Generation of New Mutations at Different Time Scales in Plants

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Wenfei Xian

aus Zhaoqing, China

Tübingen

2025

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

Tag der mündlichen Qualifikation: 06.08. 2025

Dekan: Prof. Dr. Thilo Stehle

1. Berichterstatter: Prof. Dr. Detlef Weigel

2. Berichterstatter: Prof. Dr. Daniel Huson

Table of Contents

Summary

Rapid climate change threatens both food security and biodiversity. The latter is ultimately a reflection of genetic change. Studying genetic variation in wild populations thus provides insight into how genomes change during evolution and may help explain how species adapt to different environments. While most population genetic studies have focused on the nuclear genome, organellar genomes, particularly those of mitochondria, have received much less attention, partly due to difficulties in assembly. Importantly, organellar genomes in land plants evolve more slowly than nuclear genomes and therefore tend to preserve signals of divergence and evolutionary history over deeper time scales. At the same time, there would be no complex life without mutations, which are the ultimate source of genetic diversity, but how they arise and accumulate remains poorly understood.

The development of long-read sequencing, such as PacBio high-fidelity (HiFi) reads and the most recent versions of Oxford Nanopore Technology (ONT) reads, has made it possible to assemble complete organellar genomes, but the use of these data requires a dedicated bioinformatics tool. The complete and accurate assemblies allow for more comprehensive investigation of organellar genome variation and its evolutionary significance. In parallel, long-read data also improve the quality of nuclear genome assemblies, which in turn enhances the detection of somatic mutations and then deepens our understanding of their mutation patterns and biological context.

This thesis addresses these gaps by developing a tool for assembling plant organellar genomes, applying it to more than one hundred accessions of *Arabidopsis thaliana* to investigate organellar genetic variation, and by refining mutation detection approach in a highly heterozygous oak genome to detect somatic mutations with higher accuracy in challenging genomes.

In the first chapter, I introduce TIPPo, a reference-free tool I developed for assembling plant organellar genomes using HiFi data. TIPPo addresses challenges specific to organelle genomes, including high copy number, repetitive content, and the presence of organelle-derived sequences in the nuclear genome. Through a combination of read classification and filtering strategies, TIPPo produces

high-quality assemblies that outperform existing methods. Because both the nuclear and organellar genomes were assembled from the same sample, the identification of NUPTs and NUMTs and their substitution patterns becomes more reliable.

Building on this, the second chapter presents a population-scale analysis of organellar genomes from 143 *A. thaliana* accessions. While chloroplast genomes have conserved structure and size, mitochondrial genomes are more variable and they could be grouped into two major classes based on repeat content. I found that the number of large repeats in the mitochondrial genome correlated with sampling latitude, suggesting that geographic or historical factors may have influenced mitochondrial genome structure. In addition, I identified unannotated open reading frames (ORFs) with supporting expression evidence. I found two putative horizontally transferred ORFs, one of which is associated with cytoplasmic male sterility (CMS). These findings suggest that mitochondrial genomes can acquire novel functional elements from related species, enriching their evolutionary plasticity.

The third chapter shifts focus to the origin of new genetic variation, using a long-lived oak tree to study somatic mutations. By reassembling the oak genome with HiFi data and applying a hybrid alignment strategy tailored to high heterozygosity, I recovered substantially more high confidence somatic mutations than previous study. Most of the mutations were C:G > T:A transitions, consistent with the result in mutation accumulation studies. These results provide a clearer view of how mutations accumulate over time and show that high quality assembly and careful alignment strategies are key to detecting somatic mutation in complex plant genomes.

Together, these chapters explore how genetic variation arises, accumulates, and is maintained in plant genomes. By combining tool/approach development with population-scale and individual-level analyses, this work provides new insight into both the structure of organellar genomes and the processes shaping somatic mutation patterns. The methods developed here may be broadly useful for studying genome dynamics in diverse species and offer a foundation for future investigations into the evolutionary and functional consequences of genetic variation.

Zusammenfassung

Der rasche Klimawandel bedroht sowohl unsere Ernährungssicherheit als auch die Artenvielfalt unseres Planeten. Die Artenvielfalt wiederum ist ein Ergebnis genetischer Veränderungen. Die Untersuchung genetischer Variation in natürlichen Populationen kann daher dazu beitragen, die Anpassung von Arten an unterschiedliche Lebensräume zu erklären, indem sie Erkenntnisse darüber liefert, wie sich Genome im Laufe der Evolution verändern. Während sich die meisten populationsgenetischen Studien auf das Kerngenom konzentriert haben, wurden Genome der Organellen, insbesondere die der Mitochondrien, bislang nur wenig beachtet, was zum Teil auf Schwierigkeiten bei ihrer Rekonstruktion aus Sequenzdaten zurückzuführen ist. Wichtig ist, dass sich Genome von Organellen in Landpflanzen langsamer verändern als Kerngenome und daher besser die Signale der Divergenz und Evolutionsgeschichte über längere Zeiträume bewahren. Gleichzeitig gäbe es ohne Mutationen, die letztendlich die Quelle aller genetischer Vielfalt sind, kein komplexes Leben, aber wie sie entstehen und sich in Genomen ansammeln, ist in vielen Teilen noch unverstanden.

Die Entwicklung von Sequenzierung mit *long reads*, wie PacBio High-Fidelity (HiFi) *reads* und die neuesten Versionen von Oxford Nanopore Technology (ONT) *reads*, hat die Assemblierung vollständiger Organellengenome ermöglicht, aber die Verwendung dieser Daten erfordert ein spezielles Bioinformatik-Werkzeug. Die vollständige und genaue Assemblierung zahlreicher Organellengenome ermöglicht wiederum umfassendere Untersuchung über deren Variation und die evolutionäre Bedeutung dieser Variation. Parallel dazu verbessern *long read* Daten auch die Qualität der Assemblierungen des Kerngenoms, was wiederum die Erkennung somatischer Mutationen verbessert und somit unser Verständnis zum Muster von Mutationen und deren biologischen Kontext im Genom vertieft.

Diese Arbeit schließt diese Lücken, indem sie ein Werkzeug zur Assemblierung pflanzlicher Organellengenome entwickelt, dieses auf mehr als hundert Stämme von *Arabidopsis thaliana* anwendet, um die genetische Variation von Organellen zu untersuchen, und den Ansatz zur Mutationserkennung in einem stark heterozygoten Genom einer Eiche verfeinert, um somatische Mutationen in schwierigen Genomen mit höherer Genauigkeit erkennen zu können.

Im ersten Kapitel stelle ich TIPPo vor, ein Werkzeug, das ich für die Assemblierung von pflanzlichen Organellengenomen ohne Hilfe eines Referenzgenoms unter Verwendung von HiFi-Daten entwickelt habe. TIPPo löst Herausforderungen, die spezifisch für Organellengenome sind, darunter deren hohe Kopienzahl pro Zelle, viele repetitive Sequenzen und das Vorhandensein von aus Organellen stammenden Sequenzen im Kerngenom. Durch eine Kombination aus *read* Klassifizierung und Filterstrategien erzeugt TIPPo hochwertige Assemblierungen, die bestehende Methoden übertreffen. Da sowohl das Kerngenom als auch das Organellengenom aus derselben Probe assembliert wurden, wird die Identifizierung von NUPTs und NUMTs und deren Substitutionsmuster zuverlässiger.

Auf dieser Grundlage präsentiert das zweite Kapitel eine populationsweite Analyse von Organellengenomen aus 143 *A. thaliana* Stämmen. Während Chloroplastengenome eine konservierte Struktur und Größe aufweisen, sind Mitochondriengenome variabler und lassen sich anhand ihres Gehalts an repetitiven Sequenzen in zwei Hauptklassen einteilen. Die Anzahl großer Wiederholungen im Mitochondriengenom korreliert mit dem Breitengrad der Provenienz der Stämme, was darauf hindeutet, dass geografische oder historische Faktoren die Struktur des Mitochondriengenoms beeinflusst haben könnten. Darüber hinaus habe ich nicht annotierte offene Leserahmen (ORFs), für die es auch Evidenz von RNA Daten gibt, identifiziert. Ich habe zwei mutmaßlich horizontal übertragene ORFs gefunden, von denen einer mit zytoplasmatischer männlicher Sterilität (CMS) in Verbindung steht. Diese

Ergebnisse deuten darauf hin, dass mitochondriale Genome neue funktionelle Elemente von verwandten Arten erwerben können, wodurch ihre evolutionäre Plastizität verstärkt wird.

