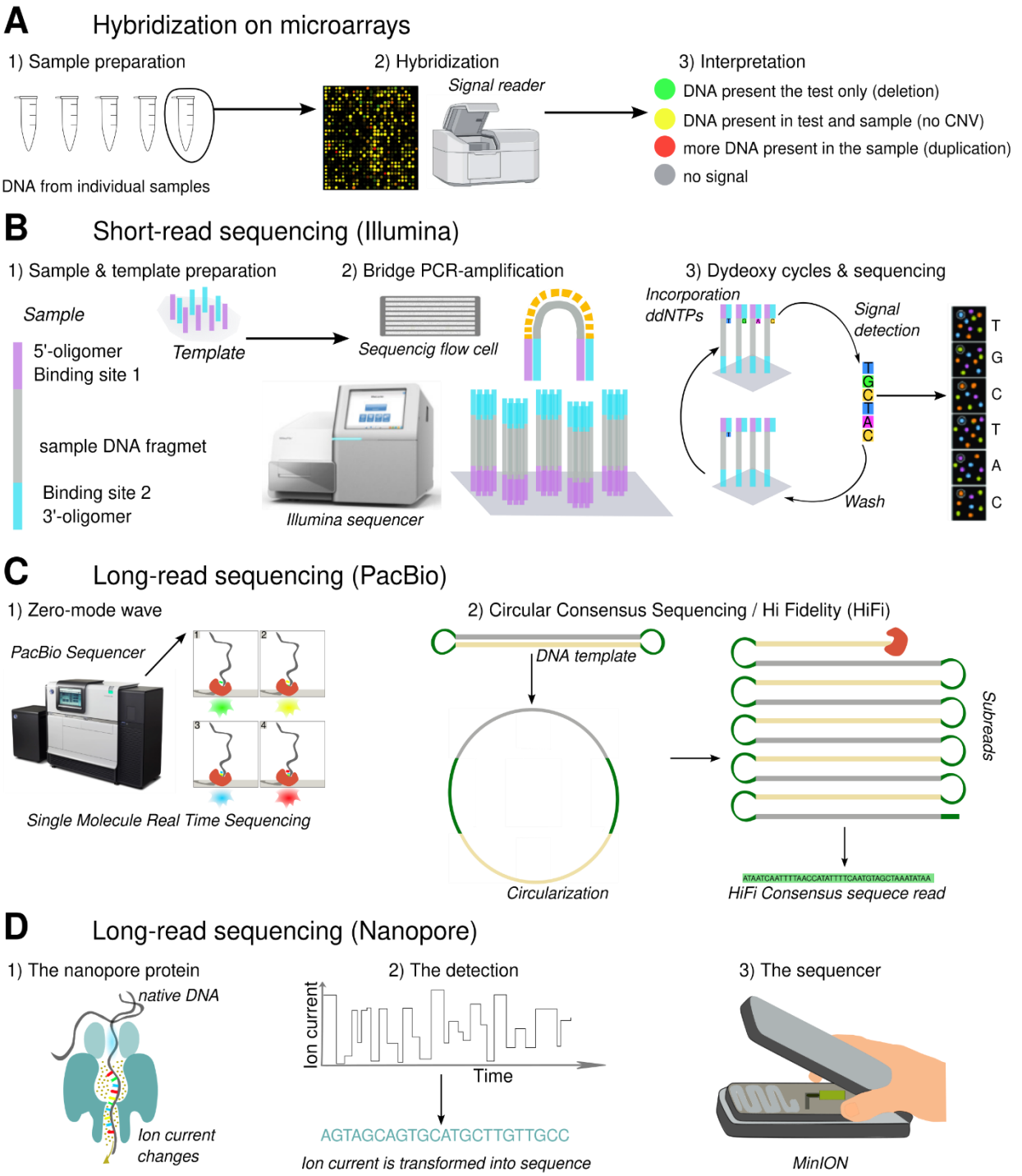


Figure 4. The principles of the methods for genome-wide SV identification. A – hybridization on microarrays; B – short-read Illumina sequencing; C – long-read PacBio sequencing; D – nanopore sequencing.

1.2.2. Short read sequencing-based methods. Sanger sequencing was fundamental for deciphering the sequence of a human genome and acquired the name of the "first-generation" sequencing method. The newer automated sequencing methods, consequently, are referred to as the "next generation" sequencing (NGS) techniques. Guided by the idea of increasing sequencing throughput, several strategies of NGS have been developed to produce large amounts of

sequencing data with relatively low costs. This aim was possible to achieve by shearing the DNA molecules into very short fragments, and these fragments are then partially sequenced from one or both sides, generating reads of 70-150 bp in length, which represent the type of data produced by NGS methods, also known as short-read sequencing. The principle of short-read sequencing can be demonstrated on the example of sequencing-by-synthesis, in the way it was implemented in the most successful NGS platform provided by Illumina company.

The NGS reaction is performed according to the following steps: sample preparation, amplification reaction, and sequencing (cycles of synthesis and fluorescence signal detection) [Metzker 2010]. It utilizes a sequencing flow cell made of glass, on which oligomers of two types, 3'- and 5'- complementary to the sequencing adapters are immobilized and function as amplification primers. The sample DNA is fragmented into very short pieces (typically 300 – 600 bp) and sequencing adapters are added to each strand of the fragment. Both sides of the prepared DNA template are annealed with the oligomers at the flow cell, followed by bridge-amplification – PCR-reaction. (Figure 4B). This way, each DNA fragment undergoes numerous amplifications forming so-called clusters represented by clones of one DNA template and spatially separated on the flow cell. The cluster formation allows strengthening the fluorescence signal during sequencing step. For example, one flow cell can contain 100-200 million clusters, making this sequencing massively parallel. The final step, sequencing, is conducted by a cycle of reactions starting with incorporation of one of the four complementary dideoxy-nucleotides (ddNTPs) to the template, and the detection of fluorescence signal (sequencing). As the incorporation of ddNTPs, each labeled with a different fluorescence marker, into DNA chain terminates its synthesis. To renew the cycle, wash and nucleotide repair steps are performed before the next sequencing cycle. The throughput of Illumina sequencing is a magnitude higher than Sanger sequencing: it can generate sequencing data that covers the genome and identify numerous variants in one sequencing experiment [Alkan 2011].

Individual short read does not carry much information when the genome of interest is many orders of magnitude longer (10^6 times in case of Arabidopsis). Therefore, various methodologies exist to analyze short-read data. The main approach for NGS data analysis is based on mapping the reads to the reference genome by sequence alignment and detection of the genomic locations with the divergence in the alignments, termed as resequencing approach. Different patterns, e.g. distance, number, and orientation of the reads indicate the variant event and might define its type, size, and breakpoints. The utilization of paired-end sequencing libraries increases the accuracy even more, as this type of libraries relies on the information from both mapped reads of a pair and the mapping distance between them – the analysis of inserts. These features enabled the generation

of Gigabases (Gb) of sequencing data and made Illumina sequencing the most widely used platform in the field.

In NGS-sequencing methodology, four main approaches exist to detect SVs from read data: read-depth, read-pair, split-read, and assembly (Figure 5) [Alkan 2011]. The simplest method, called r*ead-depth*, assumes that the distribution in read mapping coverage is uniform across whole genome sequence length and checks for deviations from this distribution: the duplication is revealed by the higher mapping coverage, while the deletion shows lower or zero coverage. The most beneficial method for paired-end reads is *read-pair* – it approaches the distance and the orientation between the reads of one pair, originated from one DNA fragment. The increased or decreased distance corresponds to the deletion or insertion, respectively, while the changes in mapping orientation might indicate an inversion.

Split-read method is designed to detect the breakpoints of SVs at high resolution, as it detects the reads that have two or more alignments to distinct regions of the genome with non-overlapping parts of the read. This method can detect the location of large SVs and is especially efficient for detecting TE-derived insertions, because some TEs often have specific patterns at their insertion flanks. Finally, the *assembly* method is based on detecting overlapping reads originating from the same genomic location and joining them into longer sequences – *contigs*. The method promises to discover the full extent of variation but in the implementation of short reads it was successful only in the case of deciphering individual variants. All the mentioned approaches allow for cost-effective and accurate detection of many variants of single-nucleotide to medium length. However, NGS is not applicable for the identification of large and complex variants, like nested SVs, and does not provide adequate information about the sequence not present in the reference genome. In addition, the approaches work best for model species for which the almost complete reference genome sequence is available.

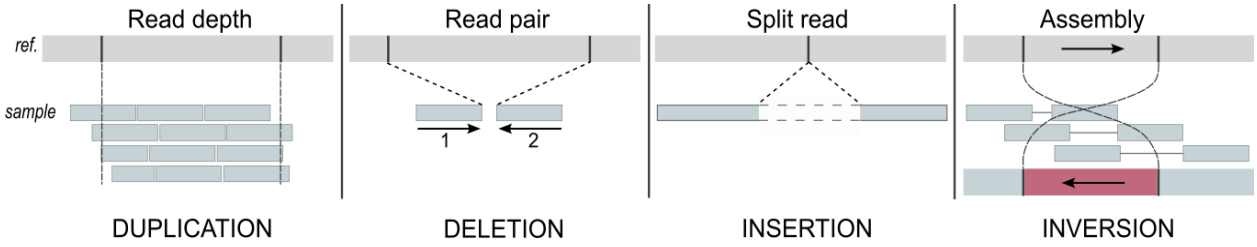


Figure 5. The approaches to infer SVs based on discordant mapping patterns and examples.

The existing software tools to call variants from NGS data usually utilize one of the four mapping analysis approaches, or a combination of them. The example SV caller based on the read-depth approach is *CNVnator* [Abyzov 2011]. Read-pair method is implied in *BreakDancer* that

measures the distance and orientation between a read and its mate [Chen 2009]. Split-read approach is utilized in *Pindel* [Ye 2009]. Other types of callers that combine two or more approaches for processing the same data set are designed to improve the sensitivity of detection. One of such methods is called *Genome STRiP*, which combines the classic read-pair mapping patterns and integrates it with coverage (read-depth) information to support the identification of SVs shared between the genomes across population [Handsaker 2015]. The assembly approach in short read data can be used to perform local assembly, for instance to confirm insertions which cause inexplicable mapping patterns [Wala 2018].

1.2.3. Long read sequencing-based methods. Despite the very high accuracy of Illumina technology, its read length remains the main drawback of NGS methods, while proper deciphering complex SVs generally requires much longer reads, ideally the ones that will overlap the whole region of the rearrangement. Sufficient elongation of sequencing reads was enabled by the introduction of the long-read sequencing technologies, also known as "third generation sequencing". The first widely popularized long-read technology was released in 2011 by Pacific Biosciences (PacBio) [Rhoads 2015]. The technique is based on the polymerase synthesis of a complementary DNA strand, with the enzyme fixed in a tiny well, through which the fluorescence signal is passed, namely zero-mode waveguide (Figure 4C). The fluorescence signal is captured during the incorporation of each nucleotide by the polymerase; thus, the sequence is read in real time. The latest release of PacBio technology is called circular consensus sequencing, or HiFi, in terms of high-fidelity. It is performed by the rolling cycle amplification and simultaneous sequencing of the same circularized template with subsequent aligning the subreads to each other, and generating the consensus sequence read. The average length of the resulting HiFi read is 15-25 kb, and the accuracy is reaching 99.9%, which is equal to the accuracy of the short-read techniques.

Another long-read sequencing technique that stands in line with PacBio technology is nanopore sequencing introduced by Oxford Nanopore Technologies (ONT) [Jain 2016]. The method uses the physical properties of a DNA molecule to cause characteristic changes of the ionic current under an applied electric field. In nanopore sequencing, the double-stranded DNA molecule is unwound and one of the strands passes through a protein pore – nanopore – and causes the changes in the electric current of the system, which are measured by a detector (Figure 4D). The method offers not only the sequencing of the DNA molecule in its native form, but also generates sequencing reads of a very long length, in theory limited only by the physical length of the DNA molecule. Moreover, the most popular nanopore sequencing device – MinION – is small in size and requires only a medium-performance computer or a laptop for basic sequencing

applications. The accessibility and read length distinguish nanopore technology from other sequencing methods. As the current study has been based on this sequencing technology, a detailed description of the method's principle and the data analysis steps are presented in the next sections.

1.2.3.1. Sequencing native DNA molecules with nanopores. The core of an ONT sequencer is a flow cell consisting of 2048 nanopore proteins, joined into 512 channels through which the electric current is conducted [Jain 2016]. Prior to loading onto nanopores, sequencing adapters are ligated to both ends of DNA molecules. The adapters provide the contact of the DNA with a helicase-like protein, which unwinds a double strand, and helps to recruit single-stranded molecules into the nanopores, by binding a special motor protein. The changes of the electric current are detected in real time as DNA is being passed through the narrowest place of the pore (active site). Because the current is disrupted by the specific chemical properties of the nucleotides and any modifications in the strand, this signal is a fluctuating "squiggle", which is then computationally transformed into nucleotide sequence by comparison with the existing (pre-trained) graphical pattern of k-mers, or a word which represents several consecutive nucleotides. The sequence information can be read in two ways: single strand reading and sequencing (1D reads) regardless the location of the two strands; or reading the strands one after another and linking the read information by the time proximity of anchoring a nanopore ($1D^2$). The 1D read type generates the highest data yield, lower accuracy, while the $1D^2$ generates so-called duplex data, in the case the reads derived from two strands can be linked by sequence complementarity. Initially $1D^2$ was not successful, however it is constantly progressing – in first releases the number of reads that created a duplex was ~8% of total, a recent version generates duplex data for 50% of the sample output [Laver 2015, Koren 2024].

The two main features that distinguish subsequent nanopore technology releases are the structure of nanopore protein, known as chemistry, and the software that transforms signals of current to sequence data, known as basecaller. Chemistry has an important function of providing DNA pass through the nanopore and signal detection; therefore, it has been regularly improved since the first release by ONT in 2014 [Loman 2015]. The first version of chemistry, delivered to the genomics community, was based on nanopores derived from the bacterial porin A (mspA). This system, referred to as R7 chemistry, generated reads of 75% of sequence accuracy, however that single flow cell output was sufficiently large to assemble the circular genome of *Escherichia coli* into a single contig [Loman 2015]. The following work on the functionality of nanopores led to the release of the new nanopore version R9, which provided the raw reads of 90% sequence accuracy. Further versions of R9 chemistry aimed to increase stability and sensitivity of nanopores, highlighting the potential of technology. The latest version of nanopore chemistry (R10) is

equipped with two active sites instead of one, thus enabling a dual read of DNA molecules, and a new motor protein for stronger control of DNA passage speed. These innovations led to improvements in simplex raw read accuracy of about 97% and increased sensitivity of sequencing homopolymers [Sanderson 2023, Sereika 2022].

1.2.3.2. DNA extraction for nanopore sequencing. The first issue to tackle when working with nanopore technology for whole-genome sequencing is sample preparation, e.g. the choice of an appropriate DNA extraction method. Long-read sequencing requires long stretches of non-fragmented genomic DNA (high-molecular weight DNA, HMW DNA) of high purity, the obtaining of which is often challenging for plant samples, as the force applied to the destruction of plant cell walls may cause undesired DNA fragmentation. Also, plant cells contain diverse metabolites, such as polysaccharides, phenols, or tannins, which reduce the purity of DNA samples, therefore using young leaves is recommended for plant HMW DNA extraction for long-read sequencing [Pucker 2022]. The standard method to obtain HMW DNA is cetrimethylammonium bromide (CTAB)-based DNA extraction, established in 1987 [Doyle 1987]. Several alternative methods of DNA extraction for long-read sequencing have been proposed as well [Mayjonade 2016, Li 2020, Kang 2023a]. Some protocols include nuclei extraction step, where the resulting HMW DNA is of high purity and free from organellar DNA, but this procedure is time-consuming and requires optimization [Li 2020]. In contrast, omitting the nuclei isolation step might require performing size selection in the subsequent steps, to deplete small DNA fragments [Mayjonade 2016].

1.2.3.3. Bioinformatic data analysis for nanopore sequencing. Long reads overcome the limitations of short reads, as they span most SVs within the length of individual reads. When compared two methods, in PacBio the read length is 20 kb on average, whereas in nanopore the reads often reach 100 and 200 kb in length. Expectedly, sequence data of that huge size occupies much more memory resources and storage space, which makes long read data computationally consuming, therefore most software tools for NGS-data were found to be inadequate for usage with long reads. Answering this need, long-read data analysis tools have been developed and are constantly appearing, at least half of which were specified on handling nanopore data [Amarasinghe 2020]. The examples of the most genome analysis tools are presented in Table 2.

The crucial step in nanopore sequencing is basecalling, a computational interpretation of ionic current signals to nucleotide sequence. Consequently, the software for basecalling is undergoing constant improvement, with the aim to better distinguish the ionic current signals between the time segments and extracting the sequence words from the known data sets. Earlier basecalling tools implemented Hidden Markov Models and recurrent neural networks, that were

set on the recognition of 6-base patterns [Boža 2017]. Later implementations were based on machine learning software, that processed ion signal as text data and trained the sample data set based on existing data sets originated from sequencing of known organisms. One of this training software, namely *Taiyaki*, was incorporated in *Guppy* basecaller (https://github.com/nanoporetech/taiyaki). The main advancement of guppy basecaller over previous tools was its capacity to use graphics processing units (GPU) for computing, which reduced computational time substantially. The latest and the most comprehensive basecalling toolkit, called *dorado*, has the functionality to process the data generated by R9 and R10 chemistry versions, call canonical and modified bases, align raw reads to the reference and perform other useful functions for data manipulations (https://github.com/nanoporetech/dorado).

Table 2. Representative software tools for nanopore sequencing and genome analysis.

Data analysis	Software tools
Base calling	**Guppy**[1](https://nanoporetech.com/software/other/guppy) no longer supported Dorado (https://nanoporetech.com/software/other/dorado)
Quality control	**NanoPack** [De Coster 2018]; **PycoQC** [Leger 2019]
Sequence mapping	**Minimap2** [Li 2018]; NGMLR[2] [Sedlazeck 2018]; Winnowmap2 [Jain 2022]
De novo assembly	**Minimap/miniasm** [Li 2016]; **Canu** [Koren 2017]; **Flye** [Kolmogorov 2019]; **HASLR** [Haghshenas 2020]; **SPAdes** [Prjibelski 2020]; wtdbg2 [Ruan 2020]; Shasta [Shafin 2020]; Verkko [Rautiainen 2023]; hifiasm [Cheng 2025]
Scaffolding	**RagTag** [Alonge 2019]
Assembly evaluation	**QUAST** [Gurevich 2013]; **BUSCO [**Simão 2015]; **Merqury** [Rhie 2020]; GenomeQC [Manchanda 2020]; **Inspector** [Chen 2021]; AssemblyQC [Rashid 2024]
Error correction	Nanopolish [Simpson 2017]; **Racon** [Vaser 2017]; Medaka (https://github.com/nanoporetech/medaka)
Mapping-based SV calling	NanoSV [Cretu Stancu 2017]; Sniffles [Sedlazeck 2018]; CuteSV [Jiang 2020]
Whole-genome comparison	**Assemblytics** [Nattestad 2016]; **MUMmer** [Marçais 2018]; **SyRi** [Goel 2019]

[1]The tools used in this study are highlighted in bald.
[2]The underlined tools were also tested during study design.

According to recent overview of software tools developed for nanopore sequencing, a high number of the tools were developed for *de novo* assembly, reflecting the fact that the genome assembly is one of the main implementations of this technology [Amarasinghe 2020]. The two most used algorithmic approaches in *de novo* assembly are Overlap-Layout-Consensus based

(OLC) and Graph based. The first one searches for overlaps between the long sequences and tries to join them into longer sequence fragments. De Brujin Graph connects long sequences (edges) by the repeated points (nodes) and tries to reproduce as longest chain as possible. Both algorithms generate draft assemblies, represented by the sets of contigs. Ideally, the number and the lengths of contigs should correspond to the number and lengths of chromosomes for a given genome. In case of linear chromosomes, the contig number of draft assemblies is bigger. The resulting draft assembly then undergoes polishing – mapping raw reads to contigs, to improve the per-base accuracy. Some algorithms for polishing also are aided by the mapping short reads, as their raw accuracy is higher – the approach called hybrid polishing. The arranging of the contigs at chromosomes is termed scaffolding. When trying to assemble the unknown species, the template for scaffolding is usually the genome of the closest species. As a result, the final assembly is represented by the set of chromosome-representing scaffolds.

De novo assemblies are generated with the aim to reconstruct sequences of chromosome length, currently near-complete but still not complete. For the linear chromosomes, the assembly evaluation is an important step to confirm that the chromosomes are reconstructed correctly. Current metrics for assembly evaluation can be done in three dimensions: contiguity, completeness, and correctness. To check the contiguity or how much the assembly length resembles the genome length, N50 metrics is used. N50 is equal to the length of the shortest contig in a set of contigs that, when sorted from the longest, represent 50% of the assembly length. The completeness of the assembly can be measured by the two parameters: BUSCO (Benchmarking Universal Single Copy Gene Orthologs) [Simão 2015] which measures the percentage of conserved gene orthologs typical for a given lineage, and k-mer completeness, which compares the percentage of set of short sequences (k-mers) between the set of reads and the assembly [Rhie 2020]. In an ideal assembly both BUSCO and k-mer percentages reach 100%. The correctness of the assembly can be measured on the per-base level (Consensus Quality Value or QV is equal to average base-level accuracy of the final sequence) and at genome-scale level, by whole-genome comparison [Goel 2019].

SV identification using long-read sequencing can be done by both genome mapping and assembly approaches. The tools to identify SVs based on mapping long reads to reference genome exist, like *Sniffles* [Sedlazeck 2018] or *CuteSV* [Jiang 2020]. However, the number of tools for assembling long reads is a magnitude higher, making this approach more promising, including not only accuracy improvement of the final sequence (consensus), but also discovery of sequence not present in the reference genome. The SV detection by *de novo* assembly is achieved by the alignment of the two genome sequences and direct comparison of positions between the reference

and the sample. The example tools for assembly-based genome alignment methods are *MUMmer* [Marçais 2018] and *SyRi* [Goel 2019]. Moreover, a genome assembly of high-quality itself carries extremely valuable information about the species, as it often becomes a template for annotations of genes, TEs, and other features in the genome, and can be integrated with transcriptomic, proteomic, or epigenomic data. For species with the known genome sequences, the genome assembly for other individuals then reference allows for identification of the whole extent of variation that exists within a given species.

The main challenge when working with nanopore technology is its accuracy. The most frequent sources of errors are low signal-to-noise ratio, nonhomogeneous speed of DNA translocation, and long word size [Rang 2018]. Therefore, basecallers often lack the capacity to accurately establish homopolymers' length, and the most frequent type of nanopore sequencing errors is deletions within homopolymers. However, homopolymer sequences remain challenging for all existing sequencing techniques [Delahaye 2021]. To deal with the issue computationally, error correction steps should be considered in the downstream analysis, based on a so-called self-correction via aligning reads to each other and generating a consensus. As an alternative, hybrid correction tools are available, for error correction of long reads by mapping short and highly accurate Illumina reads [Wang 2021]. Still, even the latter type of correction will not be useful in case of very long homopolymer stretches, exceeding the short read length.

1.2.3.4. Numerous applications of nanopore sequencing. Nanopore sequencing has multiple applications [Satyr & Żmieńko 2020]. Due to length possibilities, nanopore sequencing is being used to create draft assemblies of genomes of non-model organisms, and to close gaps in existing assembled genomes. In the case of detecting small genomes, like the human coronavirus SARS-CoV-2, nanopore technology enables its sequencing practically in real-time, by generation appropriately long reads that can be analyzed on site [Moore 2020]. Besides native DNA sequencing, nanopores also offer RNA sequencing, either by direct reading from single RNA molecules or reverse transcription and cDNA sequencing. The possibility of sequencing whole-length transcripts enables the discovery of all alternative isoforms, and improves annotating expressed genes, and non-coding RNAs. In addition to the canonical nucleotides, the signal detected by nanopores can be extended towards finding epigenetic modifications of nucleotide bases, the most common of which is 5mC. The advantage of 5mC detection by nanopore technique is that the signals can be read directly in the raw data, just by activating its "extended alphabet" functionality during the basecalling step. Moreover, recent software tools for 5mC basecalling have the capacity to process raw sequencing data generated by the older versions of nanopore chemistry, opening the possibility to re-analyze the already existing data sets. The accuracy of

5mC detection by nanopore sequencing is highly consistent with the accuracy of the gold-standard bisulfite sequencing [Yuen 2021]. Other types of DNA/RNA modifications may be detected as well, although so far with lower accuracy.

1.2.4. Beyond sequencing: alternative methods of SV detection. The new technical innovations emerge constantly and extend the opportunities of genome analysis methods; therefore, it is hard to present in detail the plethora of the approaches for SV detection. Two methods are worth mentioning in the context of this work, as they are often used to complement long-read sequencing experiments: optical mapping and Hi-C. Optical mapping relies on nicking the stretched and stained DNA molecules by endonucleases, *in silico* check of the nicking sites and physical mapping of these sites on chromosomes. The approach has been known since 90s of the previous century [Schwartz 1993], and currently it has the capacity to process ultra-long DNA molecules. Optical mapping can help to resolve very long complex DNA regions, starting from 6-8 kb, but it does not process microscopic images with base-by-base resolution [Ho 2020]. In one of the recent studies on the Arabidopsis genome, optical maps helped to identify the insertion in the genomic region overlapping three R genes, which was previously undetected by short-read mapping [Kawakatsu 2016].

Hi-C is the abbreviation for Hi-throughput 3‘C’ (chromosome conformation capture) technique. As the name suggests, the method involves the ‘capture’ of DNA strands that exist in physical proximity. Generally, the cross-linked pairs of DNA fragments are immobilized with formaldehyde, digested into smaller sets and re-ligated in the way that only covalently bound fragments can be captured, followed by sequencing [Belton 2012]. Biotin-labelled nucleotides are incorporated at the places of ligation junctions and are purified selectively prior to sequencing. The method provides information about DNA fragments that can be distant in linear genome but proximal in 3D organization of chromosomes. The chromosomal distance is evaluated by calculating contact frequency between read pairs and visualized as heatmaps. The example heatmap, performed by aligning Hi-C reads to long-read assemblies of Arabidopsis centromeres, that are organized in satellite arrays, finely demonstrated the applicability of Hi-C method for sequence validation in repeat-rich regions [Naish 2021]. Naturally, the combinations of 3C with long-read sequencing are under active development. For instance, by coupling 3C with subsequent PacBio sequencing, CiFi protocol promises the analysis of whole-genome interactions with low input requirements [McGinty 2025], while Pore-C approach is based on 3C followed by nanopore sequencing and offers not only sequencing the ligation sites, but also long DNA concatemers, which enable to calculate higher order distances between the sequences [Deshpande 2022].

1.3. SV studies in Arabidopsis: from past to present

1.3.1. Genome-wide studies of Arabidopsis intraspecies variation. A population-wide sequence variation of *A. thaliana* species has been a subject of strong research interest since the discovery of its intense natural variation. Accordingly, numerous studies aimed at understanding the genetic source of this variation have been performed after its genome had been sequenced, which nicely reflect the pace of development of the SV analysis approaches (Table 3, Figure 6). The first polymorphism data set was collected in the scale of 20 accessions, using oligonucleotide arrays [Clark 2007]. The study focused on identifying SNPs, of which more than 1 million polymorphic locations were discovered in the analyzed accessions, compared to reference genome. The authors also pointed out the unequal polymorphism frequencies between the gene families, indicating evolutionary-driven effects of polymorphisms on genes. Soon after, short read-based resequencing of 3 accessions (2 non-reference and 1 reference) was published [Ossowski 2008]. In this work, the authors mapped short reads across the reference genome and compared the mappings of two distinct accessions, Bur-0 and Tsu-1 against Col-0. They revealed ~820 thousand SNPs and almost 80 thousand small-scale variants 1-3 bp in length. The authors also found a 3 Mb genomic region rich in polymorphisms. Even though the read length of data generated in this study was 36 bp, the preponderance of resequencing methods over hybridization-based ones was highlighted.

Another microarray-based survey aimed to genotype 149 SNPs across the genome in a cohort of 5707 accessions, collected from 42 countries on 4 continents [Platt 2010]. This study described for the first time the classification of the worldwide Arabidopsis population into genetic groups. It reported isolation by distance pattern for accessions originated from North America, that were almost identical, and substantial variation among Eurasian accessions at even 1000 km distance of their habitats. The microarrays also were used to map larger than single-nucleotide deletions in 4 Arabidopsis accessions – Eil-0, Lc-0, and Sav-0, and Tsu-1 [Santuari 2010]. The authors used tiling arrays, designed to detect the absence of homologous sequence to a given tile, accompanied by short-read sequencing (paired-ended reads of 35 bp read length) to validate the positions of deletions by mapping. Most deletions, that were different between the accessions, were identified in regions encoding TEs, but also in a significant portion of the protein-coding genes. Another analysis on the topic was done by genotyping 250 SNPs on a smaller set of 1307 accessions using array panel named *RegMap* from Regional Mapping [Horton 2012]. The study confirmed the general pattern of polymorphism from the earlier study, in that the locations of the highest polymorphism frequency were found in or around TEs, intergenic regions and

pseudogenes, and the lowest polymorphism frequencies were detected in protein-coding genes. However, the authors also found examples of differentiating SNPs in the genes associated with flowering (*SHORT-VEGETATIVE PHASE*, *FLOWERING LOCUS C*, and *FRIGIDA*), that were putative targets of positive selection during the evolution differentiating plants from east Europe, Fennoscandia and Asia. The study provided an insight into the genetic basis of evolutionary adaptation of plants to different environments.

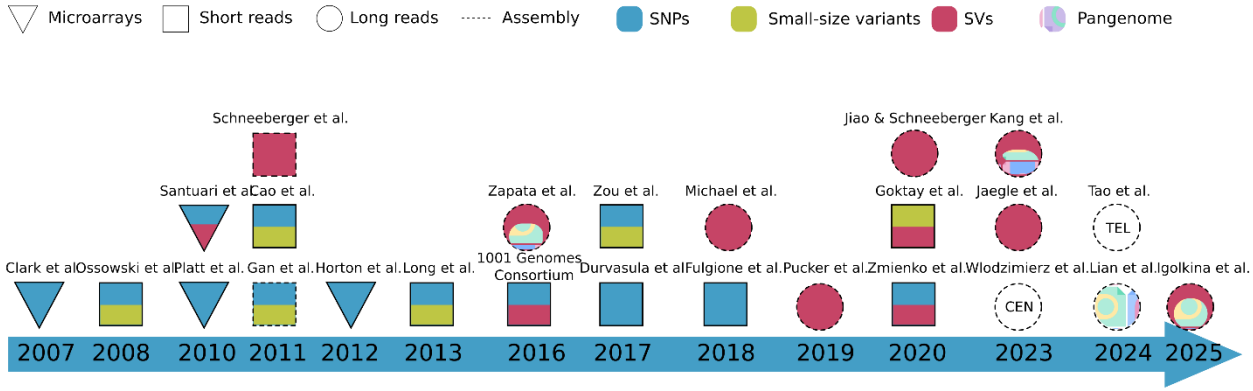


Figure 6. The studies of natural variation in Arabidopsis population by genome-wide analysis approaches.

The delivery of accessible short-read sequencing initiated the project of resequencing Arabidopsis populations to identify the extent of variation that plays a role in plant adaptation to different environments. The first phase of the initiative was realized by resequencing 80 accessions with Illumina reads [Cao 2011]. The main project publication was released in 2016, reporting the global collection of 1135 resequenced accessions [1001 Genomes Consortium 2016]. The description of single-nucleotide and small-scale variation landscape allowed for identification of genetic groups within the population, and for identification of relict accessions, that were found to be highly distinct from each other and from the rest of genetic groups, and mostly originated from Iberian Peninsula but also were found in other lands like Tanzania. The genetic groups from North America had nearly identical variation pattern, which was in line with the early observations of Platt and colleagues. Opposite, the genotypes of individuals from the British Isles were more variable, suggesting more ancient and multi-stage colonization. In general, the authors grouped the accessions into nine genetics groups, including a separate cluster for relicts. This study clarified the relationship between the genetic populations and geographic distribution of the modern Arabidopsis population, which boosted subsequent genome-wide association studies aimed to map various morphological and adaptive traits observed among the natural accessions.

Table 3. A collection of comprehensive studies of natural variation in Arabidopsis.

Reference	Accessions	Project description
Microarrays		
[Clark 2007]	20	1.05 million SNP positions in protein-coding genes identified across 20 natural accessions compared to reference genome. More than 2000 genes from 143 gene families were found to carry "major effect" polymorphisms. The highest SNP variation was reported for F-box and NB-LRR gene families.
[Platt 2010]	5707	SNP profiling across a cohort of accessions and clustering into genetic groups. The considerable variation among local groups in Eurasia continent and the isolation by distance in North American accessions. The first classification of accessions into genetic groups.
[Santuari 2010]	4	Combination of tiling hybridization and short-read sequencing. The mapping of large deletions in genes. Most deletions, distinct between the accessions, are overlapped TEs and less protein-coding genes.
[Horton 2012]	1307	Genotyping of 250 SNPs in 1307 accessions. Reported SNP frequencies were higher in TEs, intergenic regions, and pseudogenes, and lower in protein-coding genes. Several SNP hotspots that differentiate genetic groups from east Europe, Fennoscandia, and south Russia, were found to be in genes encoding proteins associated with plant development.
Short-read genome resequencing		
[Ossowski 2008]	3	The first resequencing with Illumina reads, 36-bp in length. 823 325 SNPs and 79 961 indels were identified. 3.4 Mb of sequence length was reported to be highly variable from the reference Col-0.
[Gan 2011]	18	The first short read-based assembly of 17 distinct accessions by iterative mapping and sequence extension with local *de novo* assembly. Per accession, from 497 668 to 789 187 SNPs were identified, 1.2 million indels and 100 thousand substitutions were also reported. ~17% protein-coding genes were affected by polymorphisms.
[Cao 2011]	80	I phase 1001 Genomes Project: the analysis of 80 samples of naturally inbreed strains originating from distinct geographic locations. 4.9 million SNPs and 810 467 indels identified in total; 2.68 mln SNPs and 370 258 indels were common in all genetic groups.

[Schneeberger 2011]	4	Reference-guided assembly of 4 individual strains (Ler-1, C24, Bur-0, Kro-0). 7% of the genome was covered by strongly diverse regions, affecting from 4% to 5% protein-coding genes. A substantial number of annotated genes were partially deleted, highlighting a major limitation of short read-based assembly.
[Long 2013]	180	Sequencing 180 lines from Sweden with short reads. 4.5 million SNPs and 0.6 million other variants were identified. SV was found to be a major component of genome variation. 1.3-3.3 Mb of novel sequence was assembled. Lines from south Sweden were shown as more genetically closed to European lines, while the lines from north Sweden are reported to be more distinct.
[1000 Genomes Consortium 2016]	1135	The Consortium project of resequencing 1135 distinct accessions with short reads. 10 707 430 SNPs and 1 424 879 larger variants were identified. Clustering-based classification into 9 genetic groups was made, with the separation of a group of relicts that are highly divergent. The prediction of SNP impact in coding sequence showed that at least one SNP in ~440 genes per accession may cause inactivation of its expression.
[Durvasula 2017]	78	Short read sequencing and genotyping of 78 African accessions, modern and herbarium samples. PCA-based clustering of populations from Africa and a global collection of 1135 accessions by habitat distance. African populations are shown to putatively be the most ancient.
[Zou 2017]	118	Short read sequencing and genotyping of 118 Chinese accessions. 2.66 million SNPs and 0.59 million indels were identified. PCA analysis clustered the accessions to Chinese-Yangtze river, Chinese, and Europe / North Africa populations, of which Yangtze River population was the most unique from the remaining ones. Evolutionary analysis showed that the divergence of Yangtze River population was ~60 thousand years ago.
[Fulgione 2018]	14	A comparison of variation of 14 accessions from Madeira Island with 1135 worldwide and 78 African ones with the focus on studying the evolution of the species. The diversity within Madeiran population was ~4-fold lower than in worldwide population, which was explained as the effect of isolation. By clustering, most Madeiran accessions (11) were joined to the Iberian relict's cluster.
[Zmienko 2020]	1060	Reanalysis of the global collection of >1000 genomes with the focus on large variants. 79 137 indels of length 50-499 bp and 19 003 CNVs of length 0.5 kb – 985 kb were identified. The CNV-rich regions were found to cover one third of Arabidopsis genomes and overlap with 18.3% of all protein-coding genes. The copy number status for all genes among the accessions is catalogued on the website "Arabidopsis thaliana CNV viewer".

