Insights into Mutational Processes in *Arabidopsis thaliana* from Single-Molecule Long-Read Sequencing

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Yueqi Tao

aus Anhui/China

Tübingen

2026

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

Tag der mündlichen Qualifikation:	07.05.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr.-Ing. Oliver Kohlbacher

Summary

Mutation has long been a central topic in biology, as it provides the raw material for genetic diversity and represents the ultimate source of adaptation. The genome sequences observed in present-day individuals have already undergone mutation, together with recombination and selection, over the course of evolution. Over the past decade, short-read sequencing has been widely used to study mutations and mutation rates, leading to a strong understanding of genetic variation at the scale of single-nucleotide polymorphisms (SNPs) and small insertions and deletions (indels). However, the investigation of mutations in other genomic regions, such as repetitive sequences, had to await new technologies.

In the last decade, long-read single-molecule sequencing has seen increasing application in genomics. Among these technologies, Pacific Biosciences (PacBio) High-Fidelity (HiFi) sequencing employs circular consensus sequencing, merging multiple passes of a single molecule into one high-accuracy read. Both chapters here detect mutations using HiFi sequencing, each focusing on different features of the technology.

In Chapter One, I exploit the long-read and high-accuracy characteristics of HiFi sequencing to describe telomeric repeat diversity in *Arabidopsis thaliana*. Telomeres, located at the ends of linear chromosomes, protect chromosomes from degradation. In many species, telomeric regions consist of short repeat units; in *A. thaliana*, the canonical unit is a seven base-pair (bp) sequence, TTTAGGG. While the very ends are composed of canonical repeats, more proximal regions contain variant repeat types, indicating that mutations have occurred. I comprehensively characterized sequence variation in telomeric repeat arrays at the chromosome ends across 74 genetically diverse *A. thaliana* accessions. I identified several distinct types of telomeric repeat units, uncovered evolutionary processes such as local

homogenization and higher-order repeat formation, quantified telomeric repeat number changes at both germline and somatic levels, and revealed chromosome end-specific patterns in the distribution of variant repeats. These findings provide a detailed view of telomeric repeat variation in *A. thaliana* at multiple levels, expanding our knowledge of the evolution of chromosome ends.

In Chapter Two, I leverage the features of circular consensus sequencing and amplification-free library preparation, which allow multiple passes of each DNA strand for a single molecule. This approach enables high-accuracy sequencing of both strands, allowing the detection of sequence differences between them. Such differences can reflect unrepaired errors or misrepair events during DNA replication. I present a method to identify sequence differences between the two strands within single molecules. By analyzing polymerase kinetic information, specifically, pulse width and interpulse duration, I confirmed the authenticity of these events. To validate this pipeline, I analyzed one *A. thaliana* accession with 3,747,759 molecules and identified three molecules exhibiting sequence differences greater than 50 bp. Further examination of the detailed sequence characteristics suggested that these events could result from template switching, slippage-mediated strand mispairing, or palindrome-mediated deletions. This study provides a proof-of-concept approach for capturing differences between strands immediately after new strand synthesis but prior to repair, or following erroneous repair, thereby deepening our understanding of how mutations arise.

Together, these two studies uncover mutational processes in *A. thaliana* at different scales using the latest single-molecule long-read sequencing technology. Specifically, these findings enhance our understanding of germline and somatic mutations at telomeric regions, as well as ongoing mutations during DNA replication in leaf tissue. Moreover, the approaches developed here can be applied to study mutational processes in other tissues and species.

Zusammenfassung

Mutationen sind seit langem ein zentrales Thema der Biologie, da sie das Rohmaterial für genetische Vielfalt liefern und die ultimative Quelle der Anpassung darstellen. Die in heutigen Individuen beobachteten Genomsequenzen haben bereits Mutationen durchlaufen, zusammen mit Rekombination und Selektion, im Verlauf der Evolution. In den letzten zehn Jahren wurde das Short-Read-Sequencing weit verbreitet eingesetzt, um Mutationen und Mutationsraten zu untersuchen, was zu einem tiefen Verständnis der genetischen Variation auf der Ebene einzelner Nukleotidpolymorphismen (SNPs) und kleiner Insertionen und Deletionen (Indels) geführt hat. Die Untersuchung von Mutationen in anderen genomischen Regionen, wie z. B. repetitiven Sequenzen, musste jedoch auf neue Technologien warten.

Im letzten Jahrzehnt hat das Long-Read-Single-Molecule-Sequencing zunehmend Anwendung in der Genomik gefunden. Unter diesen Technologien verwendet das High-Fidelity-(HiFi)-Sequencing von Pacific Biosciences (PacBio) das sogenannte „Circular Consensus Sequencing“, bei dem mehrere Durchläufe eines einzelnen Moleküls zu einem hochgenauen Lesedurchgang zusammengeführt werden. Beide hier vorgestellten Kapitel detektieren Mutationen mithilfe des HiFi-Sequencings, wobei jedes Kapitel unterschiedliche Eigenschaften der Technologie in den Fokus stellt.

Im ersten Kapitel nutze ich die langreichweitigen und hochgenauen Eigenschaften des HiFi-Sequencings, um die Vielfalt der telomerischen Wiederholungen in *Arabidopsis thaliana* zu beschreiben. Telomere, die sich an den Enden linearer Chromosomen befinden, schützen Chromosomen vor Abbau. In vielen Arten bestehen telomerische Regionen aus kurzen Wiederholungseinheiten; in *A. thaliana* ist die kanonische Einheit eine sieben Basenpaare (bp) lange Sequenz, TTTAGGG. Während die äußersten Enden aus kanonischen Wiederholungen bestehen, enthalten proximale Bereiche Varianten von Wiederholungseinheiten, was darauf hinweist, dass Mutationen

stattgefunden haben. Ich habe die Sequenzvariation in telomerischen Wiederholungsarrays an den Chromosomenenden über 74 genetisch diverse *A. thaliana*-Accessions umfassend charakterisiert. Dabei habe ich mehrere unterschiedliche Typen von telomerischen Wiederholungseinheiten identifiziert, evolutionäre Prozesse wie lokale Homogenisierung und Bildung höherer Wiederholungen aufgedeckt, Veränderungen der Anzahl telomerischer Wiederholungen sowohl auf Keimbahn- als auch auf somatischer Ebene quantifiziert und chromosomenspezifische Muster in der Verteilung der Varianten festgestellt. Diese Ergebnisse bieten einen detaillierten Einblick in die Variation telomerischer Wiederholungen in *A. thaliana* auf mehreren Ebenen und erweitern unser Wissen über die Evolution der Chromosomenenden.

Im zweiten Kapitel nutze ich die Eigenschaften des Circular Consensus Sequencing und der amplifikationsfreien Bibliotheksvorbereitung, die mehrere Durchläufe jeder DNA-Strangsequenz für ein einzelnes Molekül ermöglichen. Dieser Ansatz erlaubt eine hochgenaue Sequenzierung beider Stränge, wodurch Unterschiede zwischen ihnen erkannt werden können. Solche Unterschiede können unverheilte Fehler oder Fehlreparaturen während der DNA-Replikation widerspiegeln. Ich stelle eine Methode vor, um Sequenzunterschiede zwischen den beiden Strängen innerhalb einzelner Moleküle zu identifizieren. Durch die Analyse polymerasekinetischer Informationen, insbesondere Pulsbreite und Interpulsdauer, konnte ich die Authentizität dieser Ereignisse bestätigen. Zur Validierung dieses Ansatzes analysierte ich eine *A. thaliana*-Accession mit 3.747.759 Molekülen und identifizierte drei Moleküle mit Sequenzunterschieden von mehr als 50 bp. Eine genauere Untersuchung der Sequenzmerkmale deutete darauf hin, dass diese Ereignisse durch Template Switching, Schlupf-vermittelte Strangfehlpaarung oder palindromvermittelte Deletionen entstehen könnten. Diese Studie liefert einen Proof-of-Concept-Ansatz, um Unterschiede zwischen Strängen unmittelbar nach der Synthese des neuen Strangs, aber vor der Reparatur, oder nach fehlerhafter Reparatur zu erfassen, und vertieft damit unser Verständnis darüber, wie Mutationen entstehen.

Zusammen decken diese beiden Studien mutationale Prozesse in *A. thaliana* auf unterschiedlichen Skalen mithilfe der neuesten Single-Molecule-Long-Read-Sequenzierungstechnologie auf. Insbesondere erweitern diese Ergebnisse unser Verständnis von Keimbahn- und somatischen Mutationen in telomerischen Regionen sowie von laufenden Mutationen während der DNA-Replikation im Blattgewebe. Darüber hinaus können die hier entwickelten Ansätze angewendet werden, um mutationale Prozesse in anderen Geweben und Arten zu untersuchen.

Acknowledgements

I would like to thank my PhD supervisor, Detlef Weigel, for giving me enough freedom to explore different paths independently, make mistakes and learn from them. Beyond improving my project and technical skills, my time under Detlef's guidance has deeply influenced my approach to research and my understanding of academic life. I feel truly fortunate.

I thank colleagues and friends for the interactions we have shared.

I thank my family for allowing me to focus on myself.

I hold a deep memory of my maternal grandpa, Zhenchang Chen. I have been well protected for almost 30 years. Despite the grief, the memory and strength he left me will carry me through my life.

Table of Contents

1 Introduction

1.1 Mutation

Mutations are the foundation of genetic variation and evolution. They can be categorized in several ways. Depending on origin, mutations can be classified into spontaneous or induced. Based on size, they can be divided into point mutations and structural variants (SVs). Depending on the tissue in which they occur, mutations can be categorized as germline or somatic mutations.

1.1.1 Causes of mutation

For spontaneous point mutations typically arise either from chemical alterations, referred to as DNA lesions, such as deamination, or from errors introduced by DNA polymerases during replication. For chemical alterations of DNA, various factors influence the likelihood of such alterations, such as nucleosome occupancy (Chen et al. 2012). Polymerase errors are mitigated by replicative polymerase proofreading and mismatch repair. And only unrepaired lesions or errors ultimately appear as heritable mutations.

Spontaneous SVs can arise in two ways: either first occurring on a single DNA strand or directly as a double-strand break (DSB). SVs, including translocations, insertions, deletions, and duplications, can be generated by replication errors, including template switching (Chuong et al. 2024) or microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009). In these cases, the event initially occurs on a single strand, i.e., the newly synthesized strand, and is subsequently processed through repair mechanisms such as homologous recombination, eventually resulting either in restoration of the wild-type sequence or in a fixed mutation after DNA replication. Conversely, all types of

SVs, including translocations, inversions, insertions, deletions, direct or inverted duplications, and transposon insertions, can arise directly from double-strand breaks (Gómez-Herreros 2019). In addition, non-homologous end joining (NHEJ) can introduce errors during the repair process, contributing to mutagenesis (Goettel and Messing 2009).

Mutational heterogeneity can be influenced by intrinsic features of the nucleotide sequence itself (Hara and Kuraku 2025; Xie et al. 2019). Regions with high CG content tend to have higher cytosine-to-thymine transition rates because cytosines within CpG dinucleotides are frequently methylated and prone to deamination (Kiktev et al. 2018). Similarly, short tandem repeats are highly mutable due to both unequal crossing over during meiosis and strand slippage during DNA replication (Fan and Chu 2007), while certain sequences, such as specific oligonucleotides, are prone to form non-B DNA conformations (Guiblet et al. 2021), which are more susceptible to mutations via DSBs. In addition, chromatin state can influence mutational patterns, with heterochromatic regions exhibiting different mutation rates compared with open euchromatic regions (Schuster-Böckler and Lehner 2012).

Apart from spontaneous mutations, the mutation rate can be increased by the deliberate exposure of cells to physical agents, chemical agents, or viral insertions. Examples of physical mutagens include ultraviolet (UV) light and ionizing radiation. UV irradiation primarily induces DNA lesions such as cyclobutane pyrimidine dimers, leading predominantly to C to T transitions (Pfeifer 2020). Most of these lesions are subsequently repaired by nucleotide excision repair (NER). Ionizing radiation can induce single-nucleotide changes as well as double-strand breaks, which may result in insertions, deletions, and duplications (Adewoye et al., 2015). Chemical mutagens can form DNA adducts (Seo et al. 2000), or induce other modifications, such as alterations in DNA ultrastructure. Notable examples of chemically induced mutations include those

associated with human cancer (Poirier 2004). In addition, virus insertion mutagenesis, particularly using retroviruses, has been widely employed to generate large-scale mutant populations in many organisms (Golling et al. 2002).

1.1.2 Consequences of mutation

Mutations lead to genetic differences, for example, within a species, between species, and between normal and diseased individuals. In an evolutionary context, mutation primarily refers to germline mutations, which provide the raw material for genetic diversity and are the ultimate source of evolution and adaptation (Shendure and Akey 2015), but somatic mutations are important, too. In humans, they cause disease, and in plants, they generate spots that can be propagated clonally.

The genome sequences observed in any modern individuals within any species have already undergone mutation, recombination, and selection during their history. Of note, most mutations do not affect the phenotype of an individual because they are in sequences that do not have specific functions. These mutations are considered to be neutral. Some mutations, however, affect fitness. For example, small indels can cause frameshift mutations, altering the downstream amino acid sequence and potentially introducing a premature stop codon. In-frame indels can also affect protein coding by adding or deleting amino acids (Lek et al. 2016). Mutations in non-coding regulatory sequences can lead to phenotypic changes as well, and many are associated with human rare diseases (Short et al. 2018), while mutations that alter splice sites can result in truncated gene products (Delettre et al. 2001). Additionally, mutations can modify the location of recombination hotspots and cold spots within the genome (Yap and Kreuzer 1991).

1.2 DNA sequencing

1.2.1 The objective of DNA sequencing

A natural DNA sequence is a specific arrangement of four types of nucleotides: adenine, thymine, cytosine, and guanine. The primary goal of sequencing is to determine the precise order of these nucleotide bases. Beyond simply identifying the sequence, it also provides opportunities to understand why the sequence is organized in a particular way, which is closely related to many biological phenomena across all organisms. For example, once the DNA sequence is known, we can study the structure of chromosomes. In eukaryotes, long repetitive regions are found at the ends of chromosomes, which researchers have named telomeres. Sequencing also allows us to investigate patterns of mutation (Lynch 2010) by comparing sequences among individuals or species, thereby revealing how genetic diversity arises and persists.

1.2.2 Development of DNA sequencing technologies

Since the first DNA sequence was determined in the late 1970s, over 50 years have passed (Brownlee 2013), during which DNA sequencing has undergone three major revolutions. The first generation of widely used sequencing methods was Sanger sequencing, which uses chain-terminating dideoxy nucleotides to generate DNA fragments of different lengths, allowing them to be electrophoretically separated to determine the nucleotide order. After the original method was established, several improvements were made between 1980 and 1987, including the introduction of shotgun sequencing, fluorescence-based detection instead of the use of radioactively labeled chain terminators, and the development of automated sequencing machines. By this time, millions of bases had already been deposited in public datasets (Shendure et al. 2017).

From the 1990s to about 2005, most whole-genome sequencing studies were based on the Sanger method.

The second generation of widely used sequencing methods, also known as next-generation sequencing (NGS), refers to the era of short-read, high-throughput sequencing, from roughly 2006 on. NGS platforms enabled massive parallelization of sequencing reactions due to efficient in situ amplification (Heather and Chain 2016), most notably through bridge amplification, greatly increasing sequence output. Short-read sequencing made whole-genome sequencing much more accessible and significantly reduced its costs, allowing many laboratories to conduct population-level studies involving tens or even hundreds of individuals by 2018.

The third generation of widely used sequencing methods, characterized by single-molecule long-read sequencing, has become more widely used from around 2019, although early implementations date back to 2009 or earlier. Compared with the short reads of around 150 bp typical in the second-generation era, long-read sequencing now routinely produces reads with tens of kilobases in length, enabling much easier resolution of long-range genomic structures, compared to conventional short reads.

1.3 The era of long-read sequencing

The 2020s have become the era of single-molecule long-read sequencing. For the purpose of studying mutations, long-read sequencing greatly facilitates achieving the goal of comprehensive identification of mutations – especially in repetitive regions and large mutations, both of which are largely inaccessible to short-read sequencing. Moreover, long-read sequencing is typically amplification-free, ensuring that the output accurately reflects the original DNA sequence with no GC bias. Currently, two major providers dominate single-

molecule long-read sequencing products: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). These technologies have significantly increased read lengths to tens to thousands of kilobases.

1.3.1 Two representative technologies of the long-read sequencing era

1.3.1.1 PacBio: Single-molecule real-time sequencing with zero-mode waveguides

PacBio single-molecule real-time (SMRT) sequencing monitors the progress of DNA polymerase performing uninterrupted, template-directed synthesis using four distinguishable fluorescently labeled deoxyribonucleotide triphosphates, which allow real-time detection of nucleotide incorporation (Eid et al. 2009).

During library preparation, double-stranded DNA is fragmented and ligated to hairpin adapters at both ends, forming a topologically circular molecule called a SMRTbell. The molecules are then loaded into zero-mode waveguide (ZMW) nanostructure arrays in a SMRTcell, each containing millions of ZMWs, in each of which a single DNA polymerase is immobilized. In addition to nucleotide information, the raw sequencing output also includes polymerase kinetics information, specifically pulse width (PW) per base and interpulse duration (IPD) between two successive light signals. These kinetic signals can be used not only for sequence determination but also to detect biological phenomena that affect polymerase activity (Guiblet et al. 2018), such as base modifications (Tse et al. 2021) and DNA secondary structures (Eid et al. 2009).

PacBio SMRT sequencing generates two types of long reads. Continuous long read (CLR) sequencing produces reads longer than 50 kb but with a relatively

high error rate of 15%. In commercial applications, CLR has largely been replaced by high-fidelity (HiFi) sequencing from 2021 onwards. HiFi reads, generated using circular consensus sequencing (CCS), are around 15-20 kb long, with an error rate below 1%. Each DNA strand is sequenced multiple times, and the subreads are merged to form a high-accuracy consensus read.

The sequencing platforms have evolved alongside the sequencing strategies. During the CLR era, the system progressed from RS II to Sequel and then Sequel II. In the HiFi era, the platform evolved from the Sequel II system, used from 2019, to the Revio system, announced in 2022. In the Sequel II system, subreads must be processed with the CCS algorithm (https://github.com/PacificBiosciences/ccs) to generate HiFi reads. In contrast, the Revio system produces a single consensus read directly, leveraging the DeepConsensus algorithm (Baid et al. 2023). Along with system upgrades, the number of ZMWs per SMRTcell has increased, resulting in higher output (Ardui et al. 2018). Currently, the Revio system can run up to 4 SMRTcells every 24 hours, producing up to 480 Gb of HiFi reads.

1.3.1.2 ONT: Nanopore-based single-molecule sequencing

ONT sequencing infers sequence information from ionic current signals (Wang et al. 2021), generated when a DNA strand passes through a nanoscale protein pore or nanopore. A motor protein positioned above the nanopore unwinds double-stranded DNA and ratchets the single strands through the nanopore (He et al. 2024). The DNA molecule changes the electric current through the nanopore, enabling base calling and the detection of epigenetic base modifications (Kovaka et al. 2025).

The typical length of ONT long reads ranges from 10 to 100 kb, but ultra-long (UL) reads can range up to 1 Mb. UL reads are generated using a special library

preparation process. The current records exceed 4 Mb in a single continuous read.

ONT has continuously developed improved nanopore chemistries to enhance base-calling accuracy. The current R10 pore achieves around 95% accuracy, thanks to a longer barrel and a dual-reader head design (Liu et al. 2021). ONT has also tested duplex sequencing, where both strands of the same DNA molecule are sequenced consecutively to improve read accuracy, potentially reaching levels comparable to PacBio HiFi reads (Koren et al. 2024); however, this approach is not used for commercial applications.

1.3.2 Long-read sequencing beyond genomic DNA

With the continuous improvement and widespread adoption of single-molecule long-read DNA sequencing, a variety of applications beyond conventional DNA sequencing have emerged. For instance, PacBio Iso-seq, ONT direct RNA-seq, and ONT cDNA-seq enable the capture of full-length transcripts (Brooks et al. 2022). Long-read single-cell RNA sequencing further allows the identification of isoform-specific expression across different life cycle stages (Dogga et al. 2024). Moreover, the multiplexed arrays sequencing (MAS-seq) approach maximizes throughput on the PacBio CCS platform while retaining full-length transcript information as much as possible (Al'Khafaji et al. 2024).

Beyond transcriptomics, long-read sequencing technologies have also been adapted to study chromatin organization. For example, Fiber-seq utilizes kinetic information from single-molecule sequencing, combined with nonspecific DNA N6-adenine methyltransferases labeling, to profile chromatin accessibility at the individual molecule level (Stergachis et al. 2020). Similarly, scNanoATAC-seq, based on the ONT platform, enables genome-wide characterization of

chromatin accessibility, with a particular advantage in repetitive regions, such as transposons and paralogous genes (Li et al. 2025a).

1.4 Advances in genome analysis in the long-read era

1.4.1 Haplotype-phased genome assembly

Short-read–based genome assemblies are usually highly fragmented, whereas long-read assemblies can readily achieve high continuity and completeness. Haploid genome assemblies of many different species are being progressively released under the telomere-to-telomere (T2T) paradigm, including humans (Nurk et al. 2022), *Arabidopsis thaliana* (Naish et al. 2021), and mouse (Liu et al. 2024a). This improvement facilitates the accurate detection of structural variants (SVs), compared to the traditional approach of being able to examine only variant breakpoints with short-read data. For large-scale population studies or for analyses involving non-haploid genomes, combining long-read sequencing with genome graph approaches enables haplotype-phased analyses. In particular, unitigs in the graph can be used to distinguish haplotypes (Serra Mari et al. 2024) even in polyploid material such as tetraploid potato (Sun et al. 2025a).

Many studies now first generate a high-quality long-read reference genome and then map existing short-read datasets to this reference (Wei et al. 2024; Vollmer et al. 2023; Stammnitz et al. 2023; Kolora et al. 2021). This approach allows precise characterization of SVs, improving on signals from genome-wide association studies (GWAS) and more generally advancing our understanding of evolutionary processes (Gompert et al. 2025; Groh et al. 2025; Akagi et al. 2025). Long-read assemblies have also helped to validate regions with

suppressed recombination that contribute to phenotypic differences between ecotypes (Hager et al. 2022). Newly completed reference genomes are also used for phylogenetic analyses, helping to resolve previously unresolved nodes (Chen et al. 2025).

The accuracy and completeness of long-read assemblies reduces mismapping and false variant calls (Aganezov et al. 2022), making all kinds of inferences much more reliable (Gutiérrez-García et al. 2024). Consequently, long-read sequencing has become a common tool in genomic studies, preferred whenever high-quality sequence information is required (Lien et al. 2024; Satterlee et al. 2024; Nemudraia et al. 2024; Steichen et al. 2024; Reed et al. 2023; Qi et al. 2023; Shao et al. 2023; Wang et al. 2022b). Even in studies not directly focused on genome analyses, long reads have been used, for example, to efficiently detect concatemeric cDNA products in the study of de novo gene synthesis by an antiviral reverse transcriptase (Tang et al. 2024).

Long-read genomic DNA sequencing is frequently combined with other technologies, such as Iso-seq, within the same study (Arbore et al. 2024). This integration is particularly advantageous for analyses involving large sequence differences, including SVs (Koeppel et al. 2025; Zhang et al. 2024; Ebert et al. 2021), repetitive regions (Wilkinson et al. 2024; Villanea et al. 2025; Hoge et al. 2024; Pownall et al. 2023; Couger et al. 2021), or mobile elements (Widen et al. 2023).

1.4.2 Previously missing repeats

Long-read sequencing has enabled the resolution of many repetitive regions that were previously inaccessible and that were therefore missing from genome assemblies (Moss et al. 2020). Such regions are often highly mutable, meaning that much information about whole-genome variation was missing in earlier

studies. With long-read sequencing, it is now possible to detect mutational patterns in these regions for the study of their repeat structure, mutation dynamics, and evolutionary mechanisms driving their generation and diversification. Similarly, features such as chromatin accessibility (Li et al. 2025a) can now be investigated in repetitive regions using long-read sequencing.

A particularly notable example are eukaryotic centromeres, which are typically composed of satellite repeats. In humans, a centromeric repeat unit is 171 bp long; in *A. thaliana*, one unit is 178 bp. These repeat units are highly variable, and the arrays often undergo expansion or contraction. Centromeric and pericentromeric satellite arrays were largely ignored prior to the advent of long-read sequencing, and only with the development of long-read sequencing have centromeres and their evolutionary patterns become a major focus of genome research (Altemose et al. 2022). After sequencing technologies were validated in model organisms, the sequence structure of centromeric regions in diverse species has been characterized (Hu et al. 2025; Huang et al. 2024; Yan et al. 2025). Subsequent studies have included population-level analyses (Gao et al. 2024), comparisons among closely related species (Logsdon et al. 2024), and studies of spontaneous mutations (Dong et al. 2025). Notably, single-nucleotide mutation rates are significantly higher in centromeric regions compared to chromosome arms (Logsdon et al. 2024). In addition, the genome-wide distribution of centromeric and ectocentromeric repeats has recently been described with long-read human genome assemblies (Corda and Giunta 2025).

Another type of tandem repeat arrays is found at telomeres at the ends of linear chromosomes. Telomeres, which are replicated by telomerase, tend to shorten with each cell division, and excessively short or excessively long telomeres can lead to abnormal cellular phenotypes. With long-read sequencing, both telomeric repeat variation and precise telomere length measurement have

become accessible. Recent studies have shown that telomere lengths are end-specific rather than uniform across chromosomes, with end-specific telomere lengths accurately estimated based on methods that ligate molecular tags to chromosome ends prior to sequencing (Karimian et al. 2024). In species whose telomeres are shorter than typical long-read lengths, telomere length can be directly measured from long-read data (Colt et al. 2024). For telomeric repeat variants, PacBio HiFi sequencing has enabled detailed characterization of variant repeat patterns in humans (Grigorev et al. 2021) and, as shown in this thesis, in *Arabidopsis* (Tao et al. 2024). Using unique subtelomeric regions as anchors, variant patterns can be assigned to specific chromosome ends. Prior to the long-read era, only limited variant information could be obtained from short-read fragments (Choi et al. 2021), with telomere length estimates using k-mer coverage (Cook et al. 2016) being imprecise and providing only minimal biological insight.