Im dritten Kapitel verlagere ich den Schwerpunkt auf die Entstehung neuer genetischer Variationen und untersucht somatische Mutationen anhand einer langlebigen Eiche. Durch die Neuassemblierung des Genoms der Napoleon-Eiche mit HiFi-Daten und die Anwendung einer auf hohe Heterozygotie zugeschnittenen hybriden Alignment-Strategie konnte ich wesentlich mehr somatische Mutationen mit hoher Konfidenz identifizieren als es früheren Studien gelang. Die meisten Mutationen waren C:G > T:A-Transitionen, was mit den Ergebnissen von publizierten Mutationsakkumulationsstudien übereinstimmt. Diese Ergebnisse liefern ein klareres Bild davon, wie Mutationen im Laufe der Zeit akkumulieren, und zeigen, dass ein hochwertiges Genome und sorgfältige Alignment-Strategien entscheidend für den Nachweis somatischer Mutationen in komplexen Pflanzengenomen sind.

Zusammen untersuchen diese Kapitel, wie genetische Variation entsteht, wie sie akkumuliert und wie sie in Pflanzengenomen erhalten bleibt. Durch die Kombination der Entwicklung von Werkzeugen und Analysen auf Populations- und individueller Ebene liefert meine Arbeit neue Erkenntnisse sowohl über die Struktur von Organellengenomen als auch über die Prozesse, die somatische Mutationsmuster prägen. Die hier entwickelten Methoden könnten für die Untersuchung der Genomdynamik in verschiedenen Arten von Nutzen sein und eine Grundlage für zukünftige Untersuchungen zu den evolutionären und funktionellen Folgen genetischer Variation bilden.

Acknowledgements

I am deeply grateful to Detlef Weigel for giving me the opportunity to come to Tübingen and complete my PhD. This was also the first time in my life that I left my country. I sincerely thank Detlef for providing me with an exceptionally free research environment and strong support throughout my doctoral studies. His wisdom helped me recognize the limitations in many of my ideas, and for that, I am truly thankful—it saved me from taking many wrong turns.

I would also like to thank Daniel Huson and Hajk-Georg Drost for being my TAC members. Every discussion with them felt like an annual review, giving me the chance to reflect and recognize where I still need to improve.

I am grateful to many people who have supported me both in research and in life: Haim Ashkenazy, Zhigui Bao, Ilja Bezrukov, Pablo Carbonell-Bejerano, Anupam Gautam, Li He, Christa Lanz, Miriam Lucke, Andrea Movilli, Fernando A. Rabanal, Rebecca Schwab, Yueqi Tao, Sebastian Vorbrugg, Shanshan Wang, Hülya Wicher, Wei Yuan, and other current and former members of the Weigel lab.

Finally, I want to thank my family, my parents, my sister, my sister's two lovely daughters, and my girlfriend Xiaofeng, for their unconditional love and support.

Publications

Accepted manuscripts

Wenfei Xian, Ilja Bezrukov, Zhigui Bao, Sebastian Vorbrugg, Anupam Gautam, Detlef Weigel, **TIPPo: A User-Friendly Tool for De Novo Assembly of Organellar Genomes with High-Fidelity Data**, Molecular Biology and Evolution, Volume 42, Issue 1, January 2025, msae247, https://doi.org/10.1093/molbev/msae247

Wenfei Xian, Pablo Carbonell-Bejerano, Fernando A Rabanal, Ilja Bezrukov, Philippe Reymond, Detlef Weigel, **Minimizing detection bias of somatic mutations in a highly heterozygous oak genome**, G3 Genes|Genomes|Genetics, 2025;, jkaf143, https://doi.org/10.1093/g3journal/jkaf143

Preprint manuscript

Wenfei Xian, Zhigui Bao, Sebastian Vorbrugg, Yueqi Tao, Andrea Movilli, Ilja Bezrukov, Detlef Weigel, **The structure of mitochondrial genomes is associated with geography in Arabidopsis thaliana**, bioRxiv 2025.01.11.632530; doi: https://doi.org/10.1101/2025.01.11.632530

Introduction

General introduction and motivation

As global climate change accelerates, plants have to adapt to changing environments (Chaudhry and Sidhu 2022; Parmesan and Hanley 2015). Species adaptation depends on the abundance of genetic diversity present in the population (Fournier-Level et al. 2011; Hancock et al. 2011). Genetic variation is shaped by mutation, recombination, and selection. Its distribution across the genome and populations is uneven (Lynch et al. 2016; Colomé-Tatché et al. 2012; 1000 Genomes Project Consortium et al. 2015). It affects how populations respond to both short-term pressures and long-term evolutionary processes (Exposito-Alonso et al. 2019; Exposito-Alonso, Vasseur, et al. 2018). Understanding the role of genetic variation is therefore central to evolutionary biology.

Initial characterization of genome-wide variation used genotyping of a limited number of loci, mostly microsatellites and SNPs, using a variety of technologies (Weber and May 1989; Varshney, Graner, and Sorrells 2005; Nordborg et al. 2002). From around 2007 on, Illumina short-read sequencing, also known as a type of Next Generation Sequencing (NGS), became the method of choice for genome-wide characterization of sequence diversity (1001 Genomes Consortium 2016; 1000 Genomes Project Consortium et al. 2012). A shortcoming is that variation in repeat regions is difficult to access and that the methodology is largely blind to variation in sequences missing from the reference genomes, a limitation known as reference bias (Igolkina et al. 2024a; Valiente-Mullor et al. 2021). In the last decade, long-read sequencing technologies such as PacBio and Oxford Nanopore Technologies (ONT), originally referred to as third-generation methods, have been continuously improved in accuracy and throughput. Long-read sequencing now allows for the highly accurate reconstruction of near-complete genomes, including repetitive and structurally complex regions (Nurk et al. 2022; Naish et al. 2021; Wlodzimierz et al. 2023; Song et al. 2021).

One of the plant species that has played a pioneering role in the analysis of genetic variation is *Arabidopsis thaliana*, which is a widely used model for plant

biology (U. Krämer 2015). It is self-fertilizing, short-lived, and has a relatively small nuclear genome (~140 Mb), which makes it particularly suitable for large-scale genetic analyses (Woodward and Bartel 2018; Meinke et al. 1998). In the last two decades, a number of population-scale studies have been reported. Early efforts focused on population structure and linkage disequilibrium (LD) based on a small number of accessions using limited single nucleotide polymorphisms (SNPs) (Nordborg et al. 2002, 2005). Then, genome-wide identification of SNPs and small insertions/deletions followed, supporting the development of SNP arrays for large-scale genotyping (Clark et al. 2007; Sung Kim et al. 2007; Horton et al. 2012). Based on these resources, the 1001 Genomes Project sequenced over a thousand natural accessions using short reads (1001 Genomes Consortium 2016). More recently, efforts have focused on generating high-quality long-read assemblies (Jiao and Schneeberger 2020; Kang et al. 2023; Qichao Lian et al. 2024; Wlodzimierz et al. 2023; Naish et al. 2021). The ongoing 1001 Genomes Plus initiative is extending this work by collecting hundreds of accessions sequenced with long-read data, providing abundant resources for the discovery of genome diversity (Alonso-Blanco et al. 2024).

Existing population genetics studies have focused on the nuclear genome, while variation in organellar genomes is still poorly characterized. Organellar genomes, especially mitochondria, often contain large repeat sequences and complex recombination structures, which short-read data cannot handle well (Jin et al. 2020). The development of high-fidelity (HiFi) long-read sequencing has created an opportunity to overcome these limitations (Štorchová and Krüger 2024). However, a dedicated tool for the assembly of the organellar genome with HiFi data is needed. The complete organellar genome assembly allows us to explore genetic variation in a natural population and its potential relevance to plant adaptation.