[Göktay 2021]	1301	Reanalysis of >1000 genomes collection of short read sequencing data with addition of simulation data and whole-genome PacBio data for selected accession. 155 440 indels and SVs of length 50 bp – 10 kb were identified. The housekeeping genes were found to include fewest variants, whereas the defense-associated genes contained numerous variants.
Long read sequencing and de *novo* assembly		
[Zapata 2016]	1	A *de novo* assembly of the accession Ler, using a combination of Illumina short reads, PacBio long reads, and optical mapping-based scaffolding and whole-genome comparison with the reference. Contig N50 = 7.5 Mb. ~3.6 Mb of diverse sequence included 564 transpositions and 47 inversions. 40 Ler-specific genes were annotated. A 1.2 Mb-long inversion at chromosome 4 of Ler was identified.
[Michael 2018]	1	A *de novo* assembly of the accession Kbs-Mac-74, using ONT sequencing, and PacBio and optical mapping for validation. Contig N50 = 12.3 Mb. A direct sequence comparison between Kbs-Mac-74 assembly and the reference genome revealed 4280 SVs of a total length 9.5 Mb. More than half SVs were repeat contractions and expansions.
[Pucker 2019]	1	A *de novo* assembly of the accession Nd-1, using PacBio sequencing, and Illumina sequencing for error correction. Several assemblers were compared. Contig N50 = 13.2 Mb (assembled by *canu*). A direct sequence comparison between Nd-1 and the reference genome revealed 2206 SVs of length >1kb. A highly divergent region of 3.2 Mb at chromosome 2 and a 1 Mb-long inversion at chromosome 4 of Nd-1 was identified.
[Jiao 2020]	7	A chromosome-level assembly of 7 Arabidopsis genomes using PacBio Illumina sequencing. Pangenome analysis revealed 24 thousand core genes and ~1900 accession-private genes (~550 private genes per genome). Direct sequence comparison revealed sequence rearrangements affecting 13 – 17 Mb of sequence. Non-reference sequence, present in each accession, was found to reach up to 6 Mb in size. Sequence changes were found in ~5000 genes, one third of which were non-reference genes.
[Wlodzimierz 2023]	66	PacBio-based *de novo* assembly of 66 accessions with research focus on centromeres (327 out of 330 centromere arrays assembled without gaps). The structure of centromeric region is predominantly represented by 178-bp-length repeat (CEN180) and smaller fraction is represented by 159-bp CEN160. The remaining sequence of centromeric regions is presented by retrotransposons of class ATHILA.

[Jaegle 2023]	6	Using long read *de novo* assembly of 6 accessions to reveal gene duplications, that were mapped as heterozygous SNPs by short reads. 2500 genes were identified as putatively duplicated. Evidence that 90% of heterozygous SNPs, mapped by short reads, are artifacts, as Arabidopsis is a selfing plant.
[Kang 2023b]	32	PacBio HiFi sequencing and *de novo* assembly of 32 different ecotypes with subsequent analysis of the pangenome and identification of SVs. Pangenome analysis defined 21 545 genes as core genes; 3743 as softcore genes; 3929 as dispensable genes; and 2101 as accession-private. 61 322 SVs of size 50 bp and larger were identified. The size of non-reference sequence varied starting with 56 kb and ending with almost 9 Mb. SVs between individual genomes and the reference altered from 500 to 5000 genes. Evidence on SVs contributing to ecological adaptation is provided.
[Lian 2024]	69	Genome sequencing and *de novo* assembly of 72 different accessions using PacBio HiFi (48 accessions) and nanopore long-read (24 accessions) sequencing and short-read Illumina sequencing (all samples). The “core” genome, present in all 69 accessions, counted 19 721 genes; the softcore cluster had 1613 genes; 5582 genes were dispensable; and 6070 were accession-private.
[Tao 2024]	74	The analysis of telomeric repeats using publicly available data from long read-based *de novo* assemblies of 74 accessions. The telomeric repeat composition consists of canonical repeat set, a unit of 7 bp in length (TTTAGGG), variant repeat, represented by variations of 6-8 bp in length, and a stretch of degenerate repeats without any pattern, located between the telomere and the chromosome arm. The diversity of variant repeat reached saturation after analysis of 69 accessions.
[Igolkina 2025]	27	Genome sequencing and *de novo* assembly of 27 accessions using PacBio and Illumina short-reads, with subsequence whole-genome sequence comparison, identification of SVs, gene and TE annotation. 532 178 PAVs of size 15 bp – 20 kb were identified, more than the half overlap with TEs. Pangenome analysis revealed 27 435 genes to of the core; 601 genes of the softcore; 1985 dispensable genes; and 1084 accession-private genes.

The first partial assemblies of individual non-reference genomes were made for 17 accessions using paired-end data of 36 and 51 bp in length [Gan 2011]. The assembly methodology was based on iterative mapping combined with the local *de novo* assembly, resulting in N50 = 80.8 kb in average. The authors identified from 500 up to 800 thousand SNPs per accession, and a total of 1.2 million indels and 100 thousand SNPs while mapping to the reference genome. The authors also generated transcriptome data sets from plant seedlings, to annotate genes in individual plants, which allowed them to make an interesting observation that in most cases the polymorphisms in the coding sequences were compensated to save ORFs, as changed DNA sequence most frequently did not lead to their disruption. Another genome assembly study involving four non-reference Arabidopsis plants was published in 2011 [Schneeberger 2011]. The assembly approach was based on mapping short reads to the reference genomes combined with assembling *de novo* regions for which the read coverage deviated from the average.

One of the earliest surveys reports resequencing of 180 accessions from Sweden [Long 2013]. The authors used pair-end reads of 75 and 100 bp length and explored the variants within this local population, by reads mapping. They identified 4.54 million SNPs and 0.6 million other variants, of which the majority was identical with the data set of 80 accessions from Cao and colleagues [2011]. The accessions originated from north Sweden were reported to be more genetically distinct than those from south Sweden, as the latter were similar to other European accessions. Another research aimed to describe the species from the evolutionary point of view, by sequencing 78 individuals from Africa [Durvasula 2017]. The authors investigated the relationship within African samples, and between African and 1135 global samples, by distance-based clustering. They grouped African population in four clusters, of which one was sub-Saharan and three were north African. Of the latter, the Moroccan samples were concluded to be more ancient, as they harbored the highest number of private SNPs. To see the evolution of the species, the authors studied the variation in the self-compatibility locus (S-locus) and found that the sequence diversity in this locus is sufficiently higher within the African populations, than between African and Eurasian ones. The authors concluded that the split of the most ancient clades – Moroccan, Tanzanian, and Levant – was around 120-90 kya (kilo years ago), while the separation of Eurasian clade was a more recent event, at around 40 kya. In another study, a subset of 14 accessions originating from the oceanic island Madeira was sequenced and compared with the global collection of 1135 accessions and with the African samples, by principal component analysis (PCA) [Fulgione 2018]. The authors claimed that the diversity of Madeiran population is ~4 times lower than worldwide, and that most samples were nested with the cluster of relicts, indicating that the colonization of the island by Arabidopsis plants is an ancient event, and that

Madeiran accessions remained in the long-lasting isolation compared to the time of colonization on the continents. Another diverse Arabidopsis population was found to inhabit the Yangtze river basin [Zou 2017]. Sequencing of 118 Chinese accessions revealed that the divergence of this population from European as well as other Chinese accessions was a relatively recent event (~61 thousand years ago), indicating the role of polymorphisms in local adaptation.

Several studies aimed to catalogue large variations across Arabidopsis genome based on the reanalysis of short-read data from a global collection of more than 1000 genomes. One survey, published by our research group, aimed to reanalyze the short read-based data with the special focus on CNVs of size 0.5 kb and longer [Zmienko 2020]. By combining the results from 7 different SV callers and optimizing a pipeline for identification of CNVs, the colleagues found that at least one-third of Arabidopsis genome is covered by CNVs and that CNV-containing regions are represented by predominantly non-coding or TE-enriched sequences. Nevertheless, 18% of Arabidopsis protein-coding genes overlapped with CNV-regions. Among these genes, evolutionary-young and defense-related genes were most frequent. Another study of similar scale was focused on PAVs [Göktay 2021]. The authors applied three different SV callers on the short-read data from 1301 accessions and identified SVs of up to 10 kb in length. The authors verified the performance of each of the three SV callers using simulated short-read data and a long-read data set of Cvi-0 accession, which was used as an example of one of the most distinct accessions from Col-0. The authors presented evidence for purifying selection of PAVs, which was explained by low occurrence of PAVs in genic regions compared to high occurrence of PAVs in nongenic regions, which in general was consistent with the results for other types of SVs made by other groups.

Numerous variants of different ranges, from SNPs up to large-scale variation, were revealed by resequencing, however these studies relied on the comparison of sequencing reads with the reference genome. With the emerging of long-read sequencing, the process of assembling the entire genome sequence has become reference-independent. Direct sequence comparison between the genomes of individuals is a promising way to expand our understanding of the full extent of genome diversity and discover accession-specific features of the genome. Seminal long read-based studies described sequencing and assembling the genomes of individual accessions: Ler-0 by PacBio sequencing [Zapata 2016], Kbs-Mac-74 by nanopore sequencing [Michael 2018], and Nd-1 by PacBio sequencing [Pucker 2019]. The studies enabled the discovery of variation within the gene space, and annotation of accession-specific gene variants. All three studies revealed varied regions of the genome, when compared to reference accession Col-0.

Further advancement in long-read technologies made it possible to evaluate the degree of genome collinearity within the species. A whole-genome comparison of 7 *de novo* assembled genomes sequenced with the PacBio technology revealed the sequence of size 13-17 Mb in average to undergo rearrangements [Jiao 2020]. The pan-genome analysis of this dataset revealed that out of 27 445 protein-coding genes present in reference annotation, more than 24 thousand genes were present in all analyzed genomes (they were considered as "core" genes). The remaining genes were deleted from some accessions, while around 550 novel genes in each accession were private. Three recent studies describe the intraspecies genome collinearity in more detail by deciphering the structure of the Arabidopsis pangenome, based on long-read sequencing, *de novo* assembling and annotating the genomes: the sample of 27, 32 and 69 accessions, respectively [Igolkina 2025, Kang 2023b, Lian 2024] (Figure 7). Pangenome analysis of the sample set of 27 accessions revealed 24 435 core gene families, in the set of 32 genomes the core included 21 545 gene families, while the collection of 69 accessions revealed only 19 721 core gene families, indicating that the increase of the sample size reduced the proportion of core genome substantially (Figure 7). Chromosome-level assemblies allow for accurate analysis of not only the genomic differences, but also the focus on the genomic similarities, thus helping to better understand the level of genome conservation within the species.

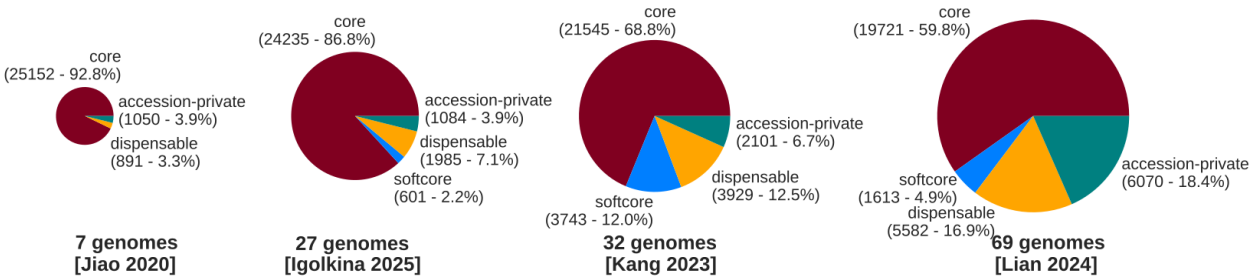


Figure 7. The structure of Arabidopsis pangenome in relation to the size of sample set.

The work on improvement of sensitivity and detection accuracy of both PacBio and nanopore technologies has led to subsequent improvement of long read quality and contig length. Recently, the analysis of repeated sequence has become possible, for example the study of centromeres and telomeres. Represented by arrays of satellite DNA, centromeres were the main reason of contig breaks [Rabanal 2022]. In a recent study, the *de novo* assemblies of 66 accessions were made, aiming to study the diversity of centromeric regions [Wlodzimierz 2023]. The authors sequenced the genomes of the selected accessions using PacBio HiFi technology with different coverage and resolved complete centromere sequences of almost all the centromeres (327 out of 330). PCA analysis of chromosome arms, based on identification of SNPs, clustered the accessions

onto four groups: Eurasian non-relicts, Iberian relicts, Iberian non-relicts, and non-Iberian relicts, which was in line with the 9 clusters previously identified by short reads. The assemblies of 74 accessions, including 66 mentioned above, were also used to perform telomere profiling [Tao 2024]. Besides the canonical unit of a telomeric repeat found in all plants, the authors also found variant repeats, highlighting the diversity of telomeres. These studies demonstrated that using long read-based *de novo* assembly, unraveling the structure of highly repeatable regions, such as centromeres and telomeres, is now possible.

1.3.2. Towards the complete sequence of Arabidopsis reference genome. The fast development of numerous genome assembly tools stimulated better assembling reference genomes, especially for model species. Historically, the first genome sequence of Arabidopsis reference accession Col-0, released in 2000, contained 115 Mb of sequence [The Arabidopsis Genome Initiative 2000], with several newer versions released at TAIR resource up to the latest reference version TAIR10, which contained 165 gaps [Lamesch 2012]. Nowadays, the possibilities to fill these gaps are becoming real. Combination of the two long-read sequencing technologies, supplemented by optical mapping and novel short-read sequencing techniques, enabled the reconstruction of the complete sequence of a human genome [Nurk 2022], from telomere to telomere (T2T) by T2T Consortium. That sequence included all the repetitive sequences as centromeres and telomeres. Due to this and other progresses, long read sequencing was admitted by the status “Method of the year” in 2022 [Marx 2023].

Subsequently, the initiative of the complete Arabidopsis reference genome assembly was supported by different research groups across the world (Table 4). The Col-CEN assembly was performed with the aim of resolving the centromere sequences using combined PacBio and nanopore reads [Naish 2021]. The Col-CEN sequence is 131.6 Mb in size, which contains ~14 Mb of novel sequence compared to TAIR10. Soon after, Col-XJTU assembly was released, with a usage of both long reads and Hi-C technology [Wang 2022], which was 2 Mb longer than Col-CEN and contained two completely assembled T2T chromosomes. Another genome version (Col-PEK) provided further improvement, by addition of Chromosome 1 assembly in T2T mode, and closing the 232.8 kb-long gap from Col-CEN and the 108 kb-long gap from Col-XJTU [Hou 2022]. Publication of alternative versions of reference genome brought the Community to reconstruct the genome including all novel sequences in one standard, that was named Community Consensus version (Col-CC) and uploaded to NCBI Genome database in January 2023 (hereafter Col-CC.v1). This genome, however, still lacked two nucleolus organizer regions (NORs) on Chromosome 2 and Chromosome 4. Subsequent assembling of these NORs finally enabled the release of the fully complete Arabidopsis reference genome sequence (hereafter Col-CC.v2) in October 2023.

Table 4. Genome assembly versions of *Arabidopsis thaliana* reference accession Col-0.

Genome version	Reference	Annotation	Size (Mb)	T2T chromosomes
2012 TAIR10	Yes	Araport11	119.1	None
2021 Col-CEN	No	No	131.6	Chr1, Chr3, Chr5
2022 Col-XJTU	No	No	133.7	Chr3, Chr5
2022 Col-PEK	No	No	133.9	Chr1, Chr3, Chr5
2023 Col-CC.v1	No	No	133.3	Chr1, Chr3, Chr5
2023 Col-CC.v2	No	TAIR12 (upcoming)	142.5	All

The first detailed reference annotation, TAIR10, was published in 2012 [Lamesch 2012]. Later reannotation effort was more comprehensive, and is being used by researchers until today, also being constantly improved [Cheng 2017]. Due to the final genome version release, the annotation of Col-CC is an ongoing project (https://phoenixbioinfo.org/enhancing-arabidopsis-thaliana-genome-annotation-a-community-effort/), in which I participated, is planned to be published soon. Though not annotated, Col-CC genome version serves as the template for sequence scaffolding and comparisons, therefore it should be noted that in this work the term "reference genome" is defined for TAIR10, the current annotated reference version approved by the scientific community. Subsequently, the term "reference annotation" is used for Araport11. Expectedly, the access to complete reference sequence will improve sequence comparisons and reduce false-positives in the identification of variants. Sequencing and assembling non-reference genomes lead to better resolution on intraspecific diversity within Arabidopsis population.

To summarize, with nanopore technology, the access to near-complete genome sequences of individual plants is now possible to get even in a small research group. While numerous studies employing comparative genomics report genome-wide variation (Table 3), only a few surveys focus on accession-specific identification of variants of immune genes [Wan de Veyer 2019]. For as important gene family as *RLP*, after its characterization in the reference genome, only one study has examined intraspecies *RLP* genes [Steidele 2021]. The detailed description of the variation status of *RLP* gene orthologs annotated individually in each of the eight genomes would contribute to better understanding of potential RLP functions.

2. RESEARCH OBJECTIVES

The aim of this work was to identify SV patterns within the *RLP* gene family in Arabidopsis by long-read sequencing and genome assembling of eight diverse accessions. I hypothesized that with the construction of chromosome-level *de novo* assemblies I will be able to investigate complex regions of the genome through direct sequence comparison and, in consequence, to identify new variants of *RLP* genes. To achieve this aim, I set the following research tasks:

1. To establish the protocol of HMW DNA extraction from plant material and subsequently use it for sample preparation and nanopore sequencing of eight selected Arabidopsis accessions.
2. To construct *de novo* genomic assemblies and thoroughly evaluate their quality using bioinformatic methods.
3. To annotate the regions spanning *RLP* gene family members in each assembly and investigate their genetic variation.

3. MATERIALS AND METHODS

3.1. Experimental procedures

3.1.1. Plant breeding. A total of eight Arabidopsis accessions were used for whole-genome sequencing. The seeds were obtained from the Nottingham Arabidopsis Stock Centre (NASC). Detailed description of plants is presented in Table 5. The seeds were sterilized with 70% ethanol (POCH basic, Poland) and 20% commercial bleach, with 3-times of washing after each step. The sterilized seeds were vernalized in 0.1% Agarose (ABO, Poland) for 5 days at 4°C. Seeds were planted into Jiffy pellets in Aracon baskets, one plant per one pot. Plants were grown in a growth chamber (PERCIVAL, USA) under long-day photoperiod (16 hours of light at 22°C and 8 hours of dark at 20°C) and 70% humidity. The plants were supplemented with 0.5x MS Plant Medium (SERVA, Germany).

Table 5. Arabidopsis accessions selected for the study.

1001 Genomes ID	Accession Name	Genetic group (origin country)	SRA ID[1]	Sample ID
N77086 (9669)	Mitterberg 2-185	central europe (Italy)	SRR1946251	MIT
N76805 (9712)	Dolna 1-40	italy balcan caucasus (Bulgary)	SRR1946290	DOL
N78946 (504)	BRR57	germany (USA)	SRR1945450	BRR
N77156 (7288)	Oy-0	admixed (Norway)	SRR1945886	OY0
N77337 (6174)	TÅD 06	north Sweden (Sweden)	SRR1945694	TAD
N77222 (7322)	Rsch-4	germany (Russia)	SRR1945896	RSH
N77196 (9887)	IP-Pun-0	relict (Spain)	SRR1946443	PUN
N76649 (9606)	Aitba-1	relict (Morocco)	SRR1946195	AIT

[1]IDs of the short-read whole-genome sequencing data [1000 Genomes Consortium 2016].

3.1.2. High-molecular weight DNA extraction. 4-week-old plants were taken for HMW DNA extraction. Rosette leaves were grinded in liquid nitrogen using mortar and pestle. 2 mg of leaf material was dissolved in 8 ml of CTAB buffer for cell lysis (EURx, Poland) with the addition of 50 µl RNase A (QIAGEN, Germany) and incubated at 65 °C for 1 hour. After 30 minutes of incubation, 80 µl proteinase K (QIAGEN or NEB, USA) was added. The lysate was centrifugated for 5 minutes at 20 000g at room temperature, and the supernatant was saved for the next step. HMW DNA was isolated using Genomic-tip 20/g columns (QIAGEN) according to manufacturer's protocol. DNA was purified on the spin-column Genomic DNA Clean & Concentrator-10 (ZYMO RESEARCH, USA). The concentration and purity of DNA samples were

measured on bioanalyzers, Qubit fluorometer (Thermo Fisher Scientific, USA) and NanoDrop spectrophotometer 2000 (Thermo Fisher Scientific, USA). The average length of fragments was measured with TapeStation (Agilent Technologies, USA) analyzer. The detailed protocol is attached in Appendix 1.

3.1.3. Preparation of sequencing libraries. 2 µg of HMW DNA was used for the preparation of two sequencing libraries per each sequencing experiment. The first library was added in the beginning of the run and the second library was added 8-16 hours after the sequencing start. Preparation of sequencing library by PCR-free adapter ligation was performed according to the manufacturer's instruction (Ligation Sequencing Kit SQK-LSK109 by ONT). Sequencing runs were performed on MinION sequencer (ONT), using flow cells of version R9.4.1. One sequencing run lasted 96 hours. One sequencing flow cell was used per one accession. Raw signals were collected with manufacturer-provided software (MinKNOW) in FAST5 format.

3.2. Bioinformatic procedures

All the bioinformatic analyses were performed using Eagle Cluster provided by Poznan Supercomputer Networking Center (PSNC) and local 16-core AMD Ryzen Threadripper 1950X Processor. Construction of draft assemblies was made by me and dr Paweł Wojciechowski from Poznan University of Technology. The annotation of TEs in the accessions were made in collaboration with prof. Wojciech Makałowski and dr Matías Rodríguez from the University of Münster in Germany, with funding by NAWA Exchange Grant. The illustrations were prepared using *Inkscape* graphics software, *R* and *Python* scripts.

3.2.1. Basecalling. The raw signals of ionic current changes generated by MinION sequencer were basecalled with *guppy* (https://nanoporetech.com/), using Graphics Processing Unit (GPU) Nvidia V100 32GB (`--device auto`) and converted to FASTQ format. The configurations for basecalling included flow cell version R9.4.1, sequencing speed 450 bp per second, and high accuracy basecalling algorithm.

3.2.2. Data filtering. Sequencing adapters were filtered out using *porechop* (https://github.com/rrwick/Porechop) with default parameters. Sequencing statistics were calculated with *NanoComp* [De Coster 2018], using `sequencing_summary.txt` file as an input. Reads of length lower than 1 kb (`--min_length 1000`) and of quality lower than 7 (`--min_q_weight 7`) were filtered out by *Filtlong* (https://github.com/rrwick/Filtlong).

3.2.3. Genome assembly. *De novo* assembly of Mitterberg 2-185 sample was performed with five assemblers: *canu* [Koren 2017], *flye* [Kolmogorov 2019], *minimap2/miniasm* [Li 2016], *HASLR* [Haghshenas 2020], and *SPAdes* [Prjibelski 2020]. *Canu* and *flye* assemblies were generated with default parameters and setting the approximate genome size as 120 Mb, as the reference genome size. *Minimap2/miniasm* was set by mapping reads to each other (`minimap2 ava-ont reads.fq reads.fq > mappings.paf`) and creating the graph in GFA format (`miniasm -f reads.gz mappings.paf > draft.gfa`). The graph was transformed into FASTA sequence with linux utility *AWK*. The same operation was done with *flye* assembly as it is also a graph-based tool. The *HASLR* and *SPAdes* -based assemblies were performed with default parameters.

The five assemblers were tested on one accession (Mitterberg 2-185), and after decision taking, the remaining genomes were assembled by two assemblers: *canu* and *flye*, independently. The draft assemblies, were then processed in the identical way. To increase the per-base accuracy of contigs, the drafts were polished by aligning reads to contigs with *minimap2* [Li 2018], and using mapping information to correct errors based on mapping coverage with one iteration of *racon* [Vaser 2017] applying the following parameters: score of match bases equal to 8 (`--match 8`), score for mismatching bases equal to -6 (`--mismatch -6`), threshold for average base quality set as 8 (`--quality-threshold 8`), and window length set as 500 (`--window-length 500`). The polished contigs were ordered into five chromosome-representing sequences (scaffolds) with *RagTag* [Alonge 2019], using the genome Col-CC.v1 as a template, as at that moment it was the most recent genome version. The chromosome locations with missing sequence were replaced with Ns of length 100. The final assemblies did not include the unplaced contigs.

3.2.4. Evaluation of the assembly quality. Basic assembly metrics were calculated using *QUAST* [Gurevich 2013], with parameters: large genome size (`--large`), eukaryote (`--eukaryote`), fragmented assembly (`--fragmented`). BUSCO Scores were calculated using *BUSCO* tool, with parameters of genome mode (`--mode genome`) for the newest lineage dataset of orthologous genes of Land Plants Embryophyta_odb10 (`--lineage embryophyta_odb10`) [Simão 2015].

The percentage of common k-mers between the datasets of short reads (the short read data downloaded from SRA database, IDs are provided in Table 5) and the assemblies was calculated with *Merqury* [Rhie 2020], by splitting the assembly sequence into a set of 18-mers and counting the frequency of k-mers present in both the assemblies and short read data sets, considering the read mapping coverage. The raw short reads were filtered with *fastp* [Chen 2018], with the default parameters and trimmed by one base from read ends, because it had the lowest quality (`fastp -i R1.fq.gz -I R2.fq.gz -o R1.filtered.fq.gz -O R2.filtered.fq.gz --trim_tail1=1`

`--trim_tail2=1`). This analysis was done for the draft assembly and for the final assembly, for each accession and each assembly.

Assembly quality index QV was calculated with *Inspector* [Chen 2021], by aligning long reads to assembled contigs using reference-free mode for nanopore data type (`--datatype nanopore`). The formula for QV calculation was: $E = 1 - (1 - \frac{Kasm}{Ktotal})^{1/k}$, $QV = -10 \log E$ in which K is equal to the number of k-mers in compared datasets, and E is equal to the probability that the base in a given position is an error, and presented as mean value across all the genome.

Structural correctness was evaluated by pairwise alignment of *canu* to *flye* assemblies and vice versa by *nucmer*/*Assemblytics* workflow. *Nucmer*, a package from *MUMmer* software, was used with the parameters of maximum match across the length, with minimum length of a single match for 100 and minimum match cluster for 500 (`--maxmatch -l 100 -c 500`) [Marçais 2018]. The outputs of *nucmer* calls in DELTA format were then processed by *Assemblytics* tool, with default parameters, with the report of sequence variants in range 100 bp to 10 kb [Nattestad 2016].

Telomere repeats were found using *Tandem Repeats Finder* (*TRF*) in command-line mode [Benson 1999]. The input for *TRF* was the set of 10-kb-sized sequences from both ends of chromosomes, extracted using `samtools faidx` [Li 2009]. *TRF* was run with default parameters, setting the maximum length of a tandem repeat to 20 (`Maxperiod` option). The output repeat patterns were then classified as canonical plant telomeres (for patterns $(TTTAGGG)_n$ or $(CCCTAAA)_n$, $n \geq 10$), derivative telomeric repeats or DTRs (the repeats of non-canonical length), or unknown. The dot plots of telomeric regions were generated using online version of *YASS* software [Noé 2005], by self-alignments of the 10 kb sequences.

Centromere repeats were found using command-line BLAST, by aligning the pattern of the canonical centromere sequence CEN180 against all the chromosome sequences. CEN180 is the consensus sequence of Arabidopsis centromere repeat monomer of length 178 bp 5'-AGTATAAGAACTTAAACCGCAACCCGATCTTAAAAGCCTAAGTAGTGTTTCCTTGTTAGAAGACACAAAGCCAAAGACTCATATGGACTTTGGCTACACCATGAAAGCTTTGAGAAGCAAGAAGAAGGTTGGTTAGTGTTTTGGAGTCGAATATGACTTGATGTCATGTGTATGATTG-3', which was annotated by Naish and colleagues [2021]. The resulting table was checked for lack of overlaps between the hits. The relative abundance of CEN180 repeats per window (10 kb) was calculated and visualized.

3.2.5. Annotations of protein-coding genes and TEs. Known genes were transferred from the reference genome to the assemblies using the homology-based tool *Liftoff* [Shumate 2021].

Two genomes including the reference and the assembly, and the reference annotation file (Araport11_GFF3_genes_transposons.Jan2021.gff) were used as an input for the analysis [Cheng 2017]. Other genes were predicted *ab initio* using *AUGUSTUS* tool, with the species parameter (`-s arabidopsis`) [Stanke 2006]. Gene models not overlapping with the *Liftoff*-mapped locations and TEs, longer than 800 bp that matched "protein-coding" description were additionally curated by homology-based search of protein sequences from the TAIR protein list (https://www.arabidopsis.org/download/file?path=Proteins/TAIR10_protein_lists/TAIR10_pep_20101214/), which were downloaded from TAIR resource, version from March 2023.

Annotation of TEs was done with *RepeatMasker* software, version 4.1.6 (https://www.repeatmasker.org/), with the parameters of RMblast search engine (`-e rmblast`), in sensitive mode (`-s`), specifying the species (`-species arabidopsis`). The output data sets were filtered for satellite DNA, simple repeats and low complexity records, and the remaining TEs were manually curated, to merge neighboring structural parts into longer and more complete TE models and to remove nested TEs, that probably appeared because misassembly.

3.2.6. Calling 5mC methylation from the nanopore data. 5mC modification marks were called from raw nanopore signals using *DeepSignal-plants* package [Ni 2019]. All the reads with the extracted 5mC marks in CG context were mapped to the respective assemblies and the average probability of 5mC was calculated per CG position in the assembly. The evaluation of 5mC level in TEs was estimated as an average 5mC probability across the lengths of individual TEs. The TEs containing at least 4 CG positions were considered for average calculation. Methylation levels for the selected regions were calculated as average 5mC probability across windows of lengths 260-300 bp.

3.2.7. Sequence comparisons of *RLP* genes. Prior to sequence analysis, each *RLP* gene was translated to protein sequence, by *AUGUSTUS* or *Emboss transeq* tool [Madeira 2024]. Global alignment for individual genes was performed using *Multalin* [Corpet 1988], using BLOSUM matrix. Domain conservation of RLP proteins was checked with *CD-Search* tool from NCBI (https://www.ncbi.nlm.nih.gov/). Phylogenetic trees for selected *RLPs* were made with *Clustal Omega* software [Madeira 2024], with Neighbor-Joining algorithm. Curated RLP proteins were retrieved from UniProt database (https://www.uniprot.org/) and TAIR (representative gene models for Araport11 annotation). To look for the expression of *RLP* genes in plant leaves, RNA-seq reads were mapped to the respective genome assemblies (the data for 6 out of 8 studied accessions were available) [Kawakatsu 2016] with STAR software [Dobin 2013], with default parameters. Visualizations of gene models in the genomes were made in *Integrative Genomics Viewer* [Robinson 2011]. The protein domain of TEs were found by extracting ORFs from their sequences

using *Emboss getorf* tool [Madeira 2024] and then scanning the found ORFs in Interpro database (https://www.ebi.ac.uk/interpro/). Auxiliary data sets to support the sequence comparisons are presented in Table 6.

Table 6. Public resources to support the analyses of *RLP* variation landscape.

Resource	Retrieved evidence	References
http://athcnv.ibch.poznan.pl/	Copy numbers of individual *RLP* genes based on short-read mappings	[Zmienko 2020]
Project ID GSE80744	RNA-seq of leaves for MIT, DOL, OY0, TAD, RSH and AIT	[Kawakatsu 2016]
https://www.araport.org/	Retrieved localization of *RLP* genes in the reference genome	[Pasha 2020]
Article	Pairwise sequence comparison of *RLPs* in the reference genome	[Wang 2008]
Article	Phylogenetic relationships between *RLP* genes within the family	[Steidele 2021]
Project ID PRJNA1033522	SV search in long read-based *de novo* assemblies of 69 accessions	[Lian 2024]
Project ID SRP056687	Short genome sequencing reads for eight accessions from Table 5.	[1001 Genomes Consortium 2016]

3.2.8. Data availability. The final scaffolds, Liftoff-based annotations, AUGUSTUS-based gene predictions, and TE annotations, generated in this work, can be downloaded under the following link: https://box.pionier.net.pl/d/da609400c6664732a99e/ for each accession, for both *canu* and *flye* assembly versions.

4. RESULTS

4.1. Nanopore sequencing and assembling de novo genomes of eight Arabidopsis accessions

4.1.1. A simple protocol of HMW DNA extraction for whole-genome long-read sequencing from plant leaves. The studied accessions were selected from the global collection of 1001 Genomes Project and originated from geographically different regions including Italy (Mitterberg 2-185), Bulgaria (Dolna 1-40), USA (BRR57), Norway (Oy-0), north Sweden (TÅD 06), Russia (Rsch-4), Spain (IP-Pun-0), and Morocco (Aitba-1). The accessions represented different genetic groups such as *central_europe* (Mitterberg 2-185), *italy_balcan_cucasus* (Dolna 1-40), *USA-germany* (BRR57 and Rsch-4), *relict* (IP-Pun-0 and Aitba-1), *north-sweden* (TÅD 06) and one admixed accession (Oy-0) (Figure 8). For the moment of selection and sequencing, no genome assembly of any of the accessions was publicly available, which was one of my selection criteria.

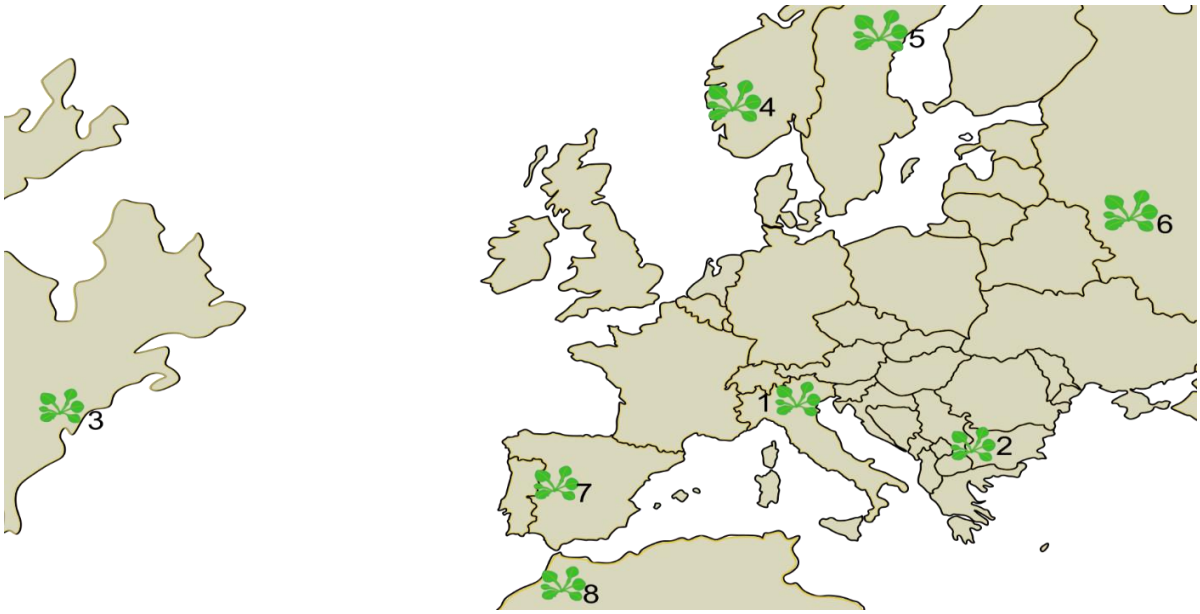


Figure 8. The geographic representation of Arabidopsis accessions selected for the study. 1 – Mitterberg 2-185, 2 – Dolna 1-40, 3 – BRR57, 4 – Oy-0, 5 - TÅD 06, 6 – Rsch-4, 7 – IP-Pun-0, 8 – Aitba-1.