Ribosomal DNA (rDNA), another type of tandem repeat arrays, typically contains hundreds of repeat units and forms the nucleolar organizer regions (NORs; Gál et al. 2024). There are two types of ribosomal RNA genes, 45S and 5S. In humans, rDNA repeat arrays are located on the short arms of five acrocentric chromosomes (Potapova et al. 2025). In *A. thaliana*, two long NORs are located on the short arms of chromosome 2 and 4 (Rabanal et al. 2017). rDNA sequences exhibit both copy number variation and sequence variation, including point mutations and small indels, which may influence a range of traits (Rodriguez-Algarra et al. 2025). Similar to centromeres, most rDNA regions were gaps in previous genome assemblies due to the limitations of short-read sequencing. The advent of long-read technologies has enabled the reconstruction of megabase-sized NORs, compared to the previous maximum recovery of only tens of repeats at the edges of NORs (Fultz et al. 2023b). Additionally, the ability of long-read sequencing to directly detect DNA

methylation is valuable for studying the potential of rDNA transcriptional activity (Hori et al. 2021).

Short tandem repeats (STRs), also known as microsatellites, are much shorter than telomeric or centromeric repeats and typically consist of 1–6 bp repeat units. Changes in STR repeat number can influence various biological processes, including gene regulation (Tanudisastro et al. 2024). In some cases, pathogenicity emerges only when repeat expansions exceed a critical size threshold (Sone et al. 2019). The high accuracy and long-range resolution required for analyzing tandem repeats often exceed the capabilities of earlier technologies, especially short-read sequencing. With the goal of leveraging long-read sequencing to resolve tandem repeat regions, several tools have been developed, such as the Tandem Repeat Genotyping Tool (TRGT; Dolzhenko et al. 2024), which can detect tandem repeats and their methylation status simultaneously from PacBio HiFi data. As a result, tandem repeat–associated haplotypes and their evolutionary histories can be easily characterized, with one example being the human *MUC19* locus, where repeat structure correlates with an introgressed haplotype (Villanea et al. 2025).

Segmental duplications (SDs) are also much more accurately described with long reads (Vollger et al. 2022). SDs are defined as genomic regions larger than 1 kb that share more than 90% sequence identity. Depending on their genomic distribution, SDs may be continuous or interspersed, and based on their orientation, they can be classified as direct or inverted duplications. SDs have now been characterized across diverse individuals and also between species (Jeong et al. 2025). The single-nucleotide mutation rate in SD regions has been found to be higher than in other regions of the genome (Vollger et al. 2023), and SD-associated regulatory effects on visible phenotypes have been elucidated (Milia et al. 2025). Gene duplication, a special form of segmental duplication, has been characterized at previously inaccessible regions, such as

the AMY amylase gene cluster, where long-read and optical genome mapping revealed unique haplotypes that span more than 200 kb (Yilmaz et al. 2024), substantially improved our understanding of the evolution of the amylase gene and its relationship to the rise of agriculture.

The analysis of transposable elements (TEs), which constitute a substantial portion of many species' genomes, ranging from 3% to 80%, also benefits from long-read sequencing (Quadrana and Henderson 2025). TEs are classified based on their transposition mechanisms. Class I elements, retrotransposons, transpose via RNA intermediates and include long interspersed elements (LINEs), short interspersed elements (SINEs), and long terminal repeats (LTRs). Class II elements, DNA transposons, transpose via DNA intermediates, typically using a cut-and-paste mechanism. In the short-read sequencing era, detection of variable TE insertions primarily relied on identifying breakpoints between TE sequences and their flanking regions by mapping reads from multiple individuals to a reference genome (Shahid and Slotkin 2020). With long reads, variable TEs can be directly identified in individual assemblies. Even when mapped to a reference, many long reads will span an entire TE insertion (Mohamed et al. 2023).

1.4.3 Genome graph approaches

The development of long-read sequencing technologies has accelerated the adoption of genome graph approaches, which had been developed for short reads (Schneeberger et al. 2009), but were only sparingly used before. Graph-based approaches are applied in two major contexts. First, graph structures can be used to represent the continuity and complexity of individual genomic sequences more fully. Single genomes consist mostly of linear paths but form dense, tangled structures in highly repetitive regions (Nurk et al. 2022). Second, genome graphs that integrate multiple genomes can substantially improve the

accuracy of variant calling, especially for SVs. In these population-level graphs, each node represents a sequence, and alternative alleles appear as bubbles, with an individual genome corresponding to an explicit path through the graph. Pangenome analyses in polyploid species (Sun et al. 2025a) integrate both aspects, enabling unbiased characterization of genomic variability. In particular, they overcome limitations of single-reference approaches, where highly divergent or missing regions lead to low read mappability and consequently cause reference bias during variant calling. Additionally, the large amount of short-read data produced before is now increasingly being aligned to pangenome graphs for variant calling (Lynch et al., 2025). A remaining challenge is sampling bias: rare haplotypes are often underrepresented in current pangenome datasets.

1.4.4 Studies leveraging long-read-specific features

1.4.4.1 Use of real-time signals

Both PacBio and ONT long-read sequencing can detect base modifications from raw sequencing signals. PacBio infers cytosine methylation from polymerase kinetic information (Kong et al. 2022), including interpulse duration (IPD) and pulse width (PW), while ONT relies on pattern changes in electric current when DNA passes through the nanopore (Liu et al. 2021). Representative work, such as the T2T-CHM13 human genome assembly, has greatly expanded the characterization of the human methylome (Gershman et al. 2022; Hoyt et al. 2022). Long-read sequencing signals can also be exploited to infer DNA secondary structure (Guiblet et al. 2018).

1.4.4.2 Use of amplification-free libraries

Because long-read sequencing is typically amplification-free, each read

represents a single DNA molecule from a single cell. This enables the detection of sequence differences at single-molecule resolution. Several studies have leveraged this approach to characterize genome-wide crossover and gene conversion events (Byun et al. 2024; Schweiger et al. 2024), to identify somatic microsatellite array variation, some of which is disease-associated (Sehgal et al. 2024), and to detect somatic transposon insertions (Movilli et al. 2025).

1.5 Genome studies in *Arabidopsis thaliana*

Arabidopsis thaliana, a model flowering plant species (Meinke et al. 1998) in the Brassicaceae family, is diploid and highly inbred. It has a relatively small genome of approximately 135 Mb distributed across five chromosomes. With a broad natural range, thousands of genetically distinct accessions have been collected over the years for many different research purposes. In plants, new genomic technologies are often first applied to *A. thaliana* (Rabanal et al. 2022), making it one of the first multicellular species with many completely reconstructed genomes (Naish et al. 2021; Wlodzimierz et al. 2023). Consequently, the development of genomic studies in *A. thaliana* reflects, to some extent, the broader progress of the genomic field.

1.5.1 Studies in the Sanger sequencing era

The first sequenced *A. thaliana* genome, published in 2000 (Arabidopsis Genome Initiative 2000), was from the Columbia (Col-0) accession, which is often a standard for genetic studies. This genome assembly, produced using a BAC-by-BAC approach and Sanger sequencing, lacked, however, large portions of the centromeres and most of the rDNA arrays on chromosome 2 and 4.

Early studies of natural sequence variation primarily relied on the amplification of PCR products, guided by the reference genome sequence (Hagenblad et al. 2004). However, some early studies had already begun conducting analyses such as linkage disequilibrium (Nordborg et al. 2002).

The *A. thaliana* reference genome was curated and periodically updated by The Arabidopsis Information Resource (TAIR; Lamesch et al. 2012). However, even in its most advanced version, TAIR10 (Lamesch et al. 2012), the assembly remained incomplete due to the limitations of Sanger sequencing reads, and it therefore continued to suffer from collapsed sequence repeats and failure to represent certain genomic elements, such as transposon copy (Kirov et al. 2021).

1.5.2 Studies in the short-read era

As costs dropped, short-read sequencing of large cohorts of natural accessions became possible (Long et al. 2013; 1001 Genomes Consortium 2016; Cao et al. 2011). The common strategy was to map reads to the Col-0 reference genome in order to identify variants. Short reads from diverse *A. thaliana* accessions, initially shorter than 50 bp, were mapped to the Col-0 reference genome with the goal of variant calling (Ossowski et al. 2008). Apart from characterizing natural genomic variation, short reads were found to be powerful tools for identifying spontaneous mutations and thus estimating the spontaneous mutation rate (Ossowski et al. 2010).

Due to their limited read length, short reads were primarily useful for detecting SNPs and small indels through alignment to a reference genome or reference guided assembly (Schneeberger 2011; Gan et al. 2011). This limitation was particularly pronounced in transposon-rich or repetitive regions, making it

difficult to achieve haplotype-level resolution for complex loci such as the S-locus (Tsuchimatsu et al. 2017) and NLR genes (Guo et al. 2011).

It is worth noting that, although short-read sequencing studies at this stage largely relied on mapping reads to a single reference genome, the issue of reference bias was already recognized as early as 2009. Researchers observed that different reference genomes could lead to differing variant calls, highlighting the need for approaches beyond a single reference, such as graph-based genome representations (Schneeberger et al. 2009) and reference-guided short-read assemblies (Schneeberger et al. 2011; Gan et al. 2011). These studies highlighted the importance of generating assemblies for each accession rather than relying solely on a single reference genome. Although this work primarily achieved good results in single-copy regions due to read length limitations, the concepts established during this stage laid the foundation for the later era of long-read sequencing.

1.5.3 Studies in the long-read era

With the advent of long-read sequencing, initial efforts generated a long-read assembly as a reference and mapped short reads to it. This was soon followed by a shift toward generating long-read sequences and performing de novo genome assemblies for multiple accessions.

Starting from 2016, studies began reporting chromosome-level de novo assemblies of single non-reference *A. thaliana* accessions (Zapata et al. 2016; Michael et al. 2018; Pucker et al. 2019). These studies demonstrated that long-read assemblies made it much easier to resolve gene isoforms, transposons, and large SVs. At this stage, assemblies were typically generated using a combination of moderate-quality long reads and short reads. In 2016, an F1 hybrid of Col-0 x Cvi-0 was sequenced, with separate assemblies of the

parental accessions, to evaluate the accuracy of haplotype phasing based on long reads (Chin et al. 2016).

Following the initial phases, several studies compared multiple chromosome-level de novo assemblies for different *A. thaliana* accessions (Jiao and Schneeberger 2020; Kileeg et al. 2024), and for different *A. thaliana* relatives (Burns et al. 2021). Some of these studies introduced the concept of a graph-based pan-genome (Lian et al. 2024; Kang et al. 2023), and others applied graph-based approaches to describe sequence diversity in specific complex regions, such as the NLR region (Teasdale et al. 2024). Currently, there are around 600 long-read assemblies, generated using PacBio HiFi, PacBio CLR, and ONT, providing a foundation for future studies (Alonso-Blanco et al. 2024).

In parallel with these assembly comparisons, a gold-standard long-read-based genome was released in 2021, primarily aimed at resolving centromeric satellite repeat regions (Naish et al. 2021). Other challenging regions in Col-0 have been resolved as well, including rDNA arrays (Fultz et al. 2023) and long mitochondrial DNA insertions (Fields et al. 2022). Population-level analyses of previously difficult-to-study regions have been reported, including sequences diversity of centromeres (Wlodzimierz et al. 2023), telomeres (Tao et al. 2024), and organellar sequences and structures (Xian et al. 2025).

Long-read studies have also supported investigations of full-length extrachromosomal circular DNA (Zhang et al. 2023a), somatic transposon insertions (Debladis et al. 2017; Movilli et al. 2025), and genome-wide crossover landscapes (Byun et al. 2024).

1.5.4 Summary of state of the art

In my thesis, I focus on *A. thaliana*, although the discussion is applicable to other species as well. Many topics in genome analysis are cyclical in nature, evolving with advances in sequencing technologies. Features such as crossovers, telomeric repeat motifs, and genome-wide copy number variants have been studied long before the advent of long-read sequencing. For example, the detection of crossovers has progressed from marker-based genotyping (Drouaud et al. 2006) to short-read sequencing (Rowan et al. 2019) and now to long-read sequencing (Byun et al. 2025). Similarly, telomeric repeats were previously analyzed using Sanger sequencing (Kuo et al. 2006), then via k-mer counting based on short reads (Choi et al. 2021), and now, long-read sequencing enables the resolution of chromosome end-specific patterns (Tao et al. 2024). Genome-wide copy number variants were also detectable with short reads (Zmienko et al. 2020), albeit with more complex workflows and lower accuracy. Importantly, with each technological advance, our understanding of genomes and their structure, evolution and function becomes more comprehensive and clearer.

1.6 Objectives

The two projects of my thesis take advantage of PacBio HiFi sequencing data to investigate different aspects of the mutational process in *Arabidopsis thaliana*.

1.6.1 Describing telomeric repeats

In the Sanger and short-read era, telomeric repeat regions were among the most challenging parts of the genome to study. For example, in the first *A. thaliana* genome published in 2000 (Arabidopsis Genome Initiative 2000), telomeric regions were significantly underrepresented. In the long-read era,

assemblies are considered complete when typical telomeric repeats are present at both ends of a chromosome (“T2T” genome assemblies). However, there are also several variant types of telomeric repeats at the chromosome ends that have not been studied in depth. These variants may be associated with G-quadruplex formation, DNA methylation, or telomere function. Previous attempts to characterize telomeric variants in *A. thaliana* relied on PCR products or short reads, which were neither comprehensive nor highly accurate. In this project, I use high-fidelity reads to comprehensively characterize telomeric repeat patterns across diverse *A. thaliana* accessions and to examine whether these patterns reflect specific evolutionary processes.

1.6.2 Detecting single-stranded structural variants

HiFi sequencing, through circular consensus sequencing, reads both forward and reverse strands multiple times within the same molecule, producing high-fidelity single-strand information. Differences between the two strands may indicate sequence changes generated during DNA replication. Such differences could reflect events occurring before repair mechanisms act or aberrant repair events. A similar approach has been applied to detect point mismatches in human material (Liu et al. 2024b). In this project, I aim to develop an approach to identify structural differences between DNA strands. The resulting pipeline is designed to be easily extendable to other datasets beyond the one examined in my thesis.

2 Atlas of telomeric repeat diversity in *Arabidopsis thaliana*

Yueqi Tao, Wenfei Xian Zhigui Bao, Fernando A. Rabanal, Andrea Movilli, Christa Lanz, Gautam Shirsekar, Detlef Weigel.

Genome Biology 25, 244 (2024). https://doi.org/10.1186/s13059-024-03388-3

Contributions

DW and YT conceived the project. WX and YT performed the analyses regarding rDNA-binding end, telomere length, and telomere cluster. ZB performed the principal component analysis. FAR and YT performed the genome assembly. YT performed all other bioinformatic analyses. AM, CL, YT, and GS performed the plant growth, HMW DNA extraction, and library preparation for three HPG1 accessions. DW provided general advice. YT and DW interpreted the data and wrote the manuscript with input from all authors.

2.1 Introduction

Telomeric repeat arrays are found at the termini of most eukaryotic chromosomes (Churikov and Price 2008). The very ends of the arrays, known as telomeres (Chan and Blackburn 2004), commonly consist of canonical units with the formula (T)x(A)y(G)z and act as functional caps that protect chromosome ends from degradation and fusion (Fulnecková et al. 2013; Verdun and Karlseder 2007). These canonical repeats are being synthesized from an RNA template by telomerase, which ensures their sequence conservation (Schrumpfová and Fajkus 2020). In contrast to these highly conserved repeats, the immediately following sequences often include degenerate and variant telomeric repeats (Wallberg et al. 2019; Vozárová et al. 2022; Richards et al. 1992; Allshire et al. 1989), which differ from the canonical unit in one or more base substitutions or small insertions and deletions (indels; Lee et al. 2014). The composition of the variant repeats displays remarkable heterogeneity within the same genetic group and among different chromosome ends (Stephens and Kocher 2024; Tham et al. 2023; Mizuno et al. 2008; Baird et al. 2000), raising questions as to the evolutionary mechanisms that generate and maintain this diversity (Mendez-Bermudez et al. 2009; Pickett et al. 2004). This telomere-adjacent region serves as a transition zone between the telomere and the rest of the chromosome that contains genes and other genetic elements (Churikov and Price 2008). Specific types of variant telomeric repeats have been implicated in determining methylation state (Farrell et al. 2022), protein binding (Wang et al. 2023), and formation of G-quadruplexes (Lee and Kim 2009). A comprehensive understanding of the evolutionary dynamics and functional significance of telomeres and telomere-adjacent regions must therefore begin with thorough knowledge of variation in the composition of telomeric repeats.

Arabidopsis thaliana has a seven-base-pair canonical unit TTTAGGG, which is the dominant telomeric unit in many other plant species as well (Peska and Garcia 2020; Richards and Ausubel 1988). The presence of variant telomeric repeats in *A. thaliana* was first established with a yeast artificial chromosome strategy (Richards et al. 1992). Subsequently, sequencing of PCR products revealed the heterogeneity of variant repeats from individual chromosome ends (Wang et al. 2010; Kuo et al. 2006). Variant repeats have also been directly observed in unassembled sequencing reads (Choi et al. 2021), and they have been identified by partially assembling four chromosome ends in the Col-0 accession from Illumina short reads (Farrell et al. 2022). However, the highly repetitive nature of telomeric regions and the presence of identical sequences shared between repeat-adjacent regions, as well as large interstitial telomeric arrays in other parts of *A. thaliana* genomes, create ambiguity when mapping reads that are only hundreds base pairs long to specific positions of the genome (Olson et al. 2023; Teano et al. 2023; Heacock et al. 2004). As a result, variation in telomeric repeat content at *A. thaliana* chromosome ends remains largely uncharacterized and has been ignored in diversity studies.

New single-molecule sequencing methods, generating reads of more than 10 kilobases (kb) in length, which exceeds the size of full-length telomeric repeat tracts and extends into unique repeat-adjacent regions, can overcome the challenges of reconstructing full telomeric sequences (Grigorev et al. 2021). However, although several *A. thaliana* genome assemblies have now been published (Hou et al. 2022; Wang et al. 2022; Naish et al. 2021), they have largely ignored the telomeric sequences apart from confirming that telomeres are structurally present at most chromosome ends. Pacific Biosciences High Fidelity (PacBio HiFi) sequencing is particularly well suited for reliable base calling in low-complexity telomeric repeats (Tan et al. 2022). In addition, the circular sequencing mode of HiFi sequencing, wherein each DNA molecule is sequenced multiple times, allows us to confidently characterize somatic

information such as repeat number variation in the telomeric regions, which is obscured in assemblies (Wenger et al. 2019; Loomis et al. 2013).

In this study, I provide a high-resolution description of telomeric repeats for all ten chromosome ends in *A. thaliana*. I identify numerous types of variant telomeric repeats and previously undescribed sequence arrangement within the telomeric region, including higher-order repeats and inter-chromosomal similarity of non-telomeric fragments. I also investigate repeat number variation of non-canonical telomeric repeat arrays at both germline and somatic levels. I illustrate chromosome end-specific and genetic group-specific patterns of repeat haplotypes along with linkage disequilibrium between telomeric repeat arrays and their adjacent non-coding regions. My findings significantly expand the collection of repeats derived from canonical telomeric repeats and telomeric sequence features in *A. thaliana*, setting the stage for a deeper understanding of the evolutionary mechanisms acting on them.

2.2 Methods

2.2.1 HiFi-based data collection

Seventy-three HiFi-based assemblies and read sets, representing 71 natural accessions (Figure 2.1), were obtained from seven public sources. The datasets of 44 accessions were from Wlodzimierz et al. (2023). 11 from Kang et al. (2024), 14 from Lian et al. (2024), the Kew-1 accession from Christenhusz et al. (2023), and three independent Col-0 datasets from Rabanal et al. (2022), Wang et al. (2022), and Naish et al. (2021).

Three HPG1 accessions (Hagmann et al. 2015) were sequenced with one SMRT Cell on the Sequel II platform (PacBio). Plant growth (Contreras-Garrido et al. 2023), DNA extraction from a single plant (Wlodzimierz et al. 2023),

preparation of a multiplexed sequencing library followed by HiFi sequencing (Rabanal et al. 2022), and genome assembly (Wlodzimierz et al. 2023) were performed as previously described.

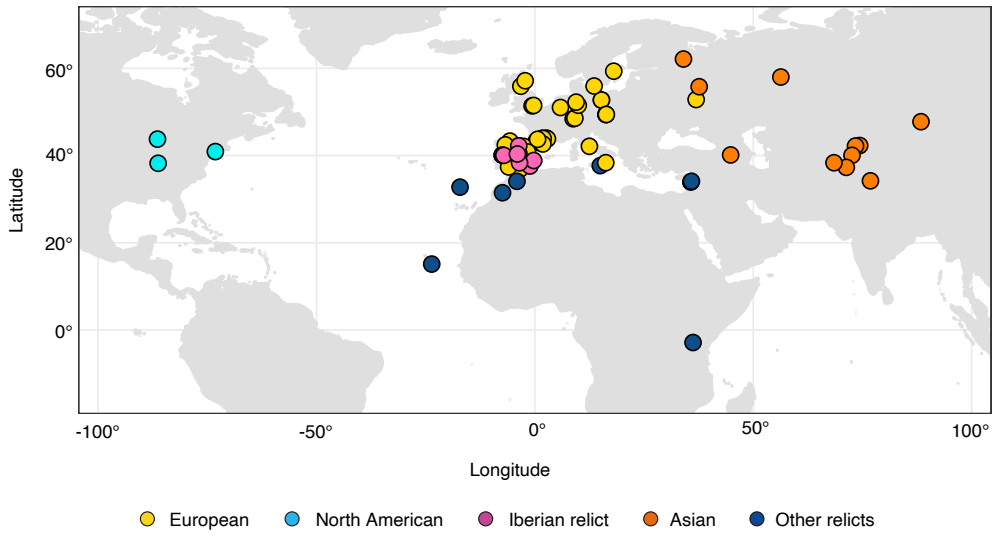

Figure 2.1 Geographic distribution of the 74 *A. thaliana* accessions. Genetic groups are indicated by colors.

2.2.2 Principal component analysis

A principal component analysis (PCA) was performed to elucidate the genetic relationship among the 74 accessions. HiFi reads from all accessions were aligned to the Col-0 Community-Consensus (Col-CC) assembly (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_028009825.2/) by minimap2 v2.26 (Li 2018) with the parameter -ax map-hifi. The output SAM files were converted to BAM format using Samtools v1.10 (Li et al. 2009) functions view -Sb, sort and index. Site depth was calculated from the aligned BAM files with mosdepth (Pedersen and Quinlan 2018). Single-nucleotide polymorphisms (SNPs) were identified using DeepVariant v1.6.0 (Poplin et al. 2018). GVCF files for each individual and each chromosome were merged into five chromosome files with GLnexus v1.4.1 (Yun et al. 2021). Sites with depth lower than 5, greater than twice the mean depth, or with a genotype quality lower than

30 were discarded. Bcftools v1.17 (Danecek et al. 2021) was used to filter SNPs with the parameter -m 2 -M 2 -i 'QUAL > 30 && MAF > 0.01 && F_missing < 0.2', to merge VCF files and to exclude repetitive regions identified by SRF (Zhang et al. 2023b) along with KMC v3.2.1 (Kokot et al. 2017). PCA was conducted using GCTA v1.94.1(Yang et al. 2011) with input BED files generated by PLINK v1.90b7.2 (Purcell et al. 2007) .

2.2.3 Extraction of telomeric sequences

In *A. thaliana*, two out of ten chromosome ends have large 45S rDNA repeat arrays adjacent to the telomeric repeats, causing most assemblies collapse and thus preventing correct mapping of telomeric sequences (Copenhaver and Pikaard 1996; Fultz et al. 2023a). Two alternative strategies were employed to extract telomeric sequences, depending on whether the sequence was adjacent to long 45S rDNA sequences.