In parallel, our knowledge about how new genetic variants arise is still limited, especially in the case of somatic mutations, with the boundary between somatic and germline mutations being fluid in plants, as the germling is not as clearly demarcated as in animals. These mutations occur during an individual's lifetime, but their low frequency makes them difficult to detect (Sangtae Kim et al. 2018; Benjamin et al. 2019). In short life-span species, the number of somatic mutations that accumulate in a single individual is often insufficient for meaningful statistical analysis (Ossowski

et al. 2010). To better study the rate and spectrum of somatic mutations, long-lived species provide a more suitable system, as mutations have more time to accumulate and can be tracked within the same individual (Plomion et al. 2018; Schmid-Siegert et al. 2017).

The following sections introduce the fundamental concepts and current research progress in these areas, laying the groundwork for a deeper understanding of the role of organellar variation and the origin of genetic variation in plant evolution.

1. Genetic diversity in *A. thaliana*

1.1 The origin and global distribution of *A. thaliana*

Arabidopsis thaliana belongs to the Brassicaceae family. Based on current molecular evolution models, it probably separated from its closest extant relative, *A. lyrata*, around five million years ago (Hu et al. 2011; Hohmann et al. 2015). But many of the traits we now associate with *A. thaliana*, such as selfing, small genome size, and five chromosomes, seem to have formed more recently, about half a million years ago (Durvasula et al. 2017). Earlier studies suggested that the species may have originated in the eastern Mediterranean or the Caucasus (Beck, Schmuths, and Schaal 2008). However, sequencing of African accessions showed that the African samples are unlikely to have been introduced by Europeans and that they have high genetic diversity (Durvasula et al. 2017). This led to the conclusion that *A. thaliana* likely originated in Africa, which is similar to what we know about human origins (1000 Genomes Project Consortium et al. 2015).

Around 120,000 to 90,000 years ago, precipitation and humidity increased in the North of Africa, and migration corridors across the Sahara became available (Osborne et al. 2008). It is thought that during this time, *A. thaliana* spread out of Africa into the Levant region, and then gradually expanded across Eurasia. When the Last Glacial Maximum arrived around 20,000 years ago, the species retreated into southern refugia, including the Iberian Peninsula, the Balkans, and parts of Central Asia. After the glaciers retreated, descendants of these populations expanded again into Eurasia. This pattern is supported by population genomic data,

which show a gradual decrease in genetic diversity from south to north (1001 Genomes Consortium 2016).

Although *A. thaliana* is now found on all continents except Antarctica, populations in the Americas and Oceania are not considered native. A previous study used 149 SNP markers to genotype a large number of accessions and found that *A. thaliana* in North America was dominated by Haplogroup1 (HPG1) (Platt et al. 2010). Later work using mutation accumulation patterns from pure HPG1 accessions estimated that this lineage arrived in North America around 400 years ago, which matches the historical timing of large-scale European colonization (Exposito-Alonso, Becker, et al. 2018). This suggests that the global spread of *A. thaliana* was likely driven by human movement during the Age of Exploration and the expansion of global trade.

1.2 Genetic variants and their functional role in *A. thaliana*

In the early stages of population genetics research in *A. thaliana*, researchers mainly used SNPs from a few regions of the nuclear genome to study population structure and LD, because sequencing was expensive (Nordborg et al. 2002, 2005). Later, microarray-based resequencing helped identify over one million SNPs and many deletions in 20 representative accessions (Clark et al. 2007). This led to the development of the 250k SNP chip for genotyping larger populations (Horton et al. 2012). Based on this progress, the 1001 Genomes project was launched to sequence the whole genomes of a thousand of natural accessions from different geographic regions using short-read sequencing (1001 Genomes Consortium 2016).

As more genetic variation data became available, research began to focus not only on population structure and LD, but also on the impact of genetic variants. Genome-wide association studies (GWAS) have identified many natural variants that are useful for both basic science and for understanding agriculturally relevant traits, such as those controlling seed dormancy, flowering time, and disease resistance (Atwell et al. 2010). In recent years, instead of relying only on traditional SNP markers, researchers have also used new methods like k-mer and graph-based genotypes (Voichek and Weigel 2020; Vorbrugg et al. 2024). These approaches help

capture different types of genetic variation and how they might influence trait diversity.

In addition to studying the genetic basis of morphological, physiological and life history traits, researchers have also looked at how natural selection shapes genetic variation. Genome-environment association (GEA) studies link genetic variants to environmental factors such as rainfall, temperature, and elevation. This helps identify alleles that might be favored in certain environments, with confounding by geography and population structure being a major issue. Early studies showed that some SNPs related to climate include non-synonymous SNPs, which suggests they may play a role in adaptation through effects on protein activity (Hancock et al. 2011). Later, common garden experiments found that alleles linked to better survival or reproduction in specific regions are also more likely to be found in populations from those regions, providing further evidence of local adaptation (Fournier-Level et al. 2011).

In larger experiments, researchers measured the survival and reproduction of many accessions under different water conditions. By combining traits with genomic data, they found that the frequency of the 5% variants have been shaped by drought-related selection (Exposito-Alonso et al. 2019). A parallel study linked drought response traits to the original climate of each accession and found that drought-tolerant alleles were more common in populations from southern Europe and Scandinavia (Exposito-Alonso, Vasseur, et al. 2018). These results suggest that some alleles in *A. thaliana* are selected by climate, and they provide insight into how environmental adaptation shapes the global distribution of genetic variation.

Notably, even though many essential functions of the organism are carried out in the mitochondria and chloroplasts, the focus has been almost exclusively on nuclear variants. The role of organellar genomes in adaptation, and the types of variation they carry, are still not well understood.

2. Organellar genomes and intraspecific variation

2.1 Characteristics of plant organellar genomes

In photosynthetic eukaryotes, two main organelles—chloroplasts and mitochondria—serve as the major sites of energy conversion on the cells. Both are believed to have originated from ancient endosymbiotic events involving α-proteobacteria and cyanobacteria, respectively (Zimorski et al. 2014). The origin of mitochondria is thought to be older, potentially dating back to the emergence of the first eukaryotic cells. Indeed, with the exception of a single parasitic lineage (Karnkowska et al. 2016), all known eukaryotes possess mitochondria. In contrast, chloroplasts are thought to have originated later, at the base of the Archaeplastida lineage. Through subsequent secondary or even tertiary endosymbiosis, many unrelated lineages, such as brown algae, also contain plastids (Keeling 2010).

Because Archaeplastida represents only a relatively recent branch of the eukaryotic tree, its chloroplast genomes are comparatively conserved in size and content. Most chloroplast genomes range between 100–200 kb, contain fewer than 250 genes and are structured as a large single-copy (LSC) region, a small single-copy (SSC) region, and two inverted repeats (IRs) (Archibald 2015). Mitochondrial genomes, however, show dramatic variation in both size and structure across different kingdoms (Butenko et al. 2024). In mammals, the mitochondrial genome has been reduced to a compact 17 kb circle DNA (Taanman 1999). In plants, particularly in land plants, mitochondrial genomes are far more variable and complex. Although green algae (chlorophytes) have small mitochondrial genomes (Vahrenholz et al. 1993), their sister group, the streptophytes, already show an increase in mitochondrial genome size and complexity (Butenko et al. 2024). As land plants diversified, this trend became stronger, with the accumulation of large repetitive sequences and alternative genome configurations (heteroplasmy) that greatly complicated assembly.

Due to these complexities, mitochondrial genomes of plants are underrepresented in public databases. While thousands of chloroplast genomes have been assembled and deposited, the number of available mitochondrial genomes remains roughly twenty times lower (Bi et al. 2024). The largest plant

mitochondrial genome reported to date exceeds 11 Mb in size, illustrating the extent of variation and structural expansion possible in this compartment (Sloan et al. 2012).