I optimized the relatively easy and cost-effective protocol of HMW DNA extraction for nanopore sequencing from Arabidopsis leaves. The protocol consists of three steps: cell lysis with CTAB, DNA extraction using gravity-flow, anion-exchange columns (Genomic tip by QIAGEN), and subsequent DNA purification by precipitation. The utilization of the gravity columns was the

key element which minimized mechanical shearing compared to spin columns, allowing efficient purification of unsheared DNA (Figure 9A). The protocol omits the laborious nuclei isolation step and does not require any harmful reagents such as chloroform or SDS (Sodium Dodecyl Sulfate). Additionally, it has been based on commercially available solutions, to ensure its reproducibility and easy handling of samples. The stepwise DNA extraction protocol is attached as Appendix 1. The main issues of successful DNA extraction include conducting all the steps in one day, using wide-bore tips, and avoiding extensive pipetting the DNA-containing solution. Special attention should be paid to the optimization of the amount of cell lysate that is loaded on the gravity columns, as its overloading blocks elution. In a recently published protocol, the authors had similar observations and recommended the reduction of grinding time and centrifugation speed [Kang 2023a].

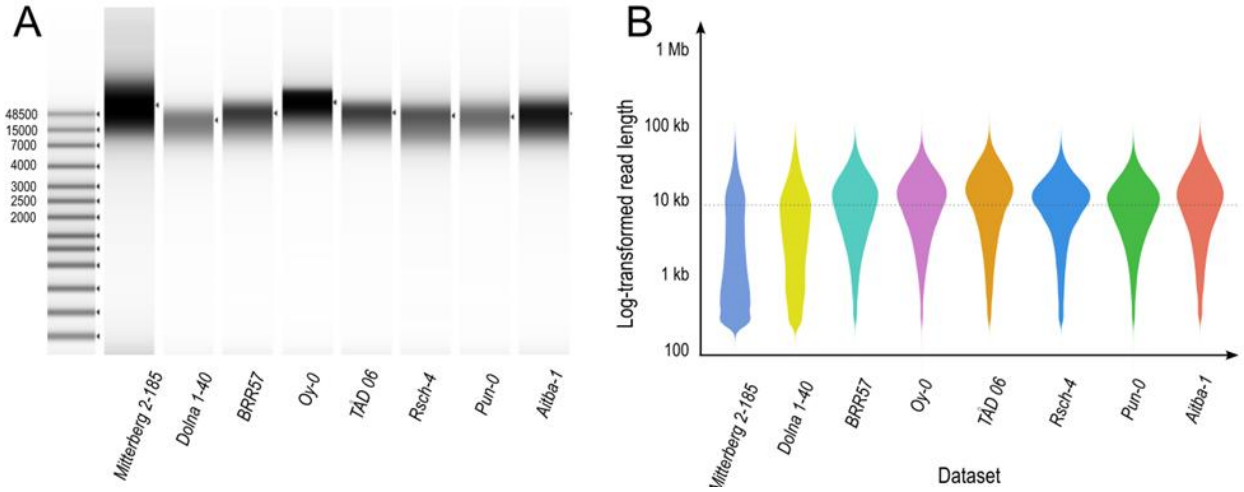


Figure 9. The length of HMW DNA and corresponding distribution of read lengths. A – the distribution of HMW DNA fragment lengths by high-resolution electrophoresis (Tape Station system 4200). B – the log-transformed distribution of read length of whole-genome sequencing datasets for corresponding samples (MinION sequencer; R9.4 flow cell).

The total yield obtained from 2 g of plant material was up to 2 µg of HMW DNA. All the samples passed the purity filters of both A260/280 and A260/230 absorbance ratios (measured by NanoDrop 2000). The average DNA fragment size prevalently ranged between 15 and 50 kb, as presented by high-resolution electrophoresis (Figure 9A). The distribution of DNA fragment lengths corresponded to the distribution of the number of long sequencing reads (Figure 9B). The substantial number of short reads (< 1 kb in length), observed at Figure 9B, can be explained by the fact that no size selection step was performed during DNA extraction. A trial of size selection was made using Short Read Eliminator Kit by Circulomics (data not shown). However, the application of size selection reduced the total DNA yield more than twice and for that reason it has been decided to pass the total DNA for sequencing. It is also worth noting that the datasets are

presented in chronological order of their preparation, which presents that simply by getting more skilled in manipulating the samples, I was able to substantially reduce the DNA fragmentation.

600 ng -1.2 µg DNA was used for preparation of sequencing libraries. To ensure as large data yield as possible, all DNA amount from the sample was used for the preparation of sequencing libraries, and one sequencing flow cell was used per one sample. Before starting the run, the flow cell was scanned for the number of active nanopores (at least 1500 / 2048 were required to start sequence run). After 6-8 hours of sequencing, the second library was loaded to ensure high yield, however the yield was variable across the runs. We explained that by the fact that nanopores have the tendency to clog up by long DNA molecules and unequal average length of DNA molecules. The sequencing data yield and quality is summarized in Table 7.

Table 7. The summary statistics of sequencing data yield and quality.

Accession	Active channels (out of 512)	Gbases sequenced	Number of reads (million)	Read length N50 (kb)	Longest read (kb)	Median read quality Q	Cumulative length of reads > Q7 (Gb)	Rate of > Q7 read data (%)
MIT	511	16.4	3.55	12.6	282.5	13.4	15.5	92.1
DOL	505	9.5	1.41	14	237.9	12.2	8.9	90.8
BRR	507	22.8	2.22	16	274.9	11.6	21.2	90.2
OY0	498	14.6	1.31	16.7	430.9	12.8	13.9	92.8
TAD	508	14.3	1.15	19.8	208.3	11.9	13.5	90.6
RSH	509	12.9	1.30	14.1	552.1	11.6	11.9	89.0
PUN	494	11.4	1.21	13.6	158.4	12.6	10.9	92.6
AIT	485	13.9	1.19	18.4	311.1	12.7	13.2	92.2

The amount of data obtained from a single sequencing flow cell was 14.5 Gb in average, with the highest of nearly 23 Gb for BRR. The data yield values were different, dependent on many factors, such as the stability of nanopores throughout run duration, length and number of DNA fragments that occupied the space on the flow cell. The resulting reads were around 12 kb in length, on average, with the highest Read N50 values in TAD (almost 20 kb) and AIT (more than 18 kb). In each data set, reads of length of 150 kb and longer were present (the longest read exceeded 500 kb), there were several such long reads per accession. The average read quality of each sequencing data set was ~12. Around 90% of reads passed the Q7 quality filter, which was applied prior to further analysis. Interestingly, there was no correlation between the read length

and quality, highlighting the fact that the error rate in the nanopore technology is independent of read length, which distinguishes it from the NGS approach, where cutting the reads to a certain length increases the overall dataset quality (Figure 10). The small subsamples of low-quality reads (<7 quality) were present in all the data sets, which were later filtered out from the analysis. The fraction of long reads (>10 kb) increased in subsequent runs as I mastered the library preparation steps. Consequently, long reads constituted the majority in most sequencing outputs (Figure 9B).

For the moment of data generation, only one report, describing the sequencing data yield from the flow cell of the same nanopore chemistry (R9.4) for Arabidopsis genome, was available [Michael 2018]. The author's data set was smaller (3.4. Gb), and the quality of data was lower (7.3 in average), while the data length distribution was comparable to our data. More recent whole-genome nanopore sequencing of Arabidopsis accession Col-0 was conducted using the large nanopore sequencer (PromethION) [Wang 2022]. The authors generated 54.5 Gb of reads and selected only the ones longer than 50 kb for the *de novo* assembly. The final length of the data set could cover the reference genome ~177 times, showing high data yield. The genome coverage by the long reads which I generated was at least 90× for all accessions except for DOL (74×), indicating the overall good performance of sequencing on a single MinION flow cell. I assumed that it will be sufficient for the purpose of generating high-quality contigs suitable for SV analysis. In conclusion, the long-read datasets of high amount and quality were generated and could be now utilized for the next step, which was *de novo* assembly.

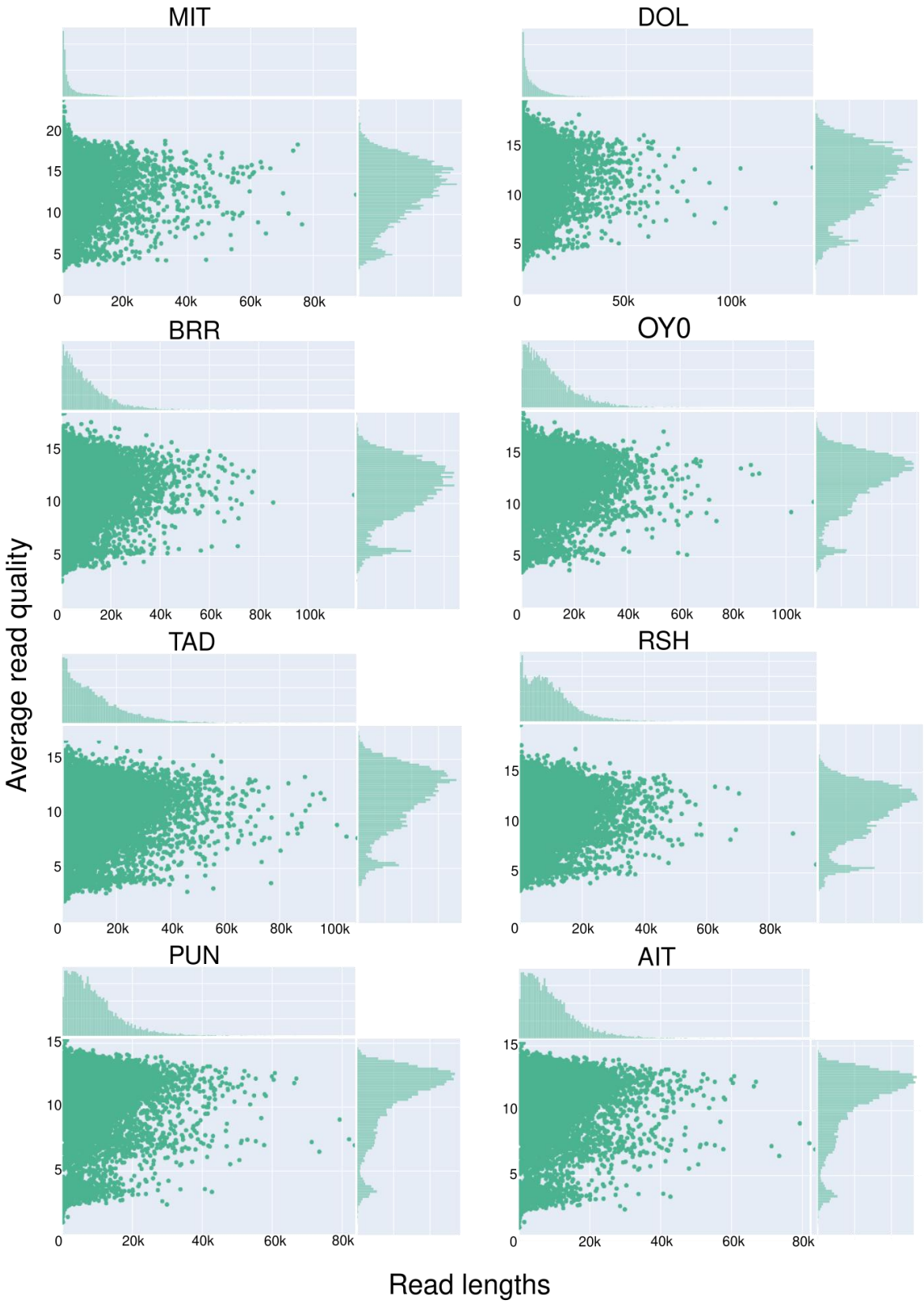


Figure 10. The scatterplot distribution of read length (X axis) and quality (Y axis). The histogram on top of each sample is the distribution of read quality and the histogram on the right is the distribution of read lengths.

4.1.2. Setting up and testing the *de novo* assembly workflow. No universal workflow for long-read sequencing and *de novo* assembly for plant genomes has been proposed, as the workflow strongly depends on the amount and type of data. The so-far best Arabidopsis assembly projects relied on a combination of sequencing data from two sources and either utilized long reads complemented by short reads [Jiao 2020] or combined long reads and linkage information [Zapata 2016, Michael 2018, Rabanal 2022]. I hypothesized that a generation of large enough data set (genome nearly 100x covered by long reads) only from the nanopore technology, will be sufficient to reconstruct near-complete chromosome-representing sequences of high quality to be able to identify and analyze SVs of regions of interest.

The first step of analysis was basecalling, or transformation of data from FAST5 (raw signal data) to FASTQ format (sequence reads). It was performed with *guppy* software, maintained by ONT (https://nanoporetech.com/software/other/guppy), version 6, which was the most recent version at the time of data generation. The generated reads underwent filtering by length and quality: the reads of length <1000 and quality <Q7 were filtered out from the analysis. The *de novo* assembly approach was tested using five different software tools using MIT dataset: *canu*, *flye*, *miniasm*, *HASLR* and *SPAdes*. *Canu* was the first assembler developed for long reads and had remained among the most widely used tools since then [Koren 2017]. The algorithm applied in *canu* is termed Overlap-Layout-Consensus (OLC) and consists of three steps: correction of read accuracy by pairwise mapping, trimming reads by the best overlap ratios, and assembly of the final chromosome sequence. *Flye*, based on de Brujin graph composition, promised the reduction of required computational resources and improvement of assembling repetitive sequence regions by the generation of disjointigs – the non-overlapping contigs – and then constructing a continuous sequence based on graph topology [Kolmogorov 2019]. The remaining assemblers were selected based on different approaches of inbuilt algorithms, and mentioning in other studies related to our scientific objectives [Li 2016, Haghshenas 2020, Prjibelski 2020]. The comparison of the results obtained with all assemblers is presented at Table 8.

Table 8. The assembly metrics for MIT assembly generated by different algorithms.

Software	All contigs[1]	>50 kb contigs	Assembly length	Longest contig	N50[2]	L50[3]	Genome (%)[4]
Canu	156	115	130 232 960	16 063 971	11 382 933	5	91.2
Flye	131	28	120 232 311	16 010 083	13 964 568	5	90.2
Miniasm	331	143	132 540 496	13 118 726	5 536 874	8	72.5
HASLR	880	478	100 925 504	1 331 800	223 475	117	81.0
SPAdes	2148	630	99 950 405	580 400	63 706	472	83.3

[1]All contigs is the number of contigs longer than 10 kb.
[2]N50 is the sequence length of the shortest contig at 50% of the total assembly length.
[3]L50 is the number of contigs which cover at least 50% of the assembly length
[4]Genome fraction (%) is the percentage of bases in the assembly aligned to the reference genome.

Using the parameters describing the assembly contiguity, I evaluated the performance of each assembler. *Canu* and *flye* outperformed other assemblers, as they generated lowest numbers of contigs (156 and 133 respectively), and these assemblies had highest N50 and lowest L50 values, indicating their high contiguity. The total assembly lengths (130 Mb and 120 Mb) were comparable with the length of the reference genome. *Miniasm*-generated assembly was of similar length (132 Mb) but was more fragmented and the resulting contigs did not align well to the reference genome, marking possible assembly errors. *HASLR* and *SPAdes* produced recognizably high numbers of contigs (880 for *HASLR* and more than 2000 for *SPAdes*), with the lowest N50 values, indicating that these assemblies were highly fragmented. Based on this evaluation, I selected *canu* and *flye* for further analyses. The differences in contigs lengths of different MIT assemblies are shown in Figure 11 (MIT).

Two assemblies, *canu*-based and *flye*-based, were prepared in parallel from the same dataset for each accession, by creating 16 draft assemblies represented by sets of contigs. To improve per-base sequence accuracy, draft assemblies underwent polishing step, by mapping raw reads to contigs and correcting errors based on read coverage, and generation a consensus sequence set. The polishing was performed with *racon* tool, one iteration [Vaser 2017]. I also tested two and three iterations of polishing, but the generated consensus sequence had increased number of contigs (data not shown), indicating that the following iteration reached saturation in improving the accuracy but broke the existing contiguity, which was expected as I operated on the same source of data. Polished contigs were ordered into five chromosome-representing sequences called scaffolds, using *RagTag* software [Alonge 2019], using Col-CC.v1 genome as a template for ordering the contigs.

4.1.3. Evaluation of the assembly quality. To ensure that the assembled sequence has few gaps, is sufficiently complete and accurate, the evaluation of each assembly's quality was performed by estimating and analyzing the parameters that claim about its contiguity, correctness, and completeness. The contiguity was calculated by the metrics describing the score of contig lengths in accordance with the reference genome: NG50 was estimated as the length of the shortest contig at 50% of length of the reference genome, LG50 was the smallest number of contigs, the cumulative length of which covers 50% of the reference genome length. For both metrics smaller values indicated higher contiguity. The completeness of the final scaffolds was checked based on their coding and non-coding information. To see the level of completeness in conserved regions, BUSCO Score was calculated, of which the highest score indicates the best completeness. To look at the non-coding regions, the content of the regions at the ends of chromosomal arms was examined: the amount and the content of repetitive sequences supposed to be present in Arabidopsis centromeric and telomeric regions were checked. Assembly correctness was assessed by estimation of QV metric that describes average accuracy of the assembled sequence and was obtained by comparison of sets of k-mers between the raw read datasets and the assembly. In addition, I took advantage of the presence of two assemblies per accession, and I compared them to each other as well to the reference genome counting the number and the length of discordant locations (assembly errors). The detailed evaluation of *canu*- and *flye*-based assemblies is presented in the next sections.

4.1.3.1. Assembly evaluation – contiguity. The assembly metrics are presented in Table 9. The letter 'c' or 'f' following assembly ID indicates the software used for generation of the draft assemblies – *canu* or *flye*, respectively. The number of contigs in the draft assemblies varied greatly, from 68 in BRRf up to 324 in DOLf, while the lengths of the longest contig in each assembly were more similar and exceeded 15 Mb, except for DOLc, OY0c, TADc, and RSHc where these contigs were shorter (11.1 Mb - 14 Mb). When the final *canu* and *flye* scaffolds were compared, drastic differences could be observed for some accessions, and they were associated with the differences in final assembly size. More specifically, only three scaffold sets (DOLc, OY0c, and RSHc) for which contigs number was highest, were also smaller in size then the reference genome. This indicated the high fragmentation of the final assemblies, additionally confirmed by distinguishingly lower NG50 values and increased LG50. Remarkably, these scaffolds were all derived from the *canu* assembler. Altogether, *flye* slightly outperformed canu in most cases. For example, when comparing the two assemblers, the assembly length was typically larger in *flye*, and the differences between them varied from almost 3 up to 11 Mb, which is a large portion of sequence. The lower contiguity of *canu* assemblies is also notable in Figure 11.

Table 9. The summary of assembly parameters that describe assembly contiguity.

Assembly	Coverage[1]	Draft assembly		Final assemblies (scaffolds)				
		Number of contigs	Length of longest contig (Mb)	Contigs included	Size (Mb)[2]	NG50	LG50	Total length of unplaced contigs (Mb)
MITc	129×	156	**16.1**	37	**122.1**	11.4	5	8.1
MITf		**131**	16.0	45	119.5	**14.0**	5	**0.7**
DOLc	74×	**195**	12.2	157	115.6	7.8	9	**2.5**
DOLf		324	**15.5**	32	**127.0**	**10.4**	**5**	5.8
BRRc	176×	95	16.2	34	120.1	**14.2**	5	4.1
BRRf		**68**	**16.6**	35	**124.2**	14.0	5	**2.9**
OY0c	115×	169	13.0	113	116.5	9.4	6	**3.6**
OY0f		**145**	**16.5**	18	**125.4**	**15.0**	**5**	6.9
TADc	112×	**101**	14.0	58	119.0	11.3	5	3.9
TADf		155	**16.7**	29	**125.5**	**12.9**	5	**3.4**
RSHc	99×	227	11.1	185	116.4	6.2	7	**2.1**
RSHf		198	**16.5**	17	**123.4**	**13.8**	**5**	9.5
PUNc	90×	**157**	16.5	42	122.8	11.7	5	5.9
PUNf		227	**17.0**	31	**127.4**	**14.4**	5	**5.7**
AITc	110×	127	16.7	44	123.3	11.5	5	7.1
AITf		**116**	**17.7**	24	**126.1**	**14.0**	5	**5.7**

[1]Coverage is the reference genome (TAIR10) coverage by raw reads.
[2]Size is the cumulative length of contigs included in the scaffolds.
Better score for a ‘c’ or ‘f’ parameter is highlighted in bold.

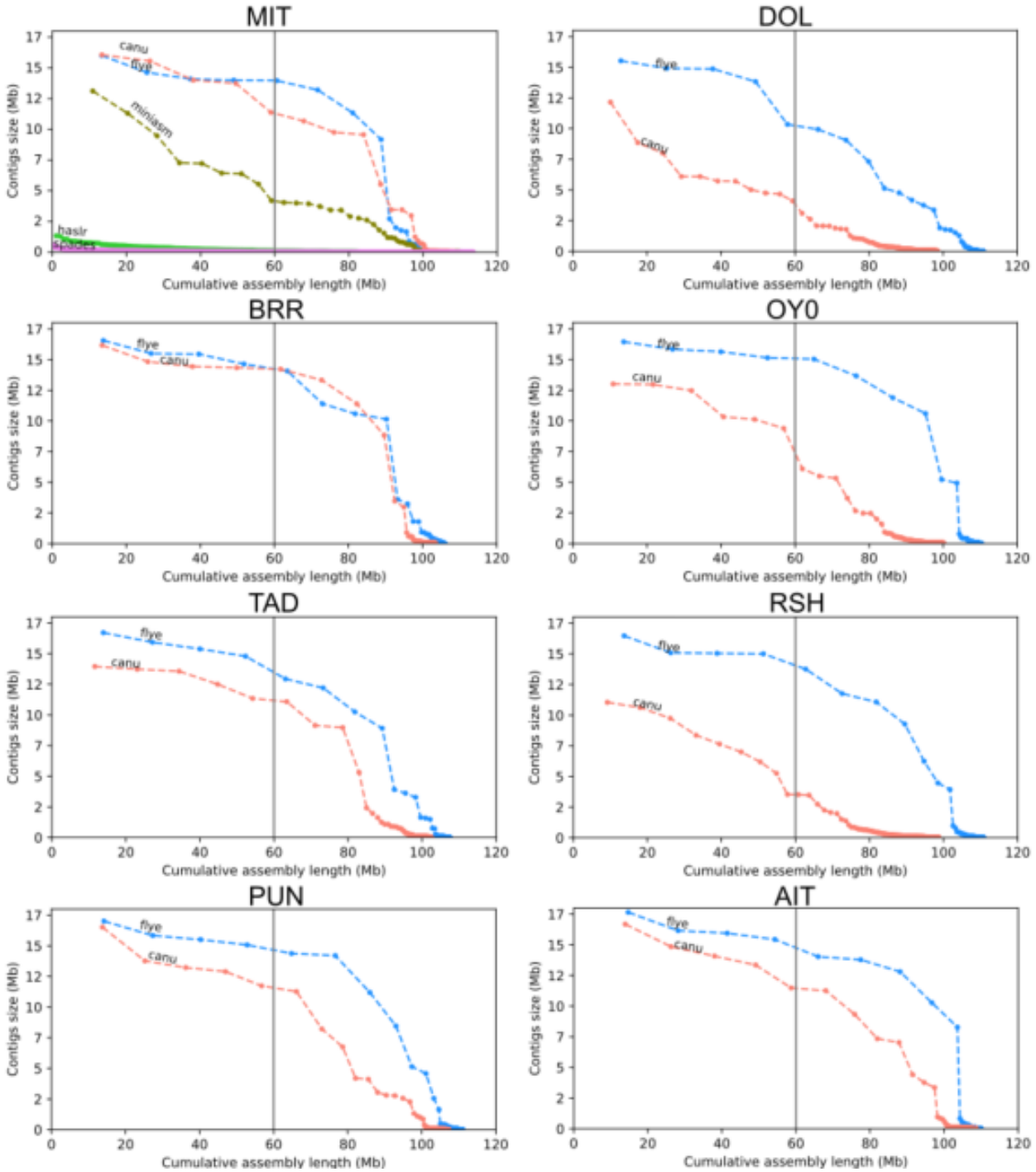


Figure 11. The comparison of contiguity by different assemblers. MIT was assembled using 5 different tools. Other samples were assembled by the two selected tools – *canu* (red) and *flye* (blue). The dots illustrate the size of contigs, beginning from the longest. NG50 marks the 50% length of Arabidopsis reference genome.

4.1.3.2. Assembly evaluation – completeness. The evaluation of completeness is critical to the assemblies generated by nanopore-only sequencing method. The level of completeness of the assembled chromosomes was checked by examination of their content. In the regions enriched in genes, the gene space completeness was estimated with calculation of BUSCO Score [Simão

2015]. This metric investigates the presence of the conserved genes, predefined by the set of known orthologous genes for the specific lineage. The estimation of BUSCO score assumes that evolutionary conserved genes are present in the investigated genome and exist in single copies. Accordingly, if many BUSCOs are missing from the genome, it can reflect the failure of the sequence assembly. Complete (C) BUSCO score includes the percentage of both Single-Copy (S) and Duplicated (D) genes found in the assembly out of total lineage gene set appropriate for analyzed species (n) and does not include Fragmented (F) alignments. The Missing (M) orthologs are present only in the lineage dataset. Here, the conserved gene set for land plants (Embryophyta), consisting of a total 1614 genes, was used for BUSCO Score calculations. The results of BUSCO Score are presented in Table 10.

The BUSCO completeness score was in the range 98.8% for *canu* assemblies and 98.9% for *flye* assemblies. The score was identical or almost identical between the two assemblies of one accession. There were no more than 10 Fragmented BUSCOs for most samples, except for RSHc, RSHf (F = 13 for both), and AITc (F = 17). The number of missing BUSCOs varied from 7 to 11 and was also similar for the two assemblies of one sample. According to BUSCO Score evaluation, the conserved gene set of the genome was assembled correctly and almost completely for each sample independently of the assembly method and the level of assembly contiguity, indicating that the fragmentation might affect noncoding, most likely repetitive regions to more extent than coding ones.

The repeat content metric was estimated using *RepeatMasker*, as the percentage of masked bases in the genome, encompassing all types of repeats such as simple, tandem, interspersed repeats, and TEs (Table 10). For comparison, the same analysis was performed on the reference genomes (20% repeat content) and its improved version Col-CC.v2 genome (over 30%). The results showed that our assemblies were more similar to TAIR10 in their repetitive regions. On average, repeat content was approximately 17.4% in *Canu* assemblies and 21% in *Flye* assemblies. Notably, except for the MIT sample, *Flye* consistently produced higher repeat content values, suggesting that the *Canu* algorithm may perform less effectively in repeat-rich regions. The higher repeat content observed in *Flye* assemblies also corresponds with their larger total assembly sizes (Table 9). Overall, these findings indicate that the main differences between *canu* and *flye* assemblies are likely located in the repetitive regions of the genome.

The gene-independent completeness calculation was made by comparison of sets of k-mers between high-accuracy short reads to the assembled sequences and calculating the percentage of common k-mers (a parameter called k-mer completeness) using the tool named *Merqury* [Rhie 2020]. To this aim, short read data of Illumina paired-ended libraries for studied accessions were

downloaded from SRA repository (see section 3.1.1). Each assembly sequence was then divided into 18-mers (k-mer size was determined automatically by *Merqury*) and the presence and the number of the same k-mers were evaluated for Illumina reads. The number of occurrences of a given k-mer in Illumina reads corresponded to genome coverage by short reads. The percentage of common k-mers between short reads and contigs in the draft assemblies, and, separately, for short reads and the final scaffolds was checked (Table 10). The k-mer completeness values were very high in all samples, within 96.7% - 99.2% range for contigs and 97.5% - 99% for scaffolds. This value was important evidence that sequence assembly accuracy was over 96% and, in some datasets, it was even higher than 99% (comparable with the basic accuracy of short reads). It is worth mentioning that this value reflects the average value for the entire genome, encompassing not only coding regions (for which the expected value is very high by BUSCO analysis), but also the repetitive regions. As the assembly of the repetitive regions is complex, it is now clear that nanopore-only data might be insufficient for the correct assembly of these regions. Then, large fraction of the repetitive regions, which had been assembled partially, lowers the k-mer completeness value.

Table 10. The summary of assembly parameters that describe assembly completeness.

Assembly	BUSCO Score[1] (%) C[S+D]	Single Copy BUSCOs[1] (S)	Duplicated BUSCOs (D)	Frag-mented BUSCOs (F)	Missing BUSCOs (M)	K-mer completeness[2] for draft assemblies (%)	K-mer completeness[2] for Scaffolds (%)	Repeat content (%)	Total size of telomere arrays (Kb)	Total size of centromere arrays (Mb)
MITc	99.1	1583	16	5	10	98.5	98.4	19.6	21.7	1.447
MITf	99.3	1587	15	4	8	98.1	97.9	18.1	13.7	0.423
DOLc	99.0	1588	10	9	7	96.6	96.7	15.2	23.7	0.005
DOLf	99.1	1591	9	7	7	98.6	98.4	22.4	19.8	6.084
BRRc	98.8	1582	12	10	10	98.7	98.6	17.7	30.6	0.389
BRRf	98.9	1584	12	9	9	99.0	98.8	20.9	27.0	4.101
OY0c	99.1	1584	15	7	8	97.5	97.5	15.9	28.3	0.098
OY0f	99.0	1582	16	8	8	99.2	99.0	21.5	25.2	3.686
TADc	99.0	1583	15	5	11	96.7	97.2	17.1	29.0	0.256
TADf	98.8	1580	14	9	11	97.6	98.0	21.7	16.1	4.601
RSHc	98.6	1580	11	13	10	97.5	97.1	15.6	34.5	0.081
RSHf	98.7	1582	11	13	8	98.9	98.7	20.4	18.2	2.627
PUNc	98.8	1582	12	9	11	98.1	98.6	18.9	25.0	0.736
PUNf	98.9	1584	12	9	9	98.6	98.8	22.2	24.6	5.420
AITc	98.3	1578	9	17	10	97.6	97.6	19.1	27.0	1.813
AITf	98.8	1584	11	10	9	98.4	98.4	21.5	10.7	4.608

[1]BUSCO Score was calculated based on *Embryophyta* lineage set of total 1614 genes.
[2]K-mer completeness was evaluated against the whole-genomic short read data set for the same accession (1001 Genomes Project).

4.1.3.2.1. Assembly completeness of telomeric regions. Correctly assembled eukaryotic chromosomes are expected to contain telomeric repeats on their ends. The putative telomeric regions were checked for presence of these repeats by extracting 10-kb sequence from both ends of the assembled chromosomes, the chromosome start sequence to represent putative p-arm telomeric region, and the chromosome end sequence to represent putative q-arm telomeric region, respectively. The chromosomes of Col-CC.v2 were checked the same way, for reference. The length of the putative telomeric region was selected to cover the expected telomere size which ranges from 2 to 9 kb among different Arabidopsis accessions [Shakirov 2004]. The identification of repeats was made using *TRF* software [Benson 1999]. They were next classified based on literature as canonical, the most abundant short repeats found in plants, or DTRs. The lengths of identified telomeric repeat arrays are presented in Table 11.

Table 11. Lengths of telomeric repeat arrays in Arabidopsis assemblies given in kb.

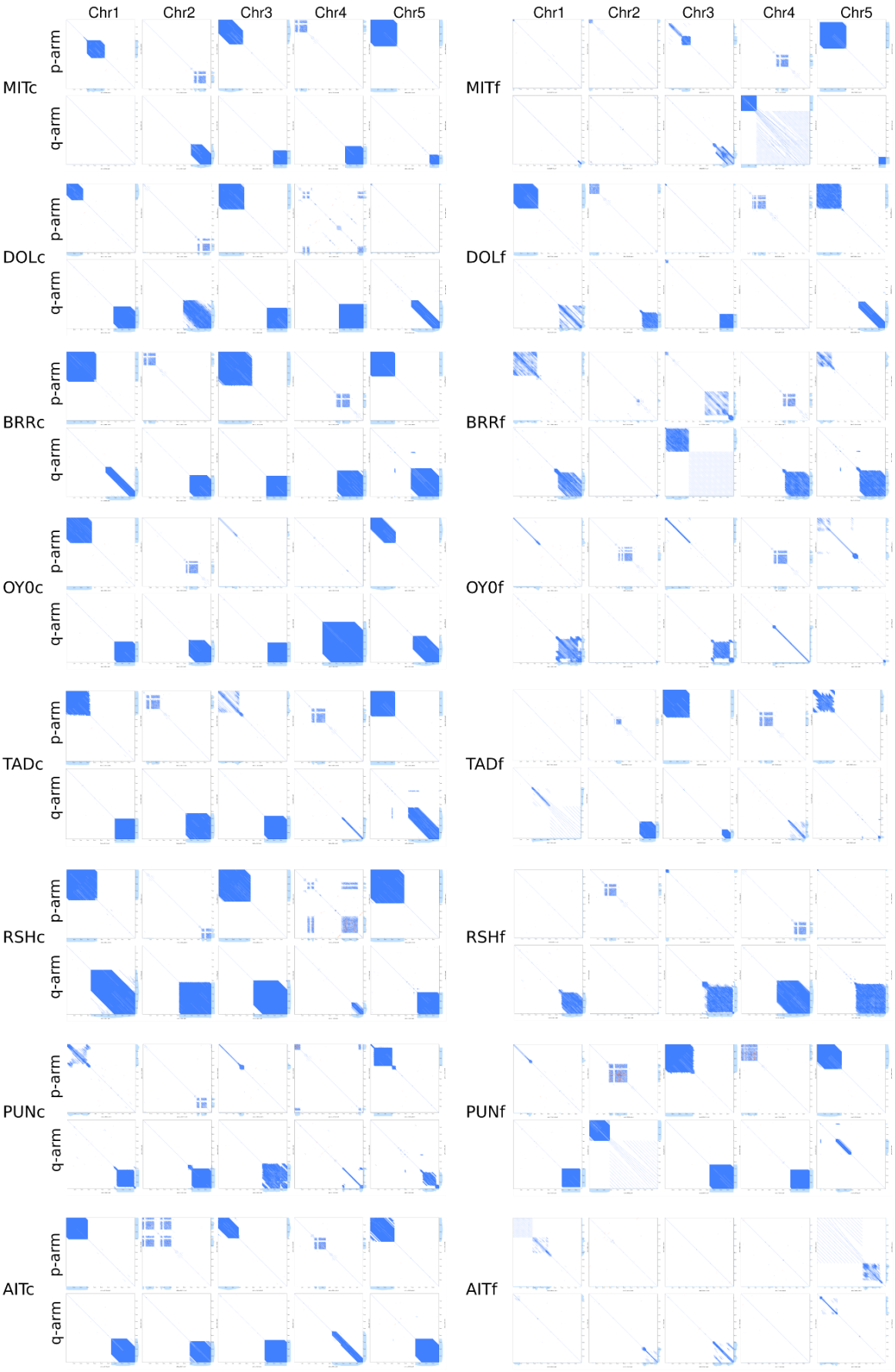
Region	Col-CC.v2	*Soft*	MIT	DOL	BRR	OY0	PUN	RSH	TAD	AIT
chr1p	**3.8**	*canu*	**2.3**	**2.2**	**4.2**	**3.4**	3.1	**4.4**	**3.3**	**3.1**
		flye	0.2	**3.6**	4	3.7	2.7	0	0	0.4
chr1q	**3.4**	*canu*	0	**3.1**	**4.3**	**2.9**	3.1	**6.4**	**2.8**	**3.4**
		flye	0.52	3.4	4	3.7	**2.7**	**3.4**	0.26	0
chr2p	0	*canu*	0	0	0	0	0	0	0	0
		flye	0	0	0	0	0	0	0	0
chr2q	**3.6**	*canu*	**2.9**	**4.1**	**3.1**	**3.2**	3.3	**4.7**	**3.6**	**2.9**
		flye	0	**2.6**	0.09	0.23	**2.9**	0.07	**2.3**	1.9
chr3p	**3.1**	*canu*	**3.4**	**3.6**	**4.7**	2.7	**3.7**	**4.5**	3.4	**2.8**
		flye	**3**	0.15	3.6	3.6	4.2	0.36	**3.9**	0
chr3q	**3.3**	*canu*	**2**	**2.9**	**3**	**2.8**	**3.6**	**4.9**	**3.3**	**3.2**
		flye	**3**	2	3.5	3.2	**3.5**	4.6	1.2	3
chr4p	0	*canu*	0	0	0	0	0	0	0	0
		flye	0	0	0	0	0	0	0	0
chr4q	**3.6**	*canu*	**2.6**	**3.5**	**3.7**	**5.9**	**3**	1.7	**3**	**4.5**
		flye	**2.2**	0	3.9	5.3	**2.7**	**4.8**	0.3	0
chr5p	**3.3**	*canu*	**3.7**	0.16	**3.5**	3.5	3.6	**4.8**	**3.5**	**3.4**
		flye	**3.7**	**3.5**	2.5	5.7	**3.5**	0.12	**3.1**	0.34
chr5q	**4**	*canu*	1.4	**4**	**4.1**	**3.8**	2.7	**3.1**	**4.4**	**3.4**
		flye	1	**3.8**	4.7	0.35	**2.2**	5	0.3	0.34

*High confidence telomeric regions are in bold. Arrays composed of DTRs are grey.
Repeat arrays larger than 2 kb with the predominance of canonical telomeric repeats were considered as high confidence regions.