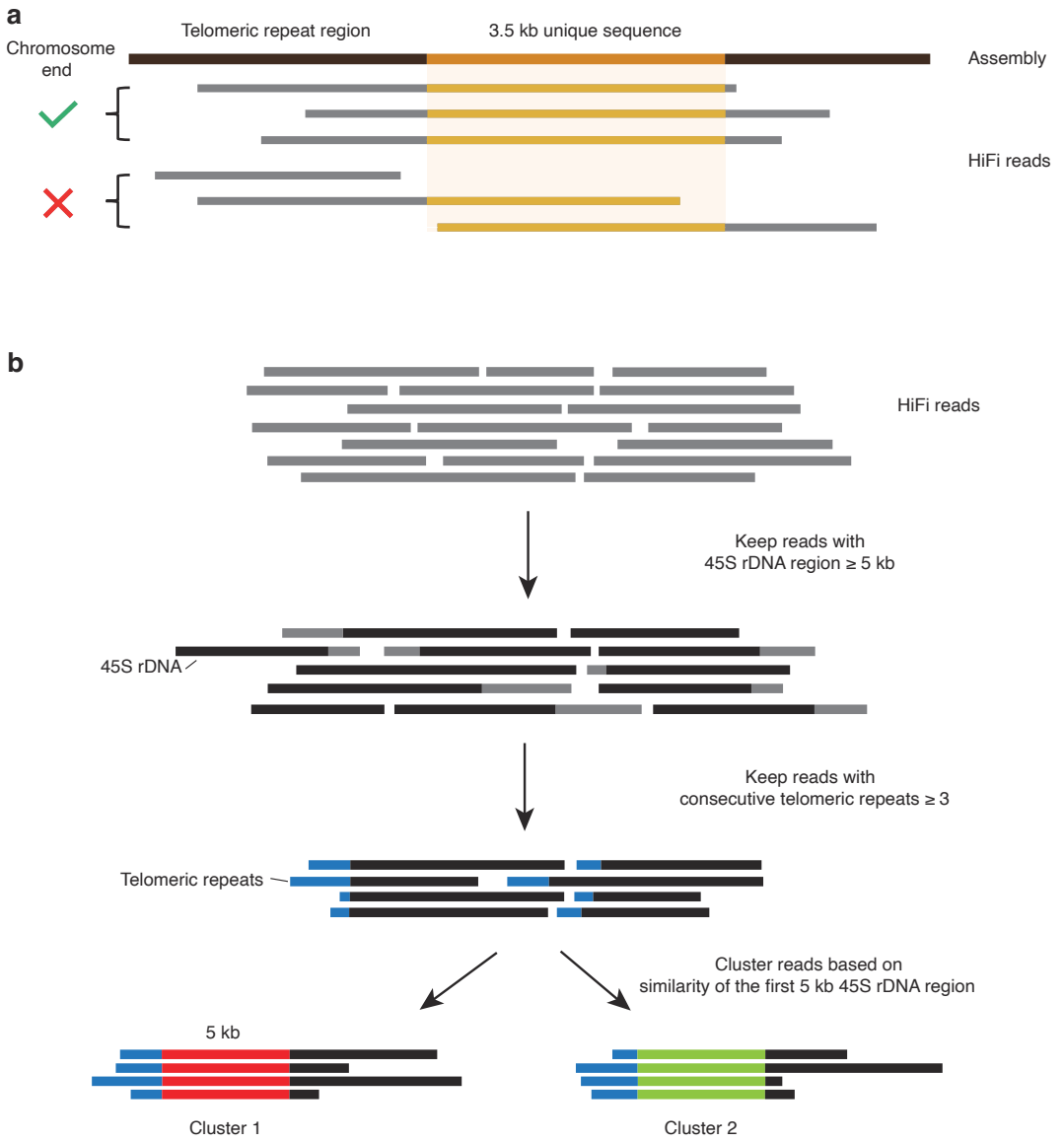

Figure 2.2 Schematic illustration of the strategies for extracting telomeric reads. **a** Strategy for the eight non- rDNA chromosome ends. Telomeric reads that contained at least 3.5-kb repeat-adjacent sequences from the relevant genome assembly were extracted. **b** Strategy for the two rDNA-binding chromosome ends. Reads containing 45S rDNA sequence and more than three consecutive telomeric repeats were extracted without the help of genome assemblies and clustered into two groups based on sequence similarity of the 45S rDNA region.

For the eight non-rDNA chromosome ends, an alignment-based approach was employed (Figure 2.2a). For each sample, HiFi reads were aligned to the

corresponding assemblies. Since the repeat-adjacent regions of different chromosome ends, which serve as markers for uniquely anchoring reads, were known to be similar in sequence (Heacock et al. 2004), all-against-all pairwise alignments of the 5 kb sequence adjacent to the telomeric repeats were performed for each chromosome end with BLAST v2.13.0 + (Altschul et al. 1990) . This resulted in a maximum alignment overlap of 3,056 bp (between chr3q and chr4q of Cvi-0). Therefore, only reads containing at least 3.5 kb of repeat-adjacent sequence were extracted with samtools view -hb -L (Farrell et al. 2022; Grigorev et al. 2021; Tan et al. 2022). BAM files were converted to FASTA format using samtools bam2fq and processed with seqtk v1.3 (https://github.com/lh3/seqtk) using option seq -A. For each accession, an all-against-all alignment was performed on the extracted reads using TIPP (https://github.com/Wenfei-Xian/TIPP). The resulting data were used to generate network graphs with R package igraph (Csardi and Nepusz 2006) to verify the accuracy of the read extraction. Potentially chimeric reads and reads containing sequencing errors were excluded after visual inspection. All reads were manually clipped to remove non-repeat sequences, retaining only the telomeric tracts. Since the irregular degenerate repeat content made the boundary between the non-repeat and repeat portions ambiguous, the start of the telomeric repeat array was arbitrarily defined as the first instance of the sequence (T)x(M)(G)y(M) (M = A or C).

For the chr2p and chr4p ends, which contain large 45S rDNA arrays, reads were directly extracted without help of the corresponding genome assembly (Figure 2.2b). Using minimap2, reads that aligned to the 45S rDNA sequence of Col-0 (Rabanal et al. 2022) were identified. Reads with at least three consecutive telomeric repeats were further retained. The 45S rDNA portions of these retained reads were aligned pairwise using BLAST. It resulted in the length of identical 45S rDNA sequences being either less than 4,800 bp or nearly the entire length of the query sequence. Reads with at least 5 kb of 45S

rDNA sequences were thus extracted and clustered into two groups, putatively from chr2p and chr4p, per accession based on sequence similarity. Based on a 45S rDNA reference sequence (Rabanal et al. 2022), RepeatMasker v4.0.9 (https://www.repeatmasker.org/) was used to mask and exclude the rDNA regions from the reads, leaving only the telomeric repeats for further analysis.

To facilitate downstream analysis, reads with telomeric repeats in the 3′-CCCTAAA-5′ orientation were first reversed to 5′-TTTAGGG-3′ using seqtk with function seq -r, followed by processing with Tandem repeats finder v4.09.1(Benson 1999) to identify repeat units (Belyayev et al. 2023; Lyčka et al. 2024). After manual curation, units were arbitrarily defined as beginning from the first T and ending with the last non-T base along the sequence. For example, the sequence TGTTTAGGGTCTGATGGG was split into the units TG TTTAGGG TCTGA TGGG.

2.2.4 Evaluation of short homopolymer errors

Because at each end of the reads, small indels rather than SNPs, particularly 1-bp deletions, often dominated the consecutive canonical TTTAGGG repeat regions, specifically TTAGGG (with two instead of three Ts) or TTTAGG (with two instead of three Gs), and these occurred at random positions. To determine whether these indels were caused by somatic mutations or sequencing errors (Lal et al. 2021), the correlation between the likelihood of errors and the occurrence of indels for each read was examined.

The likelihood of error was quantified based upon subread coverage and quality value of the HiFi reads. Samtools view -X followed by an awk command was used to extract the values of two tags, “np” (number of subreads) and “rq” (read quality), per read from the BAM files. To calculate the occurrence of indels, sequences were extracted from the read end until the variant repeat preceding

the canonical repeat array. The length of each extracted sequence was divided by seven to obtain an approximate repeat number, and a hill-climbing algorithm was used to find the nearest integer that represented the canonical repeat number in the ideal sequence (Figure 2.3), minimizing the Levenshtein distance (L-distance) between the extracted sequence (the observed string) and the ideal sequence consisting entirely of canonical repeats (the expected string), obtained with the R package stringdist. This minimized distance was further divided by the length of the extracted sequence to determine the relative distance as an indication of indel density. The correlation between the likelihood of errors and the occurrence of indels for each read was plotted using R package ggplot2, and the R function cor was used to calculate the Pearson correlation coefficient.

Figure 2.3 Example of the process for determining the expected string and calculating the L-distance, which represents the occurrence of indels. Blue blocks indicate variant repeats, while gray blocks indicate canonical repeats. ‘S’ indicates ‘string’, ‘n’ should be an integer.

2.2.5 Identification of telomeric repeat content

To visualize the degenerate and variant repeat arrays, consensus sequences were generated from two random reads with the median length of repeat array for each accession and each chromosome end. Conserved units between reads were retained, while nonconserved units were marked as “N.” The frequency of occurrence for each unit type was subsequently calculated. The

positions of the top 20 enriched unit types were then emphasized with different colors.

In addition, non-repeat sequences that disturbed the repeat arrays were manually extracted. Using BLAST, the sources of these non-repeat sequences were determined with TAIR10 transposon and organellar DNA sequences (Lamesch et al. 2012) as well as a library of *A. thaliana* rDNA and centromere sequences (Rabanal et al. 2022).

2.2.6 Identification of telomerase RNA template sequence

In *A. thaliana*, the addition of telomeric repeats is directed by a 9-bases template 3′-UCCCAAAUC-5′ in the telomerase RNA, corresponding to 3′-TCCCAAATC-5′ in the genome (Song et al. 2019). To investigate whether the variants we observed were caused by mutations in the template sequence, all 74 assemblies were searched using BLAST with the sequence of the telomerase RNA locus of Col-0 as a query (Fajkus et al. 2019). Corresponding sequences were extracted using bedtools v2.27.1 (Quinlan and Hall 2010) with function getfasta and used as input for a multiple sequence alignment with Clustal Omega (Sievers et al. 2011).

2.2.7 Estimation of telomeric repeat variants in HPG1 accessions

Three HPG1 accessions (14INRCT07, Pent-46, LI-EF-011) were sequenced. To assess the repeat number variation, the length of the sequences containing degenerate and variant repeats was calculated for each read with an awk script. The significance of the difference in length between accessions was evaluated with a two-tailed F test using the R function var.test. The length of each read was plotted using ggplot2.

2.2.8 Estimation of telomeric repeat variants in different Col-0 datasets

Three datasets of the Col-0 accession (Wang et al. 2022; Naish et al. 2021; Rabanal et al. 2022) were compared using the methods described above. The R function var.test was additionally used to assess whether different sequencing strategies (single-plant versus pooled) affected the distribution of repeat number variation of HiFi reads.

2.2.9 Haplotype structure analysis of the repeat arrays and their adjacent non-coding regions

For telomeric repeat arrays, a repeat compression approach for each sequence was used (Rautiainen et al. 2023), in order to reduce the complexity arising from repeat number variation (Figure 2.4). Pairwise L-distances between compressed arrays were calculated to estimate their similarity, and these distances were then divided by the length of the longer sequence in each pair to determine the relative distance. An F test was performed to assess whether there were significant differences in the similarity levels when comparing the same and different chromosome ends and comparing the same and different genetic groups.

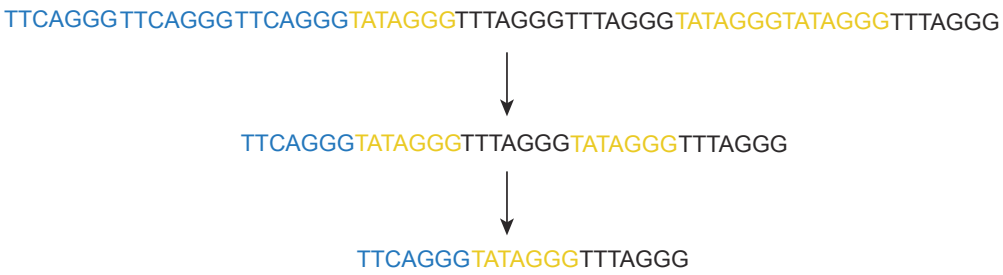

Figure 2.4 Schematic representation of the repeat compression process.

To identify the more proximal non-coding regions, Liftoff v1.6.3 (Shumate and Salzberg 2021) was used in conjunction with the TAIR10 gene set (Lamesch et al. 2012) to annotate the most terminal gene in the eight non-rDNA chromosome ends (Vrbsky et al. 2010). Subsequently, the fragment between the most terminal gene and the first telomeric repeat was extracted using bedtools getfasta. Multiple sequence alignment and NJ clustering of non-coding sequences was performed for each end with Clustal Omega, and pairwise relative distances were calculated.

To determine whether there was any correlation between variation in the telomeric repeat arrays and the non-coding regions, the relative distance values for both the repeat and the non-coding regions were merged into a square matrix. The order of accessions for each chromosome end was determined based on the NJ clustering of the non-coding regions.

2.3 Results

2.3.1 Profiling telomeric regions in *Arabidopsis thaliana*

To investigate the sequence content of telomeric regions (Figure 2.5), defined here as canonical telomeric repeats, adjacent variant and degenerate telomeric repeats as well as any unique sequences interspersed in these repeats, HiFi reads from 74 *A. thaliana* accessions of diverse geographic origins were used. Among them, 66 accessions were grouped into four main genetic clusters, with 43 non-relict accessions from Europe, 11 from Asia, 9 from Iberian relicts, and three from North America. Eight further accessions were from various relict groups (Wlodzimierz et al. 2023; 1001 Genomes Consortium 2016; Lee et al. 2017).

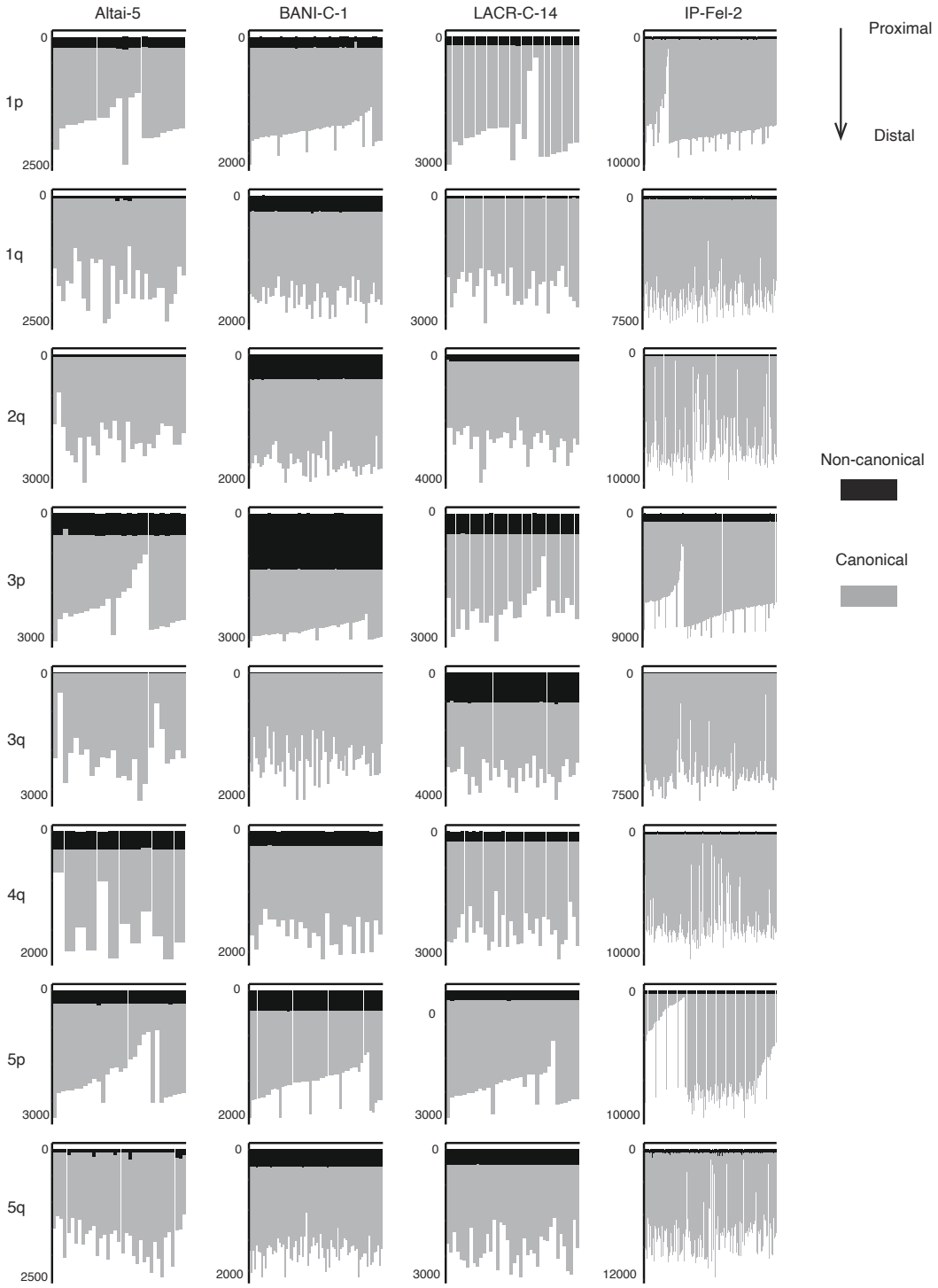

Figure 2.5 Bar plot showing the length of canonical and non-canonical (degenerate and variant) repeat arrays for each non-rDNA chromosome end in four accessions. Each vertical bar represents a single HiFi read.

For each accession, HiFi reads were unambiguously extracted for the eight non-ribosomal DNA (rDNA)-binding chromosome ends (Figure 2.6). For the ends of the p-arms of chromosome 2 and 4 (hereafter, chr2p and chr4p), which remain incompletely assembled due to large 45S rDNA tandem arrays that are immediately adjacent to the telomeres (Copenhaver and Pikaard 1996), reads could be assigned to two groups but could not be precisely assigned to chr2p or chr4p.

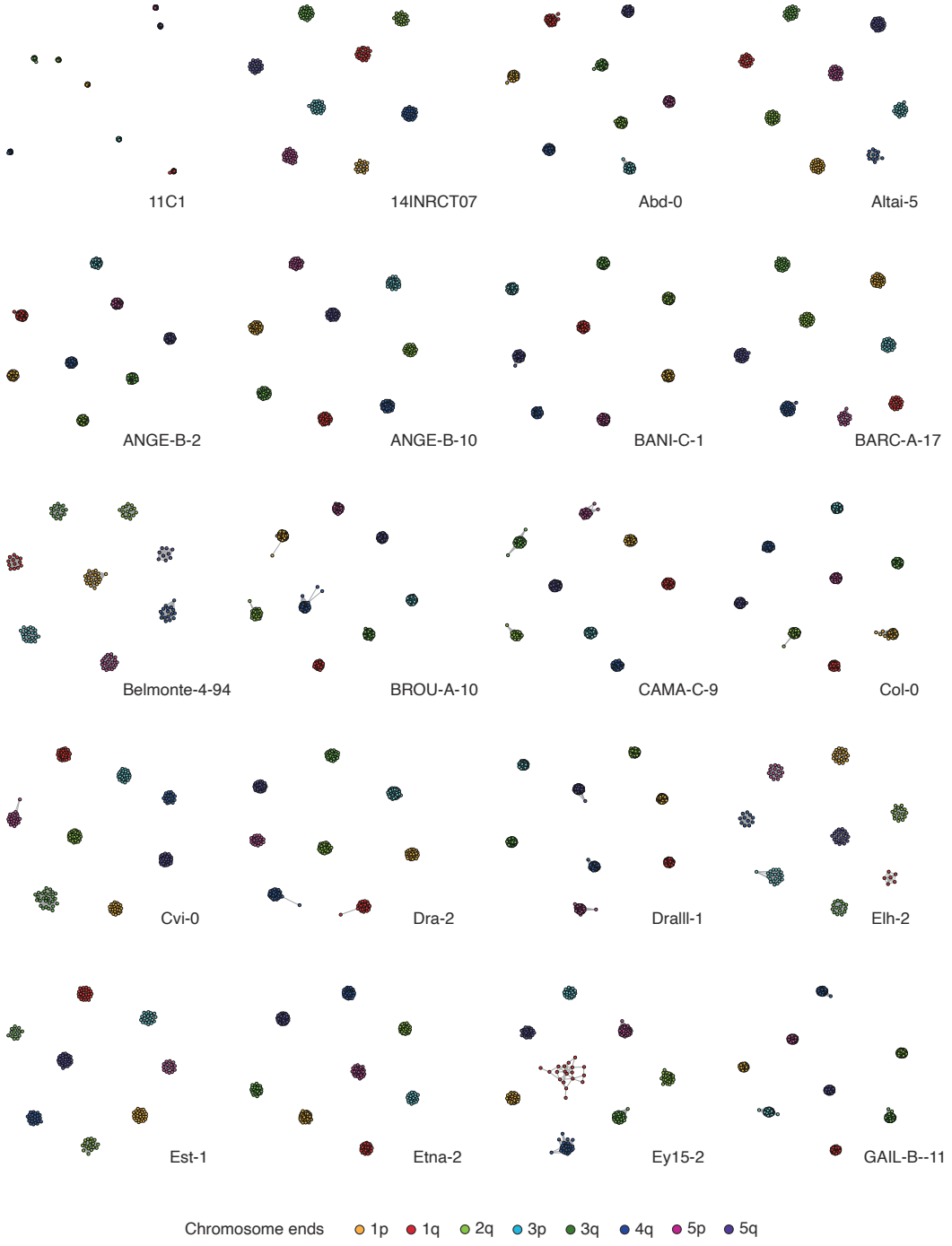

Figure 2.6 Sequence relationships of telomeric reads from non-rDNA chromosome ends of 20 accessions. Chromosome ends are indicated by colors.

Starting from the centromere-proximal side, the telomeric regions typically start with a stretch of degenerate repeats, followed by variant repeats and finally

canonical repeats, all of which were in the same head-to-tail arrangement (Figure 2.7a). The most obvious exceptions to this general pattern were chr2p and chr4p ends, where only 11 accessions had variant repeats. Additionally, 30 accessions contained non-telomeric fragments within the repeat arrays, and these are described in detail below.

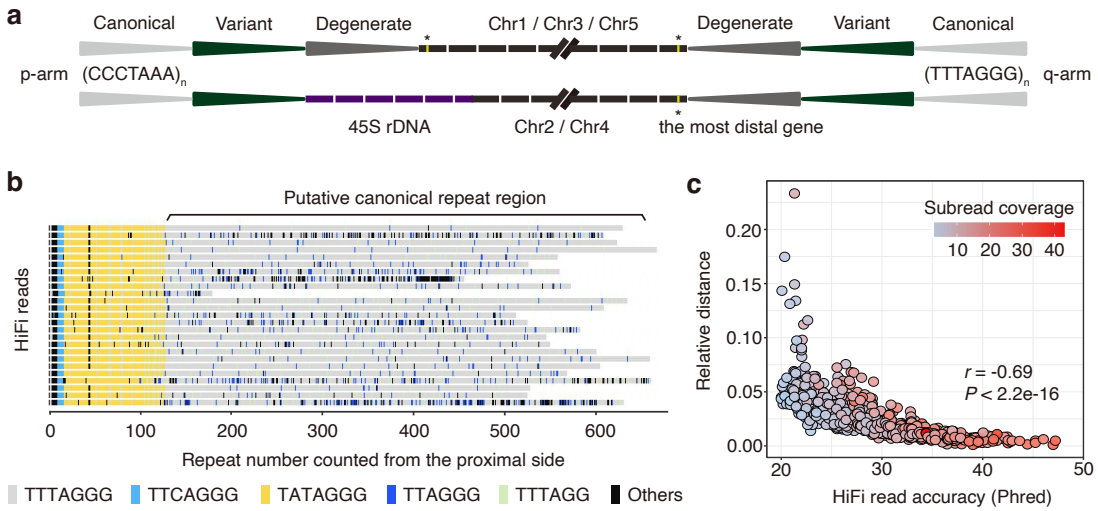

Figure 2.7 Overview of the telomeric repeat regions in *A. thaliana*. **a** Schematic representation of the different types of telomeric repeats at non-rDNA and rDNA chromosome ends. **b** Alignment of HiFi reads showing the entire telomeric repeat array in chr1q of Pent-46 accession from degenerate, variant to canonical repeats (from left to right). **c** Correlation between relative distance from expected canonical repeat sequence and entire read accuracy.

The arrays of canonical telomeric repeats at the very end were observed to harbor many indels in each read, primarily 1-bp indels, usually replacing TTTAGGG with either TTAGGG or TTTAGG (Figure 2.7b). By comparing HiFi read accuracy, the number of full-pass subreads, and the relative distance from an ideal sequence that is the entire canonical array for each read, a statistically significant negative correlation was found between relative distance and both read accuracy ($P < 2.2e-16$, Pearson's $r = -0.69$; Figure 2.7c) and subread coverage ($P < 2.2e-16$, Pearson's $r = -0.42$). Since only indels and no other mutation types were found in this region, the relative distance serves as an

indication of indel density. This result suggests that the occurrence of indels is influenced by the read quality. Different from a previous study that interpreted indels supported by a single read as genuine variants (Grigorev et al. 2021), I consider these indels to be short homopolymer run errors (e.g., TTT > TT or GGG > GG), a known issue with HiFi reads (Loomis et al. 2013; Lal et al. 2021). Therefore, the region beginning at the last conserved variant repeat until the read end was defined as the homogeneous canonical TTTAGGG repeat region. Because it was deemed to be devoid of consistent variation, this region was not further considered in the remainder of analyses.

2.3.2 Hypervariable composition of telomeric repeat arrays

Using the extracted reads, I generated consensus sequences of degenerate and variant telomeric repeats for each chromosome end in the 74 accessions. To obtain a first overview of variation, the 20 most enriched repeat types were visualized (Figure 2.8a). Sequences of accessions were ordered according to their membership in genetic groups (Figure 2.8b).