Another major challenge in assembling the plant mitochondrial genomes is the presence of organellar sequences that are integrated in the nuclear genome, where they mutate and their sequence diverges with those of the original source over time. In plants, sequences derived from mitochondria and chloroplasts that have been inserted into the nuclear genome are referred to as NUMTs (nuclear mitochondrial DNA transfers) and NUPTs (nuclear plastid DNA transfers), respectively (Michalovova, Vyskot, and Kejnovsky 2013). In practical terms, NUMTs and NUPTs pose a challenge to organellar genome assembly, especially when the DNA used as input is isolated from whole cells rather than purified nuclei. High similarity between NUPT/NUMT and genuine organellar sequences can lead to chimeric assemblies (Fields et al. 2022). Therefore, accurately distinguishing true organellar reads from nuclear copies is essential for reliable analysis, particularly in species where the rate of organelle-to-nucleus transfer is high.

HiFi sequencing has made it possible to resolve regions that short reads were unable to span, including highly repetitive areas like centromeres. The mitochondrial genome can be assembled well by using HiFi reads in theory, but there is still a need for a dedicated and efficient tool for assembling plant organellar genomes.

2.2 Current knowledge on organellar genome variation within species

Compared to the nuclear genome, organellar genomes have traditionally been considered to be more conserved and to evolve more slowly (J. Wang et al. 2024). Studies in several plant species have revealed that organellar genomes can exhibit intraspecific variation, although they have largely focussed on the chloroplast genome (Go et al. 2024; Hongfang Liu et al. 2022; N. Wang et al. 2022). In particular, mitochondrial genome variation has been linked to cytoplasmic male sterility (CMS), a maternally inherited trait that causes the abortion of pollen development while leaving female fertility unaffected (N. Wang et al. 2022). It is typically caused by novel ORFs or chimeric genes generated through mitochondrial

recombination, which disrupt normal mitochondrial function in pollen-producing tissues (Tang et al. 2017; Xiao et al. 2022; Dehaene et al. 2024). CMS has been documented across many plant species and plays a key role in hybrid breeding (Brownfield 2021). CMS is of particular interest because it provides one of the main avenues for the large-scale production of hybrid seeds for crops (Bohra et al. 2016; Hanson and Bentolila 2004).

Importantly, the expression and phenotypic effects of CMS are often dependent on interactions with nuclear-encoded restorer-of-fertility (Rf) genes (Melonek et al. 2021; Itabashi et al. 2011). This form of nuclear-cytoplasmic interaction highlights the evolutionary significance of organellar variation and contributes to the broader phenomenon of cytonuclear co-evolution (Qun Lian et al. 2024).

Beyond CMS, mitochondrial and chloroplast genomes may also harbor functional variants that do not directly cause sterility but still influence plant physiology and fitness. Because organelles are uniparentally inherited and often non-recombining, beneficial mutations can rapidly sweep to fixation if they are linked to advantageous nuclear backgrounds—a process sometimes referred to as organellar hitchhiking (Flood et al. 2016). Conversely, deleterious mutations may also persist if they are shielded by linkage to positively selected nuclear loci.

Although these dynamics have been explored in a few crop species and model systems, our understanding of population-level variation in organellar genomes of *A. thaliana* remains limited. The extent to which organellar diversity contributes to local adaptation and shapes population structure is still poorly characterized, partly due to technical difficulties in assembling and analyzing complete organellar genomes.

3. Sources of genetic variants – mutation

3.1 Mutations as the raw material of genomic variants

Mutation is the main source of genetic variation and a driver of evolution. Across different kingdoms, mutation rates over evolutionary time scales show striking

differences. In animals, mitochondrial genomes typically mutate faster than nuclear genomes. In land plants, however, the pattern is reversed. The mutation rate of the nuclear genome is about twice that of the chloroplast and six times higher than that of the mitochondrial genome (J. Wang et al. 2024). Because the mutation rates of organellar genomes are so low, studies aiming to understand the mutational landscape, underlying mechanisms, and the role of mutations in development and evolution in plants mostly focus on the nuclear genome.

Mutations can be broadly classified into three categories: single-nucleotide variants (SNVs), also known as single-nucleotide polymorphisms (SNPs), small insertions or deletions (indels), and larger-scale structural variants (SVs) such as inversions, translocations and copy number variations. Among these, SNVs are the most common and can be further divided into transitions (e.g., purine to purine) and transversions (e.g., purine to pyrimidine). Different types of mutations can have distinct functional consequences. SNVs may alter amino acid sequences, indels can cause frameshifts, and structural variants may disrupt gene structure, regulatory regions, or even whole chromosomes.

Multiple mechanisms contribute to the occurrence of mutations. Errors during DNA replication are one of the most frequent sources (Chuong et al. 2024; Fujii et al. 1999; Hasenauer et al. 2025). Although cells have high-fidelity mismatch repair systems, some errors escape correction and become fixed. Endogenous damage such as oxidative stress, deamination, or alkylation, as well as exogenous factors like UV light, radiation, or chemical exposure, can also lead to mutations—often through errors during DNA repair (Tubbs and Nussenzweig 2017; Anderson et al. 2024; Alexandrov et al. 2020; Zhivagui et al. 2023). In plants, transposable element activity is another major source of mutations, especially when it leads to recombination or interrupting local structure (Quesneville 2020; Shimada et al. 2024). Microsatellite regions (short tandem repeats), due to their repetitive structure, are prone to polymerase slippage during replication, resulting in changes in repeat number (Fan and Chu 2007). In addition, abnormal non-homologous recombination during meiosis can also cause large insertions, deletions, or chromosomal rearrangements (Gebow, Miselis, and Liber 2000).

Mutations are not evenly distributed across the genome. Their frequency is shaped by various intrinsic factors, including local sequence context, DNA methylation levels, chromatin accessibility and transcriptional activity (Monroe et al. 2022). For example, methylated cytosines are prone to deamination, leading to C>T transitions, which are often enriched in highly methylated regions (Bhagwat et al. 2016). Understanding these influences not only helps explain the non-random distribution of mutations, but also provides a theoretical framework for the formation of mutation spectra and the evolutionary consequences of mutations.

3.2 Mutation accumulation within and across generations

Mutation accumulation can occur on two distinct timescales: intragenerational somatic mutations and intergenerational germline mutations. The latter are transmitted to the next generation through germline cells and shape long-term evolutionary trajectories, while the former arise during an individual's growth and development, leading to genetic heterogeneity among different tissues within the same organism. Although both types of mutations reflect the dynamic nature of the genome, they differ substantially in terms of research strategies and technical feasibility.

For studying intergenerational mutations, the most widely used approach is the establishment of mutation accumulation (MA) lines (Schultz, Lynch, and Willis 1999). In such experiments, a single individual is selected for reproduction in each generation, minimizing the influence of natural selection and allowing spontaneous mutations to accumulate in the germline. After many generations, whole-genome sequencing is performed to identify newly arisen mutations. In *Arabidopsis thaliana*, early MA line studies estimated a mutation rate of approximately 7×10^{-9} per site per generation and revealed a mutation spectrum dominated by C>T transitions, particularly at methylated cytosine sites (Ossowski et al. 2010; Weng et al. 2019; Exposito-Alonso, Becker, et al. 2018).

In contrast, studies on mutation accumulation in plants within an individual's lifespan have mainly focused on long-lived plants, particularly old trees (Schmitt et al. 2024; Satake et al. 2024; Hanlon, Otto, and Aitken 2019; Hofmeister et al. 2020; Schmid-Siegert et al. 2017; Plomion et al. 2018). Because their tissues are

continuously produced by active meristems, it is possible to sequence different branches of a single tree to identify somatic mutations that have accumulated during development. For example, in a study sequencing different branches of a 234-year-old oak tree, only 17 somatic mutations were identified and the number was lower than expected (Schmid-Siegert et al. 2017). Some researchers have proposed that this may be due to conservative mutation-calling pipelines and the limited quality of the available reference genome (Plomion et al. 2018).

Unlike germline mutations in MA lines, which are present in all cells and thus easily detectable, somatic mutations exist only in a subset of cells and usually appear at low read frequencies, making them more difficult to distinguish from sequencing errors. What's more, a recent study suggests that some low-frequency somatic SNVs indeed be transmissible to offspring (Schmitt et al. 2024). Furthermore, somatic mutations in trees often exhibit layer specificity, and the typical sampling of entire leaves tends to dilute these signals, further reducing detection sensitivity (Goel et al. 2024).