Chromosome ends with high content of canonical repeat arrays spread across at least 2 kb were considered high-confidence telomeric regions, while the ones with the shorter repeat arrays or represented mostly by DTRs were marked as low-confidence telomeric regions. This is because originating from canonical repeats by base insertions or deletions and subsequent repeat expansions, DTRs are associated with the maintenance of telomeres in the genome and naturally accumulate in subtelomeric regions in the space between the telomere sequence arrays and the gene-rich part of the chromosome [Tao 2024]. However, in the study in which telomeric sequences in seven natural plant populations were scanned, no chromosome ends composed only of DTRs were found [Belyayev 2023]. Telomeric repeats include homopolymers, the sequencing of which is known to impose a technical limitation for nanopore sequencing. Therefore, lack of canonical telomeric repeats in some chromosome ends must be considered as assembly insufficiency.

No telomere arrays were found in p-arms of Chromosomes 2 and 4, which was however an expected result due to the presence of NORs adjacently to telomeres and high structural diversity of these chromosome arm ends. It was mentioned earlier, assembling these regions proved to be the main challenge in the multiple challenge in many efforts to obtain the complete T2T Arabidopsis genome. Due to the presence of other structurally important types of repeats in this location, these chromosome arms are known for their complex composition, which was not possible to discover by trivial search of tandem repeats. However, the canonical telomeric repeats of sufficient lengths were identified in numerous chromosome ends of the analyzed scaffolds (Table 11). The representation of telomeric repeats in the assemblies is visualized at dot plots, generated by self-aligning of 10-kb regions of p- and q- arms by *YASS* software (Figure 12).

Figure 12. The size and the structure of telomeric regions in the assembled genomes. Blocks of repeats represent the telomeric sequence, the density of the blocks reflects the level of conservation of monomers. The thickness of the squares in the plots represents the telomeric region size that was assembled. Double blocks mean that different types of repeats were found. Rectangular blocks display elongated size of monomers. Irregular shapes of blocks depict that the repeats within the arrays were not tandem.


Chr1
Chr2
Chr3
Chr4
Chr5
MITc
MITf
DOLc
DOLf
BRRc
BRRf
OY0c
OY0f
TADc
TADf
RSHc
RSHf
PUNc
PUNf
AITc
AITf
p-arm
q-arm

Most canonical repeats were found in *canu* assemblies, while in *flye* assemblies the repeats were strongly underrepresented, and their repeat patterns were often different from the canonical ones. Generally, in *canu* assemblies, the telomeric repeats were larger and represented in more chromosomes than in *flye* assemblies. A rule-like observation was that the repeat was never observed in *flye* when it was absent from *canu*, indicating poor capability of the *flye* algorithm to assemble sequence ends. The repeats found in p-arms of Chromosome 2 and Chromosome 4 of any assembly were not telomere-like. Apart from that, in *canu*, the repeat-like structure can be seen in all the plots except q-arm of Chr1 (MITc) and p-arm of Chr5 (DOLc). In *flye*, the repeat-like structure is missing in more than half of the data (Chr1p, Chr1q and Chr2q of MITf; Chr3p and Chr4q for DOLf; Chr2q for BRRf; Chr2q and Chr5q for OY0f; Chr1p and Chr5q for TADf; Chr1p, Chr2q, Chr3p, Chr5p for RSHf; Chr1q, Chr3p and Chr4q from AITf). In PUN, telomeric sequence representation was more consistent between *canu* and *flye*, than in the other accessions. The fewest number of repeats was found in AITf.

Besides canonical 7-bp repeat patterns, 12 types of non-canonical repeats were found in the assemblies. The most abundant were two 8-based units: TTTGAGGG found at q-arms of Chr2 exceptionally, and TAGGGTA found in different locations, and the remaining types occurred only once or twice. All the varieties were found to localize adjacent to the canonical repeat arrays and their repeat number was lower than the canonical repeat number within the same telomeric region. In conclusion, the assembly sets display the representation of telomere variation, however the characterization of telomeric regions based on this data is limited. Nevertheless, the presence of telomeric repeats at chromosome ends suggested successful assembling and scaffolding of the chromosome arms in the analyzed genomes.

4.1.3.2.2. Assembly completeness of centromeric regions. Being non-coding genomic elements, mostly composed by repeats, centromeres represent integral parts of chromosomes. As opposed to telomeres, centromeric repeat pattern is longer (178 bp in Arabidopsis) than telomeric unit (7 bp). Search for centromeric arrays was aimed to evaluate the performance of *canu* and *flye* in assembling repetitive sequences within chromosomes. The main building block of Arabidopsis centromeres is CEN180, a 178-bp-size repeat including variants within its sequence [Camacho 2009]. The consensus sequence for CEN180 was previously constructed based on the *de novo* assembly of all five centromeres in Col-0 accession with high resolution (Col-CEN genome version) [Naish 2021], therefore in this study the identification of centromeres in the assemblies was performed by BLAST-alignment of CEN180 consensus sequence against all scaffolds. Density of CEN180 repeats was then evaluated in 10-kb windows to reveal the presence/absence of centromeric region on each chromosome and its location.

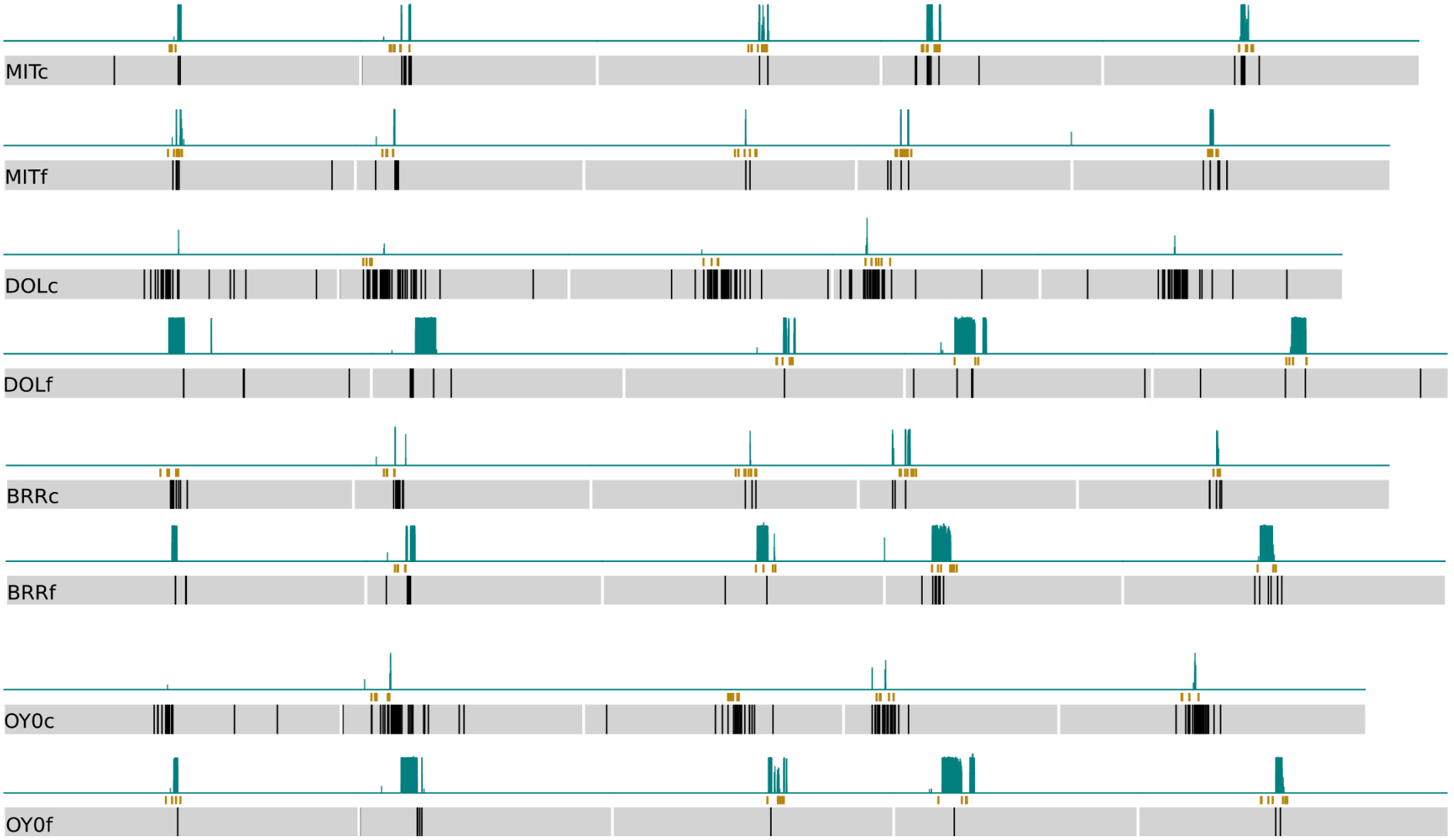
Regions of high centromeric repeats density found in the scaffolds are presented in Figure 13. To clearly observe the association of centromere completeness with the assembly contiguity, they are aligned along with contigs present in scaffolds. Additionally, the locations of ATHILA transposons, that are known to represent more than 90% of non-repetitive content of centromeres, are marked [Wlodzimierz 2023]. Centromeric regions could be detected based on the presence of either CEN180 repeats or ATHILA elements. In most cases, both types of elements could be detected, including all chromosomes of MITc, MITf and PUNc assemblies (Table 12). The width of the CEN180 repeat-rich region varied and several smaller peaks of repeats could be often visible close to each other, indcating different levels of centromeric region completeness. The relative positions of the centromeres on the chromosomes were concordant with the ones in the respective chromosomes of Col-CC.v2. However, the exact size of these regions was not taken into account due to the high level of variation in the centromere size among the accessions, despite their conservative function – a phenomenon called the centromere paradox [Wlodzimierz 2023]. In few cases, CEN180 repeats were also detected outside putative centromeres. They were found in the sites of short contig insertions into the scaffolds, mostly at chromosome ends (MITf-Chr4, BRRf-Chr4, TADf-Chr3, AITf-Chr4), indicating errors introduced during the scaffolding step.

Table 12. Presence and co-occurrence of CEN180 repeats (C) and ATHILA elements (A) in putative centromeres of scaffolded chromosomes.

Tool / Sample	*canu*			*flye*		
	C + A	Only C	Only A	C + A	Only C	Only A
MIT	5			5		
DOL	3	2		3	2	
BRR	4		1	4	1	
OY0	3	1	1	4	1	
TAD	4	1		3	2	
RSH	4		1	3	2	
PUN	5			3	2	
AIT	4	1		4	1	

In general, higher scaffold contiguity resulted in better distinction of the centromeric regions. However, it can be noted that the representation of centromeric regions in the assemblies was different between *canu*-based and *flye*-based assemblies, indicating the weak reliability of assembling repetitive sequences based on the data solely from ONT technology, with a relatively small fraction of ultra-long reads. In the case of short-size telomeric repeat, *canu* reported more results of canonical pattern, and the size of repeat-containing regions was closer to standard size of Arabidopsis telomeres (2 – 9 kb), while in *flye* these repeats were often missing. In the case of

longer centromeric repeats, the observation is the opposite: centromeric repeats were represented in fine level of completeness in *flye*, while in *canu* CEN180 repeats were less abundant and not closely clustered, likely due to substantially higher fragmentation of the *canu* assemblies around centromeric regions. On the other hand, *flye*-based assemblies lacked ATHILA annotation at least in one centromere of each scaffold, except for MITf, while the complete absence of ATHILA was rare in *canu*-based scaffolds. Overall, even though the approach of *de novo* assembly with long reads becomes increasingly efficient, the choice of the algorithm has strong impact on the sequence resolution in repeat-rich regions of the chromosome.



(continuation on the next page)

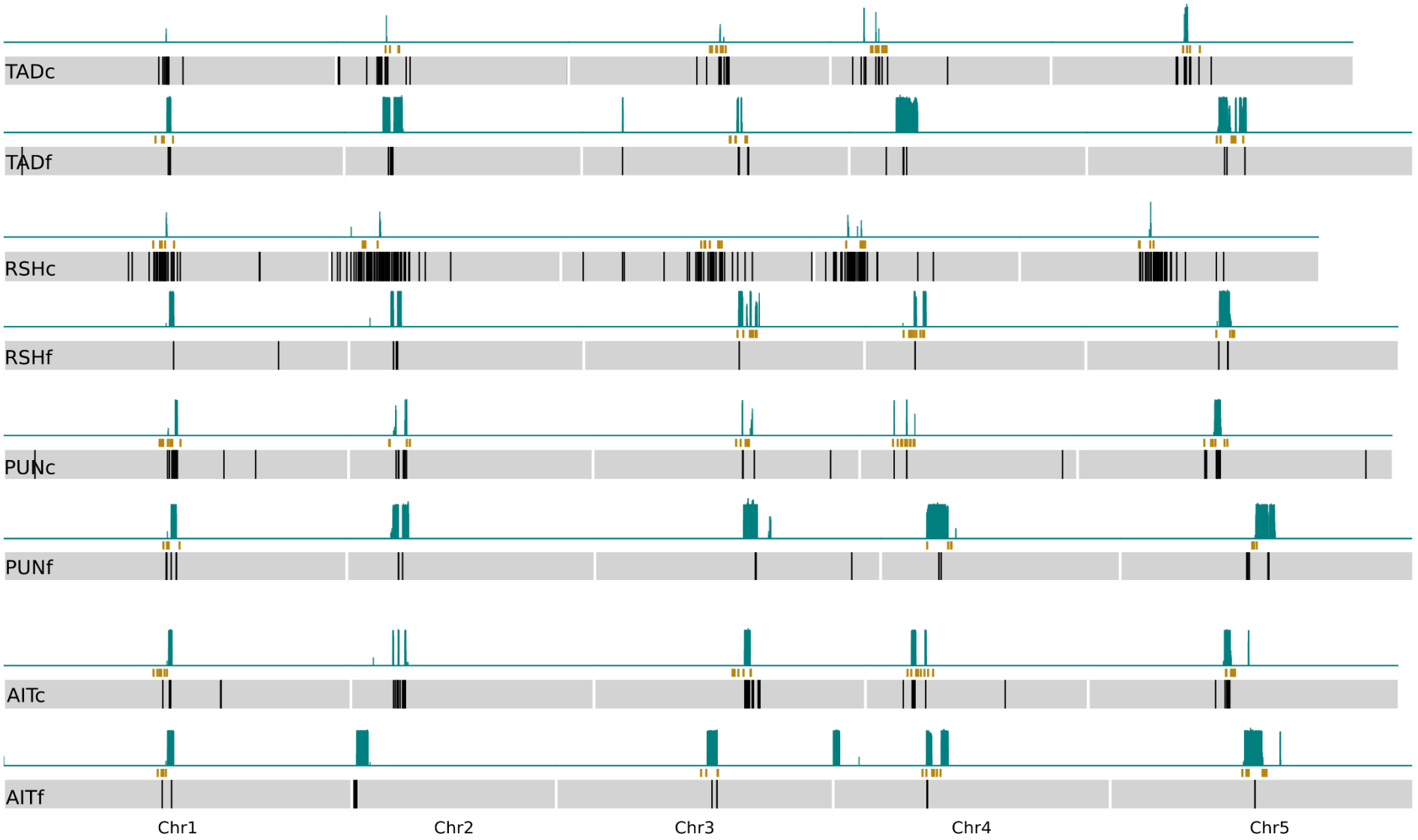


Figure 13. Centromere completeness and assembly fragmentation. The graph illustrates fraction of CEN180 repeats across chromosome lengths (green bars). The blocks below ilustrate the contigs within the scaffolds colored in grey, with the assembly breaks (colored in black). Yellow bars present the locations of ATHILA transposons.

4.1.3.3. Assembly evaluation – correctness. Overall correctness was estimated by two parameters: Assembly Quality (QV) and Error Rate (E) using *Inspector* [Chen 2021]. To this aim, unassembled long reads were mapped to polished contigs and to scaffolds, and the numbers of total and common k-mers were calculated. The average values for all the sequences in the data set are summarized in Table 13. Error Rate equals the probability of error of a given base in a given position, so the lowest E means the highest assembly accuracy. QV is logarithmically related to E. It indicates the probability that a given base in the sequence is correct. When QV = 30, the probability of correct sequence is 99.9%. Average QV was calculated for each contig set and each scaffold. In most cases, QV score was higher in scaffolds than in contigs and it never dropped which suggests that some misassembled contigs were unplaced and excluded from the scaffolds. Regardless of the assembler, the datasets with the lowest QV cores were TAD and PUN (QV < 30 for both contigs and scaffolds).

To check for assembly errors, the two assemblies for each accession were used as references against each other. *Canu* assemblies were aligned to *flye* and vice versa using *nucmer*, and variants >100 bp were called using *Assemblytics*. The results were manually curated in the following way: variants located 100 bp next to each other were joined and the resulting variants were categorized into Small (100-499 bp), Medium (500-999 bp), and Large (1 kb and longer). Due to the lack of accession-specific reference, all sequences discordant between *canu* and *flye* were marked as errors. Fewest total number of errors was observed in AIT and BRR (22-27 per assembly). These assemblies as well as MIT had smallest total length of discordant sequence (18 – 55 Kb per scaffold). On the contrary, RSH, DOL and OY0 assemblies displayed highest number of Large-size errors (30-98 per scaffold) and the largest size of discordant sequence (152 – 507 Kb per scaffold), which was likely caused by huge differences in contiguity of the *canu* and *flye* scaffolds observed for these accessions (see Table 9). Importantly, most errors reported by *nucmer* were associated with the regions identified as centromeric repeats, which was consistent with earlier observations, when the quality of the assemblies was lower in repetitive regions.

Table 13. The summary of assembly parameters that describe assembly correctness.

Assembly	QV for Contigs	Error Rate for Contigs	QV for Scaffolds	Error Rate for Scaffolds	Number of Small-Size Errors[1]	Number of Medium-Size Errors[2]	Number of Large-Size Errors[3]	Total Number of Errors	Discordant Sequence Length (Kb)
MITc	34.0	3.98×10^{-4}	34.8	3.31×10^{-4}	32	2	3	37	18
MITf	33.6	4.37×10^{-4}	35.1	3.16×10^{-4}	45	8	8	61	55
DOLc	31.3	7.41×10^{-4}	34.2	3.80×10^{-4}	58	10	75	143	400
DOLf	34.2	3.80×10^{-4}	34.2	3.80×10^{-4}	47	10	71	128	386
BRRc	32.7	5.37×10^{-4}	32.8	5.30×10^{-4}	13	1	9	23	54
BRRf	33.1	4.89×10^{-4}	33.1	4.89×10^{-4}	15	2	10	27	55
OY0c	30.5	8.91×10^{-4}	33.5	4.47×10^{-4}	40	10	38	88	179
OY0f	33.6	4.37×10^{-4}	33.7	4.27×10^{-4}	43	9	30	82	152
TADc	27.5	1.78×10^{-3}	29.1	1.23×10^{-3}	36	2	11	49	72
TADf	28.5	1.41×10^{-3}	29.4	1.14×10^{-3}	33	2	15	50	86
RSHc	32.8	5.30×10^{-4}	33.0	5.01×10^{-4}	80	14	98	192	507
RSHf	33.1	4.89×10^{-4}	33.1	4.89×10^{-4}	76	14	92	182	489
PUNc	28.2	1.58×10^{-3}	29.5	1.12×10^{-3}	34	3	18	55	111
PUNf	29.3	1.17×10^{-3}	29.9	1.05×10^{-3}	34	5	17	56	101
AITc	30.4	9.12×10^{-4}	30.5	8.91×10^{-4}	17	2	3	22	23
AITf	33.6	4.37×10^{-4}	33.6	4.37×10^{-4}	16	2	5	23	29

[1]Errors of length 100-499 bp were classified as Small.
[2]Errors of length 500 bp – 999 bp were classified as Medium.
[3]Errors of length 1 – 10 kb were classified as Large.

4.1.4. A comparison to alternative assembly: Oy-0 genome case. Another long-read assembly of Oy-0 (hereafter Oy-0_Lian) accession has been recently published in a pan-genomic study of 69 Arabidopsis natural accessions [Lian 2024]. I used this assembly for the evaluation of OY0c and OY0f assemblies based on an independent dataset (in addition to k-mer evaluation with short-read data, described above). In the mentioned work, the authors also used nanopore sequencing data to obtain Oy-0_Lian assembly: the read coverage in their data set was lower than in our work (~42x compared to 115x), but the read lengths were substantially higher (Read N50 = 22.5 Kb compared to 16.7 Kb, respectively). The final assembly was constructed by merging the draft assemblies generated with two assemblers (*SMARTdenovo* and *flye*) and polished with the use of short reads. The final scaffolds set was constructed of 50 contigs, while in this study it was 113 contigs for OY0c and 18 contigs for OY0f. The main assembly statistics of Oy-0_Lian were the following: Contig N50 = 15.3 Mb, BUSCO Score = 99.92%, QV = 46.85, genes lifted from Araport11 = 27020. While most metrics were similar as in OY0f and OY0c, the per-base sequence accuracy of the Oy-0_Lian assembly was significantly higher, as indicated by QV score. Two factors could contribute to this effect, using longer nanopore reads in the input data by Lian and colleagues as well as their using extra data source (short reads) for the polishing step.

The whole genome comparison of three assemblies of Oy-0 accession against the reference assembly Col-CC.v2 showed that none of the assemblies correctly spanned repeat-including centromeric regions. Nevertheless, the sizes of these regions on Chromosome 1 and Chromosome 5 of Oy-0_Lian were closer to Col-CC.v2 than these of OY0c and OY0f assemblies (Figure 14). The rDNA-rich regions on Chromosome 2 and Chromosome 4 were missing from all three non-reference assemblies. Pairwise comparison between Oy-0_Lian and OY0c / OY0f assemblies revealed 142 Kb / 44 Kb of discordant sequence, respectively. However, whole-genome alignments indicate that the detected variants were mainly located within poorly assembled repeat-rich regions, such as centromeres. Apart from that, there was very good consistency between the three Oy-0 assemblies. In particular, two inversions at q-arm of Chromosome 3, and one inversion at Chromosome 5 was revealed in each of them, when compared to Col-CC.v2, which strongly support the conclusion that they present real SV events.

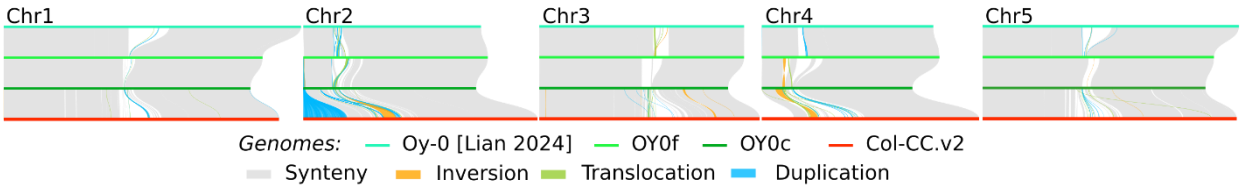


Figure 14. Whole-genome comparison of three assemblies of Oy-0 accession and the genome Col-CC.v2 (red).

Lian and colleagues merged the two assemblies with the tool called quickmerge ((https://github.com/mahulchak/quickmerge). This approach has been tested by me as well. However, for our data sets, merging led to the increased number of contigs and the drop of the BUSCO Scores. Since the purpose of my study was to focus on finely assembled gene-rich regions that were localized within the chromosome arms, I decided to keep two independent assemblies rather then combine them into a third one, which would remain incomplete (Figure 14). According to our evaluation, the differences between *canu* and *flye* assemblies in gene-rich regions are small. Moreover, comparing these two assemblies may be further exploited to mark the problematic regions requiring further sequence assessment. In the following analyses I used both assembly versions.

4.2. Genome Variation of *RLP* gene family in eight Arabidopsis accessions

4.2.1. Gene and TE annotations. The first step towards the analysis of SV within the selected gene family was to prepare genome annotations. To this end, known genes were transferred from Araport 11 into assemblies, using mapping-based tool *Liftoff* [Shumate 2021]. Of the 27 562 protein-coding genes present in the reference annotation [Cheng 2017], more than 27 thousand were found in all the studied accessions (Table 14). *Liftoff* aligns genes from the reference genome to the target genome and finds the mapping that maximizes sequence identity preserving the reference structure of exons and transcripts. Due to natural variation, the 100% identity of amino acid sequence was not required. Of all reannotated genes, almost 100% were found in the same chromosome as the reference, and more than 80% had at least one regular ORF. The gene names were transferred onto the coordinates of individual genomes, which provided much convenience in data handling while resolving the variation in selected genomic regions. However, *Liftoff* was not capable to find the highly variable genes, duplicated genes neither genes not present in the reference genome.

The *Liftoff*-based annotations were complemented by whole-genome *ab initio* gene prediction with *AUGUSTUS* software, using gene models that were trained on Arabidopsis data [Stanke 2006]. The protein sequences of the resulting models, that did not intersect with already reannotated genes neither with TEs, were aligned to TAIR10 protein sequences using BLASTP. The gene models with the unique BLASTP match were hypothetical proteins not annotated in the reference genome. The five most abundant functional descriptions of the genes predicted by *AUGUSTUS* were “Disease resistance protein (TIR-NBS-LRR class) family”; “F-box and

associated interaction domains-containing protein"; "Cysteine/Histidine-rich C1 domain family protein", "Protein kinase superfamily protein", and "RING/U-box superfamily protein" – the gene families characterized by intense variation and dynamic evolution. In summary, 27.1 thousand reannotated genes were complemented by ~300 predicted genes per accession, which together comprised the number close to the number of genes in the reference annotation (27 562 genes). The gene number in accordance with high BUSCO evaluation results (from 98.3% to 99.3% levels of completeness) indicated good annotation quality.

Table 14. The protein-coding genes annotated in the assemblies.

Assembly	Reannotated genes			Predicted models	Total number of genes
	All genes	- in the same Chr	- at least 1 valid ORF		
MITc	27 152	99.6%	88.1%	278	27 430
MITf	27 138	99.6%	89.7%	248	27 386
DOLc	27 160	99.7%	84.4%	269	27 429
DOLf	27 159	99.7%	84.9%	272	27 431
BRRc	27 163	99.7%	81.9%	294	27 457
BRRf	27 141	99.7%	82.2%	264	27 405
OY0c	27 147	99.7%	85.2%	269	27 416
OY0f	27 169	99.7%	85.7%	268	27 437
TADc	27 152	99.6%	82.9%	272	27 424
TADf	27 149	99.6%	83.6%	267	27 416
RSHc	27 182	99.7%	81.3%	261	27 443
RSHf	27 176	99.7%	82.6%	260	27 436
PUNc	27 085	99.5%	77.0%	342	27 427
PUNf	27 084	99.5%	83.6%	327	27 411
AITc	27 119	99.6%	75.4%	324	27 443
AITf	27 113	97.8%	83.6%	303	27 416

Transposon annotation in the assembled genomes was done by homology-based method, using the existing libraries of Arabidopsis TEs inbuilt in *RepeatMasker* software. In general, the representatives of all the main DNA TE and Retrotrapsposon classes were found in all the assemblies, encompassing 11-12 Mb of the assembly length (Figure 15). While the total Repeat Content occupied ~20% of the assembly size (see Table 9), which is equal to 24 Mb, half of that length seems to be represented by TEs. When compared *canu* and *flye* assemblies, the TE content

was almost identical between the two, except for MIT, BRR, and PUN. The main group which contributed to the differences in length was Tandem Inverted Repeats (TIR) – an order of DNA transposons known for its long length of the elements (7 kb on average). The remaining TE lengths were in accordance with their distribution in reference accession [Quesneville 2020].

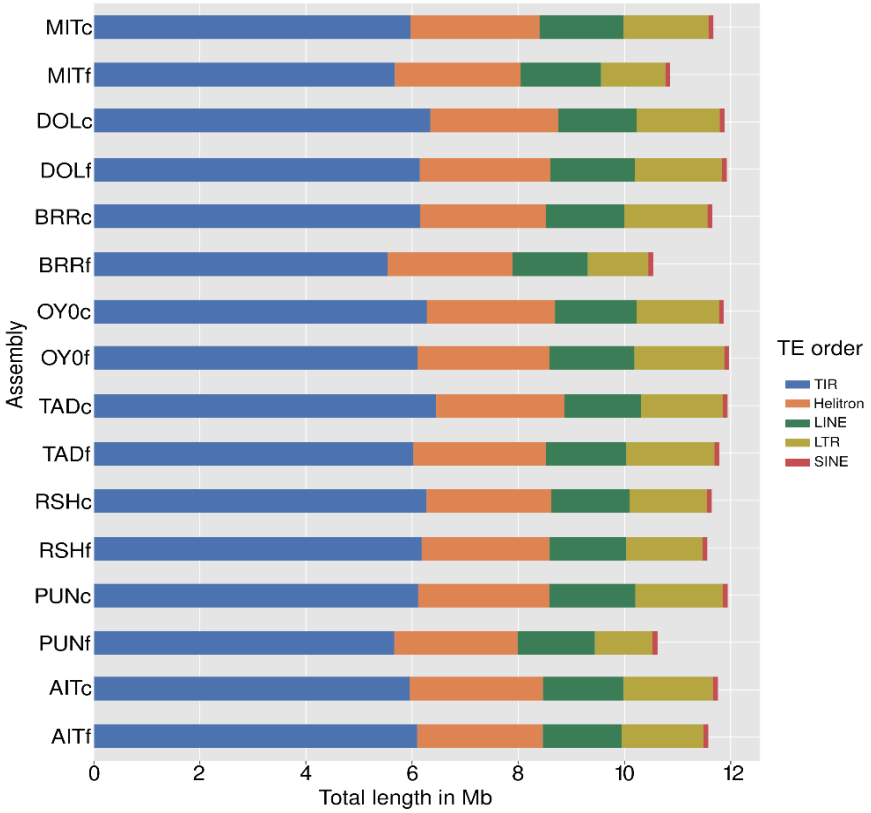


Figure 15. Main TE orders and their size annotated in the assembled genomes. The represented Retrotransposons are long terminal repeats (LTR), long interspersed nuclear elements (LINE), and short interspersed nuclear elements (SINE). The DNA transposons are represented by terminal inverted repeats (TIR) and Helitrons.

In addition, I took advantage of the possibility to extract the epigenetic mark (5mC) from nanopore data, as 5mC can be called directly from "squiggle" signals without any additional experiments. It was demonstrated that nanopore-based 5mC detection is highly correlated with the Bisulfite sequencing (corr = 0.89), which is a gold standard method for 5mC detection [Doshi 2025]. As dense DNA methylation of a genetic element is considered to be a mark of its stable silencing, I expected high levels of 5mC among TEs. The methylation levels were estimated on average per TE in CG context and are presented in Figure 16. As expected, most TEs had high 5mC levels, indicating stable repression of their activity by epigenetic mechamisms. The low methylation levels were observed in single TEs, in all groups, suggesting that some of them can be active.

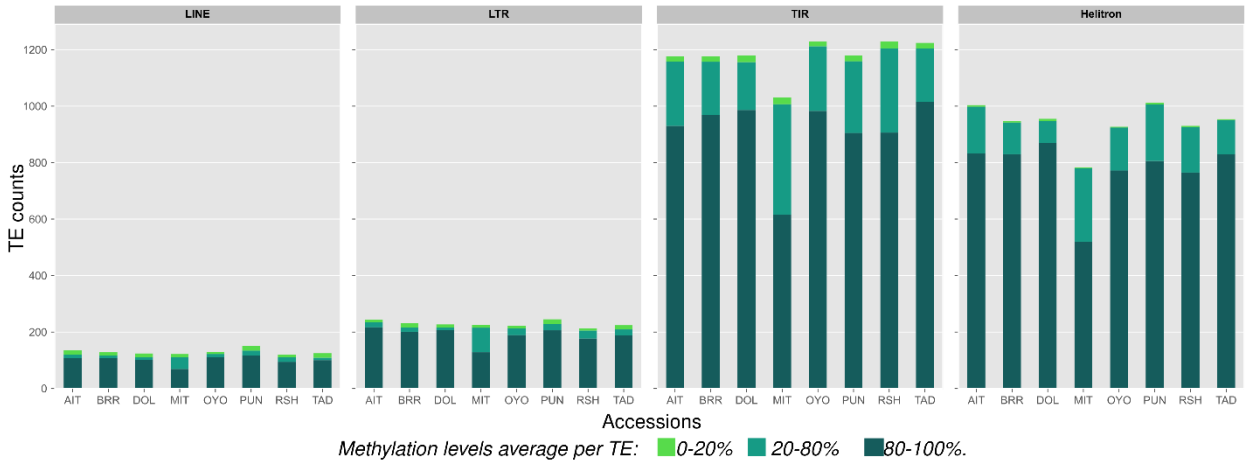


Figure 16. Methylation levels of TEs in Arabidopsis accessions. The data is presented for *flye* assemblies.

4.2.2. SV within *RLP* gene family revealed by short-read data analysis. The 57 *RLP* genes in the Arabidopsis genome were identified by Wang and colleagues [2008]. Later, *RLP8* (locus *AT1G54480*) has been merged with *RPP27* (locus *AT1G54470*), while *RLP18* and *RLP49* were categorized to pseudogenes (https://www.arabidopsis.org/), resulting in 55 *RLPs* in the current annotation. They are distributed across all five chromosomes (Figure 17). Steidele and Stam (2021) analyzed the publicly available data and suggested that *RLP* genes clustered together in the genome were more likely to be associated with the defense responses, while the non-clustered ones were more often involved in plant development. The authors analyzed the distribution of SNPs and CNVs among these two groups (clustered and non-clustered *RLPs*). No difference between the two groups was found in SNP number, while the overlap of CNV regions was significantly higher in RLP clusters. Moreover, according to Wang (2008) even clustered *RLPs* share limited sequence homology, with the highest similarity observed for genes within *RLP39-RLP42* cluster. As a follow-up to their analysis, I identified RLP clusters by grouping *RLP* genes located at <50 kb distance from each other. Based on this criterion, 12 compact-size clusters (6 to 55 kb) were assigned, each cosisting of 2 to 5 *RLPs* and up to 7 non-RLP protein-coding genes. In half of these clusters, the *RLP* genes localize adjacently to each other and are encoded at the same strand (*RLP2-RLP3*, *RLP31-RLP32*, *RLP32-RLP33*, *RLP37-RLP38* at minus strand, while *RLP11-RLP12* and *RLP13-RLP16* at plus strand). In the remaining gene clusters that included intervening genes, the orientation of genes was often the same (*RLP20-RLP21*, *RLP22-RLP23*, *RLP34-RLP35* and *RLP39-RLP42*). Interestingly, no "head-to-head" orientation of the genes in relation to each other was ever observed in *RLPs*, which was earlier shown as a common gene orientation for another immune gene family (*NLRs*) (see section 1.1.3) [Van de Weyer 2019].