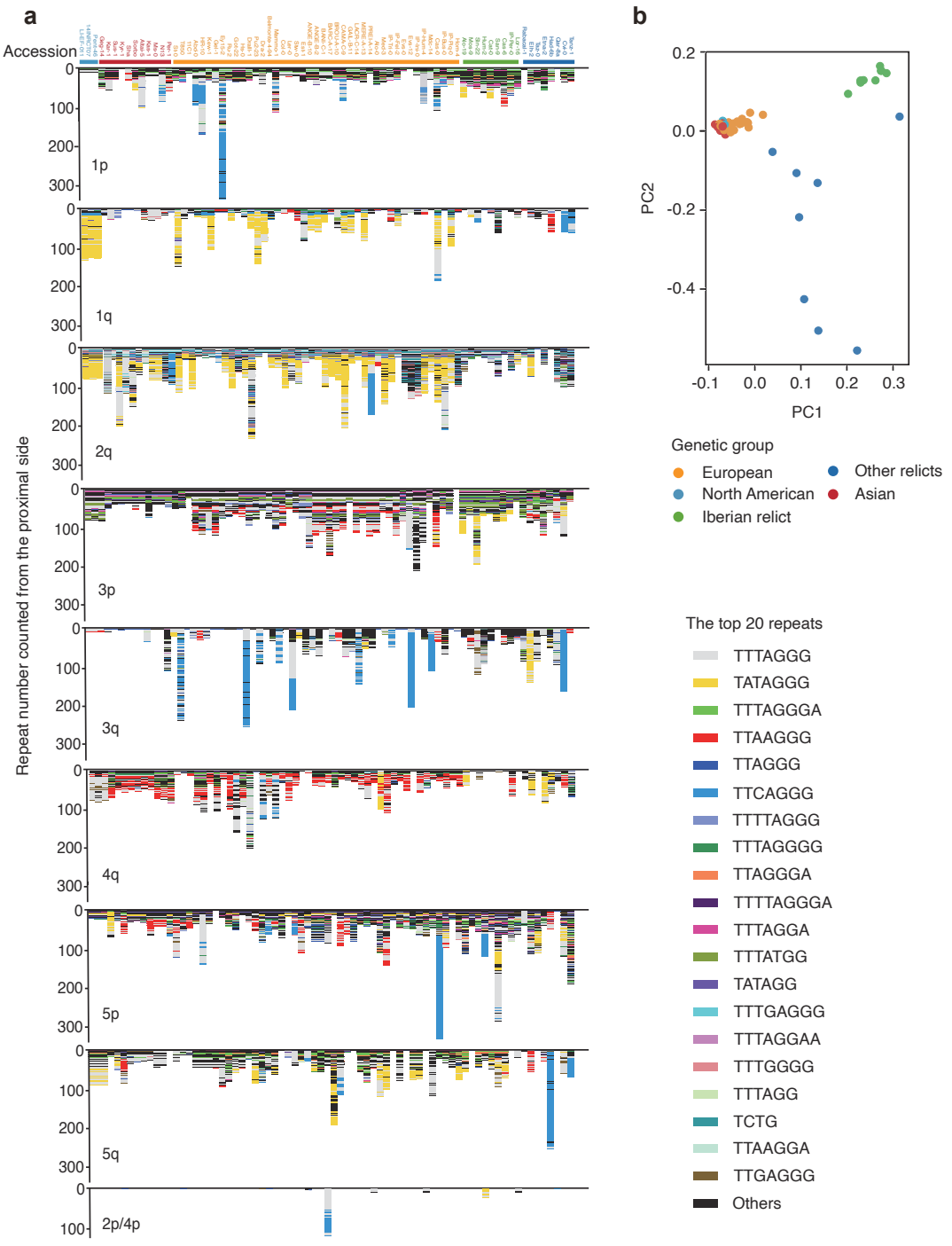

Figure 2.8 Variation of telomeric sequences in *A. thaliana*. **a** Degenerate and variant (from top down) telomeric repeat arrays at 10 chromosome ends across all 74 accessions. The top 20 most enriched units are highlighted by different colors. **b** Genetic groups of 74 accessions revealed by principal component analysis.

Of the 592 non-rDNA chromosome ends, 562 had variant repeat arrays, with lengths from 6 to 3,384 bp (chr1p of Ey15-2). Of the 148 rDNA ends, only 12 had variant repeat arrays, with lengths from 6 to 658 bp. A total of 462 distinct repeat units, ranging in size from 2 to 17 bp, were identified. The number of new repeats reached saturation with the 69th accession (Figure 2.9). Of the 462 distinct repeat units, 151 (32.7%) occurred only once. The canonical repeat, which was interspersed among arrays of variant repeats, had the highest frequency with 20.7%. It should be noted that the count of distinct repeat types greatly relies on my definition of a unit. For example, the sequence TTTAGGATTAGGG could be considered as being composed of two variant repeats, TTTAGGA and TTAGGG or TTTAGG and ATTAGGG. Therefore, I use the repeat types as a set of markers for studying the overall organization of telomeric sequences and believe that there is no need to excessively focus on the specific sequence content of individual units, especially rare ones.

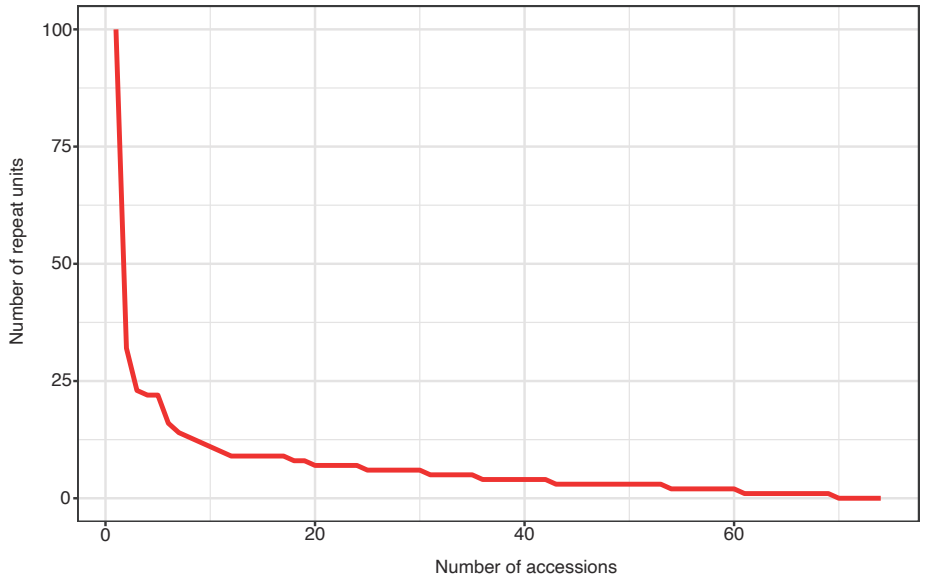

Figure 2.9 Number of new repeat units added with an increase in the number of accessions.

As an aside, the template sequence of the telomerase RNA, 5′-CUAAACCCU-3′ (Song et al. 2019), encoded on chromosome 2, was identical in all 74 accessions.

Although the sequence content of telomeric regions was highly dynamic, there were five main patterns of sequence variation and most variant sequences, 508 of 574, showed more than one of these patterns. The simplest pattern was represented by arrays in which different repeat types occurred only once, such as chr1q of Cat-0. The second pattern most likely resulted from monomer homogenization, such as chr1p of Alo-19, where a single unit, TATAGGG, was repeated consecutively 15 times (Figure 2.10a). The remaining patterns constituted higher-order repeats (HORs; Garrido-Ramos 2017). In the simplest case, such as chr3q of IP-Tri-0, two to four units made up a block that was then repeated multiple times (Figure 2.10b). A more elaborate pattern had multiple monomers (arbitrarily defined as ≥ 5 here) that were repeated several times. For example, in chr2q of IP-Per-0, five distinct units formed a block and were repeated five times, with all five blocks being identical (Figure 2.10c). The final pattern also featured HORs but with mutations distinguishing the individual HORs. For instance, in chr2q of Cvi-0, the HOR array consisted of five units repeated eight times with five of these deviating from the consensus (Figure 2.10d)

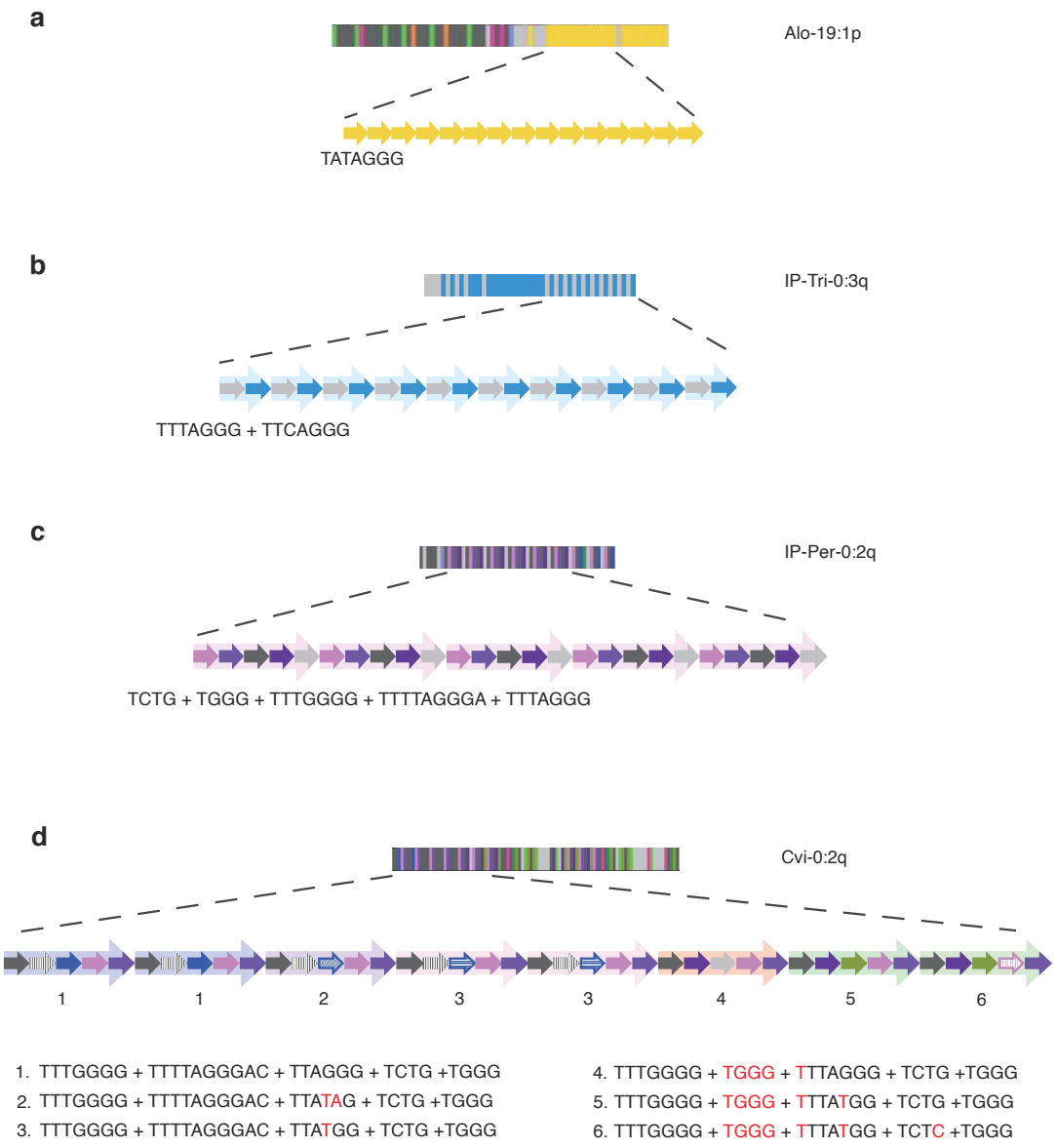

Figure 2.10 Close-up of four major types of sequence organization in the telomeric repeat arrays. Color code corresponds to that of Figure 2.8a. **a** An example of monomer homogenization. In this case, a single unit is repeated 15 times. **b** An example of simple block expansion. In this case, two units form a block and the block is repeated ten times. **c** An example of expansion of identical higher-order repeats (HORs). A HOR consisting of five distinct units is repeated five times. **d** An example of HORs with small sequence differences. Numbers indicate different HORs.

When comparing pairs of accessions, the majority of sequence differences between specific chromosome ends fell into three major categories (Fig. 2a). In

the first category, sequences were highly similar to each other, as seen in chr5p of 11C1 and HR-10 (Figure 2.11a). In the second group, sequence composition was similar, but accessions were distinguished by the number of HORs, such as chr3p of IP-Tri-0 and IP-Fel-2 (Figure 2.11b). These two categories were mainly observed with pairs from the same genetic group. The third category, sequence divergence, was observed not only in unrelated accessions but also in pairs from exactly the same local population, such as chr1p of Evs-0 and Evs-12 (Figure 2.11c).

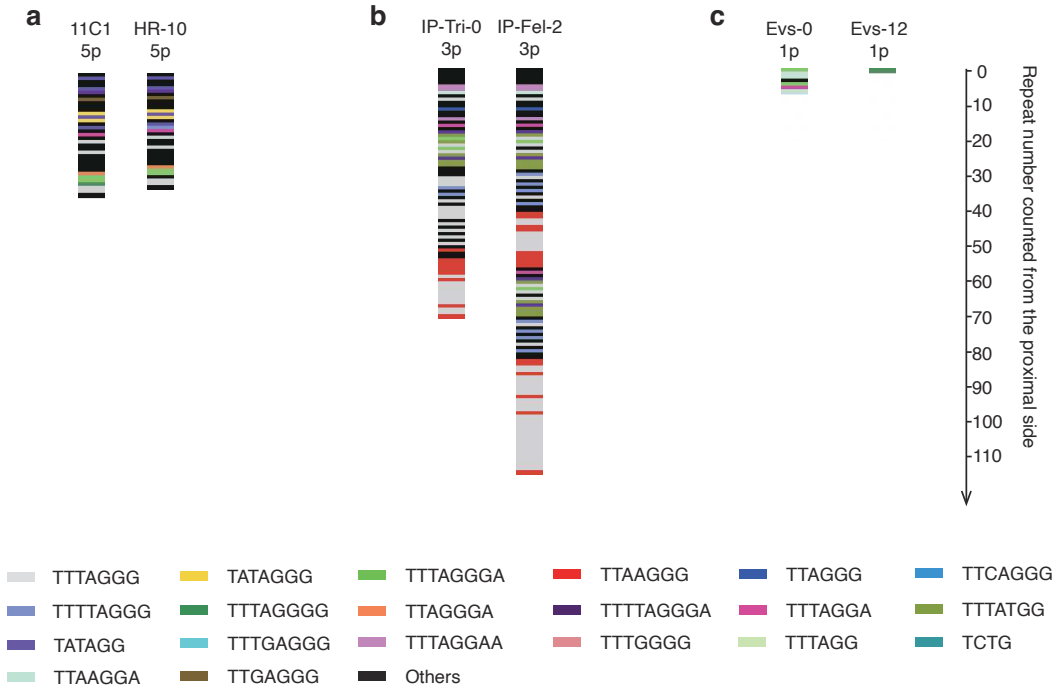

Figure 2.11 Close-up view of three categories of telomeric sequence relationships. **a** Telomeric sequences in chr5p of 11C1 and HR-10 are similar. **b** Telomeric sequence in chr3p of IP-Fel-2 has one more copy of a higher-order repeat than the sequence in chr3p of IP-Tri-0. **c** Telomeric sequences in chr1p of Evs-0 and Evs-12 are dissimilar.

Thirty accessions had non-telomeric sequences within the repeat array. Except for seven unclassified sequences ranging in length from 42 to 453 bp, the others could typically be divided into three different types. Firstly, organellar

DNA or rDNA insertions. In chr1p, 14 accessions had a 110-bp mitochondrial DNA insertion (Figure 2.12a), which has been reported previously (Kuo et al. 2006), while chr2q of Cvi-0 contained a 102-bp chloroplast DNA insertion. A 5088-bp 45S rDNA sequence was embedded in the telomeric tract in chr2q of Gel-1. In the second type, seven accessions were observed to have non-telomeric fragments that were associated with repeat array duplications. For example, chr2q of four accessions has a 244-bp sequence that forms HORs in combination with their telomeric repeats. The 244-bp fragment is identical in all HOR copies, while the repeat array exhibits a few polymorphisms (Figure 2.12b). The third type was exemplified by chr3q of Hum-2, where the repeat array was interrupted by a 495-bp non-telomeric fragment, which was identical in sequence to a fragment adjacent to the array of variant telomeric repeats of chr5q in the same accession (Figure 2.12c). The distal part of chr3q closely resembled the repeat array of chr5q.

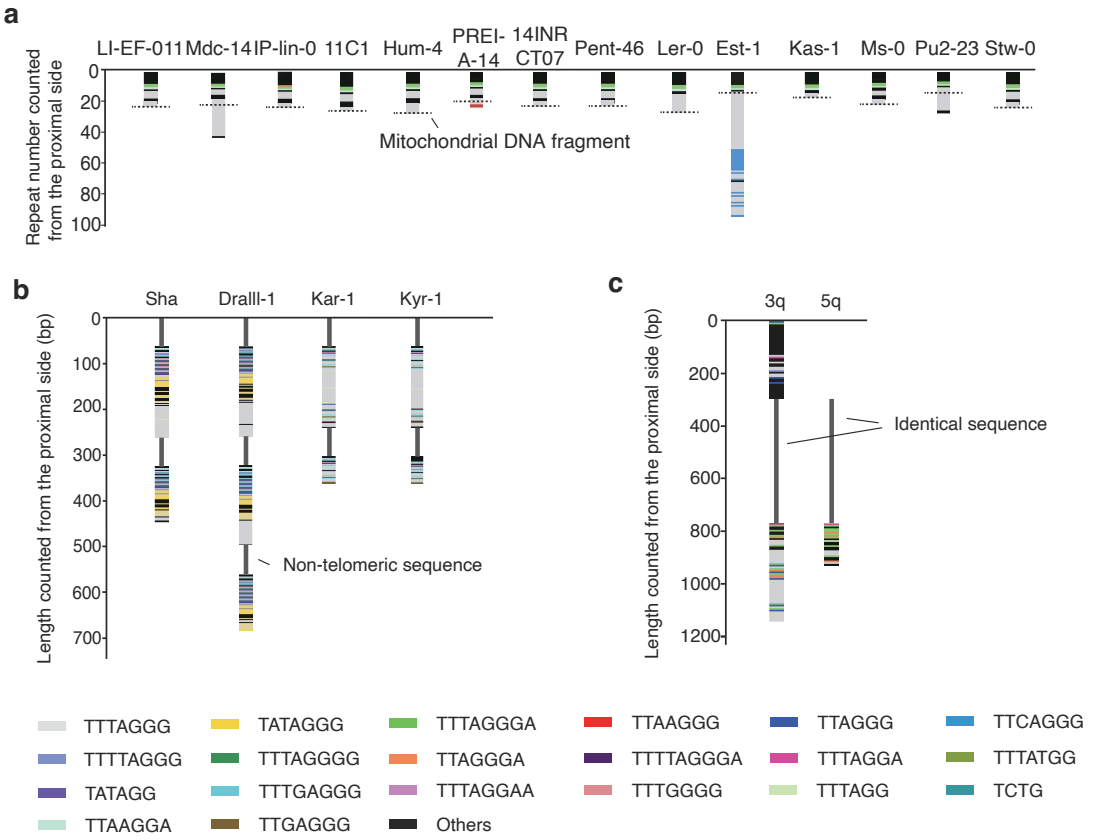

Figure 2.12 Representation of three categories of non-telomeric fragments in telomeric repeat arrays. **a** A 110-bp mitochondrial DNA insertion in 14 accessions. **b** Higher-order repeats in four accessions include a 244-bp non-telomeric fragment. **c** A 495-bp unique sequence in chr3q of Hum-2 is also found in chr5q of Hum-2.

2.3.3 Repeat number variation between closely related individuals and in somatic tissues

To examine variability in the telomere regions in a more fine-grained manner, two collections of datasets from very closely related individuals were employed. The first collection came from the lineage of North American accessions known as haplogroup-1 (HPG1), which form a clade of natural mutation accumulation lines whose common ancestor lived about 400 years ago (Exposito-Alonso et al. 2018). In parallel, three independent sequencing datasets of the Col-0 accession that had been recently published were investigated (Wang et al.

2022; Naish et al. 2021; Rabanal et al. 2022). This also offered an opportunity to examine intra-dataset variation in more detail. I therefore report not merely the most common repeat array length but present the full data for all HiFi reads.

Among the three HPG1 accessions, repeat number variation was found, but no major differences were observed in repeat type. Specifically, four of eight non-rDNA chromosome ends were significantly different in lengths of degenerate and variant repeat regions, with medians differing from 7 to 51 bp (Figure 2.13a). There was also substantial variation in repeat number among the HiFi reads from a single accession. The greatest one, from 396 to 569 bp, corresponding to approximately 25 repeat units, was observed at chr4q of 14INRCT07.

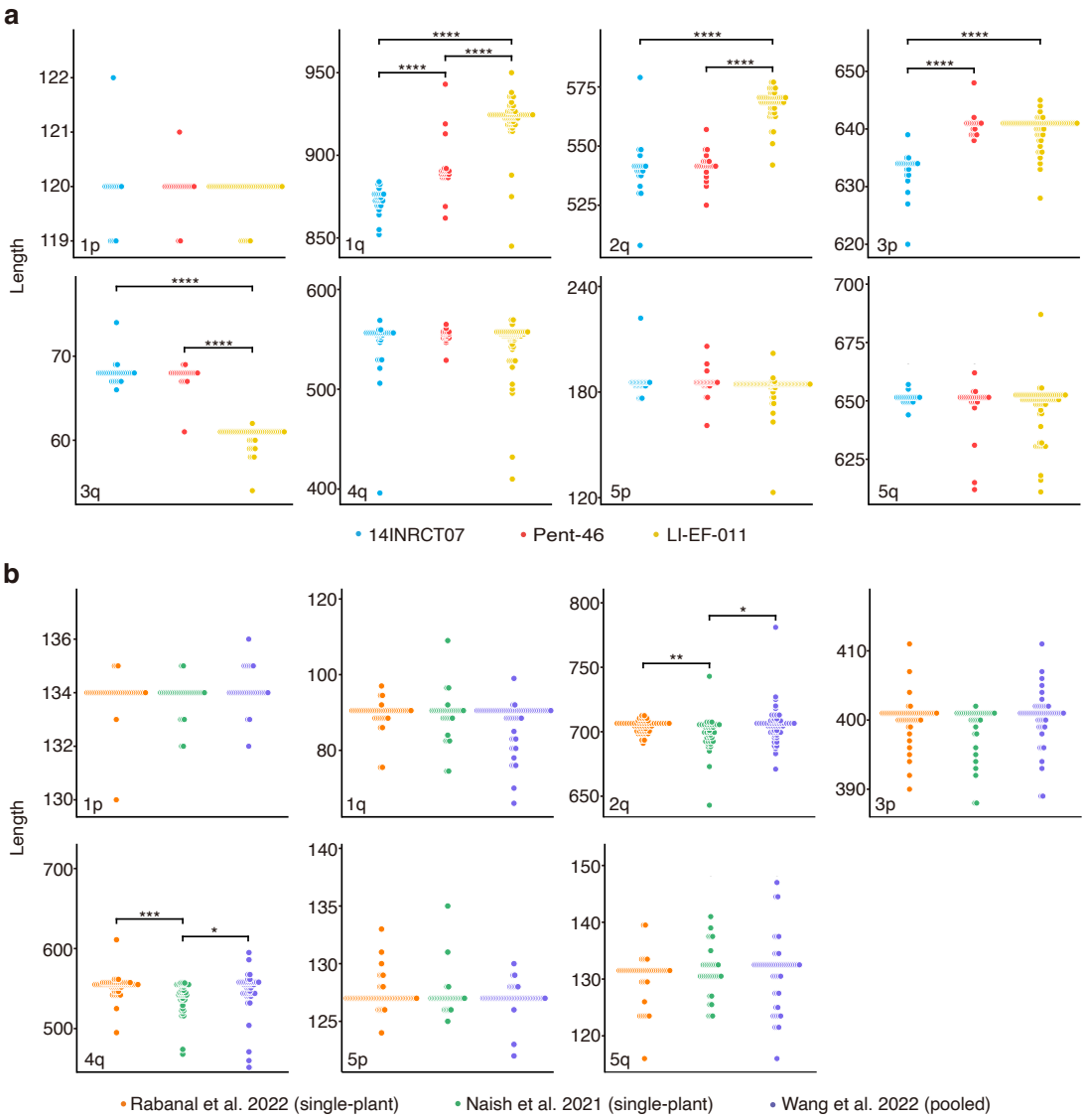

Figure 2.13 Plots showing variation in the length of degenerate and variant telomeric repeat regions in sets of three closely related samples. Dots represent individual HiFi reads. Statistically significant differences were determined by a two-tailed *F* test (****P<0.00001, ***P<0.0001, **P<0.001, *P<0.01). **a** Comparison of the three HPG1 accessions. **b** Comparison of the three Col-0 datasets.

In the Col-0 accession, the array of telomeric repeats of chr3q was found to exclusively consist of canonical repeats, and it was therefore excluded from this analysis. For the remaining seven non-rDNA chromosome ends, there was no difference in variant types. Regarding repeat number variation, at two of seven chromosome ends, one dataset differed significantly in length distribution from

the other two datasets, with median differences of 7 bp and 11 bp (Figure 2.13b). These two chromosome ends had also the longest repeat arrays. For within-dataset length variation, chr4q was the one with the greatest difference between the shortest and longest arrays of degenerate and variant repeats, at 184 bp, roughly equivalent to 26 repeats. While four of seven chromosome ends differed significantly in the degrees of variability among the three Col-0 datasets (Figure 2.14), these differences were not attributable to the pooled-sequencing dataset. Thus, differences in sequencing strategy should not affect my conclusions regarding the 74 diverse datasets I used, which had been generated by a combination of pooled and single-plant sequencing.

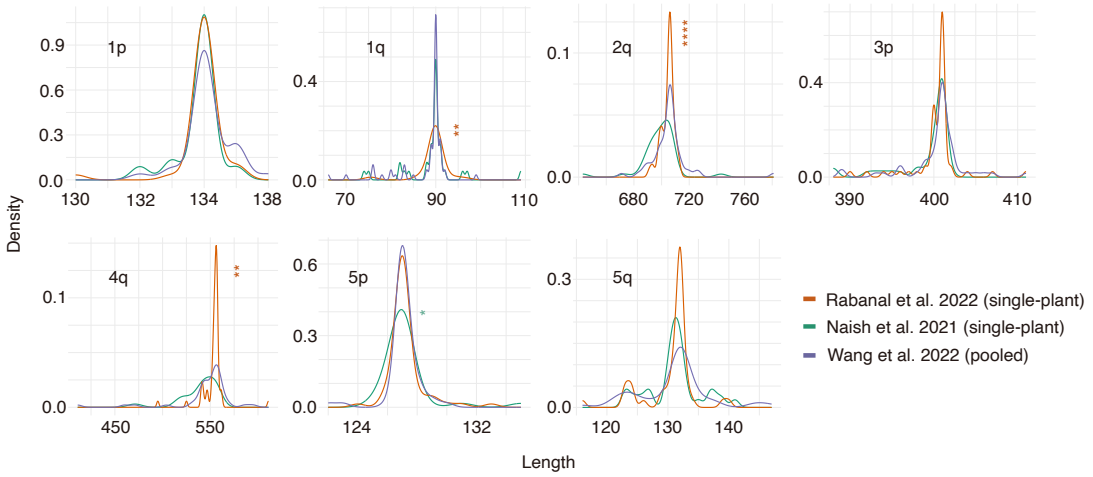

Figure 2.14 Density plot of the length distribution of degenerate and variant repeat regions at seven non- rDNA chromosome ends in three Col-0 datasets. Statistically significant differences in the deviation degree are indicated (****P<0.00001, ***P<0.0001, **P<0.001, *P<0.01, *F* test).