The quality of the reference genome plays a crucial role in the detection of somatic mutations. Even in *A. thaliana*, using a HiFi based assembly improves alignment accuracy compared to the widely used TAIR10 reference genome (Monroe et al. 2023). Unlike *A. thaliana*, which is highly inbred, most tree species are outcrossing and thus exhibit high levels of heterozygosity (Schmid-Siegert et al. 2017). This poses a considerable challenge for sequence alignment: using a haploid reference genome may fail to capture the full structural diversity, leading to unmapped regions; while using a diploid reference genome may introduce ambiguous alignments in homozygous regions.

Therefore, to reliably investigate somatic mutations in trees, it is essential to first construct a high-quality reference genome. This should be followed by the design of a robust alignment strategy, and finally, the implementation of a mutation filtering pipeline that balances sensitivity and specificity. Only by addressing these challenges can somatic mutations be accurately identified in highly heterozygous tree genomes.

4. Aims and objectives

The main aim of my thesis is to advance our understanding of genomic variants in plants, both in the nuclear and organellar genomes. I address two key questions: how to more accurately assemble and analyze organellar genomes, and how to detect somatic mutations in species with high heterozygous genomes.

In the first results chapter, I describe the development of a tool for assembling organellar genomes—both chloroplast and mitochondrial—using long-read data alone, dispensing with the need for a reference genome or information from related species. Existing tools often struggle in this context, especially with the more complex mitochondrial genomes in plants. My goal was to establish a user-friendly, broadly applicable method that still produces complete, high-quality assemblies. Beyond assembly, I also explored the evolution of organellar DNA fragments that have moved into the nuclear genome (NUPTs and NUMTs), investigating whether their sequence changes can yield insights into mutational processes or epigenetic regulation.

In the next chapter, I extend this analysis to the population level. Using *A. thaliana* as a model, I assembled organellar genomes for over a hundred natural accessions to examine structural diversity across accessions, especially in mitochondrial genomes, which are known to be more variable. I explored the role of repeat elements in shaping this diversity, the geographical distribution of particular structural types, and the presence of large structural variants. In addition, I also reassessed the coding potential of these genomes by searching for previously unannotated open reading frames and testing for their expression.

The third project of my thesis, described in the final chapter of the Results, shifts to the source of genetic variants, somatic mutations, specifically in a centuries-old oak tree. Mutation detection in highly heterozygous genomes is challenging, especially when using short-read data. My goal was to determine the extent to which the choice of reference (haploid vs. diploid) influences variant calling, and to explore methods to improve sensitivity without compromising accuracy. Ultimately, the aim was to gain a more accurate estimate of somatic mutation rates in long-lived species such as oak.

Together, these chapters deepen our understanding of how genetic variants are generated, how they evolve, and how they are maintained in plants.

Chapter One

TIPPo: A User-Friendly Tool for De Novo Assembly of Organellar Genomes with High-Fidelity Data

Content of this chapter is published as:

Xian, W., Bezrukov, I., Bao, Z., Vorbrugg, S., Gautam, A., & Weigel, D. (2025). *Molecular biology and evolution*, 42(1), msae247. https://doi.org/10.1093/molbev/msae247

See Thesis Appendix I

Abstract

Plant cells have two major organelles with their own genomes: chloroplasts and mitochondria. While chloroplast genomes tend to be structurally conserved, the mitochondrial genomes of plants, which are much larger than those of animals, are characterized by complex structural variation. We introduce TIPPo, a user-friendly, reference-free assembly tool that uses PacBio high-fidelity long-read data and that does not rely on genomes from related species or nuclear genome information for the assembly of organellar genomes. TIPPo employs a deep learning model for initial read classification and leverages k-mer counting for further refinement, significantly reducing the impact of nuclear insertions of organellar DNA on the assembly process. We used TIPPo to completely assemble a set of 54 complete chloroplast genomes. No other tool was able to completely assemble this set. TIPPo is comparable with PMAT in assembling mitochondrial genomes from most species but does achieve even higher completeness for several species. We also used the assembled organelle genomes to identify instances of nuclear plastid DNA (NUPTs) and nuclear mitochondrial DNA (NUMTs) insertions. The cumulative length of NUPTs/NUMTs positively correlates with the size of the nuclear genome, suggesting that insertions occur stochastically. NUPTs/NUMTs show predominantly C:G to T:A changes, with the mutated cytosines typically found in CG and CHG contexts,

suggesting that degradation of NUPT and NUMT sequences is driven by the known elevated mutation rate of methylated cytosines. Small interfering RNA loci are enriched in NUPTs and NUMTs, consistent with the RdDM pathway mediating DNA methylation in these sequences.

Contribution

W.X. designed the project, conducted the analyses, and wrote the first draft of the manuscript. I.B. set up the computational environment. I.B., Z.B., S.V., and A.G. tested the tool. D.W. supervised the project. W.X. and D.W. prepared the final manuscript with inputs from all authors.

Author	**Author position**	**Scientific ideas**	**Data generation**	**Analysis & interpretation**	**Paper writing**
Wenfei Xian	1st	75%	100%	72%	75%
Ilja Bezrukov	2nd	5%	-	2%	5%
Zhigui Bao	3rd	-	-	2%	-
Sebastian Vorbrugg	4th	-	-	2%	-
Anupam Gautam	5th	-	-	2%	-
Detlef Weigel	6th	20%	-	20%	20%
Status in publication process:		Paper published in *Molecular Biology and Evolution* on January 13, 2025			

Chapter Two

Organellar Pangenomes of *Arabidopsis thaliana*

Content of this chapter is published as:

Xian W., Bao Z., Vorbrugg S., Tao Y., Movilli A., Bezrukov I., Weigel D.,
bioRxiv 2025.01.11.632530; doi: https://doi.org/10.1101/2025.01.11.632530

See Thesis Appendix II

Abstract

Chloroplasts and mitochondria are the primary sites for photosynthesis and respiration, each harboring its own unique genome. Although the organellar genomes are considerably smaller compared to the nuclear genome, they are nonetheless essential for survival of the organism. A common feature of many chloroplast and mitochondrial genomes is the presence of large repeated sequences longer than 1 kb. These can be either in inverted or direct orientation, and recombination between them leads to structural heteroplasmy. To understand the intraspecific evolution of organellar genomes, we assembled chloroplast and mitochondrial genomes of 143 Arabidopsis thaliana accessions from PacBio HiFi sequencing data. We find large repeats to be associated with heteroplasmy and structural variation. Our extensive genome annotation is supported by the analysis of expression of many open reading frames (ORFs). Among newly identified ORFs, at least one appears to have been acquired via horizontal gene transfer, indicating alternative mechanisms for the diversification of plant organelles. The assembled and annotated organellar genomes constitute a rich source for future functional studies of the interaction between the three genomes of a plant.

Contribution

W.X. and D.W. designed the project. D.W. supervised the project. W.X., Z.B., S.V., Y.T., A.M. and I.B. conducted the analyses. W.X. wrote the first draft of the manuscript. W.X. and D.W. prepared the final manuscript with inputs from all authors.