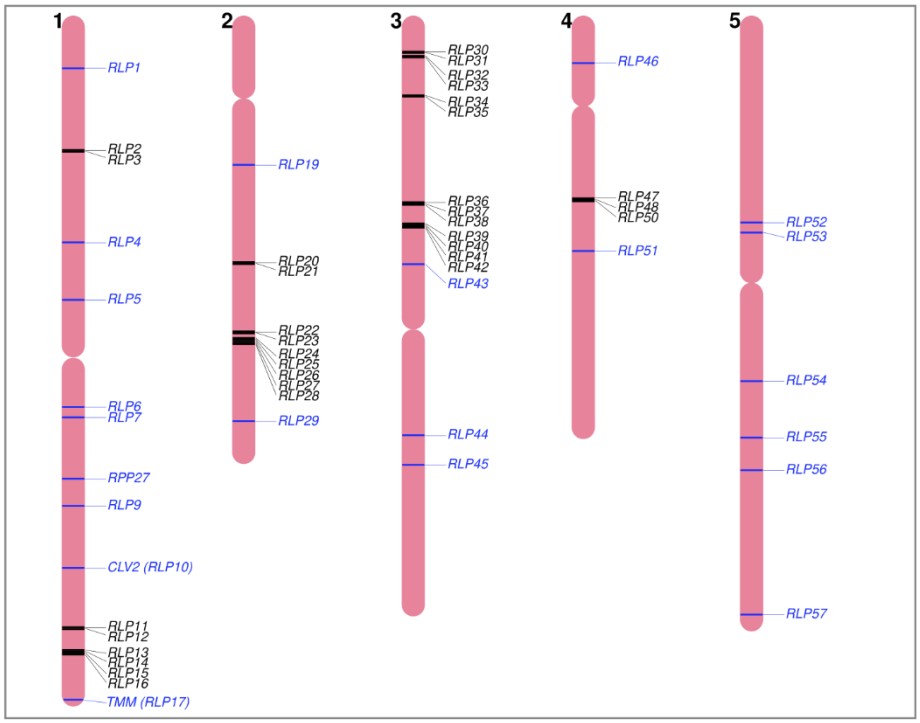


Figure 17. The distribution of *RLP* genes across Arabidopsis chromosomes. The *RLP* clusters are depicted in black, and the non-clustered genes are blue.

Next, the data from the AthCNV catalogue was scanned, to evaluate *RLP* copy number variation in large population set (1060 accessions) [Zmienko 2020]. To this aim, copy number distribution was retrieved for the 55 *RLP* genes (Figure 18). To analyse this data, I set the following paramteres: as the copy number around 2 was normal as for diploid genome, the increase for this value indicated gene duplication (≥ 3.2), while the decreased value claimed for gene deletion (≤ 0.8), respectively. According to these criteria, the non-clustered *RLPs* displayed little variability, with the lack of copy number changes in half of the genes, rare duplication events for genes *RLP7*, *CLV2*, *TMM*, *RLP52*, *RLP55* and *RLP57*, or rare deletion events for genes *RPP27*, *RLP43*, *RLP53* (Figure 18A). *RLP19* displayed classic PAV pattern, with the copy number split between 2 and 0 within population. Both increased and decreased copy numbers in single accessions were observed for *RLP5* and *RLP44*. On the contrary, the clustered RLPs show much more variation (Figure 18B). The three gene pairs (*RLP2-RLP3*, *RLP11-RLP12*, *RLP30-RLP31*) displayed numerous copy number changes, that can represent variation patterns, like PAV in *RLP11*, or a duplication of *RLP30* in a subset of accessions. The other three (*RLP20-RLP21*, *RLP22-RLP23*, and *RLP32-RLP33*) were rather nonvariable in CNV context. All the four large clusters displayed high variability. Many genes did not have clear distributions of copy number, which suggests variation events other than CNV. Based on AthCNV data, I expected to observe variation in *RLP* genes in

the analyzed accessions. Obviously, this gene family might serve as a good example to evaluate the utility of *de novo* assemblies for the analysis of the genomic variation.

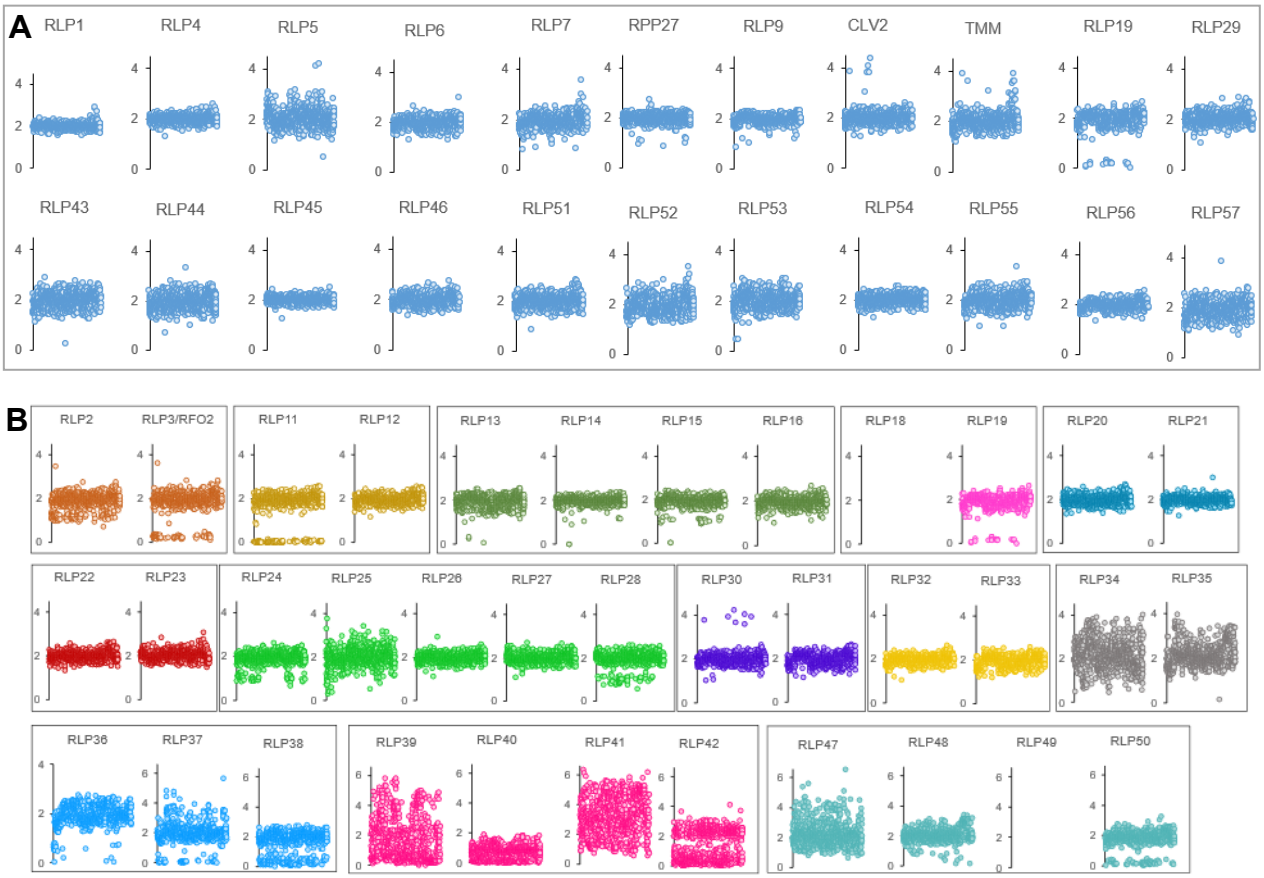


Figure 18. Copy number status of *RLP* genes in 1060 Arabidopsis accessions. A – copy numbers of non-clustered *RLPs*, B – copy numbers of *RLPs* within gene clusters. The pseudogenes *RLP18* and *RLP49* are presented to depict their location close to other *RLPs*. The data was retrieved from AthCNV catalogue [Zmienko 2020].

4.2.3. Identification of *RLP* genes in the assemblies and the analysis of their intraspecific diversity. To identify individual *RLP* genes, I performed BLASTn search against each assembly using the reference gene models as subjects. BLASTn hits of >95% sequence identity were checked for overlap with the expected *Liftoff* annotations. Next, the overlapping *de novo* gene predictions were retrieved for the same regions. To validate the accuracy of the genomic sequence, this procedure was performed for both assemblies and the results were compared. This comparison revealed that the sequence in the gene models was not always identical between *canu* and *flye*. Moreover, *de novo* predicted gene models were frequently different from the lifted ones. This is because *Liftoff* algorithm was based on the similarity to the reference and attempted to reconstitute their exon-intron structure even producing nonvalid ORFs. Therefore, it would be expected to perform correctly for the most conserved and well-characterized genes with highly confident reference models. On the contrary, *AUGUSTUS* was set to only predict genes with valid

ORFs and has been trained on Arabidopsis specific gene sets [Stanke 2006]. Based on that, *ab initio* gene annotations were used for subsequent comparative analyses of individual *RLPs*.

4.2.3.1. Variation in *RLP* genes located outside clusters. 22 non-clustered *RLPs* were expected to possess little or no variation, therefore they represent a case study to evaluate the possibility of using nanopore-only assemblies generated with R9.4 flow cell chemistry to provide the detailed resolution into gene/protein sequence. First, *ab initio* predicted proteins in *canu*- and *flye*-based assemblies were compared for each accession. Only for seven of them (RLP1, RPP27, CLV2, TMM, RLP19, RLP56 and RLP57) protein sequence was identical for all eight pairs. Of those, RLP19 in DOLf was a part of the 1.7 Mb inversion spanning genes *AT2G17270* to *AT2G14000*, which was absent from DOLc, indicating the assembly error in one of the two genomes. For seven other proteins the discrepancy between *canu* and *flye* was detected for one accession (most often it was BRR). The remaining predictions had more differences, with the discrepancy detected for two (RLP7, RLP52), three (RLP4, RLP9, RLP51), four (RLP46), six (RLP45) and seven accessions (RLP5). These observations uncover the imperfections of the *de novo* assembly based on solely nanopore reads and call for caution in relying on computational gene prediction, not supported by experimental data. On the other hand, the ability to compare gene predictions between the two independent assemblies created of the same data, validated the predicted models.

Next, non-clustered *RLPs* were compared among the accessions using sequence similarity and conserved domain prediction (to check their instraspecies conservation). For each comparison, one representative reference model was included in the analysis. *Canu-flye* discordant annotations were manually evaluated and resolved. *RLPs* with model disagreements in two or more accessions were no further analyzed, due to the lack of sequence validation capacity. The highest average pairwise sequence similarity between 9 accessions including Col-0 was revealed in the following 14 genes: *TMM* (99.9%); *RLP44* (99.8%); *RPP27*, *RLP29* and *RLP57* (99.6%); *RLP1* (99%); *RLP54* and *RLP55* (98.8%); *CLV2* (98.5%); *RLP6* (98.2%); *RLP19* (98.0%); *RLP53* (96.9%); *RLP56* (94.1%) and *RLP43* (93.4%).

4.2.3.1.1. Sequence comparisons of RLP17/TMM. In the reference genome *TMM* (*AT1G80080* locus) is the one-exon gene encoding a protein of length 496 amino acid residues (aa). It was highly conserved in the analyzed accessions (at least 99.6% pairwise sequence identity on the aa level). The predicted protein sequences were identical to the reference in 5 accessions (BRR, DOL, OY0, PUN and RSH), while in the remaining three were substitutions of one amino acid (T337M). Additional substitution – S427T was found in AIT (Figure 19). The composition of functional domains was identical for all the proteins. Due to crucial TMM function in organ

development – control of cell differentiation to stomatal pathway and orientation of stomatal cell division – the high sequence and domain conservation of this protein was expected [Geisler 2000].

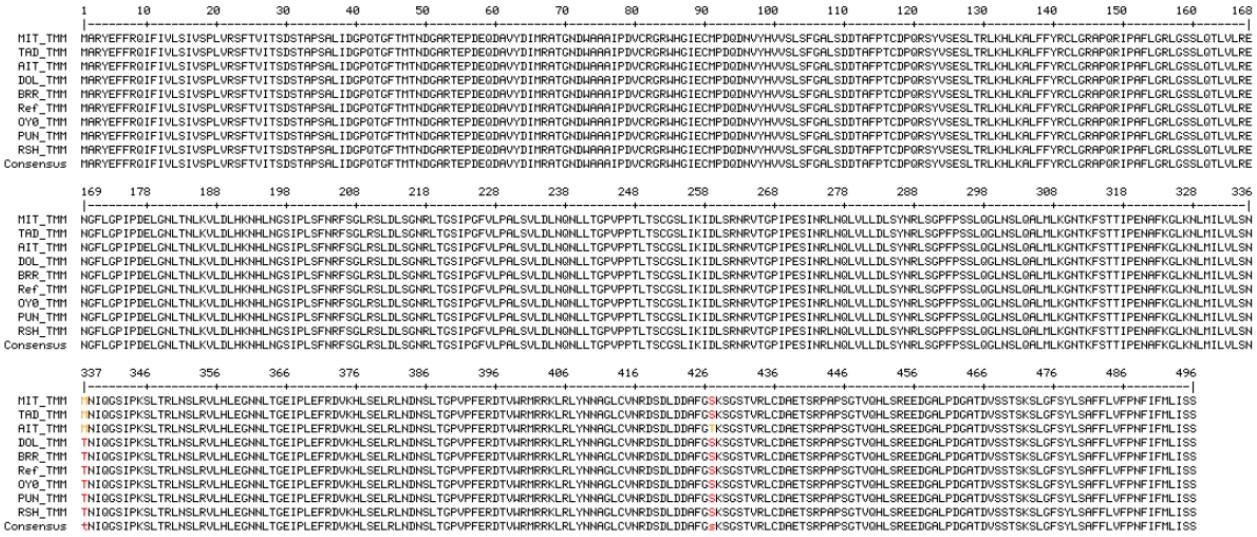


Figure 19. Multiple sequence alignment of TMM protein. The alignments generated with *MultAlin* (http://multalin.toulouse.inra.fr/multalin/).

4.2.3.1.2. Sequence comparison of RLP44. In the reference genome *RLP44* (*AT3G49750* locus) is a one-exon gene encoding a small protein of 274 aa. In non-reference accessions, this protein was observed to have similar structure. In this comparison, a 5-bp indel was found between DOLc and DOLf sequences, which resulted in the frameshift after aa 244 and the early STOP codon in DOLf. Accordingly, DOLc sequence with the valid ORF was used for sequence comparison. RLP44 protein sequences turned out to be almost identical, with the three accessions (DOL, MIT, and REF) differing from the remaining ones by a four-aa deletion in the N-end of the protein, which, however, did not affect their overall domain composition. Similarly to TMM, RLP44 is important for plant growth, as it has been shown to act as a positive regulator of brassinosteroid signaling pathway and is required for plant cell wall surveillance [Wolf 2014]. Expectedly, its genomic and protein sequence remains strongly conservative among the accessions.

4.2.3.1.3. RPP27 annotation and sequence comparison. In the reference genome *RPP27* (*AT1G54470* locus) is a multi-exon gene encoding 457 aa-long protein. It has been annotated as a Cf-like gene, where Cf comes from a fungus *Cladosporium fulvum*, as its homolog confers resistance to this fungus in tomato [Krujit 2005]. In Arabidopsis, *RPP27* has been found to confer resistance to another downey mildew-causing fungus, *Peronospora parasitica* (https://www.araport.org/). In fact, this resistance has been demonstrated in non-reference accession, Landsberg erecta (Ler), and the respective locus was mapped to Col-0 (namely *RLP8*) [Wang 2008]. However, the identification of the *RLP8 aka RPP27* locus in Col-0 has been put into

question and the original research article has been retracted. Yet, the defensive activity of *RPP27* in Ler accession was not a subject of retraction.

In eight Arabidopsis assemblies, *ab initio* prediction consistently proposed gene models that merged two reference loci: *AT1G54470* and *AT1G54475* (unknown protein). The resulting proteins were much longer than the reference: 715 aa in PUN, 956 aa in AIT (both are relicts) and 1032 aa in remaining accessions. The latter model differs from the protein sequence in Ler only by 12-aa deletion (UniProt entry Q70CT4, 1044 aa) (https://www.uniprot.org/). Notably, the curated database entry for RPP27/RLP8 in Col-0 exists and is long, as well (UniProt entry Q9SLI6, 1009 aa) and is different from Ler protein entry by a 35-aa deletion. The TAIR12 annotation of the latest Arabidopsis reference genome Col-CC.v2, in which I have participated, also reports a new gene model for the same genomic position, by merging *AT1G54470* and *AT1G54475* (Figure 20). Still, RPP27 protein sequence remains uncertain. Therefore, both UniProt models were used as reference for the comparisons. This comparison confirmed high conservation of the LRR domains and high pairwise sequence identity: 99-100% for nonrelicts, 99.3% for relicts, 98.7-98.9 for relict-nonrelict.

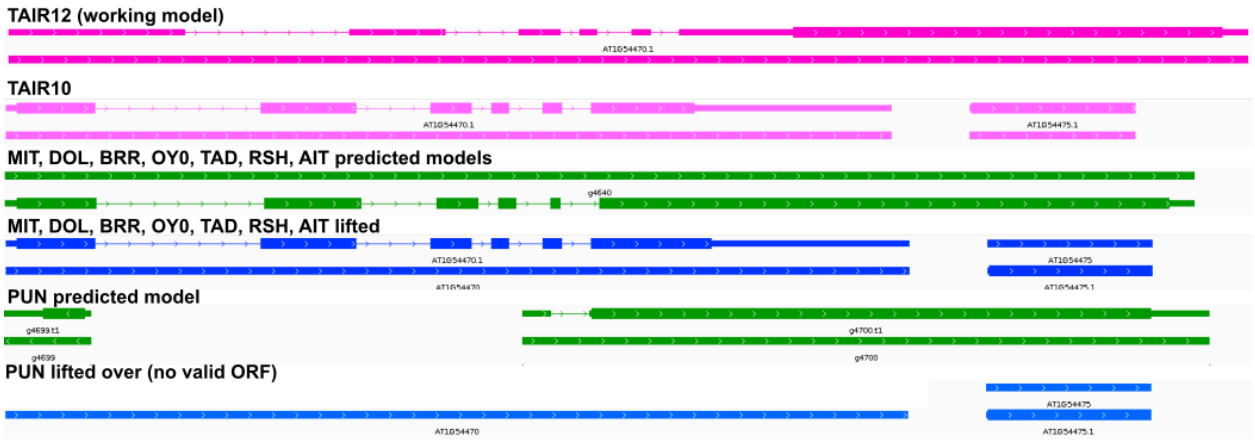


Figure 20. Two gene versions of *RPP27*. The two reference versions are pink, the lifted coding sequences are in blue, the *ab initio* predictions are in green.

4.2.3.1.4. Sequence comparison of RLP29. In the reference genome *RLP29* (*AT2G42800* locus) is a one-exon gene that encodes 462 aa-long protein. Similarly, its homologs in other accessions were of lengths 459 aa or 462 aa and were encoded in one exon. Exceptionally in BRRc, two exons were predicted due to indels associated with homopolymer sequences, however BRRf was homologous to the other sequences and it was used for sequence comparison. All *RLP29s* had very high similarity level and conserved domain organization (Figure 21). Pairwise identity ranged from 98.9% to100%.

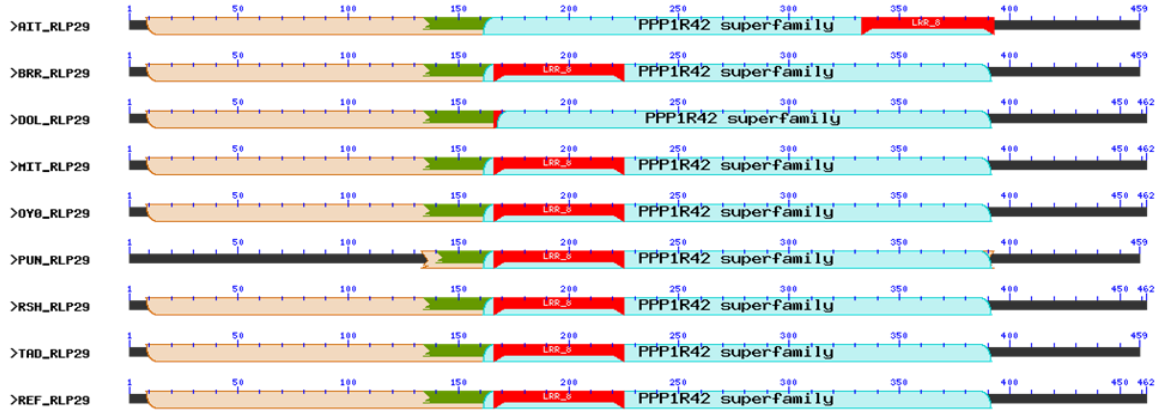


Figure 21. Conserved domain prediction for RLP29 proteins. LRR – leucine-rich domain.

4.2.3.1.5. Sequence comparison of RLP57. In the reference genome *RLP57* (*AT5G65830* locus) ia a one-exon gene encoding 279 aa-long protein. Sequence comparisons revealed that in PUN, OY0 and BRR the RLP57 homologs were identical to the reference, while in the five remaining assemblies they had two substitutions: Q44R and F127L (Figure 22). The LRR-domain conservation was saved in all the protein models.

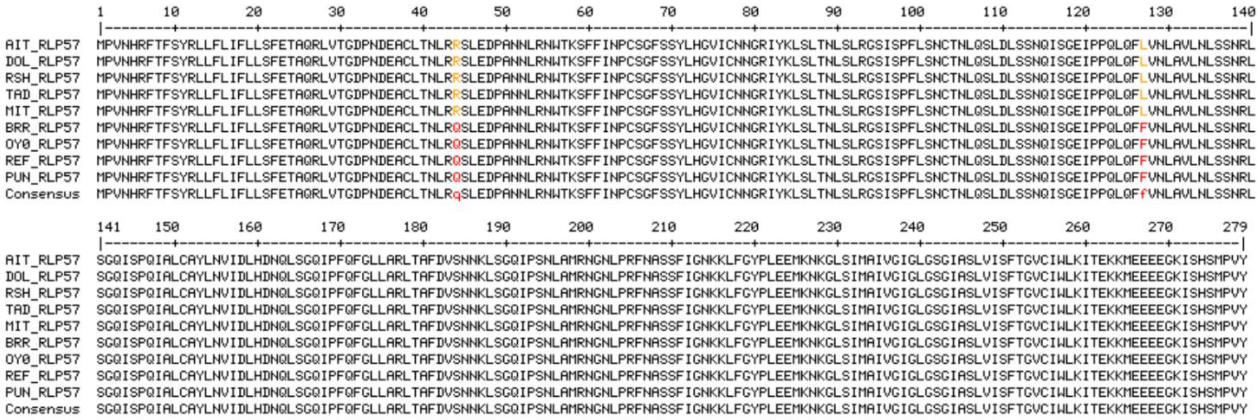


Figure 22. Multiple sequence alignment of RLP57 protein.

4.2.3.1.6. Sequence comparison of RLP1/ReMAX. In the reference genome *RLP1* (*AT1G70390* locus) is a multi-exon gene which can be expressed to numerous isoforms with the length ranging from 913 to 1083 aa. *AUGUSTUS* was set to predict only one model for each gene (the most probable), however it might be the case that leads to some differences between the proteins. Accordingly, for all accessions except BRR, the protein models were 1150 aa long and had very high sequence conservation (≤ 99.5% pairwise identity). BRR_RLP1 was of length 1103 aa and was divergent from other RLPs in the region which was also the most different among the reference protein isoforms (Figure 23A). The diversified sequence fragment layed nearby the domain containing LRR_8 motif (Figure 23B). Nevertheless, all the proteins presented highly

conserved domain organization. These results are congruent with the well-defined role of RLP1 in plant defensive mechanism to the infection by Xanthomonads [Jehle 2013].

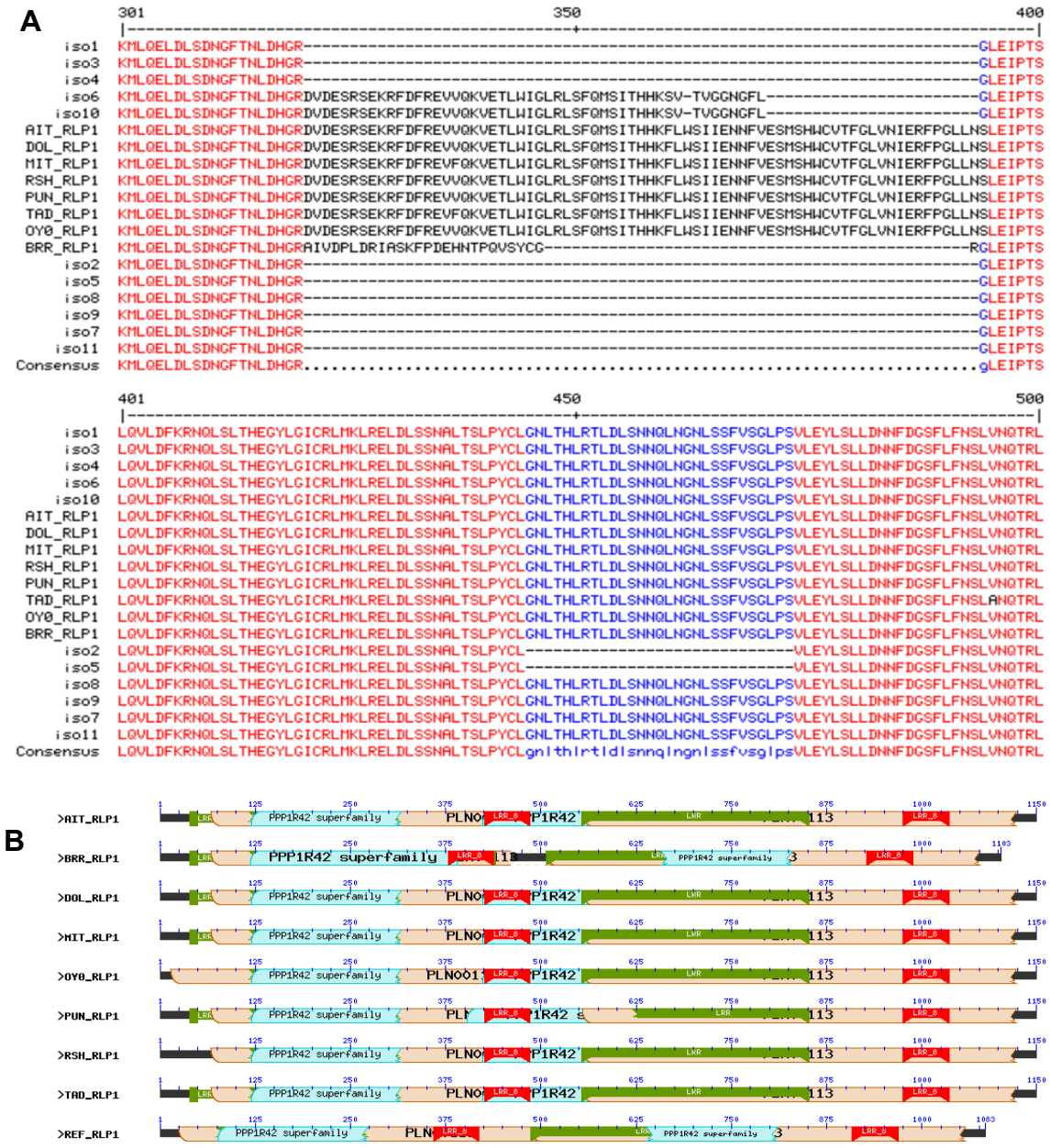


Figure 23. Sequence variants of RLP1/ReMAX protein. A – the two variable fragments between the RLP1 intraspecies variants and within reference isoforms (iso), revealed by global alignment. B – conserved domain prediction for RLP1 proteins.

4.2.3.1.7. Sequence comparison of RLP54. In the reference genome *RLP54* (*AT5G40170* locus) is a one-exon gene encoding the protein of 792 aa in length. Gene predictions revealed similarities among the accessions, except for BRRc assembly, where a 1-bp indel broke the model into two exons. Therefore, for this accession, the model BRRf was considered as more probable and was used in the subsequent comparisons. RLP54 proteins had identical lengths and altogether

had only 20 positions with different amino acids (pairwise identities of range 97.6 – 100%). Consequently, these proteins were identical in their conserved domain structure.

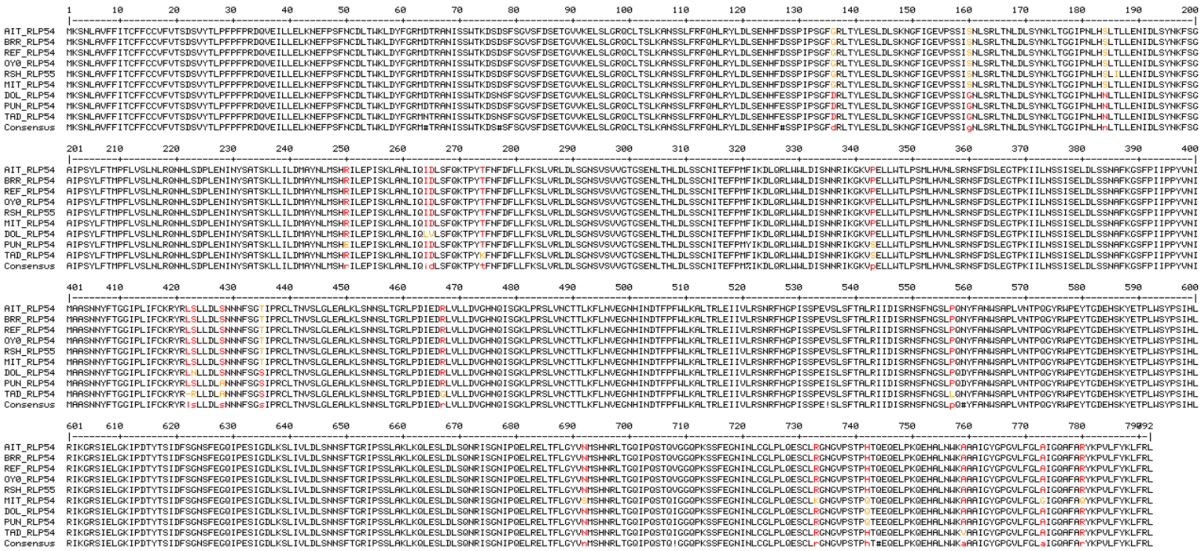


Figure 24. Multiple sequence alignment of RLP54 protein.

4.2.3.1.8. Sequence comparisons of RLP55. In the reference genome *RLP55* (*AT5G45770* locus) is a one-exon gene that encodes the protein of length 425 aa. The analysis of *ab initio* predicted models in the case of RLP55 was not possible to conduct, as in all except for BRRf assemblies, the model predictions by *AUGUSTUS* resulted in merging genes *AT5G45750-AT5G45770*, which was not likely due to the different reads mapping pictures between the two genes (Figure 25). *Liftoff*-based comparisons seemed to be more valid, as in all the assemblies they produced the valid ORF – single-gene models similar to the reference protein. One discrepancy was noticed for RSHf assembly, which was an insertion of T in a homopolymer sequence, which led to early STOP codon, therefore the RSHc protein sequence was used for further comparisons. The coding sequences were transcribed to protein sequences using *Transeq* tool (https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq) and compared. Sequence similarity was from 97.41% to 100% (RSH, OY0 and AIT were identical). As expected, all the proteins had conservative domain organization.

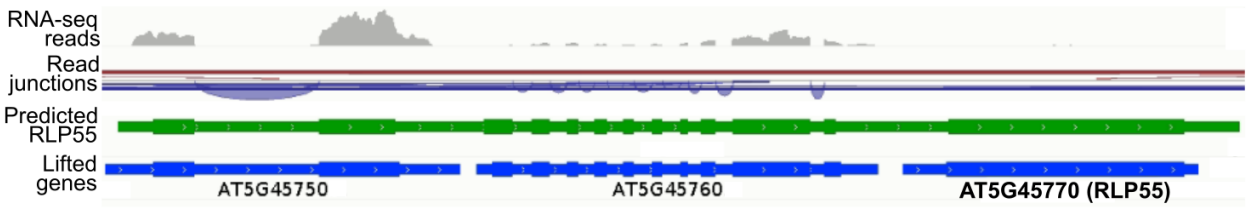


Figure 25. Mapping of RNA-seq reads to AIT genome. The read junctions demonstrate the transcripts of the two genes (*AT5G45750* and *AT5G45760*). *RLP55* (*AT5G45770*) demonstrates lack of mappings.

4.2.3.1.9. Sequence comparisons of CLV2. In the reference genome *CLV2* (*AT1G65380* locus) is a one-exon gene encoding a 720 aa-long protein. Sequence comparisons revealed the high conservation in CLV2 length and structure, and high sequence similarity: CLV2 proteins has at least 97% pairwise sequence identity among the accessions. The protein sequences of RSH_CLV2 and BRR_CLV2 and MIT_CLV2 and TAD_CLV2 were 100% identical. Consequently, all the proteins had similar domain composition (Figure 26). In OY0 the predicted LRR-including domain was slightly different. Similarly as TMM, CLV2 is the RLP that participates in regulation of plant development. CLV2 is a crucial member of CLAVATA – signaling pathway that regulates the size of stem cells in plant shoot apical meristem [Jeong 1999, Bleckmann 2010].

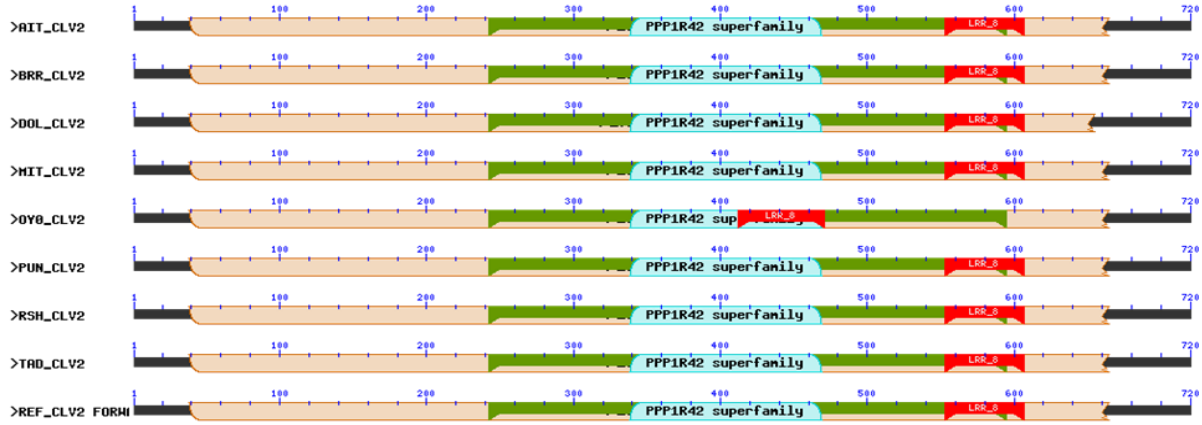


Figure 26. Conserved domain prediction for CLV2 proteins among the accessions.