2.3.4 Haplotype structure of telomeric repeat arrays and the adjacent non-coding regions

To facilitate the comparison of haplotypes across the telomeric arrays, I implemented a repeat compression process to mitigate the impact of repeat

number variation, which is likely to change more quickly than the overall arrangement and presence of variant repeats. The compressed sequences were used to perform a pairwise sequence similarity analysis based on the relative Levenshtein distance (L-distance). The result confirmed the visual impression from Fig. 2a that there is on average more similarity between the same chromosome end of different accessions than between different chromosome ends (Figure 2.15a; $P < 2.2e-308$, Wilcoxon test). The result also showed an overall lower relative L-distance within the same genetic group compared to between different genetic groups (Figure 2.15b; $P < 6.01e-59$, Wilcoxon test).

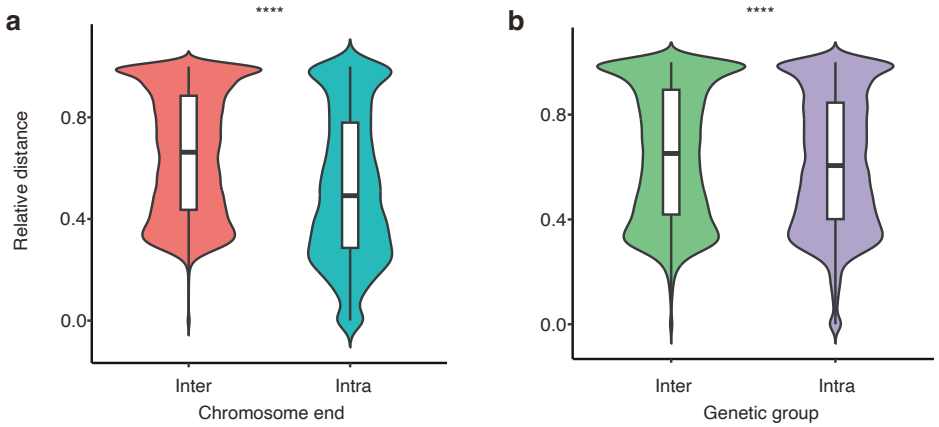

Figure 2.15 Violin plots showing the distribution of pairwise relative distances. Statistically significant differences between accessions are indicated (****P<0.00001, two-tailed F test). **a** Comparison of values within and between chromosome ends. **b** Comparison of values within and between genetic groups.

To examine whether these haplotype patterns extended beyond the telomeric repeat regions, we also looked at their adjacent non-coding regions. Non-coding sequences, which varied in length from zero to 16,542 bp, were defined as the sequence between the most distal gene and the last variant repeat of each chromosome end. Next, neighbor joining (NJ) clustering was conducted based on the multi-sequence alignment of these non-coding regions from each chromosome end. A merged matrix of repeat arrays and non-coding regions

was generated, using the accession order from the NJ exercise, to reveal the correlation between the two (Figure 2.16). Strong linkage between telomeric repeats and their adjacent non-coding regions were present at both coarse and fine resolution.

In addition to linkage disequilibrium, the matrix provided direct support for my statistical results regarding the chromosome end-specific and genetic group-specific patterns (Figure 2.16). Haplotypes from the same chromosome end clustered together, with accessions from the same genetic group typically having similar haplotypes.

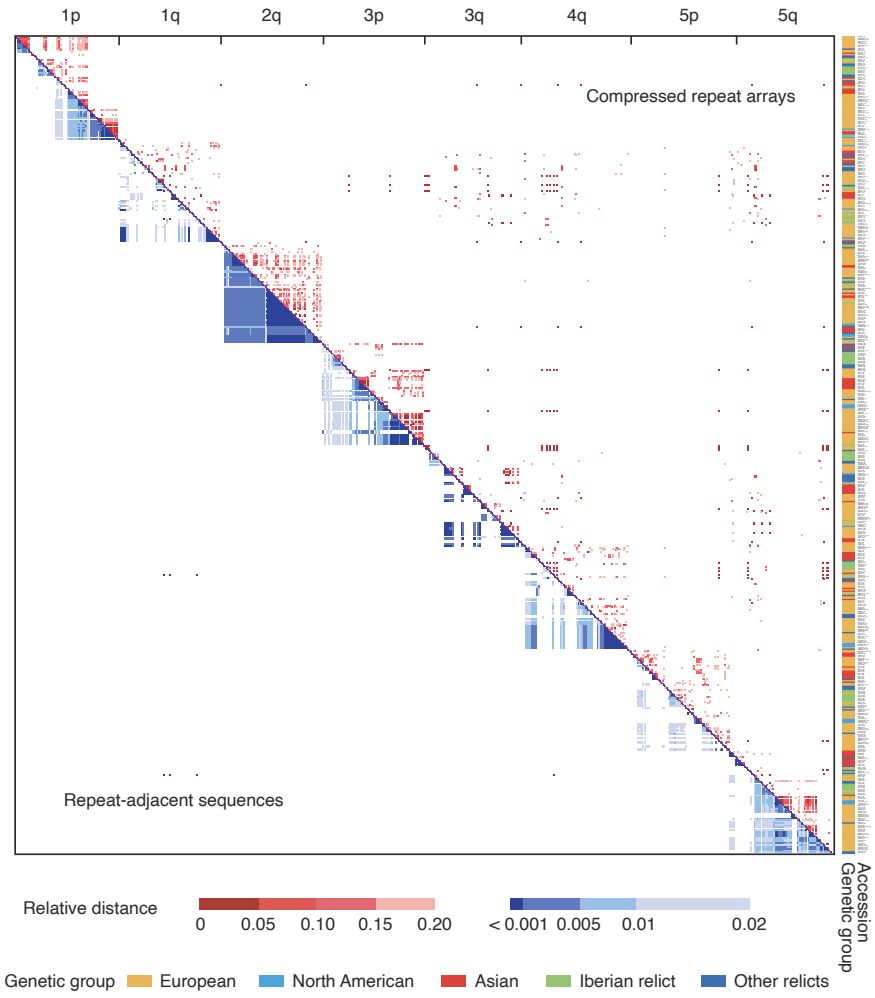

Figure 2.16 Heatmap of pairwise relative distance of compressed telomeric repeat arrays (upper triangle) and repeat-adjacent sequences (lower triangle). Membership of accessions in different genetic groups is indicated.

2.4 Discussion

My study provides a base-level view of the patterns of degenerate and variant telomeric repeats at the chromosome ends of 74 geographically diverse accessions of *A. thaliana*. The diverse sampling combined with technical advances provide a population-level view of telomeric sequence, going far beyond previously available anecdotal observations from a few common

accessions (Richards et al. 1992; Farrell et al. 2022). The superior length of PacBio HiFi reads supports unambiguous anchoring of the telomeric repeats to each chromosome end. In previous studies, total repeat abundance was reported without linking repeat location to telomeres in general, let alone to individual telomeres (Choi et al. 2021), or the focus was on only one chromosome end (Kuo et al. 2006). The superior read length mitigates the challenges arising from having multiple canonical repeats embedded within the variant repeat array, which can otherwise be taken as an erroneous indication of the chromosome end (Farrell et al. 2022). The number of variant patterns detected in my study reached saturation with the 69th accession of the 74 accessions. Therefore, I were able to detect not only mutation types of high frequency such as TTCAGGG (Richards et al. 1992) but also a much broader range of variant patterns that is likely to provide a near-complete inventory of variant types. Of note, my results overturn the previous conclusion that there are no variants at the two chromosome ends that cap the large 45S rDNA repeat arrays (Copenhaver and Pikaard 1996). In addition, my newly discovered higher-order repeats (HORs), which have before only been described in other satellite regions of *A. thaliana* such as the centromere (Naish et al. 2021). Regarding inter-chromosome similarity of the unique sequence and its subsequent telomeric tract, the only other similar example that I am aware of comes from *Caenorhabditis elegans* (Kim et al. 2019). Thus, my work greatly extends our knowledge of telomere-adjacent sequence variation up to the canonical array in this species.

There is ample evidence for local homogenization of telomeric repeats and formation of HORs, as well as repeat number variation in somatic cells and between closely related individuals, all typical characteristics of non-coding minisatellite regions (Garrido-Ramos 2017; Boán et al. 2004). The obvious scenario is that only the most distal portions of the canonical repeats, at the very ends of the chromosomes, are maintained by telomerase and thus remain

uniform (Song et al. 2019). More centromere-proximal portions are maintained by conventional DNA replication and can sustain mutations, becoming first variant repeats and eventually degenerate repeats over time (Kuo et al. 2006). In this scenario, the variant and degenerate repeats are minisatellite units of about 7 bp, and the extensive patterns of apparent repeat expansion and contraction can be explained by replication slippage and unequal crossing over (Symonds and Lloyd 2003). Two other forces shaping variant repeats have been considered in previous studies, and my analyses cannot rule out that they play a minor role as well. Variant repeats could in principle be caused by variation in the RNA template (Song et al. 2019; Fajkus et al. 2019). While I detected no sequence differences at the previously reported locus for the canonical RNA template, I cannot exclude the existence of other loci that contribute a minor amount of alternative templates (Závodník et al. 2023). Alternatively, variants could arise during reverse transcription (Gout et al. 2013), introducing variants into the newly added repeats at the most distal end of the array. Such errors will cause telomere elongation by telomerase to fail, with alternative mechanisms for telomere maintenance eventually taking over.

My haplotype analysis revealed both chromosome end-specific and genetic group-specific patterns of degenerate and variant telomeric repeat arrays. Accessions sharing the same haplotype are more likely to belong to the same genetic group (Grigorev et al. 2021), but they are not necessarily from the same local population (Kuo et al. 2006). In addition, I demonstrate that linkage disequilibrium between telomeric repeat arrays and more proximal non-coding regions, previously described for single chromosome ends in humans and *A. thaliana* (Baird et al. 2000; Kuo et al. 2006; Baird et al. 1995), as a common feature at all non-rDNA chromosome ends in *A. thaliana*. The mitochondrial DNA insertion event observed in accessions is a good example for summarizing these patterns in conjunction with the mutational process we propose. The 14 accessions, from different localities, contain a conserved mitochondrial

fragment and highly similar repeat-adjacent sequences, but the repeat arrays differ in sequence. A likely scenario is that the mitochondrial fragment was inserted before these 14 chromosome ends diverged (Kuo et al. 2006). Base substitutions in the telomeric repeat arrays then occurred stochastically in different accessions during repeat amplification.

My analysis has shown that telomeric repeats experience apparently much higher mutation rates than high-complexity sequences in chromosome arms, especially when it comes to repeat number. Telomeric repeats are therefore potentially helpful when attempting to reconstruct relationships between closely related individuals at high resolution. Information from telomeric repeats might become particularly useful if combined with genome-wide analyses of microsatellite and minisatellite mutations (Marriage et al. 2009). The substantial intra-individual variation in telomeric repeats also offers opportunities for studying the mechanisms of replication slippage and unequal crossing over of minisatellites (Smith 1976), given that the entire telomeric repeat arrays can be confidently captured by single HiFi reads.

My study leaves several open questions for future studies. One challenge will be to accurately assign telomeric reads adjacent to rDNA to specific chromosome ends, which has so far been hampered by a lack of complete assemblies of rDNA arrays across diverse genomes (Fultz et al. 2023a). Second, a few chromosome ends, including chr5p of Cas-0 as the most extreme example, had a large number of consecutive TTCAGGG repeats. The functional implication of this observation remains unknown. Lastly, I observed the sharing of the unique sequence across chromosome ends at chr2q and chr5q of Hum-2. This configuration, not yet reported in *A. thaliana*, has been proposed in a *C. elegans* study and in several reviews as evidence for chromosome healing, which involves a recombination process after a double-strand break (Kim et al. 2019, 2020; Baird 2018; Ballif et al. 2004). Further

validation of the mechanism underlying this sequence arrangement in *A. thaliana* is required.

I provide a comprehensive evaluation of nucleotide sequence polymorphisms of degenerate and variant telomeric repeat arrays at all chromosome ends in a global collection of diverse *Arabidopsis thaliana* accessions. I have greatly improved on the detection of telomeric repeat types, and report sequence arrangements including higher-order repeats and the sharing of unique fragments across chromosome ends, which to my knowledge had not been observed before in *A. thaliana*. The number of degenerate and variant telomeric repeats can vary at germline and somatic levels in otherwise isogenic accessions. Lastly, I reveal chromosome end-specific and genetic group-specific patterns of telomeric repeat haplotypes along with linkage disequilibrium between telomeric repeat arrays and their adjacent non-coding regions. Together, the findings improve our understanding of telomeric sequence diversity in plants.

3 Sequence asymmetry between strands during DNA replication in *Arabidopsis thaliana*

3.1 Introduction

Pacific Biosciences (PacBio) high-fidelity (HiFi) sequencing (Wenger et al. 2019), employs a circular sequencing strategy in which both strands of the same DNA molecule are sequenced several times. Each strand generates multiple subreads, and because library preparation does not require amplification, the resulting data reflect native DNA. The repeated passes yield not only high-accuracy information for the double-stranded DNA molecule, but also for each of the two strands, enabling the detection of differences between the two strands. It has been used to identify single-nucleotide differences that were interpreted as unrepaired point mismatch (Liu et al. 2024). Such an approach can also be used to find larger differences between the two strands, since read length is no longer a limiting factor, as reads are routinely many kilobase-pairs long.

Large differences between the two strands likely originate during DNA replication, specifically after synthesis of the new strand but before post-replicative repair. Several mechanisms can be responsible for large differences, including hairpin formation at palindromic sequences (Kurahashi et al. 2009), template switching (Ottaviani et al. 2014), and slipped-strand mispairing (Kiktev et al. 2018). These replication-association events may later be corrected by different repair mechanisms, or they may persist and become templates in the next round of DNA replication (Pearson et al. 1997), ultimately resulting in permanent large indels at the somatic level.

In this study, I present a method to detect such single-stranded, replication-associated events using unmodified HiFi sequencing data. I validate the approach using raw sequencing output from a nuclear genome sample of an inbred *Arabidopsis thaliana* accession, thereby avoiding complications from heterozygosity and haplotype phasing. I additionally use polymerase kinetic information to help distinguish biological events from sequencing artifacts possibility and characterize the sequence features associated with the detected events. Because the method relies on standard HiFi data, it can be applied broadly across species and sample types, offering a framework for studying polymerase activity and strand-specific sequence changes during DNA replication.

3.2 Methods

3.2.1 Data

The assembly and HiFi reads from leaves of *A. thaliana* accession Gel-1 are from a previous study (Wlodzimierz et al. 2023). Sequencing was performed on a PacBio Sequel II instrument, with availability of the raw sequencing output in bam format.

3.2.2 Generation and alignment of single-strand consensus reads

The raw subread bam file was processed using CCS v6.4.0 (https://github.com/PacificBiosciences/ccs) with the parameters –by-strand –chunk 30, generating separate consensus sequences for the forward and reverse strand of each read. Chunked bam files were merged using Pbmerge

v1.6.0 (https://github.com/PacificBiosciences/pbtk), and the resulting bam file was converted to fastq format with bam2fastq v1.3.1 (https://github.com/PacificBiosciences/pbtk).

Strand-specific consensus reads were aligned to the nuclear genome assembly using minimap2 v2.1 (Li 2018) with the parameter -ax map-hifi. The resulting alignment was processed with samtools v1.10 (Danecek et al. 2021) for viewing, sorting, and indexing, producing a final sorted bam file for downstream analysis.

3.2.3 Identification of read-level differences in complementary sequence

Differences in the complementary sequence between single-strand consensus reads and the assembly were identified from the sorted bam file using Sniffles v2.2 (Smolka et al. 2024) with the parameters –minsupport 1 –output-rnames –vcf. Variants supported by only a single read were extracted by filtering for the SUPPORT=1 value in the vcf file.

For each candidate event, the read (ZMW) ID was extracted from the vcf file. All subreads associated with an ID were retrieved from the raw subread bam file using samtools view. A fasta file containing the subread IDs as and their corresponding sequences was generated using an awk script and mapped back to the assembly using minimap2 with parameter -ax map-pb. The resulting alignment was converted to a sorted bam file with samtools. Finally, all events were visually inspected in IGV v2.16.0 (Robinson et al. 2011) to confirm that all subreads in one direction shared the same sequence change, and that this change was different from all subreads in the opposite direction.

3.2.4 Exclusion of heteroduplexes as cause

To evaluate whether candidate events were artifacts due to heteroduplex formation during library preparation, two filters were applied. (1) Mismapping and contamination filter. A heteroduplex could form either from two strands originating from different places in the genome or from different sources due to contamination with DNA from another accession. In this case, the strand with the mismatch should accurately map elsewhere in the assembly or to assemblies of other *A. thaliana* accessions. To test this, I extracted each candidate event along with 1 kb flanking sequence on both sides and aligned them using blastn v2.13.0 (Camacho et al. 2009) against both the focal assembly, and other assemblies generated in the same project (Wlodzimierz et al. 2023). (2) Heterozygosity filter. A heteroduplex could potentially also form from a rare mutation and rehybridization of strands of the two alleles. To test this, the entire HiFi read dataset was mapped to the site of each candidate using minimap2, followed by inspection in IGV to confirm whether any reads could be fully aligned at this position.

3.2.5 Exclusion of sequencing polymerase errors

In principle, systematic errors of the sequencing polymerase due to specific sequence contexts could generate consistent subread-specific sequence differences. In this case, the same error should be found in all reads covering this position. I therefore first examined the same region in subreads from other ZMWs. Second, I compared the polymerase kinetic information, specifically the interpulse duration (IPD), in the detected events with the IPD pattern of known sequencing errors. To this end, I mapped the raw sequencing output to the assembly to generate an aligned bam file. Variants were called at the subread level using Sniffles2 and I randomly picked variants with SUPPORT=1 tag in the vcf file, indicating an error in a single subread. I then compared IPD values

in the 100 bases flanking each candidate event on both sides with those of the reads containing apparent sequencing errors. Third, I identified shared strand-specific sequence differences that were either not shared by all subreads from a single ZMW, or that were shared by subreads from multiple ZMWs.

3.2.6 Detecting the nature of strand differences

Sequence dotplots comparing the event sequences and corresponding template sequences were generated using VectorBuilder Sequence Dot Plot (https://en.vectorbuilder.com/tool/sequence-dot-plot.html). Sequences were further analyzed using EMBOSS palindrome (https://www.bioinformatics.nl/cgi-bin/emboss/palindrome) to detect potential palindromic structures.

3.3 Results

3.3.1 Capturing single-strand mutational events

The dataset used, from a single Sequel II run of *A. thaliana* accession Gel-1, consisted of 3,747,759 reads, or 579 Gb of subreads. I identified three events where all subreads on one strand had an indel of at least 50 bases compared to the complementary sequence of the other strand and the fully assembled genome. Compared to the germline genome sequence, one event is a 144 base deletion (Figure 3.1a), one is a 99 base insertion (Figure 3.1b) and one is a 1,187 base insertion (Figure 3.1c). Individual subreads varied in length due to the high sequencing error rate at the subread level, but all subreads align cleanly to the strand-specific consensus sequences.

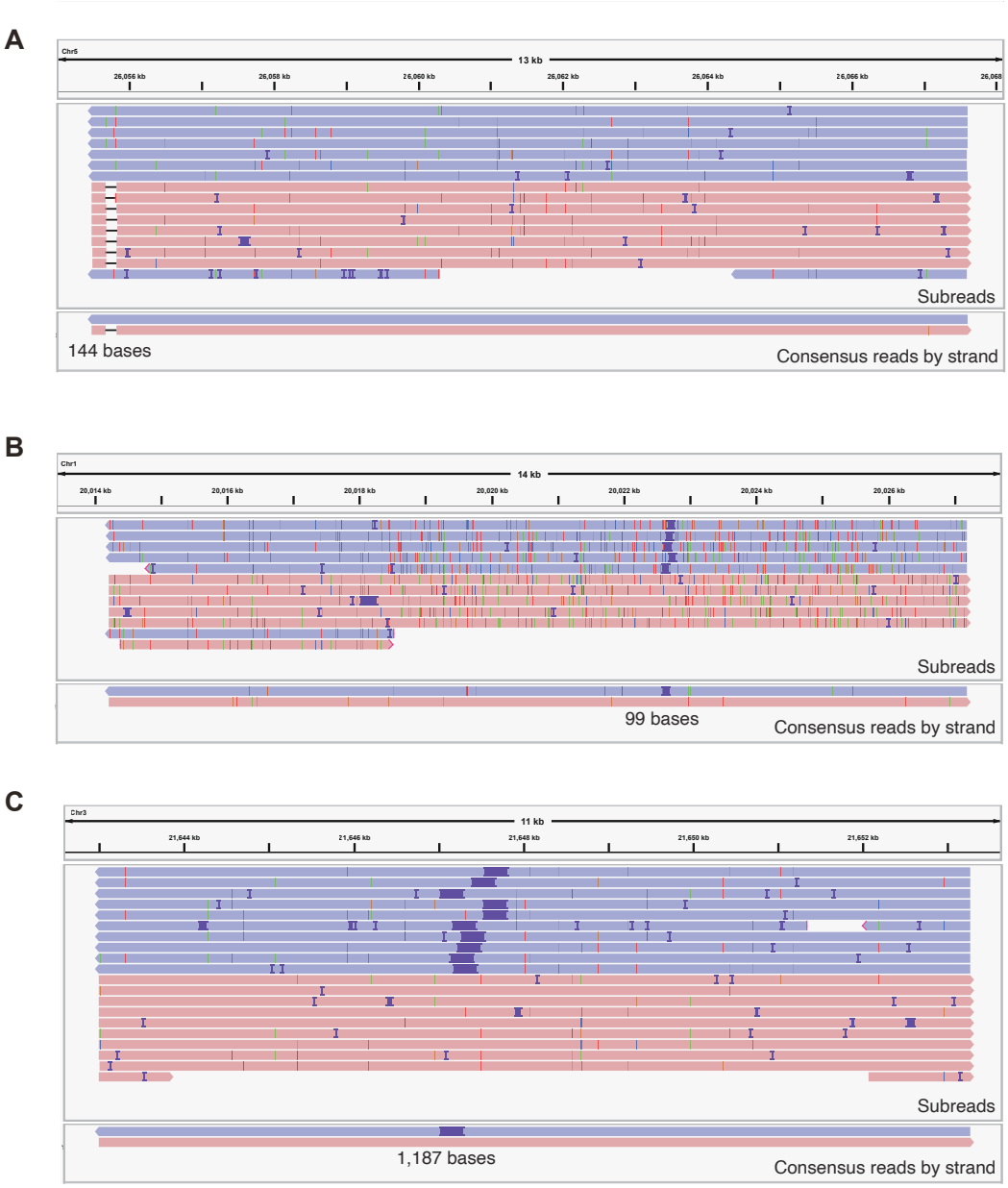

Figure 3.1 IGV screenshots of the single-strand indels. Reads are sorted and colored by strand. The top in each panel shows the alignment of all subreads from the same ZMW, the bottom shows the alignment of the two single-strand consensus reads. **a** A 144 base deletion. **b** A 99 base insertion. **c** A 1,187 base insertion.

3.3.2 Excluding false positives

I considered two scenarios that could lead to false positives: First, a read might

represent a heteroduplex artifact, with the two single strands originating from two different double-stranded molecules that hybridized during library preparation. Second, the difference between the two strands was introduced by the sequencing polymerase.

For the first scenario, the variant strands must originate either from a related region in the same genome, from a related region in another accession due to contamination during library preparation, or from a heterozygous de novo indel mutation at the site of the event. To test this, I aligned the sequence surrounding the candidate single-strand indel and its flanking regions to the whole genome of the focal accession as well as the genomes of all other accessions from the same project (Wlodzimierz et al. 2023). To exclude the possibility that the sequence originated from a heterozygous site, I aligned the sequence surrounding the candidate single-strand indel to all reads in this sample, expecting that other reads would support heterozygosity at this site. None of these analyses yielded hits that would have supported a heteroduplex artifact.

For the second scenario, I performed three independent analyses. First, I examined one of the polymerase kinetic parameters, the interpulse duration (IPD). As background, HiFi sequencing uses the engineered phi29 DNA polymerase. While the polymerase catalyzes the incorporation of fluorescently labeled nucleotides into a newly synthesized complementary DNA strand (Eid et al. 2009), the polymerase kinetics are recorded. The information has high sensitivity to detect abnormal base incorporation (Eid et al. 2009). I first inspected IPD values in two obvious polymerase errors, a 77 bp deletion and a 240 bp insertion (Figure 3.2a), each of which occurred in only a single subread. The deletion and insertion events were characterized by pronounced IPD abnormalities compared to the flanking regions (Figure 3.2b, 3.2c), with IPD

values reaching extremes such as the maximum of 255 or close to 0. In comparison, the three candidate indels had no such IPD anomalies (Figure 3.2d, 3.2e, 3.2f), with the IPD patterns across the indel regions being very similar to those in the flanking regions. Second, I examined whether subreads from other ZMWs covering the same genomic region showed evidence of indels. The rationale was that if the sequence context systematically causes aberrant behavior of the sequencing polymerase, this should lead to recurring abnormalities in subreads covering this site. However, no subreads from other ZMWs spanning these regions showed similar indel events.