Author	**Author position**	**Scientific ideas**	**Data generation**	**Analysis & interpretation**	**Paper writing**
Wenfei Xian	1st	75%	100%	70%	75%
Zhigui Bao	2nd	-	-	2%	-
Sebastian Vorbrugg	3rd	-	-	2%	-
Yueqi Tao	4th	-	-	2%	-
Andrea Movilli	5th	-	-	2%	-
Ilja Bezrukov	6th	5%	-	2%	5%
Detlef Weigel	7th	20%	-	20%	20%
Status in publication process:		Paper posted to *bioRxiv* on Jan 14, 2025			

Chapter Three

Minimizing detection bias of somatic mutations in a highly heterozygous oak genome

Content of this chapter is published as:
Xian W., Carbonell P., Rabanal F., Bezrukov I., Reymond P., Weigel D.
bioRxiv 2025.02.13.638107; doi: https://doi.org/10.1101/2025.02.13.638107
See Thesis Appendix III

Abstract

Somatic mutations are particularly relevant for long-lived organisms. Sources of somatic mutations include imperfect DNA repair, replication errors, and exogenous damage such as ultraviolet radiation. A previous study estimated a surprisingly low number of somatic mutations in a 234-year-old individual of the pedunculate oak (Quercus robur), known as the Napoleon Oak. It has been suggested that the true number of somatic mutations was underestimated due to gaps in the reference genome and too conservative filtering of potential mutations. We therefore generated new high-fidelity long-read data for the Napoleon Oak (n = 12) to produce both a pseudo-haploid genome assembly and a partially phased diploid assembly. The high heterozygosity allowed for complete reconstruction of phased and gapless centromeres for 22 of the 24 chromosomes. On the other hand, the high heterozygosity posed challenges for short-read alignments. Use of only the pseudo-haploid assembly as a reference led to potential misalignments, while use of only the diploid assembly reduced variant detection sensitivity. Since most somatic mutations are layer-specific, the fraction of reads covering a specific somatic mutation is expected to be relatively low, even where all cells in a single layer contain a specific mutation. To address this challenge, we employed a read assignment strategy, selecting the appropriate reference sequence (pseudo-haploid or diploid) based on alignment score and mapping quality. Ultimately, we identified 198 high-confidence somatic mutations, compared with 17 somatic mutations

identified before with the same set of short reads. Our approach thus increased the total estimated annual mutation rate by a factor of five.

Contribution

D.W. designed and supervised the project. P.C-B., F.A.R and P.R. prepared the leaf sample. P.C-B. and W.X. prepared the DNA library. W.X., P.C-B, and F.A.R analyzed the data. W.X., I.B. and D.W. prepared the final manuscript with inputs from all authors.

Author	**Author position**	**Scientific ideas**	**Data generation**	**Analysis & interpretation**	**Paper writing**
Wenfei Xian	1st	60%	60%	72%	75%
Pablo Carbonell-Bejerano	2nd	10%	25%	2%	-
Fernando A. Rabanal	3rd	-	5%	2%	-
Ilja Bezrukov	4th	10%	-	2%	5%
Philippe Reymond	5th	-	10%	2%	-
Detlef Weigel	6th	20%	-	20%	20%
Status in publication process:		Paper published in *G3 Genes*|*Genomes*|*Genetics* on June 21, 2025			

Discussion

This thesis was motivated by my broad interest in genetic changes, both how they arise and how they are maintained under selection. While most population genomics studies in plants focus on variation in the nuclear genome, I was particularly interested in the rest of the genome that is often overlooked, such as organellar DNA. I was also drawn to the broader question of how new variants originate at the molecular level. In that context, somatic mutations represent a direct and detectable source of genetic variation, especially in long-lived plants, and provide an opportunity to study how changes accumulate during an individual's lifetime.

Through three projects, I explored several aspects of genome dynamics in plants, developing a tool for assembling organellar genomes using long-read data, applying it to a population-scale analysis of *A. thaliana* to examine organellar diversity, and refining somatic mutation detection in highly heterozygous oak genomes. Together, these studies investigate how genomes change, through insertion, rearrangement, and mutation, and how these changes are shaped by biological context and evolutionary forces.

By approaching these questions from both a technical and biological perspective, my aim was to gain a more complete picture of how different parts of the genome evolve, and how genetic variation is generated, filtered, and, in some cases, maintained over time.

1. Development and application of a tool for organellar genome assembly

In the first chapter, I developed TIPPo, a reference-free tool for assembling plant organellar genomes using PacBio HiFi data. Using this tool, I assembled complete chloroplast genomes and highly contiguous mitochondrial genomes in multiple plant species. These assemblies are necessary for downstream analyses, including the identification and characterization of nuclear-integrated organellar fragments (NUPTs and NUMTs) and the substitution pattern after integration, which is impossible without the complete organellar genomes.

1.1 Comparison of TIPPo with other organellar assembly tools

Normal genome assemblers are optimized for input sequences with moderate sequencing depth (Cheng et al. 2021; Rautiainen et al. 2023). However, organellar genomes are high copy numbers in the cells compared to the nuclear genome, which can lead to incomplete or fragmented assemblies when using standard tools (Igolkina et al. 2024b). Therefore, separating organellar reads from nuclear reads is critical. There are two main strategies to do this: (1) directly classifying organellar reads before assembly, and (2) assembling the entire dataset first, then extracting organellar sequences.

Existing tools such as GetOrganelle (Jin et al. 2020), NOVOPlasty (Dierckxsens, Mardulyn, and Smits 2017) , and Organellar_PBA (Soorni et al. 2017) follow the first strategy by separating the organellar reads based on alignment to reference organellar genomes. While this approach works well for chloroplast genomes, which are relatively conserved and well-represented in the publicly available database, it is less effective for mitochondrial genomes due to their limited availability mitochondrial genomes and diversity among the species.

The second strategy, used in tools such as PMAT (Bi et al. 2024) and Oatk (Zhou et al. 2024), involves assembling the entire genome first and then identifying organellar contigs using a set of conserved genes. However, this approach also has limitations. Some species contain multiple mitochondrial chromosomes that lack genes, making gene-based detection ineffective. In addition, assembling the entire dataset is computationally intensive. To reduce computational time, PMAT needs to downsample the input to ultra low sequencing depth. Although organellar reads may present at relatively moderate coverage, such subsampling introduces the risk of bias, potentially excluding certain organellar regions. Increasing the depth to avoid this issue, but leads to very slow runtimes. Oatk developed an assembler using sparse k-mer graphs to reduce the time, but it still uses genes as bait.

TIPPo addresses these challenges by directly classifying reads at the read level using an existing deep learning model (Karlicki, Antonowicz, and Karnkowska 2022), because of the high accuracy of HiFi reads. This read-level approach does not rely on conserved genes or reference genomes, allowing for more complete

assembly of non-coding regions in mitochondrial genomes, especially those with multiple chromosomes.

Moreover, existing tools often ignore the presence of NUPTs and NUMTs, resulting in assemblies contaminated with nuclear-derived sequences. TIPPo identifies and removes these sequences by using a key feature: since NUPTs and NUMTs are located in the nuclear genome, their coverage is lower than true organellar sequences. By analyzing kmer count profiles, TIPPo filters out nuclear-derived reads, resulting in cleaner organellar assemblies.

In addition, mitochondrial genomes contain long repeat sequences that mediate recombination. TIPPo uses the “repeat” module in Flye to represent the complexity. While it is valuable for organelle specialists, it can be overwhelming for general users. For instance, PMAT only outputs assembly graphs in GFA format, which are difficult to interpret without specialized knowledge. To address this, TIPPo also provides simplified linear FASTA outputs to adapt a wider range of users with varying levels for the organellar genome.

1.2 The fate of NUMT/NUPT

Identification of nuclear-integrated organellar DNA (NUPTs and NUMTs) remains challenging, due to the inconsistent data in previous studies (Huang et al. 2005; Zhang et al. 2020). The nuclear genome assemblies and organellar genomes belong to different accessions. This introduces a potential source of misinterpretation: observed mismatches between the two assemblies might reflect inter-accession polymorphisms or laboratory-specific artifacts, rather than true post-integration substitutions. For example, the first *A. thaliana* mitochondrial genome was sequenced from the C24 (Unseld et al. 1997), whereas the nuclear reference genome was derived from Col-0 (Arabidopsis Genome Initiative 2000).

To overcome this limitation, our study used both the nuclear and organellar genomes from the same sample. This one-sample dataset ensured that any mismatches identified between nuclear-integrated fragments and the organellar reference were not artifacts, but instead reflect true substitution events that occurred

after insertion. This strategy improves the reliability of NUMT and NUPT substitution analyses.

I observed that NUMTs and NUPTs tend to accumulate C:G > T:A transitions, suggesting cytosine deamination is the major mechanism (Kreutzer and Essigmann 1998). Similar patterns have been reported in Arabidopsis mutation accumulation (MA) lines by Weng et al., where C:G > T:A transitions also occur at high frequency. In contrast, population-level polymorphism data, where mutations are subject to natural selection, show a lower proportion of C:G > T:A variants. This suggests that while mutations accumulate in NUMT/NUPT over time due to spontaneous processes, many of those mutations are likely neutral and not strongly affected by natural selection.