4.2.3.1.10. Sequence comparisons of RLP6. In the reference genome *RLP6* (*AT1G45616* locus) is a one-exon gene encoding 994-aa long protein. Likewise, it was predicted as a one-exon gene of similar sequence length in most accessions (BRR, DOL, MIT, RSH and TAD). In AIT and OY0, the protein models were split to two exons, and the intron was in different positions in each accession of the two, resulting in deletions of small sequence fragment. The two contrasting models were predicted for PUNc and PUNf, with one and two exons, respectively, as a result of a 1-bp indel in genomic sequence. After manual checks, PUNc was used for sequence comparisons as its model was more similar to the remaining ones. In addition, AIT_RLP6, DOL_RLP6, and PUN_RLP6 lacked 66 amino acids from N-terminus, when compared to the reference. Still, all the proteins presented high level of pairwise sequence similarity (from 96.5 to 100%) and high conservation of LRR repeat – containing domain organization.

4.2.3.1.11. Sequence comparisons of RLP19. In the reference genome *RLP19* (*AT2G15080* locus) is a one-exon gene which encodes a 983-aa long protein. *Liftoff* detected *RLP19* in all the assemblies with no sequence discrepancies between *canu* and *flye*. Similarly to the reference model, one exon was predicted by *Liftoff*, however this prediction lacked the

reference-like ORF in DOL due to a frameshift. Analysis of its genomic sequence revealed that the frameshift resulted from a 134-bp deletion. In contrast, *ab initio* prediction pointed on the valid ORF in DOL that actually has been restored by the new splice site (Figure 27). Resulting DOL_RLP19 protein is 891-aa long, while in the remaining accessions it has reference-like length. All the orthologs display 96.8% – 99.3% pairwise sequence identity and similar domain organization. Although, the biological function of RLP19 has not been elucidated yet, low intraspecies diversity supports the hypothesis about its role in plant development or adaptation. Interestingly, an interaction between RLP19 and GLR1.3 (which is a glutamate receptor) positively enhanced cold tolerance in Arabidopsis, by mediating endogenous accumulation of jasmonate [Chatr-Aryamontri 2015].

Figure 27. Sequence alignment of RLP19 orthologs that depicts the part deleted from DOL.

4.2.3.1.12. Sequence comparisons of RLP53. *RLP53* (*AT5G27060* locus) is a gene encoding the protein of length 958 aa. The protein sequence was discordant between BRRc and BRRf assemblies, that was the result of two indels in the respective genomic sequences. The indels affected the localisation of the first intron, which, in consequence, changed N-terminal part of the protein. Since BRRf had higher pairwise identity to any other *RLP53* ortholog, compared to BRRc, it was selected for subsequence comparisons. They revealed that exon-intron organization of *RLP53* genes as well as protein lengths were poorly conserved. Pairwise identity ranged from 92.3% to 99.1% with the TAD_RLP53 being the most divergent. The length and the organization of LRRs was also variable, indicating possible lack of conservation in molecular pattern recognition by these proteins (Figure 28).

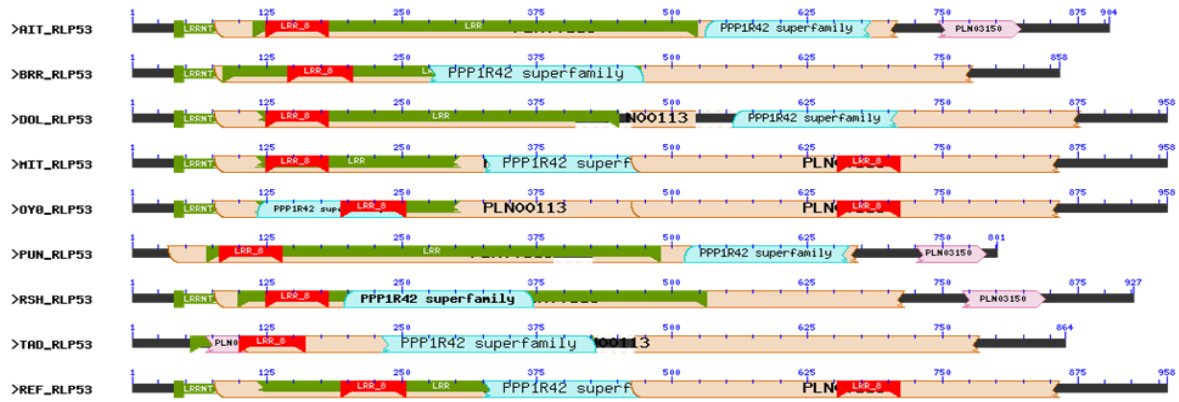


Figure 28. Conserved domain prediction for RLP53 proteins among the accessions.

4.2.3.1.13. Sequence comparisons of RLP56. In the reference genome *RLP56* (*AT5G49290* locus) is a multi-exon gene encoding numerous protein isoforms that vary in length from 742 to 937 aa. Sequence analysis revealed that the orthologs were 901-971 aa in length and formed two separate clades in phylogeny (Figure 29A). The pairwise similarity within a clade was 99.5% and higher, while between the clades lower than 90%. BRR_RLP56 and RSH_RLP56 showed the highest similarity to each other and to the reference RLP56 (all isoforms), but these two were more divergent from the other orthologs. All the proteins had conserved domain organization with constitution of many LRR-containing motifs (Figure 29B).

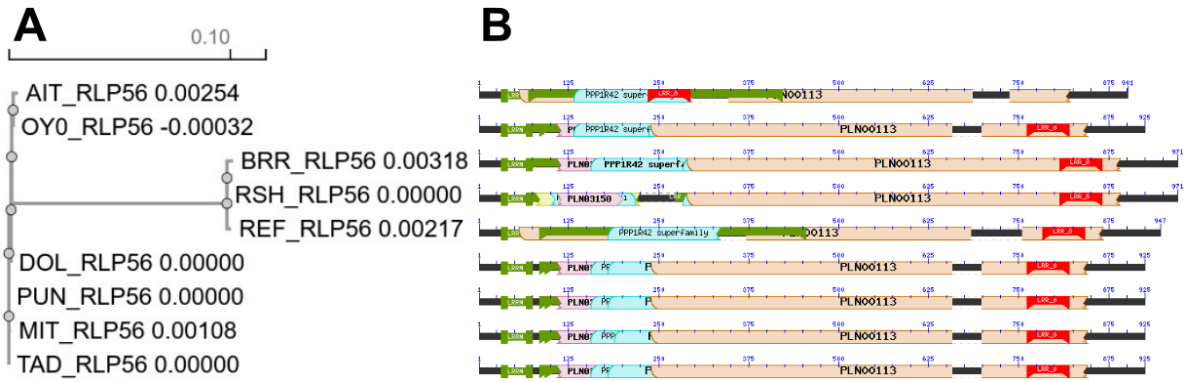


Figure 29. Sequence comparison of RLP56 orthologs. A – phylogenetic tree. B – conserved domain prediction for RLP56 proteins corresponding to their order on the tree. Phylogenetic tree was generated with *Clustal Omega* tool, using Neighbour-Joining algorithm [Madeira 2024].

4.2.3.1.14. Sequence comparisons of RLP43. In the reference genome, *RLP43* (*AT3G28890* locus) has two splicing models, but they are identical in having the same coding sequence with the respective protein of 711 aa length. Again, the BRRc-BRRf discrepancy was observed for this gene. The detailed inspection revealed that the *canu* model encoded a protein truncated on its N-terminus, due to an intron which included the reference-like START position (Figure 30A). Accordingly, BRRf model was used for the subsequent analyses, as its START position was similar to orthologs. RLP43 proteins differed substantially from the reference protein both in sequence and in length. Except for OY0_RLP43 (669 aa), they were all longer, with the lengths varied from 808 to 946 aa. The highest similarity to the reference was observed in TAD, which is is in line with the most similar length as well (Figure 30B). Pairwise similarity of RLP43 proteins was much lower compared to previously described RLPs and ranged from 89.7% to 97.0%. The detailed composition of RLP LRR-containing domains was variable, consequently (Figure 30C).

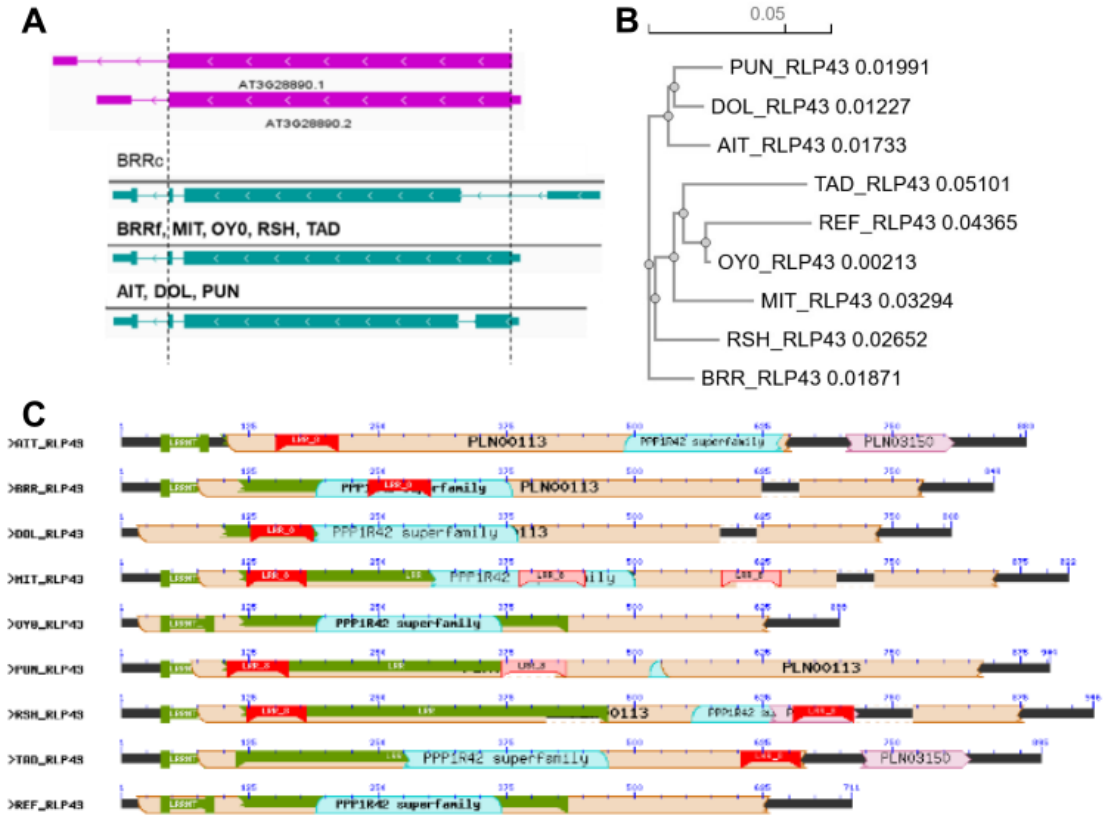


Figure 30. Sequence analysis of RLP43 orthologs. A – *ab initio* predictions of coding sequences among the accessions compared to the reference model (pink). The dashed line depicts the regions corresponding to the reference protein. B – Phylogenetic tree of RLP43 orthologs among the accessions. C – conserved domain prediction for RLP43 protein sequences.

Concluding this part, the analysis of non-clustered *RLP* genes among eight assemblies revealed limitations of the nanopore-only assemblies in providing highly-accurate single-nucleotide resolution sequence data. It is clear that the success rate in detailed characterization of individual protein sequences based on nanopore-only *de novo* assemblies was moderate. For this reason, to identify large SVs, the subsequent analysis is supplemented by the larger set of genome assemblies based on nanopore- and PacBio genomic sequencing data, referred as 69 genomes set [Lian 2024].

4.2.3.2. Variation patterns found in paired *RLPs*. Several paired *RLPs* are present in the Arabidopsis genome. Two pairs are located at Chromosome 1 (*RLP2-RLP3* and *RLP11-RLP12*), two pairs at Chromosome 2 (*RLP20-RLP21* and *RLP22-RLP23*), and three pairs at Chromosome 3 (*RLP30-RLP31*, *RLP32-RLP33* and *RLP34-RLP35*), respectively. Of them, three pairs display patterns of variation, looking at the previous data and the variation seems to affect only one gene in each of these pairs. The first, *RLP2-RLP3*, was revealed to have PAV for *RLP3*, which had been associated with plant resistance against Fusarium wilt [Shen 2013]. The second pair is *RLP11-RLP12*, with frequent deletion of the gene *RLP11* and the third pair is *RLP30-RLP31*

which displays copy number increase of *RLP30* in seven accessions, pointing on possible duplication (Figure 18B). The other three gene pairs (*RLP20-RLP21*, *RLP22-RLP23* and *RLP32-RLP33*) were not expected to undergo copy number variation. The *RLP34-RLP35* gene pair displayed high deviation from the average, due to high sequence similarity to other *RLPs*, which is challenging for mapping short reads.

4.2.3.2.1. Variation patterns in *RLP2-RLP3* across population. In the reference genome, *RLP2* (*AT1G17240* locus) and *RLP3* (*AT1G17250* locus) are both localized on the minus strand, 2.2 kb apart from each other. Sequence comparison in eight accessions revealed the complete deletion of *RLP3* in DOL and PUN, with simultaneous alteration of *RLP2* in the mentioned plants, which was assigned as *AltRLP2*, since its sequence similarity to *RLP2* is higher than to *RLP3* (Figure 31A). *AltRLP2* encodes a protein with high identity to the protein linked with compromised resistance to *Fusarium oxysporum* infection, observed in Ty-0 accession [Shen 2013]. According to AthCNV data, *RLP3* is deleted in 8% Arabidopsis polulation (86 / 1060 accessions). I additionaly checked for the presence of this variation pattern in the 69 genomes [Lian 2024], by scanning these genomes for *RLP2* and *RLP3* sequence. I found 10 accessions, with the same deletion pattern as in DOL: Are-1, Are-2, Are-6, Are-10, Bur-0, Can-0, Cant-1, Cvi-0, Rab-R1 and Sf-2. Interestingly, mapping RNA-seq reads from leaves to the respective assemblies revealed that in REF-like accessions, e.g. AIT, there is only a trace expression of *RLP2*, while *RLP3* is expressed, while the *altRLP2* is expressed in the accession with *RLP3* deletion (Figure 31B). While *RLP2* gene expression might be found in other tissues besides plant leaves, this gene is less conserved when compared to *altRLP2* and *RLP3*, which suggests that the gene dosage effect of *RLP2-RLP3* locus is preserved, regardless of the number of *RLP* genes.

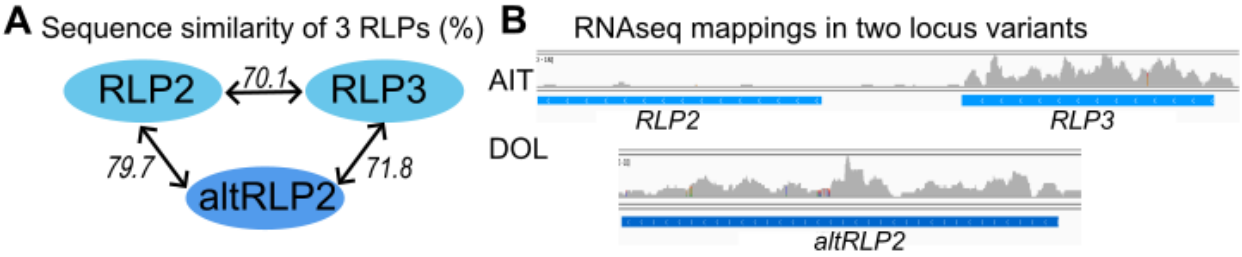


Figure 31. The two patterns of *RLP2-RLP3* gene pair. A – average sequence identity between the three proteins. B – mappings of RNA-seq reads to the two locus variants.

4.2.3.2.2. Variation patterns in *RLP11-RLP12* gene pair. In the reference genome, *RLP11* (*AT1G17390* locus) and *RLP12* (*AT1G17400* locus) are both localized adjacently to each other, on plus strand. When analyzed copy number data, *RLP11* displayed *PAV*, with copy number close to 0 in 105 accessions (10% of the population). I could not find any deletion in eight Arabidopsis accessions of this study, but scanning through 69 genomes confirmed the PAV in four

accessions: Can-0 (Spain), Co-4 (Portugal), Ice-1 (Iceland), and JEA (France). In all, the same 5142-bp region spanning *RLP11* was missing when compared to the reference genome indicating that this has originated in Arabidopsis population from a single deletion event.

4.2.3.2.3. The *RLP20-RLP21* gene pair displays little variation. In the reference genome, *RLP20* (*AT2G25440* locus) and *RLP21* (*AT2G25470* locus) are located on the reverse strand of Chromosome 2. The distance between the genes of the pair is variable as it includes the two intervening non-RLP genes, one of which (*AT2G25450*) was deleted in 32 accessions including RSH (out of 1060), and can vary from 6.1 kb in RSH, up to 10.7 kb, in accessions where both intervening genes are present.

4.2.3.2.4. Variation events in *RLP22-RLP23* gene pair. In the reference genome, *RLP22* (*AT2G32660* locus) and *RLP23* (*AT2G32680* locus) are located on minus strand, separated by one non-RLP gene (Figure 32). All the three genes did not display variation in copy numbers. The assembly-based analysis confirmed high conservation of two *RLPs* (pairwise sequence identity of protein sequences was 99.5% and 98.5% for *RLP22* and *RLP23*, respectively). Of note, the most distinct location was found to be the varied number of glutamic acids (marked as E) in position 834-855 of RLP23 protein (Figure 32A). The distance between the genes of the pair (3.5 kb in the reference genome) was increased in PUN to 5.76 kb as a result of a TE-mediated insertion (2092 bases of inserted sequence revealed by global alignment). The insertion in PUN was annotated as a DNA transposon of Helitron class, that was classified as *ATREP3* [Feschotte 2001]. 5mC methylation levels in this region in PUN was checked and an over 80% increase in methylation probability across TE length indicated it was repressed. Remarkably, its high methylation level did not affect the epigenetic status of adjacent *RLP22* gene, which was confirmed by its 5mC levels pattern in the plant without insertion – TAD (Figure 32B).

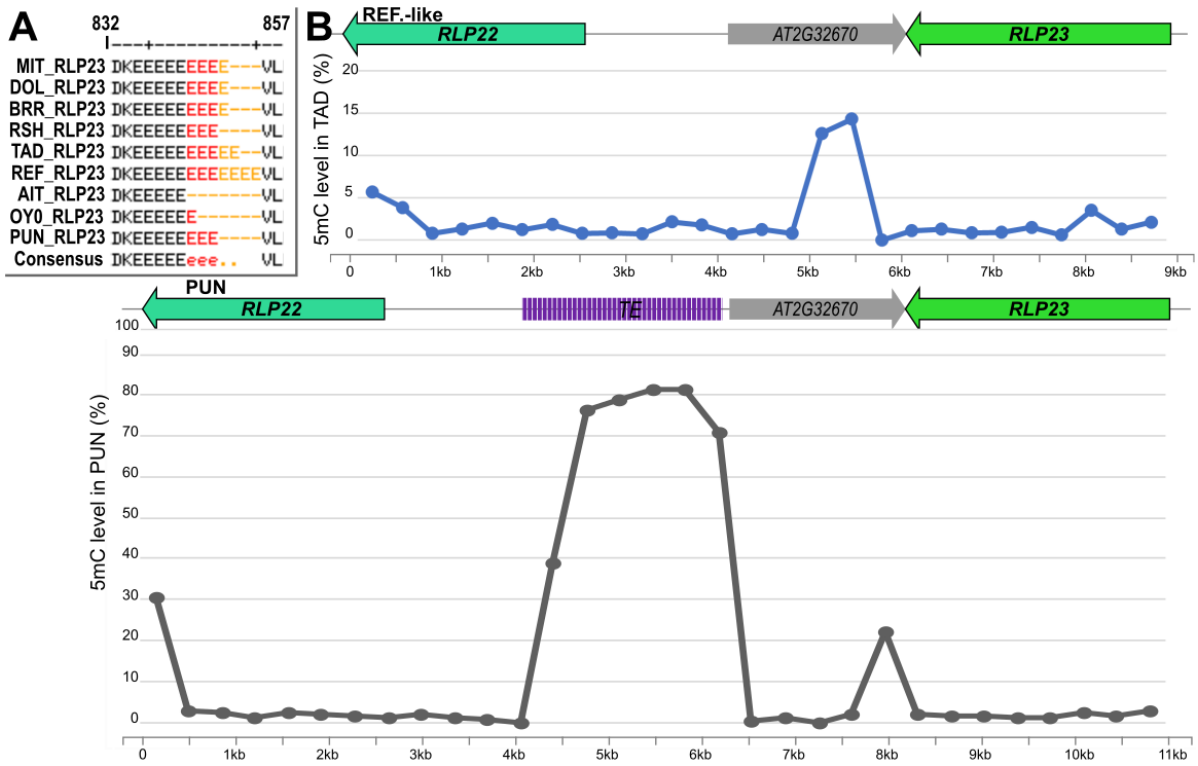


Figure 32. *RLP22-RLP23* gene pair composition. A – varied number of E amino acid among RLP23 protein sequences. B – Gene structure and methylation patterns of the reference-like pattern in TAD (on top) and TE-mediated insertion in PUN.

4.2.3.2.5. A tandem duplication in the *RLP30-RLP31* gene pair. In the reference genome, *RLP30* (*AT3G05360* locus) and *RLP31* (*AT3G05370* locus) genes were separated by a small intervening gene *AT3G05365*, located on the same strand (minus) and encoding hypothetical protein. Moreover, the reference *RLP30* gene is annotated within the longest intron of *AT3G05350* gene encoding a metallopeptidase M24 family protein and also located on the minus strand. According to AthCNV data, the entire *AT3G05350* gene, including *RLP30* sequence, has been duplicated in 7 accessions, however these accessions are different from those analyzed in the current study. The BLAST scanning of 69 genomes with *RLP30* gene sequence revealed a tandem duplication in one accession (Toufl-1), confirming low frequency of this variant in the population (Figure 33). The duplicated fragment was 9.5-kb in length and it included an extra copy of *AT3G05350-RLP30* with 99.5% pairwise sequence identity to the original sequence. Interestingly, the duplicated fragment was flanked by a 598-bp-sequence repeat on the both sides. Additionally, a 155-bp deletion was found within DOL_RLP31 genomic sequence, which removed STOP codon from it, however, *ab initio*-predicted model restored the valid ORF in this gene.

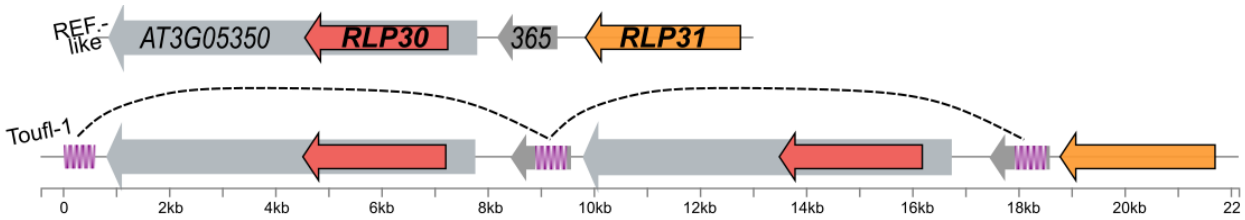


Figure 33. Variation patterns of *RLP30-RLP31* gene pair. The dashed fragment indicates duplication. The repeated sequence flanking the duplication are marked by purple rectangles.

4.2.3.2.6. A TE-derived insertion in *RLP32-RLP33* gene pair. In the reference genome, *RLP32* (*AT3G05650* locus) and *RLP33* (*AT3G05660* locus) are located directly next to each other and encoded on minus strand. Sequence comparison did not reveal structural variants in all accessions except for TAD. In this accession, *RLP32* is interrupted by an insertion of TE, which broke its ORF after 243 amino acids (Figure 34A). As the insertion was large in size (> 5 kb) but not annotated as a TE, it was translated and checked for possible features in InterPro (https://www.ebi.ac.uk/interpro/) and Dfam (https://www.dfam.org/) databases. The following motifs were found: LTRs of length 440 on both flanks, the integrase, the reverse transcriptase, and the Gag/Pol domain (Figure 34B). These findings point on the complete TE that can be classified as *LTR/COPIA* [Wicker 2007]. The missing TE in the annotation can be explained by the presence of additional ORF not assigned to any domain. The complete structure suggests that this insertion resulted from a recent transposition event, although high methylation (more than 90%) level surrounding this TE indicated its strong repression (Figure 34A). Scanning *RLP32* and *RLP33* genes across 69 genomes supported the REF-like pattern in them, concluding that this variant was found to be accession-private, which was possible to discover owing to *de novo* assembly approach.

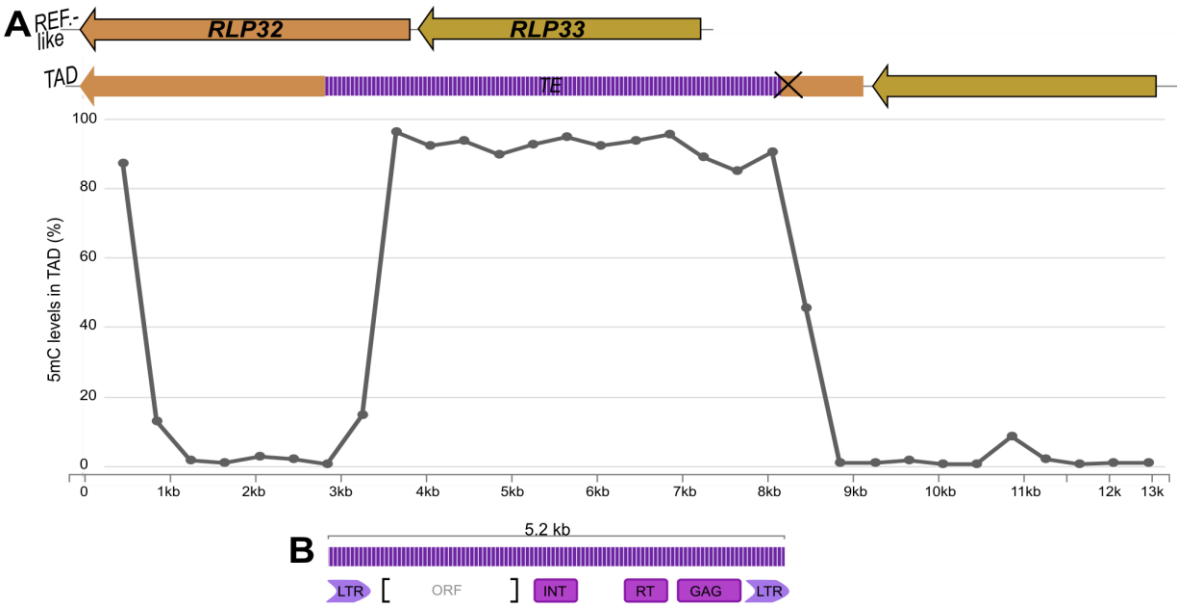


Figure 34. *RLP32-RLP33* gene pair. A – The TE-derived insertion inside *RLP32* in TAD accession and its 5mC landscape. B – The structural domains present within the inserted TE.

4.2.3.2.7. Little variation in *RLP34-RLP35* gene pair. Finally, the short read-based estimations for gene pair *RLP34-RLP35* predicted a somehow diffused pattern of copy numbers, yet without any clear duplication or deletion events. The genes are located in 16-kb distance from each other and are separated by six non-RLP protein-coding genes, one of which (*AT3G11060*) is supposed to undergo PAV within population. No large-scale sequence variation has been observed for any of the RLPs in this region, however, small-scale variants disrupted the reference-like ORFs in AIT, BRR, and OY0 and did not produce valid *Liftoff* models, while the corresponding ab initio models predicted proteins with high levels of sequence variation.

4.2.3.3. Variation patterns found in *RLP* gene clusters. The remaining *RLPs* are organized in five gene cluster in the reference genome, one at Chromosome 1 (*RLP13-RLP16*) including four *RLPs*, one at Chromosome 2 (*RLP24-RLP28*) including five genes, two at Chromosome 3 (*RLP36-RLP38* and *RLP39-RLP42*), and one at Chromosome 4 (*RLP47-RLP50*) containing three *RLPs* and a pseudogen of RLP47, respectively. All the clusters are promising to display different variation patterns, as the short read data indicated them to have PAV patterns at least in one gene of each (Figure 18B). As many genes were needed to be analyzed here, the following description is focused on structural changes.

4.2.3.3.1. Variation events in *RLP13-RLP16* gene cluster. In the reference genome, the genes *RLP13* (*AT1G74170* locus), *RLP14* (*AT1G74180*), *RLP15* (*AT1G74190*) and *RLP16* (*AT1G74200*) form a compact-size uninterrupted cluster of around 18 kb in size, with all the genes localized on minus strand. Nevertheless, all the cluster members display limited similarity to each

other. RLP13 is most similar to RLP15 and RLP16 with pairwise identity of 70% and 65% respectively, while the identity of the remaining proteins to each other was below 60%. Analyzing short read data, deviations from the normal copy number were observed for the genes, suggesting possible deletion of *RLP13*, *RLP14* or *RLP15* genes in several accessions originated from south Sweden. However, in the set of eight accessions analyzed in this study no gene deletion has been detected, instead the two independent TE-insertions were found. One of them was found in AIT, with the length equal to 1.1 kb and localized between *RLP14* and *RLP15*, which was annotated as a TIR order. In PUN, *LTR/COPIA36* was inserted between *RLP13* and *RLP14*. None of the two inserts affected the entire structure of *RLP* genes. Further analysis of 69 genomes revealed an identical pattern with PUN in another accession from Spain (Cant-1) and a complete deletion of *RLP16* in an accession from Tanzania Tanz-1 (Figure 35).

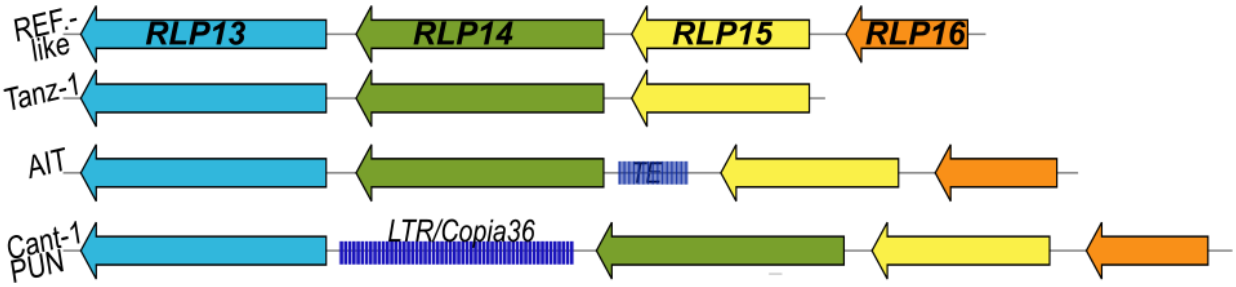


Figure 35. Structural variants of *RLP13-RLP16* gene cluster found among accessions.

4.2.3.3.2. Variation patterns in *RLP24-RLP28* gene cluster. *RLP24-RLP28* cluster is localized on Chromosome 2 and is composed of five *RLPs* and two intervening genes (*AT2G33040* annotated as a subunit of mitochondrial ATP synthase and *AT2G33070* that encodes an enzyme responsible for the formarion of nitrile). Short read data analysis indicated copy number alteration for the three genes of the cluster. *RLP25* copy number estimates were noisy and did not separate into copy number groups. In addition, for *RLP24* and *RLP28* copy number estimates were slightly decreased in some accessions including AIT (*RLP24*) and OY0 (*RLP28*). Small descrease in copy numbers might result from partial deletion of the gene sequence or its high divergence that prevented short reads from mapping to the reference genome in that region. Indeed, in OY0, a 1.53 kb deletion was detected which overlapped the reference *RLP28* gene by 1357 bp and covered 55% of the protein length from the N-terminus. In AIT, no sequence deletion spanning *RLP24* was detected, however in this accession *RLP24* had strickingly lower similarity to the reference gene, both at the DNA and the protein levels.

Finally, the BLAST scanning of 69 genomes uncovered new evidence that supported the short read-based and assembly-based findings, thus allowing me to define variation patterns within

the cluster. A 2-kb deletion spanning *RLP24* and *RLP25* was identified in 11 accessions (Eln-2, Etna-2, Mr-0, Pyl-1, Qar-8a, Are-1, Are-2, Are-6, Are-10, Bur-0, and Rab-R1 thus assigning this subset to PAV pattern. Interestingly, the six latter accessions also displayed PAV in *RLP2-RLP3* gene pair (see section 4.2.3.2.1). The two partial deletions altering *RLP28* were found: OY0-like variant in three more accessions (Abd-0, Anz-0, Co-4), and a 1.8-kb deletion that took the C-terminal part of the gene in Bla-1 accession. *RLP26* and *RLP27* remained unchanged across the 69 genomes. In summary, these results confirm the short read-based data pattern and complement it by the extra variant of *RLP28* in one accession (Figure 36).

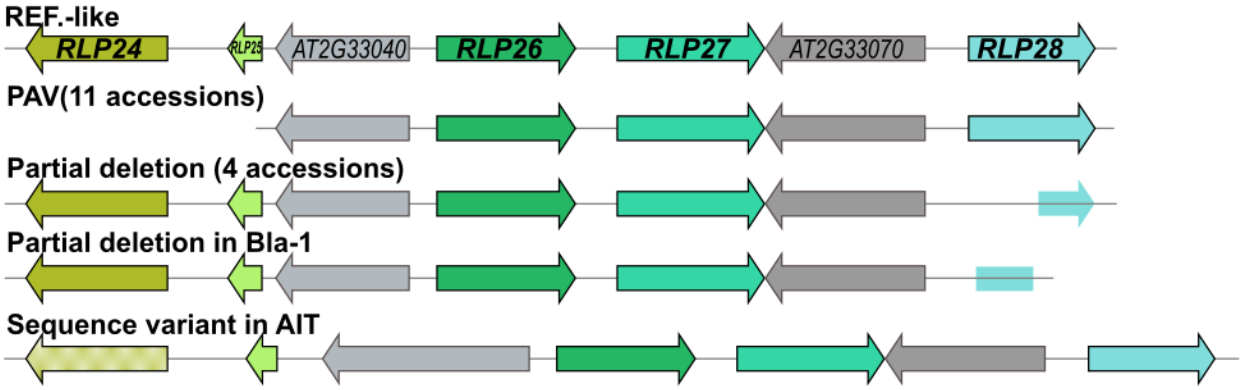


Figure 36. Variants of structural organization of *RLP24-RLP28* gene cluster.

4.2.3.3.3. Variation events in *RLP36-RLP38* gene cluster. The next cluster under analysis is composed of a pair of *RLP37-RLP38* genes, located next to each other and separated from the upstream *RLP36* gene by 43.8 kb distance, where 6 interrupting genes are localised. *RLP36* is entirely covered by copy number variants and is partially or entirely deleted from some accessions, according to short read data, including PUN (copy number = 0) and AIT (copy number = 0.8). Analysis of PUN assembly confirmed complete deletion of *RLP36* gene, but no deletion was observed in AIT. However, AIT_RLP36 had much lower similrity to its orthologs from other accessions and the reference, both at the DNA and amino acid level. Simlarly to RLP24 described above, this probably resulted in poor mapping of short reads to the reference genome. Two other *RLPs* in the cluster did not display large-scale SV, despite the fact that, according to AthCNV data, deletion of *RLP38* in MIT might have been expected (copy number = 0.2). This gene was however sucesfully annotated in the expected position, both by *Liftoff* and *AUGUSTUS*, in MITc and MITf assemblies.