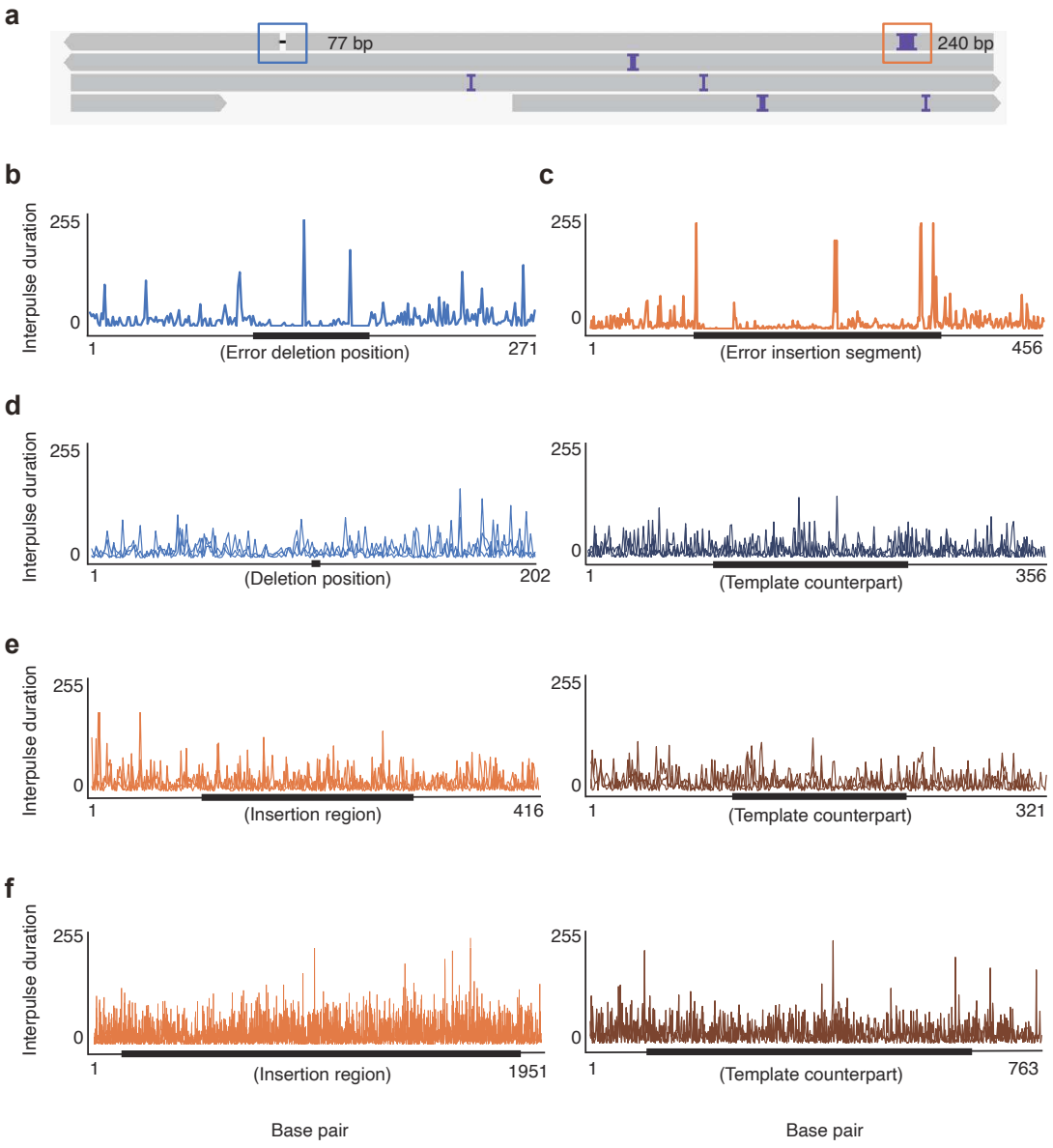

Figure 3.2 Polymerase kinetic information. **a** IGV screenshot of a ZMW containing typical sequencing errors that occur in only one subread. Each line represents one subread. b-c Line plots showing interpulse duration (IPD) values for regions with indel sequencing errors and 100 bases up-and downstream. **b** Deletion error. **c** Insertion error. d-f Line plots showing IPD values of true indel events supported by all subreads from the same strand (light color) with their corresponding template regions (dark color), including 100 bases up- and downstream. **d** The 144 base deletion. **e** The 99 base insertion. **f** The 1,187 base insertion.

Extending the logic of this second analysis, that there might be regions

associated with systematically aberrant behavior of the sequencing polymerase, I surveyed all ZMWs alignable to the nuclear genome to assess whether any genomic position showed (Figure 3.3): (i) indels occurring in multiple but not all subreads from one strand of a ZMW (Figure 3.3a), (ii) indels occurring in all subreads from one strand of a ZMW and also in subreads from the opposite strand of the same ZMW (Figure 3.3b), or (iii) indels occurring in all subreads from one strand in a ZMW and also appearing in other ZMWs at the same site (Figure 3.3c). I detected such events only in genomic regions containing tandem repeats. Similarly, I found no instances where different ZMWs had evidence for the same single-strand indel within any subset of subreads. Taken together, these analyses make it very unlikely that the observed indels supported by all subreads from one strand but not by any of the subreads from the other strand are due to polymerase errors during sequencing.

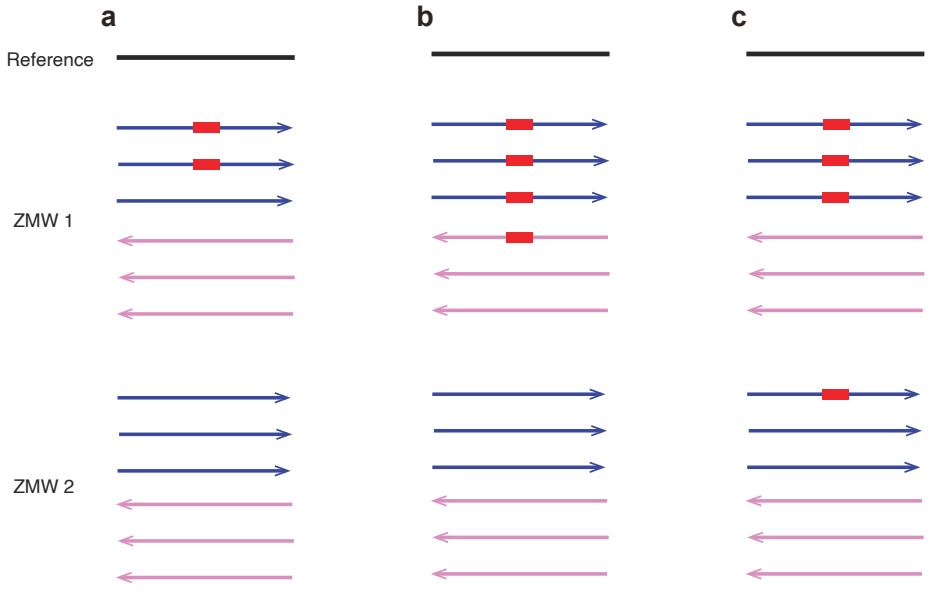

Figure 3.3 Schematic diagram showing alignment patterns if events were caused by systematic errors. Red rectangles indicate events of interest. Lines with arrows represent subreads, and the arrow direction and color indicate the two strand orientations. **a** Indels occurring in multiple but not all subreads from one strand of a ZMW. **b** indels occurring in all subreads from one strand of a ZMW and also in subreads from the opposite strand of the same ZMW. **c** indels occurring in all subreads

from one strand of a ZMW and also appearing in other ZMWs.

3.3.3 Sequence features of single-strand indels

The three true single-strand indels are (i) loss of a sequence that is unique in that region (Figure 3.4a), (ii) a direct duplication (Figure 3.4b), and (iii) a triplication, in which the second copy is inverted (Figure 3.4c).

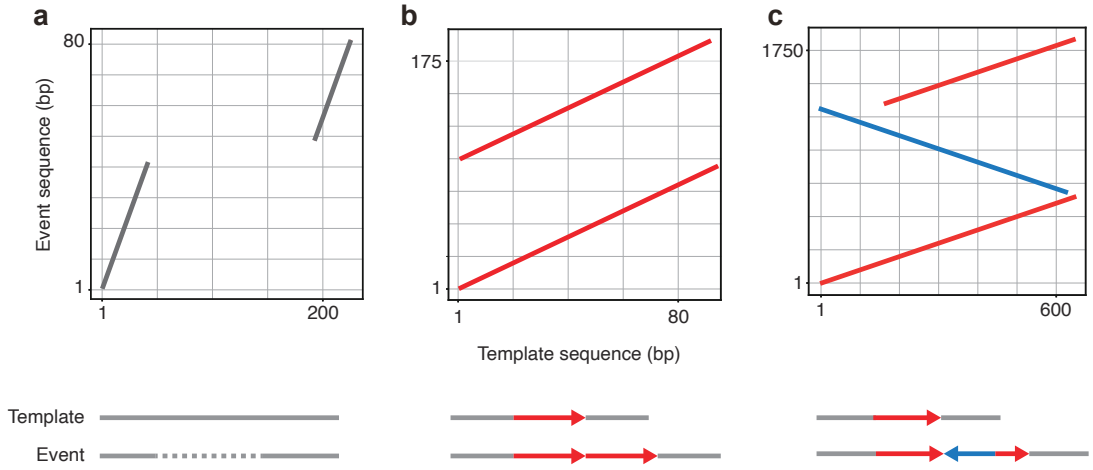

Figure 3.4 Dotplots and schematic representations of single-strand indels compared to their template strand. **a** The 144 base deletion, including 100 bases up- and downstream. **b** The 99 base insertion corresponding to an additional copy of the preceding 99 bases in the same orientation. **c** The 1,187 base insertion corresponding to two partial additional copies of the preceding 693 bases, with the middle copy in an inverted direction. Gray lines indicate flanking regions, red lines indicate sequences in the same direction as the template, and blue lines denote sequences in the opposite direction. Configurations shown schematically at the bottom.

The loss of the 144 base unique segment could be caused by slipped-strand mispairing (Tan et al. 2010; Kiktev et al. 2021), as the first 9 bases of the deleted sequence, TTTTAAAAA, are identical to the first 9 bases immediately downstream of the deletion (Figure 3.5). This single-strand deletion thus likely corresponds to a bulge that formed in the template strand during DNA replication.

TTTTACACTAAAAAAATATCTAAAATCTCACCAATTTACTTTTCAATTTTTTTT
AAAAACAATGTTATGAAATTTGAATAACAGTGGATTTGATATGATTTTTATAA
ATTCATGATTGAATAACATGAAAGTTGAAATGATTTGAAATATTTTTTCATAAA
TTCAGTTTTAATAACTTGAAATTTGAGAAAATATTTTAAAAATCACTAATTGAA
TAACGGGTGATTTAACAAAAAAAAAAACC

Figure 3.5 Sequence of the 144 base deletion event and the 50 base flanking region on the template strand. The deleted fragment is highlighted in red. Microhomology sequences potentially contributing to the event are shown in yellow.

For the 99 base insertion, this could also result from slipped-strand mispairing (Tan et al. 2010; Shi et al. 2013). The first two bases (TT) may serve as a microhomology-mediated anchor, as the same TT sequence occurs immediately after the duplicated segment (Figure 3.6). This single-strand insertion thus likely corresponds to a bulge that formed in the newly synthesized strand during DNA replication.

TAAATTGTAAACAAAAATTAGACAAAATTTATATTTAACTATATCAAATATT
ATATGAAGATTTATAACTATTAAATAGAAGTTATTTAAAACACACATAAAAT
CGTTTACAATACTTGACAACATATTTTAAATGTCGGCAATTGATTTTTTTTT
CCGATTCAATTCAGTAATTTTTGTTCAATTGCCTCCTATCACTC

Figure 3.6 Sequence of the 99 base insertion and the 50 base flanking region on the template strand. The newly duplicated fragment is highlighted in red. Microhomology sequences potentially contributing to the event are shown in yellow.

For the 1,187 base insertion, the two extra copies are not full-length, but contain deletions of 26 bases and 171 bases, respectively, relative to the 693 base preceding sequence (Figure 3.7). The features align with the mechanism of template switching mediated by palindromes (Reams and Roth 2015; Reams et al. 2012; Chuong et al. 2025).

AACTATTAGAAAAATACAATGTCTTATTAAATAAAACTAACTAATAGAAAAAT
ACATTATTTTACTAAATATGGTAAAGTAGCACTTTTTTAATTGGTTAGTAGG
AGAGACCATTTTTATTTGTATATACTGTAAATCTTGTTTATTTCCATCTAAAA
TATTATCCTTTTTGTTGTTTATAATAGTTTTTTGCTCTCTGCGTAGCATTAAG
TGGTAGAGAGGTGAATAAGACTAGAAGAACGGTCAATGGTGTCGATCACA
AAGATTTCTTTCGCGATGGAAAAGTTGGTGACTGGAAGAATCATCTGAGTG
TGACTCTTGAAACGGAGAACAAAATTGATATGACCATCAAGGAGAAATTTC
AAGGGTCAGGAACTCAGGATTGAAATTTTGAGTTAAATATGTATGTTGGTTT
GTGTGCTTTCTTTCATGCTTTGAAATAAATGTACCTTTCGTTGTGCCGTTGT
TATGTTCCTATTATTAATGGTTCTAGTTTGAATCGAGTGAGTGTCATTGGGG
GGTTTTGTTTCAAACGGTTACTTTGACTATAATATTTGTATCAAACTGTTCG
GAATATTTGGATAGAAACATTTATTAGGATCCGTTCGGGTTTAATAACTAAT
CGATTCGGTTTCAGAATCGGACACTTTTTATTTGGTTCCGGTTCGGTTTCC
AGGTTTTGCATTTTATGCGCAAACCTGGAAACCGAACCGGAACCAAATAAA
AAGTGTCCGATTCTGAAACCGAATCGATTAGTTATTAAACCCGAACGGATCC
TAATAAATGTTTCTATCCAAATATTCCGAACAGTTTGATACAAATATTATAGTC
AAAGTAACCGTTTGAAACAAAACCCCCCAATGACACTCACTCGATTCAAAC
TAGAACCATTAATAATAGGAACATAACAACGGCACAACGAAAGGTACATTTA
TTTCAAAGCATGAAAGAAAGCACACAAACCAACATACATATTTAACTCAAAA
TTTCAATCCTGAGTTCCTGACCCTTGAAATTTCTCCTTGATGGTCATATCAAT
TTTGTTCTCCGTTTCAAGAGTCACACTCAGATGATTCTTCCAGTCACCAACT
TTTCCATCGCGAAAGAAATCTTTGTGATCGACACCATTGACCGTTCTTCTAG
TCTTATTCACCTCTCTACCACTTAATGCTACGCAGAGAGCAAAAAACTATTAT
AAACAACAAAAAGGATAATATTTTAGATGGAAATAAACAAGATTTACAGTATAT
ACAAATAAAAATGGTCTCTCCTACTAACCAATTAAAAAAGTGCTACTTTACCAT
ATTTAGTAAAATAATGTATTTTTCTATTAGTTAGTTTTATTTAATAAGACATTG
TATTTTTCTAATAGTTTTTTGCTCTCTGCGTAGCATTAAGTGGTAGAGAGGTG
AATAAGACTAGAAGAACGGTCAATGGTGTCGATCACAAAGATTTCTTTCGCG
ATGGAAAAGTTGGTGACTGGAAGAATCATCTGAGTGTGACTCTTGAAACGGA
GAACAAAATTGATATGACCATCAAGGAGAAATTTCAAGGGTCAGGAACTCAG
GATTGAAATTTTGAGTTAAATATGTATGTTGGTTTGTGTGCTTTCTTTCATGCT
TTGAAATAAATGTACCTTTCGTTGTGCCGTTGTTATGTTCCTATTATTAATGGT
TCTAGTTTGAATCGAGTGAGTGTCATTGGGGGGTTTTGTTTCAAACGGTTACT
TTGACTATAATATTTGTATCAAACTGTTCGGAATATTTGGATAGAAACATTTAT
TAGGATCCGTTCGGGTTTAATAACTAATCGATTCGGTTTCAGAATCGGACACT
TTTTATTTGGTTCCGGTTCGGTTTCCAGGTTTTGCATTTTATGCGCAAACCT

Figure 3.7 Sequence of the 1,187 base insertion and the 693 base template sequence, which is twice partially duplicated. The two new copies are highlighted in red. Palindromic sequences potentially contributing to the event are shown in yellow.

In the original 693 base template sequence, the first and last 6 bases of the 26 bases missing in the first extra copy are AGGTTT and AAACCT, which can form a palindrome, causing DNA synthesis to snap back and thereby initiating the inverted copy. In the original 693 base template sequence, the first 8 bases of

the 171 sequence missing in the third copy is AACTATTA, and the following 8 bases (positions 172–179) are TAATAGTT. The formation of the third copy is likely due to TAATAGTT matching ATTATCAA in the template, complementary to TAATAGTT in the first copy. The 1,187 bp insertion thus likely corresponds to a set of hairpins that formed in the newly synthesized strand (Figure 3.8).

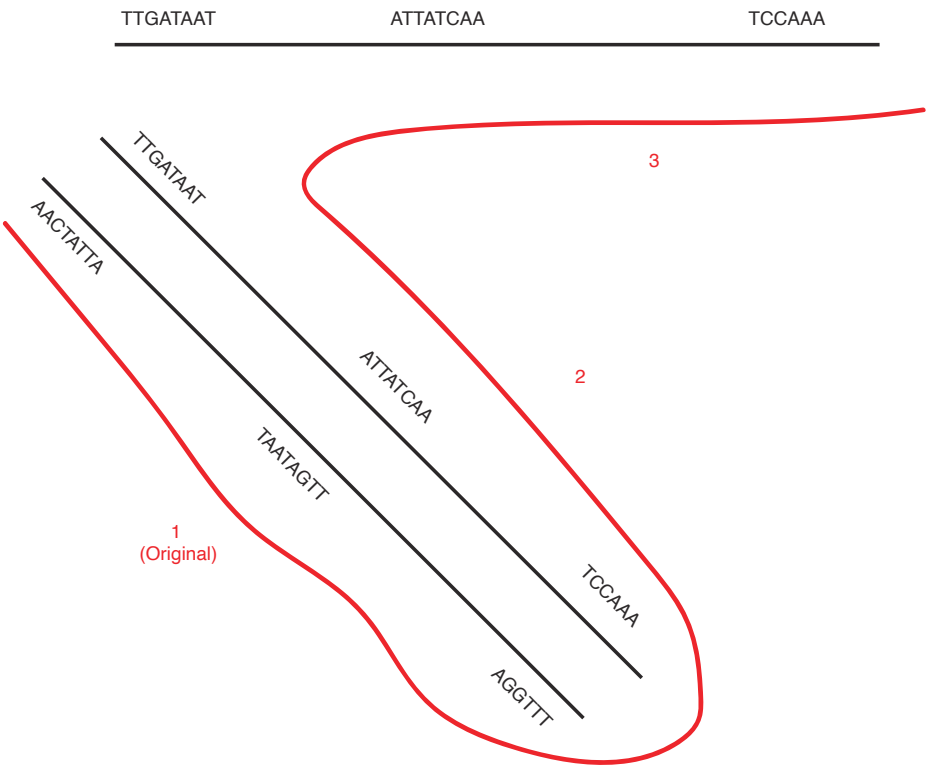

Figure 3.8 Schematic diagram showing the possible polymerase path (red line) for the palindrome-stimulated template-switching event, which leads to the generation of the triplicated sequence.

To look for potential stable secondary structures, I analyzed polymerase kinetics using IPD values (Guiblet et al. 2018; Eid et al. 2009). No significant changes in IPD were observed, making stable hairpins or G-quadruplex structures unlikely.

3.4 Discussion

To my knowledge, this is the first report documenting strand-specific sequence differences of >50 bases captured directly using standard PacBio HiFi sequencing. These events likely represent replication-associated structural changes that occurred after synthesis of the nascent strand but before repair, or alternatively reflect errors introduced by misrepair. Building on these findings, I am currently extending the analysis to events shorter than 50 bases, which are inherently more challenging to evaluate because shorter features are more susceptible to sequencing or alignment error.

I also plan to extend this approach to additional *A. thaliana* accessions. Compared with point mutations identified using a similar strategy (Liu et al. 2024), larger strand differences provide a clearer signal and reduce concerns about insufficient subreads per strand, minimizing both false positives and false negatives. Consequently, no special pre-experiment treatments are required for library preparation.

Overall, this proof-of-concept study demonstrates that circular consensus sequencing can be used not only to detect unrepaired single-nucleotide mismatches but also larger, replication-associated structural discrepancies between DNA strands. As raw subreads remain available for Sequel II–based datasets, the approach provides a practical means for examining strand-specific replication errors across a wide range of biological contexts.

4 Discussion

4.1 Insights from the two projects

Mutation is the ultimate source of genetic diversity and an important driver of evolution. By studying different layers of mutation, including germline mutations, somatic mutations, and unrepaired mutations, we can gain a clearer understanding of the processes of mutation, repair, and selection. Because mutations are so important to our understanding of evolution, both indirect and direct approaches for detecting mutations have a long history, beginning with measurement of lethal and fitness effects (Muller 1928; Mukai 1964), moving on to phenotypic observation (Ohad et al. 1996), protein and DNA blot analyses (Vidal et al. 1999), histochemical staining (Lucht et al. 2002), Sanger sequencing (Polyak et al. 1998), and short-read sequencing (Sloan et al. 2012). These approaches, however, provided only partial access to the complexity of mutational landscapes, particularly in repetitive or structurally challenging genomic regions, or in regions where small mutations do not produce a non-sequence readout such as an altered phenotype or gene expression changes.

The development of long-read sequencing has fundamentally changed this situation. Single-molecule, high-accuracy PacBio HiFi sequencing now enables continuous reads through long repeats, permits accurate resolution of complex structural variation, and, uniquely, allows multiple independent observations of each DNA strand.

The two projects in my PhD thesis leverage complementary aspects of this technology. The telomere project exploits the long read length and high accuracy of HiFi reads in repetitive regions to fully resolve previously largely

inaccessible sequences. The single-strand structural variant project, by contrast, takes advantage of circular sequencing, which produces multiple passes for each strand, to characterize transient, replication-associated sequence differences between individual strands of the same DNA molecule. Together, these studies illuminate mutational processes at two distinct scales: the longer-term accumulation of changes across generations and the immediate sequence states present during ongoing DNA replication.

4.1.1 Telomere diversity: conclusions and outlook

The telomere project provides the first genome-wide, population-level characterization of degenerate and variant telomeric repeats in a multicellular eukaryote. Except for the two chromosome ends adjacent to the long 45S rDNA arrays, regions that remain challenging even with current long-read technologies, I were able to reconstruct in detail the repeat composition at all chromosome ends.

My work clarifies multiple aspects of telomere biology. I revealed structural features that are conserved across the same chromosome end in different individuals, as well as extensive diversity between and across chromosome ends. The genetic-group-specific signatures that I discovered point to local evolutionary dynamics, including homogenization, higher-order repeat formation, and potential recombination of repair-mediated processes influencing telomere evolution.

Beyond the telomeres, my work draws attention to interstitial telomeric repeats, particularly those located in pericentromeric regions of *Arabidopsis thaliana*. This situation is reminiscent of centromeric repeats, with interstitial centromeric

repeats having only recently attracted attention (Corda and Giunta 2025). *A. thaliana* contains extended telomeric repeat regions within the pericentromeric regions of chromosome 1 (Naish et al. 2021) and chromosome 4 (Teano et al. 2023), with lengths of 355 kb and 72 kb, respectively. My unpublished results indicate that pericentromeric regions display alternating patterns of centromeric repeats, telomeric repeats, and transposons, and the mechanisms and functional significance of these patterns are still under investigation. In addition to repeats in pericentromere regions, there are other telomeric repeats scattered throughout the genome. The mechanisms generating these composite structures remain unknown, but long-read assemblies now make them accessible for systematic study.

This project also opens the door to studying telomeric repeat variants in other plants. Arabidopsis-like telomeric repeats are dominant in flowering plants (Peska and Garcia 2020; Adams et al. 2001), found in around 80% of species. The remaining species have distinct repeat types, such as TTTTTTAGGG in *Cestrum* (Peška et al. 2015), and CTCGGTTATGGG in *Allium* (Fajkus et al. 2016). Systematic genome-wide analysis of variant telomeric repeats in these species, using the framework that I established, would be of considerable interest.

While my studies did not focus on this aspect, several recent studies have employed long-read sequencing to directly measure telomere lengths (Colt et al. 2024). Previously, telomere length could only be estimated by k-mer counts from short reads, which was confounded by interstitial repeats and which could not resolve telomere lengths at individual chromosome ends (Choi et al. 2021). The most accurate approach involves the ligation of specific adaptors prior to sequencing to ensure reads extend to chromosome ends (Karimian et al. 2024).

In *A. thaliana*, telomere lengths typically range from 2 kb to 9 kb (Fulcher et al. 2015; Shakirov and Shippen 2004), which is shorter than standard long reads. Therefore, conventional long-read sequencing is generally sufficient for *A. thaliana*. In contrast, for species such as *Zea mays*, where average telomere lengths can reach 26 kb (Chen et al. 2023), direct measurement methods are less suitable.

4.1.2 Single-stranded structural variants: conclusions and outlook

My second project demonstrates that HiFi sequencing can be used to detect indel sequence differences between the two strands of the same native DNA molecule. Such differences likely represent events generated during replication before repair or errors introduced during misrepair. Although only three clear cases were found among 3.7 million molecules, which is consistent with the expectation that such events are rare, they provide direct evidence that circular consensus sequencing can capture transient, strand-specific structural alterations. The observed structures are consistent with template switching, slippage-mediated mispairing, or palindromic deletions—mechanisms known from studies in other organisms (e.g., Ottaviani et al. 2014; Kiktev et al. 2018; Kurahashi et al. 2009).