Interestingly, I also detected siRNA enriched in NUMTs and NUPTs, suggesting that these foreign sequences are recognized and targeted by host genome defense mechanisms. Similar phenomena have been reported in the context of transposable element silencing, where small RNAs play a key role in post-transcriptional regulation and heterochromatin formation (X. Wang, Weigel, and Smith 2013). Our findings raise the possibility that nuclear-integrated organellar DNA is subject to similar silencing processes, reflecting a broader host strategy for managing exogenous DNA.

1.3 Broadening the application of TIPPo and future directions

One direction for improving TIPPo is the use of methylation signals in the PacBio HiFi reads to the read classification process. Because *A. thaliana* Col-0 mitochondrial genomes lack endogenous methylation (Zhong et al. 2025), whereas nuclear-integrated organellar fragments often exhibit methylation, particularly in CG and CHG contexts (Fields et al. 2022), the methylation profile could be a useful feature between genuine mitochondrial reads and NUMT/NUPT. This approach would be especially valuable in samples with a lot of recently inserted organelle-to-nucleus DNA.

Beyond plant systems, TIPPo has the potential to be extended to non-plant data, including animals, fungi, and protists. While animal mitochondrial genomes are traditionally considered compact and circular, recent studies have challenged this

viewpoint. Such as, snails have been reported to possess recombining mitochondrial structures (Sharbrough et al. 2023), which may require the assembler to represent the diversity of genome structure. Applying TIPPo to these cases could reveal hidden complexity existing assumptions about organellar genome evolution. In metagenomic or environmental samples, where multiple organellar genomes may coexist, the ability of TIPPo to distinguish organellar components without reference genomes could facilitate microbiome profiling (Qing et al. 2024).

2. Population diversity and structural dynamics of organellar genomes

In the second chapter, I assembled the chloroplast and mitochondrial genomes of 143 *A. thaliana* accessions using PacBio HiFi data. These assemblies revealed organellar genomic diversity at the population level. Compared to chloroplast genomes, mitochondrial genomes showed higher levels of structural variation, driven primarily by large repeat contents. I also annotated open reading frames (ORFs) in all the organellar assemblies and identified several cluster-specific ORFs not present in the Col-0 reference. Using publicly available RNA-seq data, I found some of these ORFs appeared to be transcribed. And two ORFs were likely transferred from other Brassicaceae species.

2.1 Organellar genetic diversity revealed by assembly alignment

Before the widespread use of long-read data, studies of genetic diversity in both nuclear and organellar genomes primarily relied on mapping short reads to a reference genome (1001 Genomes Consortium 2016). However, single nucleotide polymorphisms only capture a small part of genetic diversity (Igolkina et al. 2024b). While linkage between variants makes SNPs useful for identifying causal loci (Uffelmann et al. 2021). As long-read sequencing became more available, structural variants in nuclear genomes have been increasingly explored, but organellar genomes, particularly mitochondria, remain uncharacterized well.

By assembling the organellar genomes of 143 *A. thaliana* accessions, I found that the two large repeat pairs present in the mitochondrial reference genomes (Col-0 and C24) are not conserved (Unseld et al. 1997). Based on the number of

large repeats, the accessions could be broadly divided into two groups: one group retains both pairs of large repeats, consistent with the reference genomes; the other group contains only one pair. Interestingly, accessions sampled from the Yangtze River basin in China almost all contain both large repeat pairs, whereas in Central Asian accessions, one large repeat pair is more common. However, it remains unclear whether this pattern reflects local adaptation or simply historical divergence (Budar and Roux 2011).

Previous experimental studies have shown that knocking out one copy of a large repeat in the chloroplast genome does not have obvious phenotypic changes, suggesting that natural variation in repeat copy number may be tolerated without severe consequences (C. Krämer et al. 2024). In this experiment, plants responded to the deletion by increasing the copy number of chloroplast DNA. I was curious whether accessions with only one repeat copy might also show a higher copy number of mitochondrial genome. However, since datasets came from different labs, I observed that the organellar genome copy number was more correlated with the lab than with repeat copy number, limiting the ability to draw conclusions.

2.2 Expression potential beyond conserved genes in organellar genomes

After chloroplasts and mitochondria evolved from ancient bacterial endosymbionts, they became specialized organelles responsible for energy conversion in photosynthetic eukaryotes cells. Their genomes still retain key components of energy metabolism, such as genes involved in photosynthesis and respiration (J. Wang et al. 2024). Since chloroplasts and mitochondria retain their own translation systems, their genomes still encode a subset of rRNAs and tRNAs necessary for protein translation within the organelle. As a result, current genome annotations of organelles have largely focused on these well-known, conserved genes. It remains an open question whether other protein-coding regions exist beyond these canonical genes.

A recent study used ribosome profiling (Ribo-seq) to capture ribosome-protected mRNA fragments, revealing widespread translation signals in intergenic regions, more than half of which were located in organellar genomes

(H.-Y. L. Wu et al. 2024). This unexpected enrichment of ribosome footprints in non-coding regions suggests that the organellar genomes of *A. thaliana* may harbor additional protein-coding potential beyond the annotated set of conserved genes.

Consistent with the ribosome profiling results, I found that some of the non-conserved ORFs indeed show evidence of transcription, even under a stringent expression threshold. While our study does not yet confirm whether these transcripts are functional or simply represent transcriptional noise (Ng et al. 2022), it nonetheless provides new insight into the uncharacterized coding potential of plant organellar genomes. These results highlight the need for further investigation and suggest that the functional landscape of organelles is broader than previously appreciated.

2.3 HGT contributes to mitochondrial genome innovation

In our organellar ORF clustering analysis, I identified two orthogroups (OGs) that were specific to subsets of *A. thaliana* accessions. According to sequence similarity , I found both OGs probably came from other Brassicaceae species through horizontal gene transfer (HGT). This suggests that genetic diversity in the *A. thaliana* mitochondrial genome came not only from its own mutations and rearrangements, but also from genes transferred from other species.

HGT had been reported in plant mitochondrial genomes, particularly in parasitic species where physical contact helps DNA transfer (Garcia et al. 2021; Richardson and Palmer 2007). But finding HGT in *A. thaliana*, which is not a parasite, means it might be more common than we thought. These observations align with growing evidence indicating that plant mitochondria are unusually permissive to foreign DNA, often acquiring and retaining non-native sequences over evolutionary timescales (Bergthorsson et al. 2004).

Closer check of the insertion sites revealed short repeat sequences flanking at least one of the horizontally acquired sequences. This suggests that integration may have been mediated by microhomology-mediated repair pathways, a potential mechanism of structural rearrangement and foreign DNA incorporation in plant mitochondria (Roulet et al. 2024). Repeat sequences might help repair broken DNA and, in the process, let outside DNA slip into the mitochondrial genome.

Notably, one of the HGT-derived ORFs was annotated as orf117, a gene previously implicated in cytoplasmic male sterility (CMS) (Dehaene et al. 2024). CMS is a mitochondrially encoded trait that can affect reproductive success and has practical importance in hybrid breeding systems. The acquisition of a CMS-related gene via HGT raises the possibility that these transferred sequences may have functional consequences and are not merely neutral genomic passengers.

Together, our findings point to HGT as a significant and underexplored contributor to mitochondrial genome innovation in *A. thaliana*. This gives us a new way to think about how mitochondrial genes evolve. Other species might actually contribute to shaping. To figure out whether these HGT events are just random or might actually help plants adapt, we'll need more studies, like comparing mitochondrial genomes, gene expression and wet lab experiments.

3 Detecting somatic mutations in a long-lived plant

In the final project of my thesis, I assembled a high-quality genome for the Napoleon Oak (*Quercus robur*) using PacBio HiFi data and developed a strategy to detect somatic mutations in a highly heterozygous genome. Previous studies had estimated a low mutation rate in the same tree, but we suspected that technical limitations, including the quality of reference genome and mutation calling strategies, might have led to underestimation. To improve detection sensitivity without sacrificing specificity, I created both a pseudo-haploid and a diploid genome assembly and developed a hybrid alignment approach that assigns short reads to the appropriate reference based on alignment score and mapping quality.