4.2.3.3.4. Complex variation in *RLP39-RLP42* gene cluster. In the reference genome, *RLP39-RLP42* gene cluster is composed of four *RLP* genes located on minus strand and occupy a compact-size region of 20 kb. The cluster also contains intervening genes including two hypothetical proteins: *AT3G24929* on plus strand that overlaps with *RLP40* (*AT3G24982* locus)

and *AT3G25012* that overlaps with *RLP41* (*AT3G25010* locus), respectively. The region is expected to display complex CNV, as was shown in short-read data. The picture of copy number estimates of individual *RLPs* was distorted as a result of high sequence similarity between the four cluster members [Wang 2008]. Especially high sequence identity was observed for the genes *RLP39* (*AT3G24900* locus) and *RLP41* (91.3%). Accordingly, copy number estimate was interpretable only for *RLP42* (*AT3G25020* locus), which is the most distinct from the remaining three, and it resembled PAV pattern (Figure 18B). Besides noisy copy number distribution, the *Liftoff*-based annotation was also biased (Figure37A). For example, in AIT all four ORFs of the *RLP* cluster members were annotated in the same position at the genome, while the fact that three genes were expected to be deleted from this accession (Figure 37B).

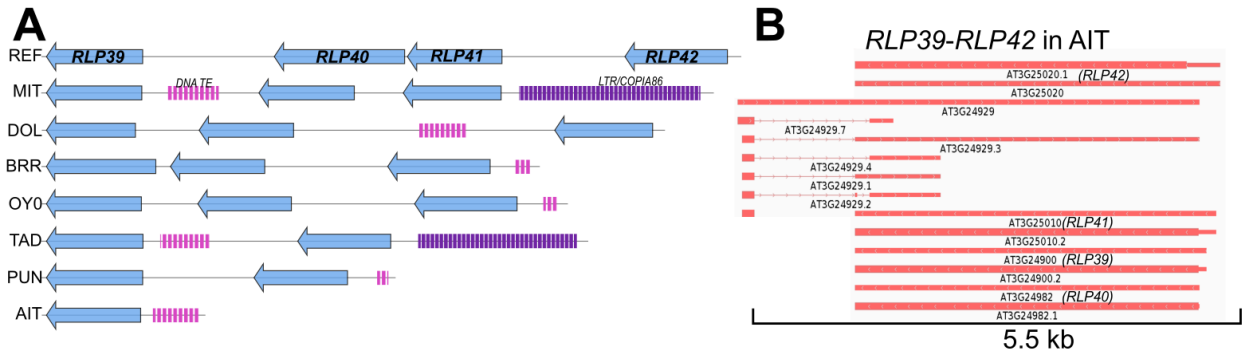


Figure 37. Organization of *RLP39-RLP42* gene cluster among the accessions. A – putative localization and size of *RLP* cluster genes and TEs. B – *Liftoff*-based annotation of similar *RLP* genes. RSH accession is not shown due to the assembly error. All the *RLP* cluster members are marked in blue. Non-RLP genes are not shown for simplicity.

To resolve complex variation in this cluster across the assemblies, its borders were first identified, by checking the location of non-variable genes at both flanks (*AT3G24880* and *AT3G25030*). Next, copy numbers for *RLP* genes of the cluster in eight accessions were retrieved from AthCNV catalogue [Zmienko 2020]. Finaly, *ab initio* predicted protein models within the region were used as queries against Arabidopsis reference protein set, and the names of best hits were assigned to them. Some consistency was present between the two resources (Table 15). For example, in MIT, a duplication in *RLP41* resulted in finding two corresponding proteins and the deleted *RLP40* was also missing from the predictions. The non-RLP protein was annotated as hypothetical protein. In DOL, the increased copy numbers for highly similar genes resulted in the duplicated hit of *RLP41* and one *RLP39*, while the deletion of the other two was confirmed by protein annotation. Consequently, the deletion of three genes out of four in AIT was equivalent to a single BLAST hit as well as to the smallest region size.

Apart from the protein-coding genes, the 5-kb TE of *LTR/COPIA86* order were found in MIT and TAD (Figure 37A). Both elements have the complete structure given for these

transposons and display 87% sequence identity when aligned to each other (MIT had a 700-bp insert in its coding part compared to TAD). The TE fragments from the DNA class were found in all the accessions, the fragment size ranged from 350 bp in PUN to 1.5 kb in MIT. The presence of numerous TE traces in the cluster and the variation in their distribution claim that the high variation may be associated with the TE activity at least partially.

Table 15. Copy numbers of subsequent *RLP* genes of the cluster retrieved from short reads and the assigned protein annotations in the assemblies.

Copy numbers by AthCNV					Best hits retrieved from BLAST in respective order				Size (kb)
RLP-	*39*	*40*	*41*	*42*					
MIT	1.2	0.7	3.7	2.7	RLP41	RLP39	RLP41	NP_187936	22.6
DOL	4	0.8	5.3	0.2	RLP41	RLP41	RLP39		21.5
BRR	0.2	0.5	2.3	1.3	RLP41	RLP41	RLP42		17.1
OY0	1.6	0.5	2.1	2.9	RLP41	RLP41	RLP42		18.1
TAD	3.3	0.3	1.2	0.1	RLP39	RLP41	NP_001328700		18.8
PUN	1.3	1.0	0.9	0.4	RLP41	RLP40			13.2
AIT	0.5	0.1	2.1	0.7	RLP41				6.1

*RSH is excluded from the analysis due to the assembly error in this location.
**Gene duplications are marked in green and and deletion in red, subsequently (see section 4.2.2).

On the contrary, the two BLAST hits of *RLP41* in BRR and OY0 did not reflect on its copy number estimate, that can be explained by its extremely high gene sequence identity with *RLP39.* Consequently, the cumulative copy number estimate for *RLP39* and *RLP41* in TAD (4.5) indicates the presence of two genes in one copy, the proteins for which were successfully annotated. This ilustrates that numerous erroneous mappings of short reads, known as a common problem in mapping-based approaches which rely on the single reference. Mapping reads to the reference genome resulted in discordant coverage and many SNPs that likely appeared because the lack of the adequate target in the reference sequence [Jaegle 2023, Marszalek-Zenczak 2023]. When compared with mapping to own genomes, it was observed that the reads mapped more evenly and there were few SNPs (Figure 38). Many SNPs were present exclusively in the location of *LTR/COPIA86_MIT*, in which an aforementioned insertion was found. On this example it is clearly seen that the assembly approach gave more information about the cluster structure.

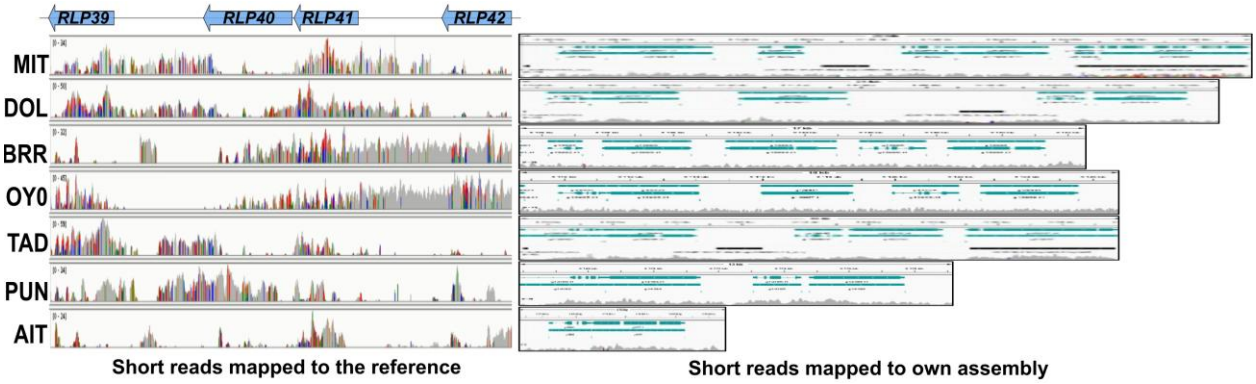


Figure 38. Comparison of reference-based approaches to assembly approach on the example of *RLP39-RLP42* gene cluster.

4.2.3.3.5. Complex variation in *RLP47-RLP50* gene cluster. *RLP47-RLP50* cluster is one of the largest clusters (42 kb in the reference genome) spanning three *RLP* genes: *RLP47* (*AT4G13810* locus), *RLP48* (*AT4G13880* locus) and *RLP50* (*AT4G13920* locus) and one pseudogene of *RLP47*, *RLP49* (due to their high similarity). The cluster also includes seven non-RLP protein-coding genes. The analysis of short read data mappings demonstrates the occurrence of complex variation events in this cluster (Figure 18B).

RLP50, localized on the right end of the cluster, is the most distinct from the other two (average sequence identity 58%) [Wang 2008], displays PAV pattern. BLAST scanning across 69 genomes revealed its complete deletion in 8 accessions (Altai-5, Are-10, Etna-2, Rab-R1, Sha, Sus-1, Toufl-1, Zin9) and partial deletion in one accession (700-bp deletion from gene start in Zal-1). Interestingly, *RLP50* was categorized as a dispensable gene in pangenome analysis [Lian 2024] (Figure 7). However, *RLP50* was present in all eight accessions analyzed here (Figure 39). Its protein sequence was quite conserved, except for the C-terminus in BRR, which deviated from the reference sequence. In addition, in PUN the *ab initio* predicted model merged RLP50 with the adjacent genomic sequence which led to the extension of the protein on N-terminus (1166 aa compared to 891 aa in RLP50_REF) and notably differed from the other RLP50 models.

RLP47, located at the left end of the cluster, was annotated in eight accessions. According to AthCNV data, its copy number varies in the population without the discrimination of copy number groups (Figure 18B). The unclear variation pattern is a result of of high similarity between *RLP47*, *RLP48*, and its pseudogene *RLP49*, as mentioned above. In AthCNV data, the copy number of *RLP47* was the highest in OY0 (6.5), suggesting its duplication. Indeed, BLAST search against *RLP47* gene sequence found one hit at Chromosome 1 absent from other accessions (Figure 39A). I found the hit analogous to OY0 in two more genomes from the set of 69 genomes by Lian and colleagues [2024] (Edi-0 abd JEA) and in three more accessions from the pangenome of 32 plants [Kang 2023b] (Kelsterbach-2, Pra-6, IP-Cum-1), which was an additional evidence for SV.

But overall BLAST results were very noisy as well as the copy number picture so I analysed this gene in more detail.

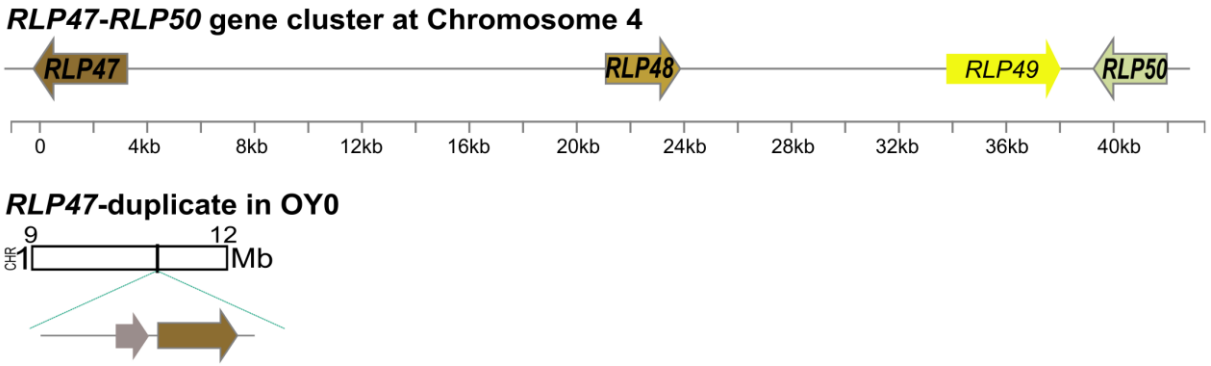


Figure 39. *RLP47-RLP50* gene cluster organization and the duplication of *RLP47* in OY0. The non-RLP genes are not shown for simplicity.

To find the respective proteins annotated in the assemblies, I used the genomic localizations of the predicted gene models equivalent to Liftoff-based positions of intervening genes and their distances to *RLPs*. I identified *RLP47*, *RLP48*, and *RLP49* genes in the expected positions, except OY0, in which *Liftoff* wrongly assigned the *RLP47* locus to *RLP49*. Moreover, *AUGUSTUS* predicted one or two valid ORFs at the expected localization of *RLP49*. One ORF (RLP49_a) was 795 to 1008 aa, and the other, present only in OY0, BRR, and DOL (RLP49_b) was 502 - 602 aa. Phylogenetic analysis of the proteins revealed high similarity of most RLP49_a sequences to RLP47s, while RLP49_b sequences were more related to RLP48s (Figure 40). Altogether, Altogether, high variation of the RLP49-like genomic sequences indicates that they may not be functional or essential for the plant. The duplicate of *RLP47* at Chromosome 1 in OY0 was also predicted to have protein-coding capacity, being most similar to RLP47_REF of all the analyzed proteins.

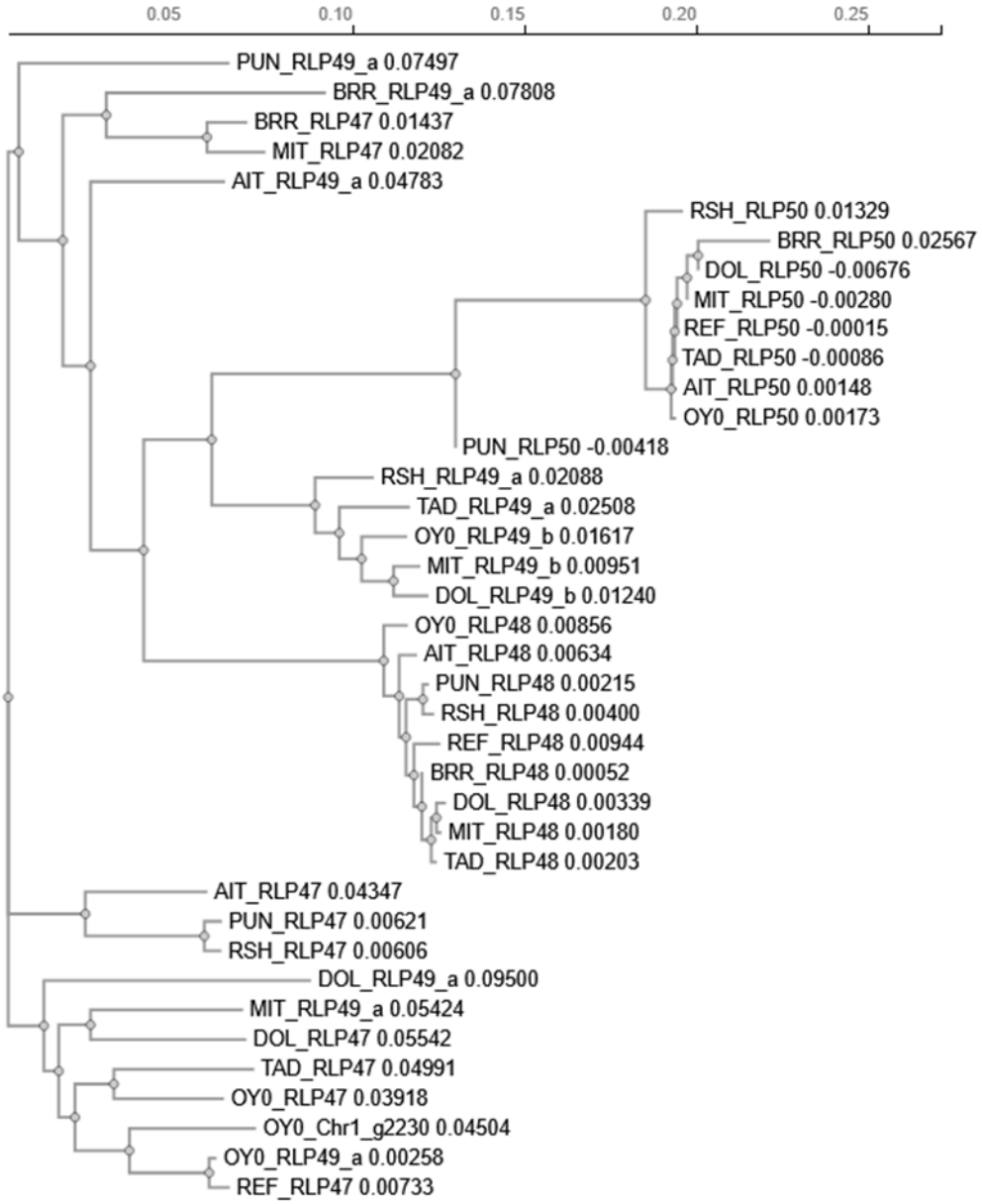


Figure 40. Phylogenetic relationships between the gene variants found in the cluster.

Besides intense variation altering genes, *RLP47-RLP50* cluster included two TE-derived insertions in BRR, which extended the region size to 55 kb (Figure 41). The first one, annotated as RNA transposon *LINE/L1* (Long interspersed nuclear element - 1), occupied almost 6 kb and localized between *RLP47* and *RLP48.* The second one, located between *RLP48* and *RLP50*, was annotated as *LTR/COPIA86* (Figure 41A). The detailed search of the stuctural elements revealed that both TEs were complete as they included all the elements typical for given TE orders but the ORFs contained many STOP codons which claimed that the TEs were not active. 5mC profiling confirmed these observations (over 90% on average for *LINE/L1* and over 80% on average for

LTR/COPIA86) (Figure 41B). The presence of polyA tail in *LINE/L1* together with its hypermethylation suggests that this insertion is evolutionary recent.

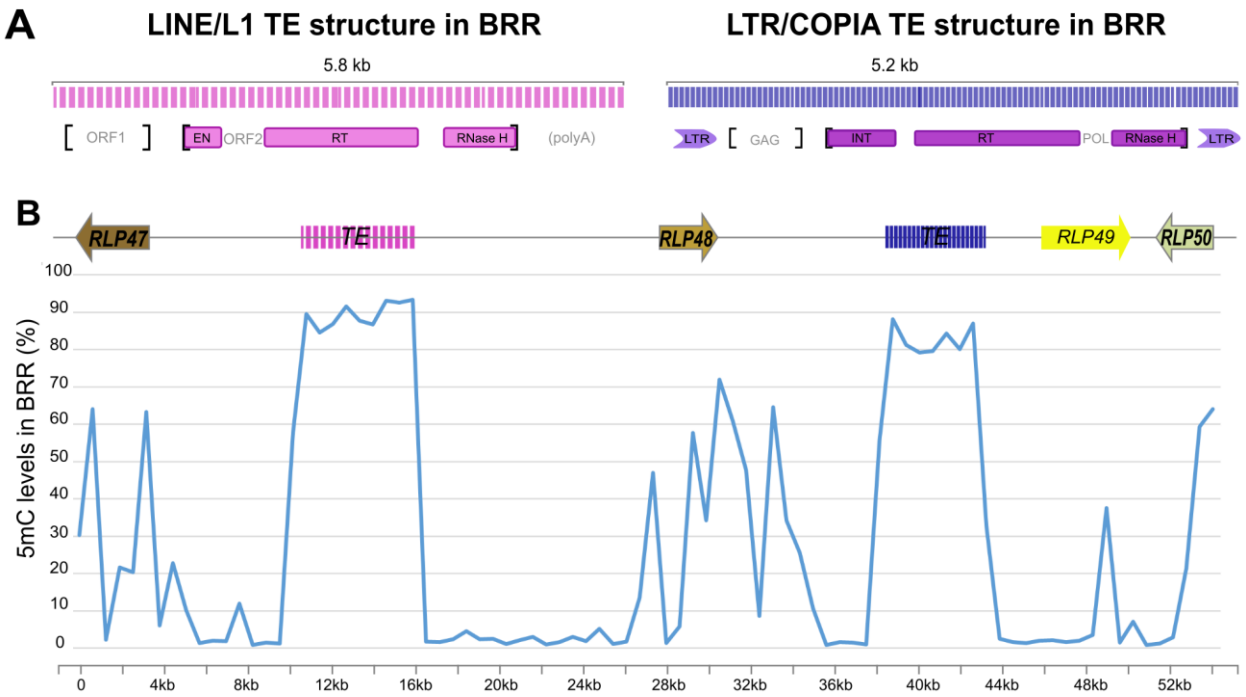


Figure 41. TE-derived insertions in *RLP47-RLP50* gene cluster in BRR accession. A – the structural elements of TEs within the cluster. B – 5mC level across the cluster in windows of 650 bp in length. Intervening genes are not shown. The found domains are: EN – endonuclease, RT – reverse transcriptase, INT – integrase.

Overall, sequence comparison of the assembled genomes revealed that *RLP47-RLP50* gene cluster displayed numerous variation events, at sequence level of individual gene models intraspecifically and at a genome scale. *RLP47* was found to possess a copy number distribution of 2 and 4, tending to duplications, while *RLP50* expressed PAV. *RLP49*, known as pseudogene, was found to include coding sequence likely being an active gene. In one accession TE-derived insertions were found between the genes, however they were highly methylated. These findings are promising in the context of potential role of gene variants within the cluster in plant adaptation. Associating of *RLP48* with abiotic stress responses by Stetter and colleagues [2015] strengthens the idea of a potential functional investigation into the role of *RLP47-RLP50* genes in plant adaptive mechanisms.

5. DISCUSSION

5.1. Benefits and challenges of using long read-based assemblies for characterizing complex structural variations

Aiming to tackle the novel long-read sequencing technique, in this study I prepared the high-quality genome assemblies of eight natural Arabidopsis accessions. Prior to our study, there were no genomic assemblies for the above accessions in the public databases, but throughout the project, *de novo* assembly of Oy-0 accession has been released in the larger *A. thaliana* pangenome study [Lian 2024]. The remaining accessions have not been sequenced so far. This assembly set will contribute to the broader initiative of a global collection of genomes representation for the upcoming 1001+ Genomes Project [The 1001 Genomes Plus Consortium 2024 biorXiv].

I designed the relatively simple and fast HMW DNA extraction protocol and obtained long and non-fragmented DNA of high purity (Appendix 1). The protocol describes the extraction of the total DNA from plant material by omitting nuclei isolation step, however this step proved to be not critical as the organellar DNA was filtered out during the preparation of genome assemblies. The procedure avoids work with hazardous reagents and demonstrates that careful handling and column-based extraction can result in minimal DNA fragmentation. The protocol was later tested on another model plant in our laboratory, *Medicago truncatula*, with similar results.

The DNA sequencing on MinION device generated high coverage (>90× for most accessions), reads of high average length (~12 kb, even with longest reads exceeding 500 kb), and consistent quality, highlighting the capacity of nanopore flow cells to produce data sets suitable for the downstream *de novo* assembly. When compared to another Arabidopsis single-flow-cell nanopore sequencing and genome assembly by Michael and colleagues [2018], our data sets demonstrated improvements in sequencing yield, read length, and overall quality, positioning as a promising resource for future genomic and comparative study. However, the fraction of short reads was still large, which can be explained by the fact that although the chemistry version R9 allows for obtaining ultra long reads, it also collects the signals from shorter reads that easier enter the nanopore active sites. For comparison, in the study published by Wang and colleagues [2022] in which the authors generated 55 Gb of data (more than twice over our largest yield for BRR) and enriched the final read set for the reads longer than 50 kb, they achieved substantially higher data quality which resulted in better assembly metrics. Yet, the decision to filter out so much data also

is dependent on the budget of the project and in the mentioned study the authors sequenced only one accession – Col-0.

It should be noted that the aim of this study was not to obtain ultra long reads (in this case it is not recommended to use column-based extraction, furthermore, short read elimination step may be added) but the set of sufficiently long reads in high amount that enables us to perform *de novo* assembly with the simultaneous application for many genomes. Considering high interest in resolving intraspecies variation [Saintenac 2011, Chia 2012, Fuentes 2019], the method tested on the genome of model organism has a chance to be broadened for genomes of other organisms, particularly the ones with the high genome size. The subsequent increase of read length and quality will facilitate the simplification of genome sequencing and will stimulate the projects of genome sequencing. Still, the assembly of the complete genome sequence requires collecting huge amounts of data and advanced bioinformatic analyses [Zhao 2020, Lian 2023, Kang 2023, Igolkina 2025], which limits the dissemination of long read sequencing for many laboratories. Therefore, the implementation of simplified solutions that preserve the benefits of using long-read sequencing with simultaneous reduction of costs and simplification of the methodology (as in this case) may prove to be a practical approach to opening the way to SV analysis on an even broader scale than is currently.

Next, *de novo* assembly was performed. Previously, *de novo* assemblies combined long reads with either short reads or linkage information to achieve high-quality chromosome-level sequences [Zapata 2016, Michael 2018]. I hypothesized that sufficiently high-coverage nanopore long-read data (~100× genome coverage) would alone be adequate to generate near-complete, chromosome-representing sequences. As no universal assembly strategy exists for my data amount and type, I first tested five different assembly algorithms on one accession. The two of them, *canu* and *flye*, showed comparable parameters in length and contig number so I generated parallel draft assemblies for each sample and processed them up to obtaining the final chromosome-representing sequences, to evaluate the performance of both in more detail and to keep the two versions for future verification tests.

I performed a comprehensive evaluation of the assemblies under three categories: contiguity, completeness and correctness. Evaluation of assembly contiguity revealed the variability between the assemblies. The number of contigs ranged widely, from 68 in BRRf to 324 in DOLf, while the longest contigs generally exceeded 15 Mb, with a few exceptions, where they were shorter (11 - 14 Mb). Despite such a broad range of contig numbers in drafts, the longest of them were placed in the chromosome arms locations. The three most fragmented samples were all assembled with *canu*. Looking at overall genome length, in most cases *flye* produced slightly

longer assemblies and higher contiguity metrics (Table 9). These results suggest that, while both assemblers were capable of producing sequences that represent large chromosomal parts, *flye* generally resulted in more contiguous assemblies.

The assembly completeness was evaluated at both gene and genome levels. BUSCO gene set indicated near-complete content of single-copy orthologs, with completeness scores of 98.8% - 98.9% for both *canu* and *flye* assemblies, and minimal differences between assembly methods or accessions. Gene-independent, k-mer completeness, was assessed with using short-read Illumina data of high overall accuracy and this parameter was also high (96.7% - 99.2% for contigs and 97.5% - 99% for scaffolds). Analysis of gene-poor repeat content revealed that *flye* assemblies captured a higher proportion of repetitive regions (~21%) compared with *canu* assemblies (~17%), which was consistent with their larger total assembly sizes, suggesting that assembly algorithm choice influences representation of repeat-rich regions. To evaluate the contents of the repetitive part, the repetitive elements of the genome, telomeres and centromeres, were checked. Telomeric repeats were generally better represented in *canu* assemblies, with larger and more consistent arrays of canonical 7-bp repeats at chromosome ends, whereas *flye* assemblies often underrepresented these sequences. In contrast, centromeric regions, composed largely of 178-bp repeats and ATHILA transposons, were more completely captured in *flye* assemblies, while *canu* assemblies showed less clustered centromeric repeats, likely reflecting local fragmentation. These observations indicate that, although nanopore-based *de novo* assembly can reconstruct chromosome arms effectively, repetitive regions remain challenging, and the assembler algorithm critically shapes the completeness and accuracy of repeat-rich genomic elements.

The assembly correctness was evaluated using both Assembly Quality as QV and Error Rate as E. QV scores, which reflect the probability of correct base calls, were generally higher in scaffolds than in contigs, suggesting that misassembled contigs were effectively excluded during scaffolding. Most datasets achieved QV ≥30, corresponding to >99.9% base accuracy, although TAD and PUN had lower values. Whole-genome sequence comparisons between *canu* and *flye* assemblies indicated that discordant regions, particularly large variants >1 kb, were concentrated in repeat-rich areas, especially centromeres. For comparison, in the mentioned study of Wang and colleagues, QV parameters were all higher than 62, but the data sources were originated from two long-read technologies, ONT and PacBio [2022]. In the Col-CEN genome assembly the QV scores were of 41 - 43 range, as there the nanopore reads were collected from both R9 and R10 chemistries [Naish 2021].

Although the constant improvement of nanopore technology is being done due to the increase of accuracy metrics with the newer chemistry versions, it should be clearly stated that

using nanopore-only data is not yet sufficient for accurate sequence comparisons. In the case of R9 chemistry utilized in this study, sequencing errors mostly originated from homopolymeric or short-repeat regions, though occasional long insertions are also observed [Delahaye 2021], which caused complications by frequently occurring discrepancies between *canu* and *flye* assembly versions. Remarkably, the discrepancies were often observed in BRR, which was the largest data set with the highest genome coverage by reads (176×, Table 9). The sequence accuracy remains the main problem of both nanopore and PacBio long read techniques, but in PacBio the introducing of the consensus high fidelity (PacBio HiFi) reads improves the accuracy of individual reads. Yet, nanopore technology would require additional data sources. It can be noted that on the long-read Arabidopsis genome sequencing studies published recently, nanopore-based ones are in minority [Naish 2021, Wang 2022, Lian 2024].

Despite limitations and complications, the genome assemblies I generated were acceptable for the identification of large structural genomic rearrangements. The evidence of this comes at least from the average assembly size, which was comparable with the size of the current reference genome version TAIR10 (Table 9). Also, the fragmented parts are accumulated in the centromeres (Figure 13) which were challenging to resolve in the reference genome and in the new reference assembly version [Naish 2021]. The *de novo* assemblies allowed for annotation of genes in reference-independent manner, which may be more complete, as the reference genome, no matter if its genome assembly is complete, still is the genome of a single plant. The assembly approach was especially beneficial for genes with highly similar sequences, as I observed in the case of *RLP39-RLP42* gene cluster, in which the mapping of short reads to the reference did not allow for identification of the region size but mapping the same reads to individual assemblies only confirmed region lengths (Figure 38). In the previous study from my group, the RNA-seq reads mapping to individual assemblies with high resolution allowing for characterization of tissue-specific expression of duplicated genes *BARS1* and *BARS2* [Marszalek-Zenczak 2023]. In this study I showed that expression of *altRLP2* is preserved even with the lack of *RLP3* though when both original *RLP2* and *RLP3* are present in the gene pair *RLP2-RLP3*, the expression of the former is silenced (Figure 31). These examples highlight the capacity to resolve the genomic regions that could not be properly aligned to the reference and were though as “dark” genes before [Sedlazeck 2018, Ebbert 2019]. To this end, constantly lowering costs and simplifying the methodology promises nanopore-based genomic assemblies to become a routine solution for studying such regions [Michael 2019]. In fact, one of the assemblies generated in this study, MIT, has been used to provide the evidence for a rare duplication of the acyltransferase gene *AT5G47980* in the thalianol biosynthesis cluster, while the analysis of DOL assembly in the same study helped to

reveal the accession-private deletion of *AT5G36130* gene in the tirucalladienol biosynthesis gene cluster [Marszalek-Zenczak 2023].

5.2. *RLP* genes display numerous variation patterns

SV is the pivotal source of genomic diversity in plant genomes, with the duplications, deletions, insertions and inversions altering protein-coding genes partially or entirely. These genomic rearrangements might affect gene function by changing its structure, gene expression or gene function. Notably, duplications can lead to the formation of novel genes, or pseudogenization, and gene deletions lead to function losses, which altogether contribute to the adaptability of plant genomes to environmental forces [Burssed 2022]. Here, I systematically characterized the variation of *RLP* gene family that includes the genes associated with plant development and defense responses, in which I expected to observe different variation events. As functions of many *RLPs* have not been characterized so far (Table 1), the description of their intraspecies variation events would lead to selecting potential candidates for future functional studies.

As many *RLP* genes encode receptors that recognize molecular patterns of primarily conserved structure, *RLP* genes were also considered as conserved ones, and many of them were even believed to be functionally redundant [Krujit 2005]. More recent studies highlighted the categorization of *RLP* genes onto two groups – development-associated and biotic-associated [Wang 2008], and these biotic-associated were thought to undergo diversification under selective forces, for example pathogens. In one survey, the authors showed little variability between the two groups but they analyzed publicly available data, mostly based on mapping to the reference genome [Steidele 2021]. Here, the combination of *Liftoff*-based gene transfer, *ab initio* predictions of gene models and encoded proteins with *AUGUSTUS*, and TE annotation in the individual genomes provided a comprehensive view on the variation of *RLPs*, especially on the structure of *RLP* gene pairs and clusters, with the patterns findable even on the small set of eight accessions. However, in case finding the variation patterns across population, the screening of a given variant was extended to the assembly set of 69 individuals [Lian 2024], and in all cases the potential variant was confirmed.

Indeed, development-associated *RLPs* were shown to be highly conserved among eight accessions. This conservation includes known examples like *TMM*, *CLV2* and *RLP44* all of which are localized outside gene pairs or clusters. Some clustered RLPs also showed strong conservation, for instance, *RLP20-RLP21* and *RLP34-RLP35* gene pairs. To my knowledge, the functional roles

of any of these genes remain unknown. In contrast, several non-clustered genes (e.g., *RLP53* and *RLP56*) displayed indels and sequence differences, which could simply reflect natural variation, as no major structural changes were detected. However, in the case of *RLP53*, a few additional clues may cause attention. In a recent study, rlp53 mutants showed high susceptibility to viral and fungal pathogens [Chen 2022]. Protein-protein interaction database indicates the interaction of RLP53 protein with two putative transmembrane proteins encoded by *AT1G67860* and *AT5G52420* (https://thebiogrid.org/). In addition, *RLP18* is annotated as a pseudogene of *RLP53* (https://www.araport.org/). These findings suggest that of non-clustered genes, *RLP53* is potential candidate for functional studies.

The *RLP* gene pairs displayed variation patterns such as PAV, gene duplication, and TE-derived insertion. The PAV was found to be driven by various mechanisms in the case of *RLP2-RLP3* gene pair and *RLP11-RLP12*, as the deletion of *RLP3* was followed by the formation of *altRLP2*, likely to preserve the locus functionality, while in the *RLP11-RLP12* gene pair no substantial alteration of *RLP12* was observed in the accessions with deleted *RLP11*. Another interesting case was the tandem duplication of the genomic region spanning *RLP30* and *AT3G05350* (its overlapping gene on plus strand), which was thought to be accession-private (see section 4.2.3.2.5). As *RLP30* encodes receptor recognizing PAMP derived from fungus *Sclerotinia sclerotiorum* [Zhang 2013], it would be interesting to compare its expression in response to Sclerotinia infection in the accessions with duplication and the single copy. Although I also found the duplication only in one accession, AthCNV data predict the occurrence of this duplication in at least 4 more accessions (Figure 18B), raising the possibility of its functional effect.

Complex variation was particularly dynamic in multi-gene clusters. *RLP13-RLP16* gene cluster was found to have accession-specific variation events (complete deletion of *RLP16* in Tanz-1 and TE-derived insertion in two accessions from Spain). *RLP24-RLP28* displayed variation events in a subset of accessions (Figure 36). *RLP39-RLP42*, in turn, displayed challenges in short-read mapping due to high sequence identity among cluster members, reinforcing the advantage of assembly-based approaches. Finally, *RLP47-RLP50* gene cluster exemplified extensive large-scale variation, including gene duplications, PAV, and TE insertions, with additional evidence that previously annotated *RLP49* may retain coding potential. Numerous variation events indicate their contribution to evolution of Arabidopsis adaptive responses to environmental challenges.

The most intense variability was observed for *RLP47-RLP50* gene cluster. All three *RLP* cluster members displayed variation patterns: *RLP47* was found to have duplication in 6 accessions and *RLP50* had PAV in a subset of accessions (see section 4.2.3.3.5). Additionally, RLP47 protein was found to interact with RLK (CRK33, CYSTEINE-RICH KINASE 33)

(https://thebiogrid.org/), which suggests possible formation of RLP-RLK complex similarly to RLP23-SOBIR1-BAK1 [Albert 2015] and RLP53-SOBIR1-BAK1 [Chen 2022]. Thus, *RLP47* is the next candidate for functional studies. *RLP48* likely possessed sequence variation yet in our data sets it was not possible to resolve due to high sequence similarities between *RLP48*, *RLP47* and *RLP49*. To note, *RLP48* was found to be involved in response of Arabidopsis roots to low concentration of phosphate, a survey conducted by Stetter and colleagues [2015], and these responses displayed natural variation among 166 accessions.