My method opens a path toward quantifying and characterizing early replication-associated events genome-wide. While previous studies have primarily focused on point mutations (Liu et al. 2024), my findings substantially broaden our understanding of mutational processes occurring during DNA replication.

In mature leaves of *A. thaliana*, only a few cells still longer actively divide, and thus there is limited DNA replication. It is plausible that in other tissues with higher rates of cell division, such as meristems, considerably more events are found than in the material that I examined. Furthermore, the methodology developed in this study could be highly relevant for other biological contexts, including cancer cells, where active replication and repair processes continuously generate new mutations.

The single-stranded SV events were first discovered accidentally while analyzing somatic SVs, i.e., double-stranded SVs. Because I had access to subread-level data from the PacBio Sequel II system, I observed that some SVs appearing only in some subreads of a single strand could be represented as an SV in the HiFi read, which merges all subreads into a consensus sequence. In contrast, the newer PacBio Revio system no longer outputs raw subreads directly, as it integrates DeepConsensus during data processing.

4.2 Long-read sequencing studies in other species

Long-read sequencing is being widely applied across the tree of life, accelerating the discovery of variation in genome structures in many different species In other plant species, for example, Iso-seq-based transcriptome analysis has been performed in *Zea Mays* (Wang et al. 2016), pangenome studies in *Oryza sativa* (Zhang et al. 2022), and investigations of F-box gene duplications in *Petunia*. Beyond plants, long-read sequencing has improved assembly completeness in algae (Li et al. 2025b), protozoa (Sun et al. 2025b), fungi (Caron et al. 2025), bacteria (Abbot et al. 2025), and metagenomes (Cuscó et al. 2025), and has even supported the discovery of new species

within metagenomic samples (Sereika et al. 2025). Overall, these studies exploit the same key features of long-read sequencing: superior read length, direct signal detection, amplification-free library preparation, and high accuracy.

4.3 Remaining challenges for long-read sequencing

Although long-read technologies have solved many longstanding problems, several challenges remain. These include the detection of large inversions, centromeric regions, high sequence-identity segmental duplications (SDs), and multiallelic variable number tandem repeats (VNTRs; Schloissnig et al. 2025), as well as structural variant (SV) calling in metagenomes (Agustinho et al. 2024). More specific examples include the acrocentric regions in humans (Porubsky et al. 2025) and 45S rDNA regions in *A. thaliana*. These regions are generally much longer than current long-read lengths. Nevertheless, the field is moving toward unified frameworks combining genomics, epigenomics, and chromatin conformation capture using long reads, and workflows for high-accuracy cross-modality integration are being continuously improved.

The excitement surrounding third-generation sequencing is not limited to DNA; it also extends to RNA and protein sequencing. Emerging single-molecule protein sequencing technologies, based on nanopores (Ritmejeris et al. 2024; Brinkerhoff et al. 2021; Motone et al. 2024), or other methods (Reed et al. 2022), aim to read not just four nucleotides but twenty amino acids. Although not yet commercialized, the ability to sequence individual long protein molecules could revolutionize proteomics, revealing long-range protein structures and post-translational modifications, just as single-molecule long-read DNA sequencing has transformed genome studies.

4.4 Future directions: Understanding mutational processes with the help of artificial intelligence

As elsewhere, artificial intelligence (AI) is emerging as a transformative force in genomics. Many long-read sequencing tools now leverage AI methods such as deep learning (DL) or machine learning (ML). For example, DeepSomatic (Park et al. 2025) and DeepVariant (Poplin et al. 2018) use convolutional neural networks to detect variant bases from pileup images representing different sequencing features of aligned reads. DeepConsensus (Baid et al. 2023) employs a transformer model on base-calling features to achieve high accuracy of sequences.

In addition to tools designed for processing sequencing data, a more recent direction involves applying natural language processing (NLP) concepts to genomic sequences, referred to as genome language models (GLMs). The core function of GLMs is to predict a masked base or the next base in a sequence (Brixi et al. 2025), as DNA can be treated as a language: sequences of nucleotides correspond to sentences, while k-mers or individual nucleotides serve as “words” or tokens (Nguyen et al. 2023). We are now in an era of abundant public sequencing data, such as those generated by the Earth BioGenome project (Lewin et al. 2018), which provides strong opportunities for the application of GLMs. Representative tools include Evo 2 (Brixi et al. 2025), Nucleotide Transformer (Dalla-Torre et al. 2025), and DNABERT (Ji et al. 2021). In the plant genomics field, several GLM tools have emerged in the past five years, including PlantCAD2 (Zhai et al. 2025) and GPN (Benegas et al. 2023).

Many applications of these tools are related to the study of mutational effects, such as the detection of long-range nucleotide dependencies (Tomaz da Silva

et al. 2025), which can be influenced by secondary structure. The latent features learned by GLMs enable both the identification of mutational signals across species and the discovery of previously uncharacterized mutational features within the same species, for example, the identification of previously unannotated enhancers that are functionally similar yet share little sequence similarity.

Beyond analytical applications, AI may also contribute to hypothesis generation (O'Brien et al. 2024). Generative models could be used to design synthetic telomeres or centromeres that could be tested for functionality in vivo. They might also be able to predict the minimum functional size of various repeat arrays, such as centromeres, telomeres and rDNA arrays, and CRISPR-mediated targeted deletions could be used to test these predictions in vivo (Thomson et al. 2025).

In summary, AI will almost certainly come to play a central role in the interpretation of genome structure and function. AI-based models have the potential to bridge the gap between biochemical experiments performed in vitro and processes occurring in vivo, by accurately identifying at scale rare, but highly informative events such as strand-specific SVs described in this thesis.

References

1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166: 481–491.

Abbot B, Field S, Carneal L, White RA III, Buchan A, West C, Lee L, Carter ME. 2025. Comparative genomics reveals multipartite genomes undergoing loss in the fungal endosymbiotic genus Mycetohabitans. bioRxiv. http://dx.doi.org/10.1101/2025.06.12.659383.

Adams SP, Hartman TP, Lim KY, Chase MW, Bennett MD, Leitch IJ, Leitch AR. 2001. Loss and recovery of Arabidopsis-type telomere repeat sequences 5'-(TTTAGGG)(n)-3' in the evolution of a major radiation of flowering plants. Proc Biol Sci 268: 1541–1546.

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. Science 376: eabl3533.

Agustinho DP, Fu Y, Menon VK, Metcalf GA, Treangen TJ, Sedlazeck FJ. 2024. Unveiling microbial diversity: harnessing long-read sequencing technology. Nat Methods 21: 954–966.

Akagi T, Fujita N, Shirasawa K, Tanaka H, Nagaki K, Masuda K, Horiuchi A, Kuwada E, Kawai K, Kunou R, et al. 2025. Rapid and dynamic evolution of a giant Y chromosome in Silene latifolia. Science 387: 637–643.

Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, Gatzen M, Sarkizova S, Schwartz MA, Blaum EM, et al. 2024. High-throughput RNA isoform sequencing using programmed cDNA concatenation. Nat Biotechnol 42: 582–586.

Allshire RC, Dempster M, Hastie ND. 1989. Human telomeres contain at least three types of G-rich repeat distributed non-randomly. Nucleic Acids Res 17: 4611–4627.

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 23: 258.

Alonso-Blanco CC, Ashkenazy H, Baduel P, Bao Z, Becker C, Caillieux E, Colot V, Crosbie D, De Oliveira L, Fitz J, et al. 2024. The 1001G+ project: A curated collection ofArabidopsis thalianalong-read genome assemblies to advance plant research. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2024.12.23.629943.

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. 2022. Complete genomic and epigenetic maps of human centromeres. Science 376: eabl4178.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215: 403–410.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Arbore R, Barbosa S, Brejcha J, Ogawa Y, Liu Y, Nicolaï MPJ, Pereira P, Sabatino SJ, Cloutier A, Poon ESK, et al. 2024. A molecular mechanism for bright color variation in parrots. Science 386: eadp7710.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 46: 2159–2168.

Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, et al. 2023. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. Nat Biotechnol 41: 232–238.

Baird DM. 2018. Telomeres and genomic evolution. Philos Trans R Soc Lond B Biol Sci 373. http://dx.doi.org/10.1098/rstb.2016.0437.

Baird DM, Coleman J, Rosser ZH, Royle NJ. 2000. High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. Am J Hum Genet 66: 235–250.

Baird DM, Jeffreys AJ, Royle NJ. 1995. Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. EMBO J 14: 5433–5443.

Ballif BC, Wakui K, Gajecka M, Shaffer LG. 2004. Translocation breakpoint mapping and sequence analysis in three monosomy 1p36 subjects with der(1)t(1;1)(p36;q44) suggest mechanisms for telomere capture in stabilizing de novo terminal rearrangements. Hum Genet 114: 198–206.

Belyayev A, Kalendar R, Josefiová J, Paštová L, Habibi F, Mahelka V, Mandák B, Krak K. 2023. Telomere sequence variability in genotypes from natural plant populations: unusual block-organized double-monomer terminal telomeric arrays. BMC Genomics 24: 572.

Benegas G, Batra SS, Song YS. 2023. DNA language models are powerful predictors of genome-wide variant effects. Proc Natl Acad Sci U S A 120: e2311219120.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences.

Nucleic Acids Res 27: 573–580.

Boán F, Blanco MG, Quinteiro J, Mouriño S, Gómez-Márquez J. 2004. Birth and evolutionary history of a human minisatellite. Mol Biol Evol 21: 228–235.

Brinkerhoff H, Kang ASW, Liu J, Aksimentiev A, Dekker C. 2021. Multiple rereads of single proteins at single-amino acid resolution using nanopores. Science 374: 1509–1513.

Brixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, Gonzalez GA, King SH, Li DB, Merchant AT, et al. 2025. Genome modeling and design across all domains of life with Evo 2. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2025.02.18.638918.

Brooks AN, Hughes AL, Clauder-Münster S, Mitchell LA, Boeke JD, Steinmetz LM. 2022. Transcriptional neighborhoods regulate transcript isoform lengths and expression levels. Science 375: 1000–1005.

Brownlee GG. 2013. Frederick Sanger (1918-2013). Curr Biol 23: R1074–6.

Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, Novikova PY, Nordborg M. 2021. Gradual evolution of allopolyploidy in Arabidopsis suecica. Nat Ecol Evol 5: 1367–1381.

Byun D, Son N, Kim H, Kim J, Park J, Park S-J, Kim H, Kim J, Kim J, Lee S, et al. 2024. COmapper: High-resolution mapping of meiotic crossovers by long-read sequencing inArabidopsis. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2024.10.21.619347.

Byun D, Son N, Kim H, Kim J, Park J, Park S-J, Kim H, Kim J, Kim J, Lee S, et al. 2025. COmapper: high-resolution mapping of meiotic crossovers by long-read sequencing in Arabidopsis. New Phytol 247: 1942–1957.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43: 956–963.

Caron T, Crequer E, Le Piver M, Le Prieur S, Brunel S, Snirc A, Cueff G, Roueyre D, Place M, Chassard C, et al. 2025. Identification of quantitative trait loci (QTLs) for key cheese making phenotypes in the blue-cheese mold Penicillium roqueforti. PLoS Genet 21: e1011669.

Chan SRWL, Blackburn EH. 2004. Telomeres and telomerase. Philos Trans R Soc Lond B Biol Sci 359: 109–121.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18: 170–175.

Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, et al. 2023. A complete telomere-to-telomere assembly of the maize genome.

Nat Genet 55: 1221–1231.

Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, Shi S, He X. 2012. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science 335: 1235–1238.

Chen Z, Baeza JA, Chen C, Gonzalez MT, González VL, Greve C, Kocot KM, Arbizu PM, Moles J, Schell T, et al. 2025. A genome-based phylogeny for Mollusca is concordant with fossils and morphology. Science 387: 1001–1007.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13: 1050–1054.

Choi JY, Abdulkina LR, Yin J, Chastukhina IB, Lovell JT, Agabekian IA, Young PG, Razzaque S, Shippen DE, Juenger TE, et al. 2021. Natural variation in plant telomere length is associated with flowering time. Plant Cell 33: 1118–1134.

Christenhusz MJM, Twyford AD, Hudson A, Royal Botanic Gardens Kew Genome Acquisition Lab, Royal Botanic Garden Edinburgh Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. 2023. The genome sequence of thale cress, Arabidopsis thaliana (Heynh., 1842). Wellcome Open Res 8: 40.

Churikov D, Price CM. 2008. Telomeric and subtelomeric repeat sequences. Encyclopedia of Life Sciences. http://dx.doi.org/10.1002/9780470015902.a0005065.pub3.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215–219.

Colt K, Petrus S, Abramson BW, Mamerto A, Hartwick NT, Michael TP. 2024. Telomere length in plants estimated with long read sequencing. bioRxiv. http://dx.doi.org/10.1101/2024.03.27.586973.

Contreras-Garrido A, Galanti D, Movilli A, Becker C, Bossdorf O, Drost H-G, Weigel D. 2023. Transposon dynamics in the emerging oilseed crop Thlaspi arvense. bioRxiv 2023.05.24.542068. https://www.biorxiv.org/content/10.1101/2023.05.24.542068v1 (Accessed October 13, 2023).

Cook DE, Zdraljevic S, Tanny RE, Seo B, Riccardi DD, Noble LM, Rockman MV, Alkema MJ, Braendle C, Kammenga JE, et al. 2016. The genetic basis of natural variation in Caenorhabditis elegans telomere length. Genetics 204: 371–383.

Copenhaver GP, Pikaard CS. 1996. RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of Arabidopsis thaliana adjoin the telomeres on chromosomes 2 and 4. Plant J 9: 259–272.

Corda L, Giunta S. 2025. Chromosome-specific centromeric patterns define the centeny map of the human genome. Science 389: eads3484.

Couger MB, Roy SW, Anderson N, Gozashti L, Pirro S, Millward LS, Kim M, Kilburn D, Liu KJ, Wilson TM, et al. 2021. Sex chromosome transformation and the origin of a male-specific X chromosome in the creeping vole. Science 372: 592–600.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJ Complex Syst 1695: 1–9.

Cuscó A, Duan Y, Gil F, Chklovski A, Kruthi N, Pan S, Forslund S, Lau S, Löber U, Zhao X-M, et al. 2025. Capturing global pet dog gut microbial diversity and hundreds of near-finished bacterial genomes by using long-read metagenomics in a Shanghai cohort. bioRxiv. http://dx.doi.org/10.1101/2025.09.17.676595.

Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, Dallago C, Trop E, de Almeida BP, Sirelkhatim H, et al. 2025. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nat Methods 22: 287–297.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10. http://dx.doi.org/10.1093/gigascience/giab008.

de Almeida BP, Dalla-Torre H, Richard G, Blum C, Hexemer L, Gélard M, Mendoza-Revilla J, Tang Z, Marin FI, Emms DM, et al. 2025. Annotating the genome at single-nucleotide resolution with DNA foundation models. Nat Methods. http://dx.doi.org/10.1038/s41592-025-02881-2.

Debladis E, Llauro C, Carpentier M-C, Mirouze M, Panaud O. 2017. Detection of active transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing technology. BMC Genomics 18: 537.

Denyer T, Ma X, Klesen S, Scacchi E, Nieselt K, Timmermans MCP. 2019. Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing. Dev Cell 48: 840–

852.e5.

Dogga SK, Rop JC, Cudini J, Farr E, Dara A, Ouologuem D, Djimdé AA, Talman AM, Lawniczak MKN. 2024. A single cell atlas of sexual development in Plasmodium falciparum. Science 384: eadj4088.

Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol 42: 1606–1614.

Dong X, Jiao W-B, Campoy JA, Rabanal FA, Ton J, Smith LM, Weigel D, Schneeberger K. 2025. The mutational dynamics of the Arabidopsis centromeres. bioRxiv. http://dx.doi.org/10.1101/2025.06.02.657473.

Drouaud J, Camilleri C, Bourguignon P-Y, Canaguier A, Bérard A, Vezon D, Giancola S, Brunel D, Colot V, Prum B, et al. 2006. Variation in crossing-over rates across chromosome 4 of Arabidopsis thaliana reveals the presence of meiotic recombination “hot spots.” Genome Res 16: 106–114.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372: eabf7117.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138.

Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, et al. 2018. The rate and potential relevance of new mutations in a colonizing plant lineage. PLoS Genet 14: e1007155.

Fajkus P, Peška V, Sitová Z, Fulnečková J, Dvořáčková M, Gogela R, Sýkorová E, Hapala J, Fajkus J. 2016. Allium telomeres unmasked: the unusual telomeric sequence (CTCGGTTATGGG)n is synthesized by telomerase. Plant J 85: 337–347.

Fajkus P, Peška V, Závodník M, Fojtová M, Fulnečková J, Dobias Š, Kilar A, Dvořáčková M, Zachová D, Nečasová I, et al. 2019. Telomerase RNAs in land plants. Nucleic Acids Res 47: 9842–9856.

Farmer A, Thibivilliers S, Ryu KH, Schiefelbein J, Libault M. 2021. Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. Mol Plant 14: 372–383.

Farrell C, Vaquero-Sedas MI, Cubiles MD, Thompson M, Vega-Vaquero A, Pellegrini M, Vega-Palas MA. 2022. A complex network of interactions

governs DNA methylation at telomeric regions. Nucleic Acids Res 50: 1449–1464.

Fields PD, Waneka G, Naish M, Schatz MC, Henderson IR, Sloan DB. 2022. Complete Sequence of a 641-kb Insertion of Mitochondrial DNA in the Arabidopsis thaliana Nuclear Genome. Genome Biol Evol 14. http://dx.doi.org/10.1093/gbe/evac059.

Fulcher N, Teubenbacher A, Kerdaffrec E, Farlow A, Nordborg M, Riha K. 2015. Genetic architecture of natural variation of telomere length in Arabidopsis thaliana. Genetics 199: 625–635.

Fulnecková J, Sevcíková T, Fajkus J, Lukesová A, Lukes M, Vlcek C, Lang BF, Kim E, Eliás M, Sykorová E. 2013. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biol Evol 5: 468–483.

Fultz D, McKinlay A, Enganti R, Pikaard CS. 2023a. Sequence and epigenetic landscapes of active and silenced nucleolus organizers in Arabidopsis. bioRxiv 2023.06.07.544131. https://www.biorxiv.org/content/10.1101/2023.06.07.544131v3 (Accessed October 13, 2023).

Fultz D, McKinlay A, Enganti R, Pikaard CS. 2023b. Sequence and epigenetic landscapes of active and silent nucleolus organizer regions in. Sci Adv 9: eadj4509.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477: 419–423.

Gao S, Zhang Y, Bush SJ, Wang B, Yang X, Ye K. 2024. Centromere landscapes resolved from hundreds of human genomes. Genomics Proteomics Bioinformatics 22. http://dx.doi.org/10.1093/gpbjnl/qzae071.

Garrido-Ramos MA. 2017. Satellite DNA: An evolving topic. Genes (Basel) 8: 230.

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, et al. 2024. Building pangenome graphs. Nat Methods 21: 2008–2012.

Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. 2022. Epigenetic patterns in a complete human genome. Science 376: eabj5089.

Gigante S, Gouil Q, Lucattini A, Keniry A, Beck T, Tinning M, Gordon L, Woodruff C, Speed TP, Blewitt ME, et al. 2019. Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Res 47:

e46.

Goel M, Sun H, Jiao W-B, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 20: 277.

Gompert Z, Feder JL, Parchman TL, Planidin NP, Whiting FJH, Nosil P. 2025. Adaptation repeatedly uses complex structural genomic variation. Science 388: eadp3745.

Gout J-F, Thomas WK, Smith Z, Okamoto K, Lynch M. 2013. Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci U S A 110: 18584–18589.

Grandi FC, Modi H, Kampman L, Corces MR. 2022. Chromatin accessibility profiling by ATAC-seq. Nat Protoc 17: 1518–1552.

Grigorev K, Foox J, Bezdan D, Butler D, Luxton JJ, Reed J, McKenna MJ, Taylor L, George KA, Meydan C, et al. 2021. Haplotype diversity and sequence heterogeneity of human telomeres. Genome Res 31: 1269–1279.

Groh JS, Vik DC, Davis M, Monroe JG, Stevens KA, Brown PJ, Langley CH, Coop G. 2025. Ancient structural variants control sex-specific flowering time morphs in walnuts and hickories. Science 387: eado5578.

Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky E, Eckert K, Chiaromonte F, Makova KD. 2018. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res 28: 1767–1778.

Guo Y-L, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D. 2011. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. Plant Physiol 157: 757–769.

Gutiérrez-García K, Aumiller K, Dodge R, Obadia B, Deng A, Agrawal S, Yuan X, Wolff R, Zhu H, Hsia R-C, et al. 2024. A conserved bacterial genetic basis for commensal-host specificity. Science 386: 1117–1122.

Hagenblad J, Tang C, Molitor J, Werner J, Zhao K, Zheng H, Marjoram P, Weigel D, Nordborg M. 2004. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two Arabidopsis thaliana flowering time loci. Genetics 168: 1627–1638.

Hager ER, Harringmeyer OS, Wooldridge TB, Theingi S, Gable JT, McFadden S, Neugeboren B, Turner KM, Jensen JD, Hoekstra HE. 2022. A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. Science 377: 399–405.

Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, Schneeberger

K, Fitz J, Altmann T, Bergelson J, et al. 2015. Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS Genet 11: e1004920.

Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. 2022. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res 50: D316–D325.

Heacock M, Spangler E, Riha K, Puizina J, Shippen DE. 2004. Molecular analysis of telomere fusions in Arabidopsis: multiple pathways for chromosome end-joining. EMBO J 23: 2304–2313.

Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107: 1–8.

He S, Liu Y, Fang S, Li Y, Weng T, Tian R, Yin Y, Zhou D, Yin B, Wang Y, et al. 2024. Solid-State nanopore DNA Sequencing: Advances, challenges and prospects. Coord Chem Rev 510: 215816.

Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Human Pangenome Reference Consortium, Marschall T, Li H, Paten B. 2024. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol 42: 663–673.

Hoge C, de Manuel M, Mahgoub M, Okami N, Fuller Z, Banerjee S, Baker Z, McNulty M, Andolfatto P, Macfarlan TS, et al. 2024. Patterns of recombination in snakes reveal a tug-of-war between PRDM9 and promoter-like features. Science 383: eadj7026.

Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. 2022. A near-complete assembly of an Arabidopsis thaliana genome. Mol Plant 15: 1247–1250.

Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, Limouse C, Halabian R, Wojenski L, Rodriguez M, et al. 2022. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science 376: eabk3112.

Huang G, Bao Z, Feng L, Zhai J, Wendel JF, Cao X, Zhu Y. 2024. A telomere-to-telomere cotton genome assembly reveals centromere evolution and a Mutator transposon-linked module regulating embryo development. Nat Genet 56: 1953–1963.

Hu G, Wang Z, Tian Z, Wang K, Ji G, Wang X, Zhang X, Yang Z, Liu X, Niu R, et al. 2025. A telomere-to-telomere genome assembly of cotton provides insights into centromere evolution and short-season adaptation. Nat Genet 57: 1031–1043.

Jeong H, Dishuck PC, Yoo D, Harvey WT, Munson KM, Lewis AP, Kordosky J, Garcia GH, Human Genome Structural Variation Consortium (HGSVC),

Yilmaz F, et al. 2025. Structural polymorphism and diversity of human segmental duplications. Nat Genet 57: 390–401.

Jiao W-B, Schneeberger K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun 11: 989.

Ji Y, Zhou Z, Liu H, Davuluri RV. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37: 2112–2120.

Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, Liu W, Chen C, Song Y, Tan L, et al. 2023. The pan-genome and local adaptation of Arabidopsis thaliana. Nat Commun 14: 6259.

Karimian K, Groot A, Huso V, Kahidi R, Tan K-T, Sholes S, Keener R, McDyer JF, Alder JK, Li H, et al. 2024. Human telomere length is chromosome end-specific and conserved across individuals. Science 384: 533–539.

Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al. 2016. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell 166: 492–505.

Kileeg Z, Wang P, Mott GA. 2024. Chromosome-Scale Assembly and Annotation of Eight Arabidopsis thaliana Ecotypes. Genome Biol Evol 16. http://dx.doi.org/10.1093/gbe/evae169.

Kim C, Kim J, Kim S, Cook DE, Evans KS, Andersen EC, Lee J. 2019. Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in C. elegans. Genome Res 29: 1023–1035.

Kim C, Sung S, Kim J, Lee J. 2020. Repair and reconstruction of telomeric and subtelomeric regions and genesis of new telomeres: Implications for chromosome evolution. Bioessays 42: e1900177.

Kirov I, Merkulov P, Dudnikov M, Polkhovskaya E, Komakhin RA, Konstantinov Z, Gvaramiya S, Ermolaev A, Kudryavtseva N, Gilyok M, et al. 2021. Transposons Hidden in Genome Assembly Gaps and Mobilization of Non-Autonomous LTR Retrotransposons Unravelled by Nanotei Pipeline. Plants (Basel) 10. http://dx.doi.org/10.3390/plants10122681.

Koeppel J, Ferreira R, Vanderstichele T, Riedmayr LM, Peets EM, Girling G, Weller J, Murat P, Liberante FG, Ellis T, et al. 2025. Randomizing the human genome by engineering recombination between repeat elements. Science 387: eado3979.

Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33: 2759–2761.

Kolora SRR, Owens GL, Vazquez JM, Stubbs A, Chatla K, Jainese C, Seeto K, McCrea M, Sandel MW, Vianna JA, et al. 2021. Origins and evolution of extreme life span in Pacific Ocean rockfishes. Science 374: 842–847.

Kong Y, Cao L, Deikus G, Fan Y, Mead EA, Lai W, Zhang Y, Yong R, Sebra R, Wang H, et al. 2022. Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. Science 375: 515–522.

Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, Jenike KM, Lucas J, McNulty B, Park J, Rautiainen M, et al. 2024. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. bioRxiv. http://dx.doi.org/10.1101/2024.03.15.585294.

Kovaka S, Hook PW, Jenike KM, Shivakumar V, Morina LB, Razaghi R, Timp W, Schatz MC. 2025. Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment. Nat Methods 22: 681–691.

Kuo H-F, Olsen KM, Richards EJ. 2006. Natural variation in a subtelomeric region of Arabidopsis: implications for the genomic dynamics of a chromosome end. Genetics 173: 401–417.

Laitinen RAE, Schneeberger K, Jelly NS, Ossowski S, Weigel D. 2010. Identification of a spontaneous frame shift mutation in a nonreference Arabidopsis accession using whole genome sequencing. Plant Physiol 153: 652–654.

Lal A, Brown M, Mohan R, Daw J, Drake J, Israeli J. 2021. Improving long-read consensus sequencing accuracy with deep learning. bioRxiv 2021.06.28.450238. https://www.biorxiv.org/content/10.1101/2021.06.28.450238v3 (Accessed October 13, 2023).

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40: D1202–10.

Lee C-R, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, Alonso-Blanco C, Weigel D, Nordborg M. 2017. On the post-glacial spread of human commensal Arabidopsis thaliana. Nat Commun 8: 14458.

Lee JY, Kim DS. 2009. Dramatic effect of single-base mutation on the conformational dynamics of human telomeric G-quadruplex. Nucleic Acids Res 37: 3625–3634.

Lee M, Hills M, Conomos D, Stutz MD, Dagg RA, Lau LMS, Reddel RR, Pickett HA. 2014. Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. Nucleic Acids Res 42:

1733–1746.

Levenshtein VI. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl 10: 707–710.

Levy B, Xu Z, Zhao L, Kremling K, Altman R, Wong P, Tanner C. 2022. FloraBERT: cross-species transfer learning withattention-based neural networks for geneexpression prediction. Research Square. http://dx.doi.org/10.21203/rs.3.rs-1927200/v1.

Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, Mercier R. 2024. A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. Nat Genet 56: 982–991.

Lien Y-W, Amendola D, Lee KS, Bartlau N, Xu J, Furusawa G, Polz MF, Stocker R, Weiss GL, Pilhofer M. 2024. Mechanism of bacterial predation via ixotrophy. Science 386: eadp0614.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. Genome Biol 21: 265.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li M, Jiang Z, Xu X, Wu X, Liu Y, Chen K, Liao Y, Li W, Wang X, Guo Y, et al. 2025a. Chromatin accessibility landscape of mouse early embryos revealed by single-cell NanoATAC-seq2. Science 387: eadp4319.

Liu J, Li Q, Hu Y, Yu Y, Zheng K, Li D, Qin L, Yu X. 2024a. The complete telomere-to-telomere sequence of a mouse genome. Science 386: 1141–1146.

Liu MH, Costa BM, Bianchini EC, Choi U, Bandler RC, Lassen E, Grońska-Pęski M, Schwing A, Murphy ZR, Rosenkjær D, et al. 2024b. DNA mismatch and damage patterns revealed by single-molecule sequencing. Nature 630: 752–761.

Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, Foox J, Mason C, Carroll M, Cheng A, et al. 2021. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol 22: 295.

Liu Z-Y, Berthel A, Czech E, Stitzer M, Hsu S-K, Pennell M, Buckler ES, Zhai J. 2025. GeneCAD: Plant genome annotation with a DNA foundation model.

bioRxiv. http://dx.doi.org/10.1101/2025.10.31.685877.

Li X, Li Y-L, Zhong C, Li J, Su L, Liu J-X, Pang S. 2025b. Chromosome-level genome assembly for the ecologically and economically important alga Saccharina japonica. Sci Data 12: 290.

Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, Porubsky D, Mao Y, Yoo D, Rautiainen M, et al. 2024. The variation and evolution of complete human centromeres. Nature 629: 136–145.

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V, et al. 2013. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat Genet 45: 884–890.

Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ. 2013. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res 23: 121–128.

Lucht JM, Mauch-Mani B, Steiner H-Y, Metraux J-P, Ryals J, Hohn B. 2002. Pathogen stress increases somatic recombination frequency in Arabidopsis. Nat Genet 30: 311–314.

Lyčka M, Bubeník M, Závodník M, Peska V, Fajkus P, Demko M, Fajkus J, Fojtová M. 2024. TeloBase: a community-curated database of telomere sequences across the tree of life. Nucleic Acids Res 52: D311–D321.

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 107: 961–968.

Marriage TN, Hudman S, Mort ME, Orive ME, Shaw RG, Kelly JK. 2009. Direct estimation of the mutation rate at dinucleotide microsatellite loci in Arabidopsis thaliana (Brassicaceae). Heredity (Edinb) 103: 310–317.

Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M. 1998. Arabidopsis thaliana: a model plant for genome analysis. Science 282: 662, 679–82.

Mendez-Bermudez A, Hills M, Pickett HA, Phan AT, Mergny J-L, Riou J-F, Royle NJ. 2009. Human telomeres that contain (CTAGGG)n repeats show replication dependent instability in somatic cells and the male germline. Nucleic Acids Res 37: 6225–6238.

Mendoza-Revilla J, Trop E, Gonzalez L, Roller M, Dalla-Torre H, de Almeida BP, Richard G, Caton J, Lopez Carranza N, Skwark M, et al. 2024. A foundational large language model for edible plant genomes. Commun Biol 7: 835.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel

D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun 9: 541.

Milia S, Leonard AS, Mapel XM, Bernal Ulloa SM, Drögemüller C, Pausch H. 2025. Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. Genome Res 35: 1041–1052.

Mizuno H, Wu J, Katayose Y, Kanamori H, Sasaki T, Matsumoto T. 2008. Chromosome-specific distribution of nucleotide substitutions in telomeric repeats of rice (Oryza sativa L.). Mol Biol Evol 25: 62–68.

Mohamed M, Sabot F, Varoqui M, Mugat B, Audouin K, Pélisson A, Fiston-Lavier A-S, Chambeyron S. 2023. TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. Genome Biol 24: 63.

Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 38: 701–707.

Motone K, Kontogiorgos-Heintz D, Wee J, Kurihara K, Yang S, Roote G, Fox OE, Fang Y, Queen M, Tolhurst M, et al. 2024. Multi-pass, single-molecule nanopore reading of long protein strands. Nature 633: 662–669.

Movilli A, Sushko S, Rabanal FA, Weigel D. 2025a. Long-read detection of transposable element mobilization in the soma of hypomethylated Arabidopsis thaliana individuals. Genome Biol 26: 231.

Movilli A, Sushko S, Rabanal F, Weigel D. 2025b. Long-read detection of transposable element mobilization in the soma of hypomethylatedArabidopsis thalianaindividuals. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2025.02.07.637047.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, et al. 2021. The genetic and epigenetic landscape of the centromeres. Science 374: eabi7489.

Nemudraia A, Nemudryi A, Wiedenheft B. 2024. Repair of CRISPR-guided RNA breaks enables site-specific RNA excision in human cells. Science 384: 808–814.

Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, Patel A, Rabideau C, Massaroli S, Bengio Y, et al. 2023. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. ArXiv. https://www.ncbi.nlm.nih.gov/pubmed/37426456.

Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, Zhao H, Zou Y, Huang Y, Li J, et al. 2023. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nat Commun 14: 4054.

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, et al. 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. Nat Genet 30: 190–193.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, et al. 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat Genet 48: 1077–1082.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. Science 376: 44–53.

Ohad N, Margossian L, Hsu YC, Williams C, Repetti P, Fischer RL. 1996. A mutation that allows endosperm development without fertilization. Proc Natl Acad Sci U S A 93: 5319–5324.

Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. 2023. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet 24: 464–483.

Ono Y, Hamada M, Asai K. 2022. PBSIM3: a simulator for all types of PacBio and ONT long reads. NAR Genom Bioinform 4: lqac092.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18: 2024–2033.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92–94.

Ou S, Scheben A, Collins T, Qiu Y, Seetharam AS, Menard CC, Manchanda N, Gent JI, Schatz MC, Anderson SN, et al. 2024. Differences in activity and stability drive transposable element variation in tropical and temperate maize. Genome Res 34: 1140–1153.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20: 275.

Park J, Cook DE, Chang P-C, Kolesnikov A, Brambrink L, Mier JC, Gardner J, McNulty B, Sacco S, Keskus AG, et al. 2025. Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. Nat Biotechnol. http://dx.doi.org/10.1038/s41587-025-02839-x.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669–680.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34: 867–868.

Peška V, Fajkus P, Fojtová M, Dvořáčková M, Hapala J, Dvořáček V, Polanská P, Leitch AR, Sýkorová E, Fajkus J. 2015. Characterisation of an unusual telomere motif (TTTTTTAGGG)n in the plant Cestrum elegans (Solanaceae), a species with a large genome. Plant J 82: 644–654.

Peska V, Garcia S. 2020. Origin, diversity, and evolution of telomere sequences in plants. Front Plant Sci 11: 117.

Pfeifer GP. 2020. Mechanisms of UV-induced mutations and skin cancer. Genome Instab Dis 1: 99–113.

Pickett HA, Baird DM, Hoff-Olsen P, Meling GI, Rognum TO, Shaw J, West KP, Royle NJ. 2004. Telomere instability detected in sporadic colon cancers, some showing mutations in a mismatch repair gene. Oncogene 23: 3434–3443.

Polyak K, Li Y, Zhu H, Lengauer C, Willson JK, Markowitz SD, Trush MA, Kinzler KW, Vogelstein B. 1998. Somatic mutations of the mitochondrial genome in human colorectal tumours. Nat Genet 20: 291–293.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36: 983–987.

Pownall ME, Miao L, Vejnar CE, M'Saad O, Sherrard A, Frederick MA, Benitez MDJ, Boswell CW, Zaret KS, Bewersdorf J, et al. 2023. Chromatin expansion microscopy reveals nanoscale organization of transcription and chromatin. Science 381: 92–100.

Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B. 2019. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLoS One 14: e0216233.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.

Qi X-G, Wu J, Zhao L, Wang L, Guang X, Garber PA, Opie C, Yuan Y, Diao R, Li G, et al. 2023. Adaptations to a cold climate promoted social evolution in Asian colobine primates. Science 380: eabl8621.

Quadrana L, Henderson IR. 2025. The natural history of transposons in plant pangenomes and panepigenomes. Curr Opin Plant Biol 88: 102818.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rabanal FA, Gräff M, Lanz C, Fritschi K, Llaca V, Lang M, Carbonell-Bejerano P, Henderson I, Weigel D. 2022. Pushing the limits of HiFi assemblies reveals centromere diversity between two Arabidopsis thaliana genomes. Nucleic Acids Res 50: 12309–12327.

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol 41: 1474–1482.

Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. Nat Commun 13: 1948.

Reed BD, Meyer MJ, Abramzon V, Ad O, Ad O, Adcock P, Ahmad FR, Alppay G, Ball JA, Beach J, et al. 2022. Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. Science 378: 186–192.

Reed J, Orme A, El-Demerdash A, Owen C, Martin LBB, Misra RC, Kikuchi S, Rejzek M, Martin AC, Harkess A, et al. 2023. Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree. Science 379: 1252–1264.

RepeatMasker Home Page. http://www.repeatmasker.org. (Accessed June 6, 2025b).

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21: 245.

Richards EJ, Ausubel FM. 1988. Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. Cell 53: 127–136.

Richards EJ, Chao S, Vongs A, Yang J. 1992. Characterization of Arabidopsis thaliana telomeres isolated in yeast. Nucleic Acids Res 20: 4039–4046.

Ritmejeris J, Chen X, Dekker C. 2024. Single-molecule protein sequencing with nanopores. Nat Rev Bioeng. https://www.nature.com/articles/s44222-024-00260-8.

Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. 2019. An Ultra High-Density Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. Genetics 213: 771–787.

Satterlee JW, Alonso D, Gramazio P, Jenike KM, He J, Arrones A, Villanueva G, Plazas M, Ramakrishnan S, Benoit M, et al. 2024. Convergent evolution of

plant prickles by repeated gene co-option over deep time. Science 385: eado1663.

Schiff Y, Kao C-H, Gokaslan A, Dao T, Gu A, Kuleshov V. 2024. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. arXiv [q-bioGN]. https://doi.org/10.48550/arXiv.2403.03234.

Schloissnig S, Pani S, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, Asparuhova M, et al. 2025. Structural variation in 1,019 diverse humans based on long-read sequencing. Nature 644: 442–452.

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D. 2009. Simultaneous alignment of short reads against multiple genomes. Genome Biol 10: R98.

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N, et al. 2011. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proc Natl Acad Sci U S A 108: 10249–10254.

Schrumpfová PP, Fajkus J. 2020. Composition and Function of Telomerase-A Polymerase Associated with the Origin of Eukaryotes. Biomolecules 10. http://dx.doi.org/10.3390/biom10101425.

Schweiger R, Lee S, Zhou C, Yang T-P, Smith K, Li S, Sanghvi R, Neville M, Mitchell E, Nessa A, et al. 2024. Insights into non-crossover recombination from long-read sperm sequencing. bioRxiv. http://dx.doi.org/10.1101/2024.07.05.602249.

Sereika M, Mussig AJ, Jiang C, Knudsen KS, Jensen TBN, Petriglieri F, Yang Y, Jørgensen VR, Delogu F, Sørensen EA, et al. 2025. Genome-resolved long-read sequencing expands known microbial diversity across terrestrial habitats. Nat Microbiol 10: 2018–2030.

Serra Mari R, Schrinner S, Finkers R, Ziegler FMR, Arens P, Schmidt MH-W, Usadel B, Klau GW, Marschall T. 2024. Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. Genome Biol 25: 26.

Shahid S, Slotkin RK. 2020. The current revolution in transposable element biology enabled by long reads. Curr Opin Plant Biol 54: 49–56.

Shakirov EV, Shippen DE. 2004. Length regulation and dynamics of individual telomere tracts in wild-type Arabidopsis. Plant Cell 16: 1959–1967.

Shao Y, Zhou L, Li F, Zhao L, Zhang B-L, Shao F, Chen J-W, Chen C-Y, Bi X, Zhuang X-L, et al. 2023. Phylogenomic analyses provide insights into primate evolution. Science 380: 913–924.

Shendure J, Akey JM. 2015. The origins, determinants, and consequences of

human mutations. Science 349: 1478–1483.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature 550: 345–353.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. Bioinformatics 37: 1639–1643.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods 14: 407–410.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10: e1001241.

Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. Science 191: 528–535.

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol 42: 1571–1580.

Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, et al. 2019. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat Genet 51: 1215–1221.

Song J, Logeswaran D, Castillo-González C, Li Y, Bose S, Aklilu BB, Ma Z, Polkhovskiy A, Chen JJ-L, Shippen DE. 2019. The conserved structure of plant telomerase RNA provides the missing link for an evolutionary pathway from ciliates to humans. Proc Natl Acad Sci U S A 116: 24542–24550.

Stammnitz MR, Gori K, Kwon YM, Harry E, Martin FJ, Billis K, Cheng Y, Baez-Ortega A, Chow W, Comte S, et al. 2023. The evolution of two transmissible cancers in Tasmanian devils. Science 380: 283–293.

Steichen JM, Phung I, Salcedo E, Ozorowski G, Willis JR, Baboo S, Liguori A, Cottrell CA, Torres JL, Madden PJ, et al. 2024. Vaccine priming of rare HIV broadly neutralizing antibody precursors in nonhuman primates. Science 384: eadj8321.

Stephens Z, Kocher J-P. 2024. Characterization of telomere variant repeats using long reads enables allele-specific telomere length estimation. BMC Bioinformatics 25: 194.

Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science 368: 1449–1454.

Sun H, Tusso S, Dent CI, Goel M, Wijfjes RY, Baus LC, Dong X, Campoy JA, Kurdadze A, Walkemeier B, et al. 2025a. The phased pan-genome of tetraploid European potato. Nature 642: 389–397.

Sun J, Chen Y, Wei Y, Zhang K, Fu Y, Yan Z, Zhu X, Zhang S, Zhang L, Li J. 2025b. A chromosome-scale genome assembly of Giardia duodenalis by long-read sequencing of ten trophozoites. Sci Data 12: 1079.

Su W, Yang Y, Zhao Y, Yuan S, Xie X, Hao Y, Zhang H, Ye D, Lyu H, Lin H. 2025. iPro-MP: a BERT-based model to predict multiple prokaryotic promoters. Genome Biol 26: 353.

Symonds VV, Lloyd AM. 2003. An analysis of microsatellite loci in Arabidopsis thaliana: mutational dynamics and application. Genetics 165: 1475–1488.

Tang S, Conte V, Zhang DJ, Žedaveinytė R, Lampe GD, Wiegand T, Tang LC, Wang M, Walker MWG, George JT, et al. 2024. De novo gene synthesis by an antiviral reverse transcriptase. Science 386: eadq0876.

Tan K-T, Slevin MK, Meyerson M, Li H. 2022. Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. Genome Biol 23: 180.

Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. 2024. Sequencing and characterizing short tandem repeats in the human genome. Nat Rev Genet 25: 460–475.

Tao Y, Xian W, Bao Z, Rabanal FA, Movilli A, Lanz C, Shirsekar G, Weigel D. 2024. Atlas of telomeric repeat diversity in Arabidopsis thaliana. Genome Biol 25: 244.

Teano G, Concia L, Wolff L, Carron L, Biocanin I, Adamusová K, Fojtová M, Bourge M, Kramdi A, Colot V, et al. 2023. Histone H1 protects telomeric repeats from H3K27me3 invasion in Arabidopsis. Cell Rep 42: 112894.

Teasdale LC, Murray KD, Collenberg M, Contreras-Garrido A, Schlegel T, van Ess L, Jüttner J, Lanz C, Deusch O, Fitz J, et al. 2024. Pangenomic context reveals the extent of intraspecific plant NLR evolution. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2024.09.02.610789.

Tham C-Y, Poon L, Yan T, Koh JYP, Ramlee MK, Teoh VSI, Zhang S, Cai Y, Hong Z, Lee GS, et al. 2023. High-throughput telomere length measurement at nucleotide resolution using the PacBio high fidelity

sequencing platform. Nat Commun 14: 281.

Tomaz da Silva P, Karollus A, Hingerl J, Galindez GST, Wagner N, Hernandez-Alias X, Incarnato D, Gagneur J. 2025. Nucleotide dependency analysis of genomic language models detects functional elements. Nat Genet 57: 2589–2602.

Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, Chan SL, Poon LCY, Leung TY, Chan KCA, et al. 2021. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. Proc Natl Acad Sci U S A 118: e2019768118.

Tsuchimatsu T, Goubet PM, Gallina S, Holl A-C, Fobis-Loisy I, Bergès H, Marande W, Prat E, Meng D, Long Q, et al. 2017. Patterns of Polymorphism at the Self-Incompatibility Locus in 1,083 Arabidopsis thaliana Genomes. Mol Biol Evol 34: 1878–1889.

Verdun RE, Karlseder J. 2007. Replication and protection of telomeres. Nature 447: 924–931.

Vidal R, Frangione B, Rostagno A, Mead S, Révész T, Plant G, Ghiso J. 1999. A stop-codon mutation in the BRI gene associated with familial British dementia. Nature 399: 776–781.

Villanea FA, Peede D, Kaufman EJ, Añorve-Garibay V, Chevy ET, Villa-Islas V, Witt KE, Zeloni R, Marnetto D, Moorjani P, et al. 2025. The MUC19 gene: An evolutionary history of recurrent introgression and natural selection. Science 389. http://dx.doi.org/10.1126/science.adl0882.

Voichek Y, Weigel D. 2020. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet 52: 534–540.

Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, Rozanski AN, Lucas J, Asri M, Human Pangenome Reference Consortium, et al. 2023. Increased mutation and gene conversion within human segmental duplications. Nature 617: 325–334.

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, et al. 2022. Segmental duplications and their variation in a complete human genome. Science 376: eabj6965.

Vollmer SV, Selwyn JD, Despard BA, Roesel CL. 2023. Genomic signatures of disease resistance in endangered staghorn corals. Science 381: 1451–1454.

Vozárová R, Wang W, Lunerová J, Shao F, Pellicer J, Leitch IJ, Leitch AR, Kovařík A. 2022. Mega-sized pericentromeric blocks of simple telomeric repeats and their variants reveal patterns of chromosome evolution in ancient Cycadales genomes. Plant J 112: 646–663.

Vrbsky J, Akimcheva S, Watson JM, Turner TL, Daxinger L, Vyskot B, Aufsatz W, Riha K. 2010. siRNA-mediated methylation of Arabidopsis telomeres. PLoS Genet 6: e1000986.

Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics 20: 275.

Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun 7: 11708.

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al. 2022a. High-quality Arabidopsis thaliana genome assembly with nanopore and HiFi long reads. Genomics Proteomics Bioinformatics 20: 4–13.

Wang C-T, Ho C-H, Hseu M-J, Chen C-M. 2010. The subtelomeric region of the Arabidopsis thaliana chromosome IIIR contains potential genes and duplicated fragments from other chromosomes. Plant Mol Biol 74: 155–166.

Wang L-B, Li Z-K, Wang L-Y, Xu K, Ji T-T, Mao Y-H, Ma S-N, Liu T, Tu C-F, Zhao Q, et al. 2022b. A sustainable mouse karyotype created by programmed chromosome fusion. Science 377: 967–975.

Wang S, Xu Z, Li M, Lv M, Shen S, Shi Y, Li F. 2023. Structural insights into the recognition of telomeric variant repeat TTGGGG by broad-complex, tramtrack and bric-à-brac - zinc finger protein ZBTB10. J Biol Chem 299: 102918.

Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39: 1348–1365.

Wei X, Chen M, Zhang Q, Gong J, Liu J, Yong K, Wang Q, Fan J, Chen S, Hua H, et al. 2024. Genomic investigation of 18,421 lines reveals the genetic architecture of rice. Science 385: eadm8762.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37: 1155–1162.

Widen SA, Bes IC, Koreshova A, Pliota P, Krogull D, Burga A. 2023. Virus-like transposons cross the species barrier and drive the evolution of genetic incompatibilities. Science 380: eade0705.

Wilkinson ME, Li D, Gao A, Macrae RK, Zhang F. 2024. Phage-triggered reverse transcription assembles a toxic repetitive gene from a noncoding

RNA. Science 386: eadq3977.

Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, Scott A, Mandáková T, Gorringe N, Tock AJ, Holland D, et al. 2023. Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature 618: 557–565.

Xian W, Bao Z, Vorbrugg S, Tao Y, Movilli A, Bezrukov I, Weigel D. 2025. The structure of mitochondrial genomes is associated with geography inArabidopsis thaliana. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/2025.01.11.632530.

Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, Li X, Tian S, Wang L, Wang Z. 2023. Long-read sequencing identifies novel structural variations in colorectal cancer. PLoS Genet 19: e1010514.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76–82.

Yan H, Han J, Jin S, Han Z, Si Z, Yan S, Xuan L, Yu G, Guan X, Fang L, et al. 2025. Post-polyploidization centromere evolution in cotton. Nat Genet 57: 1021–1030.

Yilmaz F, Karageorgiou C, Kim K, Pajic P, Scheer K, Human Genome Structural Variation Consortium, Beck CR, Torregrossa A-M, Lee C, Gokcumen O, et al. 2024. Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. Science 386: eadn0609.

Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. 2021. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics 36: 5582–5589.

Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, Patel V, Velikkakam James G, Koornneef M, Ossowski S, 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: E4052–60.

Závodník M, Fajkus P, Franek M, Kopecký D, Garcia S, Dodsworth S, Orejuela A, Kilar A, Ptáček J, Mátl M, et al. 2023. Telomerase RNA gene paralogs in plants - the usual pathway to unusual telomeres. New Phytol 239: 2353–2366.

Zhai J, Gokaslan A, Hsu S-K, Chen S-P, Liu Z-Y, Marroquin E, Czech E, Cannon B, Berthel A, Romay MC, et al. 2025a. PlantCAD2: A long-context DNA language model for cross-species functional annotation in angiosperms. bioRxiv. http://dx.doi.org/10.1101/2025.08.27.672609.

Zhai J, Gokaslan A, Schiff Y, Berthel A, Liu Z-Y, Lai W-Y, Miller ZR, Scheben A, Stitzer MC, Romay MC, et al. 2025b. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language

model. Proc Natl Acad Sci U S A 122: e2421738122.

Zhang F, Xue H, Dong X, Li M, Zheng X, Li Z, Xu J, Wang W, Wei C. 2022. Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. Genome Res. http://dx.doi.org/10.1101/gr.276015.121.

Zhang P, Mbodj A, Soundiramourtty A, Llauro C, Ghesquière A, Ingouff M, Keith Slotkin R, Pontvianne F, Catoni M, Mirouze M. 2023a. Extrachromosomal circular DNA and structural variants highlight genome instability in Arabidopsis epigenetic mutants. Nat Commun 14: 5236.

Zhang X, Meng W, Liu D, Pan D, Yang Y, Chen Z, Ma X, Yin W, Niu M, Dong N, et al. 2024. Enhancing rice panicle branching and grain yield through tissue-specific brassinosteroid inhibition. Science 383: eadk8838.

Zhang Y, Chu J, Cheng H, Li H. 2023b. De novo reconstruction of satellite repeat units from sequence data. Genome Res 33: 1994–2001.

Zheng T, Zhang X, Liu T, Liu X, Bowler C, Lin X. 2025. The roles of transposable elements and gene family dynamics in shaping diversity and evolution in diatoms. bioRxiv. http://dx.doi.org/10.1101/2025.10.24.684338.

Zhou L, Qiu Q, Zhou Q, Li J, Yu M, Li K, Xu L, Ke X, Xu H, Lu B, et al. 2022. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. Nat Commun 13: 2563.

Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, Karlowski WM, Figlerowicz M. 2020. AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome. Plant Cell 32: 1797–1819.