The *RLP* gene loci were denoted to their wealth in TE insertions. Two gene pairs, and three gene clusters contained TE-insertions between *RLP* genes, both of RNA and DNA TE class. The insertions of TEs happened between the genes in all cases except *RLP32* in TAD, in which the insert broke the coding sequence of the gene. RLP32 has recently been shown to function as PRR [Fan 2022] by recognizing proteobacterial PAMP (IF1) and by formation complexes with SOBIR1 and BAK1, similarly to RLP23-SOBIR1-BAK1 [Albert 2015]. Also, RLP32 has been associated with natural variation in response of Arabidopsis accessions to Ralstonia treatment. In *RLP47-RLP50* gene cluster the TE-derived insertions did not touch the structure of the genes, yet the presence of all the domains and even the preservation of polyA tail in *LINE/L1* insert (Figure 41) between *RLP47* and *RLP48* only strengthens the dynamic changes in this locus. As Arabidopsis is a plant with relatively small genome fraction of TEs, their abundance in relatively small gene family, which is *RLP*, states for their contribution to variation events.

DNA methylation at cytosines (5mC) is a well-known signature of TE silencing in many organisms. In plants, this modification helps keep TEs inactive and thereby protects genome integrity, and the regulatory system that maintains this silencing is considered highly stable throughout development [Hure 2025]. Beyond their own repression, TEs can also influence the expression of nearby genes through cis-regulatory effects [Quadrana 2016]. Having the raw nanopore signals was a great advantage to extract the signals of one epigenetic mark, 5mC, and check methylation levels in CG context in the TEs overlapping with *RLPs*. The findings indicated high methylation levels across all the TEs of high completeness (>80% on average 5mC level per TE of length 5 kb and longer). Although the gene methylation levels were not studied in this work, with these findings I wanted to highlight that the same data source for generation of genome assemblies and estimation of 5mC with high accuracy [Ni 2019] in specific locations across the genomes.

In summary, these results provide a comprehensive landscape of variability of *RLP* genes as a whole family. The results show that in contrast to previous hypothesis about gene conservation and, therefore, functional redundancy of *RLP* genes, this was true only for those genes associated

with plant development (Figure 19). These studies can lay a foundation for further functional studies inter alia checking for the expression of the varied *RLP* genes. The first candidates for functional studies are the two *RLP* gene clusters that display dynamic variation patterns. *RLP39-RLP42* is worth seeing for its expression in order to estimate the gene dosage effect of three genes that share sequence similarity and *RLP47-RLP50* needs to be evaluated for differential expression between the accessions that possess duplication of *RLP47* and the ones with one *RLP47* copy.

Overall, the integration of high-quality assemblies, gene annotation, and population-level CNV data demonstrates that *RLP* genes are altered by a spectrum of variation events. This stays in contrast with the long-hold "Zigzag" model of plant-pathogen interactions, where PRR role is restricted to detecting microbial PAMPs and thus they are considered to be highly conserved [Jones 2006]. Rather, it supports more and more acknowledged crosstalk between the PTI and ETI systems [Wu 2018, Lu 2021]. Of note, it has been pointed out, that plant cellular receptors, including RLPs may also be involved in more specific resistances, for example by recognizing non-cell penetrating apoplastic pathogens [Kanyuka 2019]. For such interactions, substantial level of plant receptor protein variation would be expected, due to the evolutionary arms race between plants and their pathogens. Non-clustered *RLP* genes are frequently conserved, often due to critical developmental and defense roles, whereas paired and clustered *RLPs* show substantial diversity, likely reflecting selective pressures associated with resistance or adaptation. These findings highlight the utility of *de novo* assembly approaches for resolving complex genomic regions and uncovering numerous SVs typical for plant receptor gene families.

6. CONCLUSIONS

1. A simple and cost-effective protocol for HMW DNA from Arabidopsis leaves enabled long read nanopore sequencing with high yield, by generation sequencing data sets suitable for *de novo* genome assembly.
2. Among several assemblers, *canu* and *flye* were proved to be the best draft assemblies, so two assembly versions were generated with each tool for each genome.
3. Gene-rich regions were assembled nearly completely, indicating reliable reconstruction of functional genomic elements across accessions. Telomeric and centromeric sequences were assembled only partially, with *canu* better resolving short telomeres and *flye* better capturing long centromeric repeats.
4. Comparison with published Oy-0 assemblies confirmed accuracy in gene-rich regions but reinforced difficulties in assembling repeat-dense regions; keeping independent *canu* and *flye* assemblies allowed better identification of problematic genomic areas.
5. The quality of sequence assemblies was not sufficient to resolve sequence variation of RLP genes at single nucleotide resolution.
6. Non-clustered *RLPs* showed limited variation, while the paired and clustered ones exhibited higher variability, including PAV and gene duplications, reflecting their dynamic evolution.
7. Analysis of *RLP* gene family indicated substantial level of their variability, especially the paired and the clustered ones, suggesting that the functional role of RLP proteins in plant immunity should be revisited.
8. *RLP* gene pairs displayed SVs including PAV, tandem duplication, and TE insertions. These events were sometimes accession-specific but could impact gene dosage and plant resistance traits. Some pairs like *RLP20-RLP21* and *RLP22-RLP23* were largely conserved.
9. *RLP* gene clusters showed high structural complexity and their intraspecies landscape contained deletions, duplications, and TE insertions. The genomic regions overlapping TE-insertions were strongly methylated.
10. The *RLP47-RLP50* cluster revealed complex variation events and even unexpected coding potential for *RLP49* (previously thought to be a pseudogene) and highlighted possible functional relevance of gene variants for plant adaptation.

BIBLIOGRAPHY

1. 1001 Genomes Consortium. (**2016**) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. The Cell. 166(2):481-491.
2. Abyzov A, Urban AE, Snyder M, Gerstein M. (**2011**) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21(6):974-984.
3. Albert I, Böhm H, Albert M, Feiler CE, Imkampe J, *et al*. (**2015**) An RLP23-SOBIR1-BAK1 complex mediates NLP-triggered immunity. Nat Plants. 1:15140.
4. Alkan C, Coe BP, Eichler EE. (**2011**) Genome structural variation discovery and genotyping. Nat Rev Genet. 12(5):363-376.
5. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, *et al*. (**2019**) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20:224.
6. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie M, *et al.* (**2020**) Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21(1):30.
7. Baumgarten A, Cannon S, Spangler R, May G. (**2003**) Genome-Level Evolution of Resistance Genes in *Arabidopsis thaliana*. Genetics. 165(1):309-319.
8. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, *et al.* (**2012**) Hi–C: A comprehensive technique to capture the conformation of genomes. Methods. 58(3):268-276.
9. Belyayev A, Kalendar R, Josefiová J, Paštová L, Habibi F, *et al*. (**2023**) Telomere sequence variability in genotypes from natural plant populations: unusual block-organized double-monomer terminal telomeric arrays. BMC Genomics. 24:572.
10. Benson G. (**1999**) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573-580.
11. Bleckmann A, Weidtkamp-Peters S, Seidel CA, Simon R. (**2010**) Stem cell signaling in Arabidopsis requires CRN to localize CLV2 to the plasma membrane. Plant Physiol. 152(1):166-76.
12. Bož̌a V, Brejová B, Vinař T. (**2017**) DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One. 12(6):e0178751.
13. Bridges CB. (**1936**) The Bar "Gene" a Duplication. Science. 83(2148):210-211.
14. Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. (**2022**) Mechanisms of structural chromosomal rearrangements formation. Mol Cytogenet. 15(1):23.
15. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al.* (**2009**) BLAST+: architecture and applications. BMC Bioinformatics. 10:421.
16. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, *et al.* (**2011**) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 43(10):956-963.
17. Carter NP. (**2007**) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 39(7 Suppl):S16-21.
18. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, *et al*. (**2015**) The BioGRID interaction database: 2015 update. Nucleic Acids Res. 43(Database issue):D470-8.
19. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, *et al.* (**2009**) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 6(9):677-681.
20. Chen R, Sun P, Zhong G, Wang W, Tang D. (**2022**) The RECEPTOR-LIKE PROTEIN53 immune complex associates with LLG1 to positively regulate plant immunity. J Integr Plant Biol. 64(9):1833-1846.
21. Chen S, Zhou Y, Chen Y, Gu J. (**2018**) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884-i890.

22. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. (**2021**) Accurate long-read *de novo* assembly evaluation with Inspector. Genome Biol. 22(1):312.
23. Cheng C, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, *et al*. (**2017**) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 89(4):789-804.
24. Cheng H, Qu H, McKenzie S, Lawrence KR, Windsor R, *et al*. (**2025**) Efficient near telomere-to-telomere assembly of Nanopore Simplex reads. BioRxiv. doi: https://doi.org/10.1101/2025.04.14.648685.
25. Chia J-M, Bradbury PJ, Costich D, de Leon N, Doebley J, *et al*. (**2012**) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 44:803-807.
26. Cho H, Lee J, Oh E. (**2023**) Leucine-Rich Repeat Receptor-Like Proteins in Plants: Structure, Function, and Signaling. J Plant Biol. 66:99-107.
27. Clauw P, Coppens F, Korte A, Herman D, Slabbinck B, *et al*. (**2016**) Leaf Growth Response to Mild Drought: Natural Variation in Arabidopsis Sheds Light on Trait Architecture. The Plant Cell. 28(10):2417-2434.
28. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, *et al.* (**2007**) Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana.* Science. 317(5836):338-342.
29. Corpet F. (**1998**) Multiple sequence alignment with hierarchical clustering. Nucl Acid Res. 16(22):10881-10890.
30. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, *et al*. (**2017**) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 8:1326.
31. Cross JM, von Korff M, Altmann T, Bartzetko L, Sulpice R, *et al*. (**2006**) Variation of Enzyme Activities and Metabolite Levels in 24 Arabidopsis Accessions Growing in Carbon-Limited Conditions. Plant Physiol. 142(4):1574-1588.
32. Cui H, Tsuda K, Parker JE. (**2015**) Effector-triggered immunity: from pathogen perception to robust defense. Annu Rev Plant Biol. 66:487-511.
33. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. (**2018**) NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 34(15):2666-2669.
34. Delahaye C, Nicolas J. (**2021**) Sequencing DNA with nanopores: Troubles and biases. PLoS One. 16(10):e0257521.
35. Deshpande AS, Ulahannan N, Pendleton M, Dai X, Ly L, *et al*. (**2022**) Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol. 40:1488-1499.
36. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, *et al*. (**2013**) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21.
37. Doody E, Zha Y, He J, Poethig RS. (**2022**) The genetic basis of natural variation in the timing of vegetative phase change in *Arabidopsis thaliana*. Development. 149(10):dev:200321.
38. Doshi R, Kinnear E, Chatterjee S, Guha P, Liu Q. (**2025**) Reliable investigation of DNA methylation using Oxford nanopore technologies. Sci Rep. 15:15900.
39. Doyle JJ, Doyle JL. (**1987**) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bulletin. 19(1): 11-15.
40. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, *et al*. (**2017**) African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 114(20):5213-5218.
41. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, *et al*. (**2019**) Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 20(1):97.

42. Elliott L, Kalde M, Schürholz AK, Zhang X, Wolf S, *et al.* (**2024**) A self-regulatory cell-wall-sensing module at cell edges controls plant growth. Nat Plants. 10:483-493.
43. Fan L, Fröhlich K, Melzer E, Pruitt RN, Albert I, *et al.* (**2022**) Genotyping-by-sequencing-based identification of *Arabidopsis* pattern recognition receptor RLP32 recognizing proteobacterial translation initiation factor IF1. Nat Commun. 13(1):1294.
44. Feschotte C, Wessler SR. (**2001**) Treasures in the attic: Rolling circle transposons discovered in eukaryotic genomes. Proc. Natl. Acad. Sci. U.S.A. 98(16):8923-8924.
45. Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, *et al*. (**2019**) Structural variants in 3000 rice genomes. Genome Res. 29(5):870-880.
46. Fulgione A, Koornneef M, Roux F, Hermisson J, Hancock AM. (**2018**) Madeiran *Arabidopsis thaliana* Reveals Ancient Long-Range Colonization and Clarifies Demography in Eurasia. Mol Biol Evol. 35(3):564-574.
47. Fultz D, McKinlay A, Enganti R, Pikaard CS. (**2023**) Sequence and epigenetic landscapes of active and silent nucleolus organizer regions in Arabidopsis. Sci Adv. 9(44):eadj4509.
48. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, *et al*. (**2011**) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature. 477(7365): 419-423.
49. Geisler M, Nadeau J, Sack FD. (**2000**) Oriented asymmetric divisions that generate the stomatal spacing pattern in arabidopsis are disrupted by the *too many mouths* mutation. The Plant Cell. 12(11):2075-2086.
50. Goel M, Sun H, Jiao WB, Schneeberger K. (**2019**) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 20(1):277.
51. Göktay M, Fulgione A, Hancock AM. (**2021**) A New Catalog of Structural Variants in 1,301 *A. thaliana* Lines from Africa, Eurasia, and North America Reveals a Signature of Balancing Selection at Defense Response Genes. Mol Biol Evol. 38(4):1498-1511.
52. González R, Butković A, Rivarez MPS, Elena SF. (**2020**) Natural variation in *Arabidopsis thaliana* rosette area unveils new genes involved in plant development. Sci. Rep. 10(1):17600.
53. Gurevich A, Saveliev V, Vyahhi N, Tesler G. (**2013**) QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072-1075.
54. Haghshenas E, Asghari H, Stoye J, Chauve C, Hach F. (**2020**) HASLR: Fast Hybrid Assembly of Long Reads. iScience. 23(8):101389.
55. Hailemariam S, Liao CJ, Mengiste T. (**2024**) Receptor-like cytoplasmic kinases: orchestrating plant cellular communication. Trends Plant Sci. 29(10):1113-1130.
56. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, *et al*. (**2015**) Large multiallelic copy number variations in humans. Nat Genet. 47(3):296-303.
57. Ho SS, Urban AE, Mills RE. (**2020**) Structural variation in the sequencing era. Nat Rev Genet. 21(3):171-189.
58. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, *et al*. (**2012**) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet. 44(2):212-216.
59. Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. (**2022**) A near-complete assembly of an *Arabidopsis thaliana* genome. Mol Plant. 15(8):1247-1250.
60. Hure V, Piron-Prunier F, Yehouessi T, Vitte C, Kornienko AE, *et al*. (**2025**) Alternative silencing states of transposable elements in Arabidopsis associated with H3K27me3. Genome Biol. 26(1):11.
61. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, *et al.* (**2004**) Detection of large-scale variation in the human genome. Nat Genet. 36(9):949-951.
62. Igolkina AA, Vorbrugg S, Rabanal FA, Liu H-J, Ashkenazy H, Kornienko AE, *et al*. (**2025**) A comparison of 27 *Arabidopsis thaliana* genomes and the path toward an unbiased characterization of genetic polymorphism. Nat Genet. Aug 19 (Published online).

63. International Human Genome Sequencing Consortium. (**2001**) Initial sequencing and analysis of the human genome. Nature. 409:860-921.
64. Jaegle B, Pisupati R, Soto-Jiménez LM, Burns R, Rabanal FA, *et al.* (**2023**) Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. Genome Biol. 24(1):44.
65. Jain C, Rhie A, Hansen NF, Koren S, Phillippy A. (**2022**) Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods. 19:705-710.
66. Jain M, Olsen HE, Paten B, Akeson M. (**2016**) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17(1):239.
67. Jameson N, Georgelis N, Fouladbash E, Martens S, Hannah LC, *et al.* (**2008**) *Helitron* mediated amplification of cytochrome P450 monooxygenase gene in maize. Plant Mol Biol. 67(3):295-304.
68. Jamieson PA, Shan L, He P. (**2018**) Plant cell surface molecular cypher: Receptor-like proteins and their roles in immunity and development. Plant Sci. 274:242-251.
69. Jehle AK, Lipschis M, Albert M, Fallahzadeh-Mamaghani V, Fürst U, *et al.* (**2013**) The Receptor-Like Protein ReMAX of *Arabidopsis* Detects the Microbe-Associated Molecular Pattern eMax from *Xanthomonas*. Plant Cell. 25(6):2330-2340.
70. Jeong S, Trotochaud AE, Clark SE. (**1999**) The Arabidopsis CLAVATA2 gene encodes a receptor-like protein required for the stability of the CLAVATA1 receptor-like kinase. Plant Cell. 11(10):1925-1934.
71. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, *et al*. (**2020**) Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 21:189.
72. Jiao WB, Schneeberger K. (**2020**) Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 11(1):989.
73. Joly-Lopez Z, Bureau TE. (**2014**) Diversity and evolution of transposable elements in Arabidopsis. Chromosome Res. 22:203-216.
74. Jones J, Dangl J. (**2006**) The plant immune system. Nature. 444:323-329.
75. Kang M, Chanderbali A, Lee S, Soltis DE, Soltis PS, *et al.* (**2023**a) High-molecular-weight DNA extraction for long-read sequencing of plant genomes: An optimization of standard methods. Appl Plant Sci. 11(3):e11528.
76. Kang M, Wu H, Liu H, Liu W, Zhu M, *et al.* (**2023**b) The pan-genome and local adaptation of *Arabidopsis thaliana*. Nat Commun. 14(1):6259.
77. Kanyuka K, Rudd JJ. (**2019**) Cell surface immune receptors: the guardians of the plant's extracellular spaces. Curr Opin Plant Biol. 50:1-8.
78. Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, *et al;* 1001 Genomes Consortium. (**2016**) Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions. Cell. 166(2):492-505.
79. Kim J, Jang H, Huh SM, Cho A, Yim B, *et al.* (**2024**) Effect of structural variation in the promoter region of *RsMYB1.1* on the skin color of radish taproot. Front Plant Sci. 14:1327009.
80. Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. (**2016**) A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. Plant J. 88(6):1058-1070.
81. Kolmogorov M, Yuan J, Lin Y, Pevzner P. (**2019**) Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5): 540-546.
82. Koornneef M, Alonso-Blanco C, Vreugdenhil D. (**2004**) Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol. 55:141-172.
83. Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, *et al.* (**2024**) Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. Genome Res. 34(11):1919-1930.

84. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.* (**2017**) Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. 27(5): 722-736.
85. Kruijt M, DE Kock MJ, de Wit PJ. (**2005**) Receptor-like proteins involved in plant disease resistance. Mol Plant Pathol. 6(1):85-97.
86. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C. *et al*. (**2012**) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40(Database issue):D1202-1210.
87. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, *et al.* (**2015**) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 3:1-8.
88. Leger A, Leonardi T. (**2019**) PycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software. 4(34):1236.
89. Li H. (**2016**) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 32(14):2103-2110.
90. Li H. (**2018**) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094-3100.
91. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* (**2009**) 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078-2079.
92. Li Z, Parris S, Saski CA. (**2020**) A simple plant high-molecular-weight DNA extraction method suitable for single-molecule technologies. Plant Methods. 16:38.
93. Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C. *et al*. (**2024**) A pan-genome of 69 Arabidopsis *thaliana* accessions reveals a conserved genome structure throughout the global species range. Nat Genet. 56:982-991.
94. Liang SC, Hartwig B, Perera P, Mora-García S, de Leau E, *et al*. (**2015**) Kicking against the PRCs – A Domesticated Transposase Antagonises Silencing Mediated by Polycomb Group Proteins and Is an Accessory Component of Polycomb Repressive Complex 2. PLoS Genet. 11(12):e1005660.
95. Liu P, Cuerda-Gil D, Shahid S, Slotkin RK. (**2022**) The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond. Annu Rev Genet. 30(56):63-87.
96. Loman NJ, Quick J, Simpson JT. (**2015**) A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat Methods. 12(8):733-735.
97. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, *et al*. (**2013**) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat Genet. 45:884-890.
98. Lu Y, Tsuda K. (**2021**) Intimate Association of PRR and NLR-Mediated Signaling in Plant Immunity. MPMI. 34(1):3-14.
99. Lye ZN, Purugganan MD. (**2019**) Copy Number Variation in Domestication. Trends Plant Sci. 24(4):352-365.
100. Manchanda N, Portwood JL, Woodhouse MR, Seetharam AS, Lawrence-Dill CJ, *et al*. (**2020**) GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics. 21:193.
101. Madeira F, Madhusoodanan N, Lee J, Eusebi A, Niewielska A, *et al*. (**2024**) The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. Nucleic Acids Research. 52(W1):W521-W525.
102. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, *et al*. (**2018**) MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 14(1):e1005944.
103. Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, *et al*. (**2023**) Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. Front Plant Sci. 14:1104303.

104. Marx V. (**2023**) Method of the year: long-read sequencing. Nat Methods. 20(1):6-11.
105. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, *et al*. (**2016**) Extraction of High-Molecular-Weight Genomic DNA for Long-Read Sequencing of Single Molecules. Biotechniques. 61(4):203-205.
106. McGinty SP, Kaya G, Sim SB, Corpuz RL, Quail MA, *et al*. (**2025**) CiFi: Accurate long-read chromatin conformation capture with low-input requirements. BiorXiv. DOI: https://doi.org/10.1101/2025.01.31.635566.
107. McVey M, Lee SE. (**2008**) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 24(11):529-38.
108. Metzker M. (**2010**) Sequencing technologies – the next generation. Nat Rev Genet. 11(1):31-46.
109. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, *et al*. (**2018**) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 9(1):541.
110. Moore SC, Penrice-Randal R, Alruwaili M, Randle N, Armstrong S, *et al*. (**2020**) Amplicon-Based Detection and Sequencing of SARS-CoV-2 in Nasopharyngeal Swabs from Patients With COVID-19 and Identification of Deletions in the Viral Genome That Encode Proteins Involved in Interferon Antagonism. Viruses. 12(10):1164.
111. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, *et al*. (**2021**) The genetic and epigenetic landscape of the Arabidopsis centromeres. Science. 374(6569):eabi7489.
112. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, *et al*. (**2009**) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 461:1130-1134.
113. Nattestad M, Schatz MC. (**2016**) Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 32(19):3021-3023.
114. Ni P, Huang N, Zhang Z, Wang DP, Liang F, *et al*. (**2019**) DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics. 35(22):4586-4595.
115. Noé L, Kucherov G. (**2005**) YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 33(Web Server issue):W540-543.
116. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, *et al*. (**2022**) The complete sequence of a human genome. Science. 376(6588):44-53.
117. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, *et al*. (**2008**) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res. 18(12):2024-2033.
118. Padmanabhan M, Cournoyer P, Dinesh-Kumar SP. (**2009**) The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. Cell Microbiol. 11(2):191-198.
119. Pasha A, Subramaniam S, Cleary A, Chen X, Berardini T, *et al*. (**2020**) Araport Lives: An Updated Framework for Arabidopsis Bioinformatics. The Plant Cell. 32:(9)2683-2686.
120. Plantegenet S, Weber J, Goldstein DR, Zeller G, Nussbaumer C, *et al*. (**2009**) Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance. Mol Syst Biol. 5:242.
121. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, *et al*. (**2010**) The Scale of Population Structure in Arabidopsis thaliana. PLoS Genet. 6(2):e1000843.
122. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. (**2020**) Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics. 70(1):e102.
123. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, *et al*. (**2019**) A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLoS One. 14(5):e0216233.
124. Pucker B, Irisarri I, de Vries J, Xu B. (**2022**) Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quantitative Plant Biology. 3:e5.

125. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, *et al*. (**2016**) The Arabidopsis thaliana mobilome and its impact at the species level. eLife. 5:e15716.
126. Quesneville H. (**2020**) Twenty years of transposable element analysis in the Arabidopsis thaliana genome. Mobile DNA. 11:28.
127. Rabanal FA, Gräff M, Lanz C, Fritschi K, Llaca V, *et al*. (**2022**) Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. Nucleic Acids Res. 50(21):12309-12327.
128. Rang FJ, Kloosterman WP, de Ridder J. (**2018**) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19(1):90.
129. Rashid U, Wu C, Shiller J, Smith K, Crowhurst R, *et al*. (**2024**) AssemblyQC: a Nextflow pipeline for reproducible reporting of assembly quality. Bioinformatics. 40:(8):btae477.
130. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, *et al*. (**2023**) Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 41:1474-1482.
131. Rhie A, Walenz BP, Koren S, Phillippy AM. (**2020**) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21(1):245.
132. Rhoads A, Au KF. (**2015**) PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 13(5):278-289.
133. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, *et al*. (**2011**) Integrative Genomics Viewer. Nat Biotechnol. 29:24-26.
134. Rowe HC, Kliebenstein DJ. (**2008**) Complex Genetics Control Natural Variation in Arabidopsis thaliana Resistance to Botrytis cinerea. Genetics. 180(4):2237-2250.
135. Ruan J, Li H. (**2020**) Fast and accurate long-read assembly with wtdbg2. Nat Methods. 17:155-158.
136. Saintenac C, Jiang D, Akhunov ED. (**2011**) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 12:R88.
137. Sameer AS, Nissar S. (**2021**) Toll-Like Receptors (TLRs): Structure, Functions, Signaling, and Role of Their Polymorphisms in Colorectal Cancer Susceptibility. Biomed Res Int. 2021:1157023.
138. Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, *et al*. (**2023**) Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. Microb Genom. 9(1):mgen000910.
139. Sanger F, Nicklen S, Coulson AR. (**1977**) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74(12):5463-5467.
140. Santuari L, Pradervand S, Amiguet-Vercher AM, Thomas J, Dorcey E, *et al*. (**2010**) Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. Genome Biol. 11:R4.
141. Satyr A, Żmieńko A. (**2020**) Sekwencjonowanie nanoporowe i jego zastosowanie w biologii. Postępy Biochemii. 66(3):193-204.
142. Saxena RK, Edwards D, Varshney RK. (**2014**) Structural variations in plant genomes. Brief Funct Genomics. 13(4):296-307.
143. Schmalenbach I, Zhang L, Ryngajllo M, Jiménez-Gómez JM. (**2014**) Functional analysis of the Landsberg erecta allele of FRIGIDA. BMC Plant Biol. 14:218.
144. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, *et al*. (**2011**) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proc Natl Acad Sci U S A. 108(25): 10249-10254.
145. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. (**1993**) Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping. Science. 262(5130):110-114.

146. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, *et al*. (**2018**) Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 15(6):461-468.
147. Sereika M, Kirkegaard RH, Karst SM, Michaelsen T, Sørensen EA, *et al*. (**2022**) Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 19:823-826.
148. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, *et al*. (**2020**) Nanopore sequencing and the Shasta toolkit enables efficient de novo assembly of eleven human genomes. Nat Biotechnol. 38(9):1044-1053.
149. Shakirov EV & Shippen D. (**2004**) Length Regulation and Dynamics of Individual Telomere Tracts in Wild-Type Arabidopsis. The Plant Cell. 16:1959-1967.
150. Shao ZQ, Xue JY, Wu P, Zhang YM, Wu Y, *et al*. (**2016**) Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. Plant Physiology. 170(4):2095-2109.
151. Shen Y, Diener AC. (**2013**) Arabidopsis thaliana RESISTANCE TO FUSARIUM OXYSPORUM 2 Implicates Tyrosine-Sulfated Peptide Signaling in Susceptibility and Resistance to Root Infection. PLoS Genet. 9(5): e1003525.
152. Shumate A, Salzberg S. (**2021**) Liftoff: accurate mapping of gene annotations. Bioinformatics. 37(12):1639-1643.
153. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (**2015**) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210-3212.
154. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, *et al*. (**2017**) Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 14(4):407-410.
155. Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, *et al*. (**2007**) Boron-Toxicity Tolerance in Barley Arising from Efflux Transporter Amplification. Science. 318(5855):1446-1449.
156. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. (**1999**) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature. 400(6745): 667-671.
157. Stanke M, Schöffmann O, Morgenstern B, Waack S. (**2006**) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 7:62.
158. Steidele CE, Stam R. (**2021**) Multi-omics approach highlights differences between RLP classes in Arabidopsis thaliana. BMC Genomics. 22(1):557.
159. Stetter MG, Schmid K, Ludewig U. (**2015**) Uncovering Genes and Ploidy Involved in the High Diversity in Root Hair Density, Length and Response to Local Scarce Phosphate in Arabidopsis thaliana. PLoS One. 10(3): e0120604.
160. Tan X, Meyers BC, Kozik A, West MAL, Morgante M, *et al*. (**2007**) Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis. BMC Plant Biol. 7:56.
161. Tao Y, Xian W, Bao Z., Rabanal FA, Movilli A, *et al*. (**2024**) Atlas of telomeric repeat diversity in Arabidopsis thaliana. Genome Biol. 25(1):244.
162. The 1001 Genomes Plus Consortium. (**2024**) The 1001G+ project: A curated collection of Arabidopsis thaliana long-read genome assemblies to advance plant research. BiorXiv. Published 29 December 2024. DOI: 10.1101/2024.12.23.629943.
163. The Arabidopsis Genome Initiative. (**2000**) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 408:796-815.
164. Uauy C, Nelissen H, Chan RL, Napier JA, Seung D, *et al*. (**2025**) Challenges of translating Arabidopsis insights into crops. The Plant Cell. 37(5):koaf059.

165. Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, *et al*. (**2019**) A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. Cell. 178(5):1260-1272.e14.
166. Vaser R, Sović I, Nagarajan N, Šikić M. (**2017**) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5):737-746.
167. Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, *et al*. (**2007**) Epigenetic Natural Variation in Arabidopsis thaliana. PLoS Biol. 5(7):e174.
168. Varshney RK, Barmukh R, Bentley A, Nguyen HT. (**2024**) Exploring the genomics of abiotic stress tolerance and crop resilience to climate change. Plant Genome. 17(1):e20445.
169. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, *et al*. (**2018**) SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 28(4):581-591.
170. Wang B, Yang X, Jia Y, Xu Y, Jia P *et al*. (**2022**) High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. Genomics Proteomics Bioinformatics. 20(1):4-13.
171. Wang G, Ellendorff U, Kemp B, Mansfield JW, Forsyth A, *et al*. (**2008**) A Genome-wide functional investigation into the roles of receptor-like proteins in Arabidopsis. Plant Physiol. 147(2):503–517.
172. Wang Y, Xiong G, Hu J, Jiang L, Yu H, *et al*. (**2015**) Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nat Genet. 47(8):944-948.
173. Wang Y, Zhao Y, Bollas A. Wang Y, Au KF. (**2021**) Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 39(11):1348-1365.
174. Weigel D. (**2012**) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiol. 158(1):2-22.
175. Weigel D, Mott R. (**2009**) The 1001 genomes project for Arabidopsis thaliana. Genome Biol. 10(5):107.
176. Wicker T, Sabot F, Hua-Van A.Bennetzen JL, Capy P, *et al*. (**2007**) A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973-982.
177. Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, *et al*. (**2023**) Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature. 618(7965):557-565.
178. Wolf S, van der Does D, Ladwig F, Sticht C, Kolbeck A, *et al*. (**2014**) A receptor-like protein mediates the response to pectin modification by activating brassinosteroid signaling. Proc Natl Acad Sci U S A. 111(42):15261-15266.
179. Wu C-H, Derevnina L, Kamoun S. (**2021**) Receptor networks underpin plant immunity. Science. 360(6395):1300-1301.
180. Xiao H, Jiang N, Schaffner E, Stockinger EJ, Van der Knaap E. (**2008**) A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. Science. 319:1527-1530.
181. Yan FH, Zhang LP, Cheng F, Yu DM, Hu JY. (**2020**) Accession-specific flowering time variation in response to nitrate fluctuation in Arabidopsis thaliana. Plant Divers. 43(1):78-85.
182. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. (**2009**) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 25(21):2865-2871.
183. Yuen ZWS, Srivastava A, Daniel R, McNevin D, Cameron J, *et al*. (**2021**) Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat Commun. 12:3438.
184. Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, *et al*. (**2016**) Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci U S A. 113(28):E4052-4060.

185. Zhang L, Kars I, Essenstam B, Liebrand TWH, Wagemakers L, *et al*. (**2014**) Fungal Endopolygalacturonases Are Recognized as Microbe-Associated Molecular Patterns by the Arabidopsis Receptor-Like Protein RESPONSIVENESS TO BOTRYTIS POLYGALACTURONASES1. Plant Physiol. 164(1):352-364.
186. Zhang W, Fraiture M, Kolb D, Löffelhardt B, Desaki Y, *et al*. (**2013**) Arabidopsis RECEPTOR-LIKE PROTEIN30 and Receptor-Like Kinase SUPPRESSOR OF BIR1-1/EVERSHED Mediate Innate Immunity to Necrotrophic Fungi. The Plant Cell. 25(10):4227-4241.
187. Zhang Y, Yang Y, Fang B, Gannon P, Ding P, *et al*. (**2010**) Arabidopsis snc2-1D Activates Receptor-Like Protein-Mediated Immunity Transduced through WRKY70. The Plant Cell. 22(9):3153-3163.
188. Zhang Z, Mao L, Chen H, Bu F, Li G, *et al*. (**2015**) Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. The Plant Cell. 27(6):1595-1604.
189. Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, *et al*. (**2020**) AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome. The Plant Cell. 32(6):1797-1819.
190. Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. (**2014**) Copy number polymorphism in plant genomes. Theor Appl Genet. 127(1):1-18.
191. Zmienko A, Samelak-Czajka A, Kozlowski P, Szymanska M, Figlerowicz M. (**2016**) Arabidopsis thaliana population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. BMC Genomics. 17(1):893.
192. Zou YP, Hou XH, Wu Q, Chen JF, Zi-Wen Li ZW, *et al*. (**2017**) Adaptation of Arabidopsis thaliana to the Yangtze River basin. Genome Biol. 18(1):239.

SUPPLEMENTARY DATA

Appendix 1. HMW DNA extraction from plant leaves.

General remarks :

- Use at least 1 g of material for efficient extraction.
- Before starting, heat CTAB buffer to 67 °C.
- Use wide-bore tips whenever it is possible.

1. Homogenize 2 g of plant material with mortar and pestle in liquid nitrogen.
2. Pour CTAB lysis buffer, which was previously heated to 65 °C. The proportion of buffer to the material should be: 400 µl / 100 mg of material.
3. Portion out the lysis solution into 2 ml tubes.
4. Add RNase A in proportion 1 mg / 100 mg of material. Mix by flicking the tube gently.
5. Put the solution in the incubator, gently shaking, and incubate at 65 °C.
6. After 30 mins of incubation, add Proteinase K in proportion 0.8 mg / 1 ml of the solution. Incubate 30 mins more. The general time of incubation is: 60 min.
7. Centrifugate the lysate 5 min in 20,000g at Room Temperature.
8. (Optional) Prepare filling buffer (120 mmol MOPS): weigh 2.51 g of MOPS and dissolve it in 100 ml of sterile water.
9. Transfer the supernatant to the new tube. Use wide-bore tips, however, try not to touch the cell debris (precipitate). If the supernatant contains wisps, pass it across the DNA Shredder column (QIAGEN). If the supernatant is dense and intensively green, dilute it with the filling buffer from step 8 by adding 0.5 vol of filling buffer to DNA-containing supernatant. The clear and liquid supernatant is now ready for HMW DNA extraction.
10. Equilibrate the Genomic-tip 20/G column (QIAGEN) with 1 ml of QBT Buffer.
11. Load the solution on the column in 1-ml portions. After each portion wait till the solution penetrates the column totally.
12. Wash the column with 1 ml of QC Buffer and discard the flow-through. Repeat the washing step 3 times.
13. (Optional) Preheat the QF Buffer to 50 °C, to make the elution more effective.
14. Elute HMW DNA with 1 ml QF Buffer, to 5 ml tube. Repeat the elution 2 times.
15. Concentrate HMW DNA with Genomic DNA Clean & Concentrator-10 (ZYMO RESEARCH) according to the manufacturer's instruction.

LIST OF FIGURES AND TABLES

Być doktorantem (in Polish).

Kto pracuje w ICHaBie,
ten ciągle siedzi w labie,
noce i dni artykuły studiuje
do seminarium przygotowuje.
Profesor zadań daje szczodrze
przed weekendem mówi mądrze:
żeby nauka miała smaczek
zrób do sesji sprawozdawczej plakacik,
wniosek grantowy napisz wybitnie,
by był oceniony na pozytywnie.
Te transpozony też musisz zbadać,
i na spotkaniach nam opowiadać.
A publikacja już obrażona
leży na biurku niezakończona.
Staże, praktyki i konsultacje
Oczywiście, zamiast wakacji.
Hirsze, impakty ciągną życia nić
jak doktorantem trudno jest być.

18.05.2023