3.1 Correcting somatic mutation detection bias in highly heterozygous genomes

When we use short-read data for mutation detection, read alignment is a critical step. While most previous work has focused on the quality and filtering of the reads, the choice of reference genome is often overlooked, especially in high heterozygous trees (Schmitt et al. 2024; Duan et al. 2022). Under such conditions, using a collapsed assembly that represents only one haplotype cannot fully represent the diploid nature of the genome. Because the sequencing reads came from both

haplotypes, reads originating from highly divergent regions in the unrepresented haplotype often fail to align correctly, leading to missed mutations (Goel et al. 2019; Jaegle et al. 2023).

We can assemble the complete diploid genome using HiFi data. The assembler tools will produce two phases as long as at least a single germline SNP is present within the long read (~20 kb) (Cheng et al. 2021; Rautiainen et al. 2023; Bankevich et al. 2022). This means that long homozygous regions are present twice, in both haplotypes, causing ambiguous alignment in these identical regions. It results in mapping quality of zero, which negatively affects the detection of somatic mutations. This issue was also noted in a recent study of apricot using a trio-based diploid assembly (Goel et al. 2024). Despite generating a high-quality phased genome, the researchers still used only one haplotype as the reference during variant calling, thus returning to the fundamental limitation of single-reference representation.

Our approach combines the use of diploid and haploid assembly for a better alignment, which substantially improves alignment accuracy and enhances the sensitivity of mutation detection. We believe this approach holds broad applicability for the analysis of somatic mutations in other highly heterozygous species, such as grapevine and citrus (Calderón et al. 2024; G. A. Wu et al. 2018).

Further improvements in mutation detection could come from the adoption of graph-based genome representations (Schneeberger et al. 2009). Unlike linear references, genome graphs can incorporate sequence diversity across multiple haplotypes or individuals, allowing for more accurate read alignment in regions with high allelic divergence (Garrison et al. 2018, 2024). This has been proved to mitigate the mapping bias resulting from the limitations of a single reference and improve variant calling sensitivity, especially in genetically diverse regions and ancient DNA samples (Sirén et al. 2021; Martiniano et al. 2020).

Long reads may offer unique advantages in detecting complex mutation types, such as SVs and short tandem repeat (STR) expansions or contractions, which are often difficult to resolve using short-read technologies due to their limited read length and challenges in spanning repetitive or structurally complex regions (Willems et al. 2017; Readman et al. 2021). Notably, the mutation rate of STRs is

nearly three orders of magnitude higher than that of SNVs, resulting often in multiple alleles segregating in the population (Fan and Chu 2007; Steely et al. 2022).

The integration of STR analysis with graph-based genome representations improves genotyping accuracy and detecting copy number variation in multiple assemblies. First, graph-based references can encode multiple STR alleles at the same locus, providing alternative paths for read alignment and thereby reducing mapping bias and improving both sensitivity and specificity. Second, pan genome graph facilitates the direct approach of repeat copy number variation. In ideal cases, where a repeat motif is represented in a single node but duplicated multiple times. It has already been successfully applied to characterise copy number variation in the amylase gene cluster in humans (Bolognini et al. 2024).

However, current genome graph construction methods often do not consider individual STR units, but instead include the entire arrays as nodes, because the repeat units are typically too short to be explicitly represented within the graph. This design limitation reduces the resolution of STR-specific variation and constrains their accurate representation in graph-based frameworks. A promising solution, as implemented in tools like ExpansionHunter, is to incorporate short single-node loops that can model tandem repeat expansions within sequence graphs (Dolzhenko et al. 2019). This approach enables more accurate representation of STR variation and facilitates their integration into graph-based variant calling workflows.

DeepVariant has begun to incorporate graph-based information with the aim to improve alignment accuracy, but it still relies on projecting (“surjecting”) reads back onto a linear reference for downstream genotyping. For example, DeepVariant extracts pileup images from a linearized coordinate system before applying a convolutional neural network model to classify variants (Asri et al. 2025). This limits the ability to fully exploit the structural diversity encoded in the genome graph, especially in highly divergent regions.

Taken together, these developments underscore the need for fully integrated frameworks that use graph-based genome representations not only as references, but also as a foundation for the entire mutation detection pipeline including alignment, variant calling, and genotyping.

3.2 The detection of low-frequency mutations relies on high base quality

Low-frequency mutations are often filtered during variant calling to reduce false positives (Schmid-Siegert et al. 2017). However, this strategy may increase the risk of false negatives, particularly in somatic mutation detection. Recent high-throughput studies have demonstrated that many somatic mutations in plants are layer-specific. They occur in only a subset of the meristematic layers (such as L1, L2, or L3) (Goel et al. 2024; Sun et al. 2024). As a result, these mutations may be present at low variant allele frequencies (VAFs) in bulk tissue samples. Such low-frequency mutations can still be biologically meaningful and heritable, especially if they originate from layers (L2) that contribute to gametogenesis (Schmitt et al. 2024).

The existence of genuine low-frequency somatic mutations shows not only the importance of accurate read alignment but also the need for high base accuracy of sequencing data. One existing approach to improve bases accuracy is duplex sequencing (Kennedy et al. 2014), which uses unique molecular identifiers (UMIs) to mark the same DNA molecule through multiple rounds of amplification. While effective, this method is limited to targeted sequencing regions (Kennedy et al. 2014; Waneka et al. 2024).

In contrast, PacBio HiFi sequencing offers an alternative for accurate genome-wide detection of low-frequency mutations. HiFi reads are generated by circular consensus sequencing, in which both strands of a single DNA molecule are sequenced multiple times (Wenger et al. 2019). This also avoids strand bias. Recent human studies have shown that as few as five passes per strand are sufficient to achieve the base accuracy needed for confident detection of ultra-rare mutations (M. H. Liu et al. 2024).

3.3 Somatic mutation spectrum and implications for mutational mechanisms

Consistent with findings in other species, the somatic mutation spectrum in oak is dominated by C:G > T:A transitions (Cagan et al. 2022). This pattern has often been estimated to be the result of spontaneous deamination of methylated cytosines, as

suggested in early mutation accumulation (MA) study in *A. thaliana* (Ossowski et al. 2010), which showed mutated cytosine having higher methylation levels. However, in CpG contexts, transitions at G:C sites without known methylation do not occur more frequently than expected by chance, suggesting that 5-methylcytosine deamination is not the only driver of C:G > T:A transitions.

Yeast MA lines studies have also reported a predominance of C:G > T:A transitions, despite yeast genomes being unmethylated (Haoxuan Liu and Zhang 2021; Serero et al. 2014). Cytosines in single-stranded DNA are over 100 times more likely to deaminate than in double-stranded DNA, which likely explains the pattern (Beletskii and Bhagwat 1996).

This insight brings attention to the possible involvement of ssDNA-exposed regions, such as transcripted regions, R-loops and G-quadruplexes (Wulfridge and Sarma 2024). While transcribed regions benefit from transcription-coupled repair (TCR) on the template strand, the non-template strand lacks protection (Selby et al. 2023). Nonetheless, gene bodies are likely under strong evolutionary constraints, and may have other mechanisms to suppress mutations (Monroe et al. 2022; Quiroz et al. 2024). In contrast, non-coding regions, including those prone to forming secondary structures like G4 and R-loops, may be more vulnerable. Indeed, a study has reported elevated germline mutation rates in these regions (Guiblet et al. 2018), but their role in somatic mutagenesis remains largely unexplored.

Finally, replication errors may be another possible contributor to the observed mutation spectrum. Early studies noted that C:G > T:A mutations frequently occur at dipyrimidine sites, which are prone to damage by UV light and undergo imperfect repair, leading to mutations (Ossowski et al. 2010; Serero et al. 2014; D. Wang, Kreutzer, and Essigmann 1998). More direct evidence came from a recent work in humans, which proved DNA polymerase ε is a key driver of C:G > T:A mutations, further highlighting the complex interplay among multiple mutational processes (Tomkova et al. 2024). Together, these observations suggest that the prevalence of C:G > T:A transitions among oak somatic mutations may result from a combination of cytosine deamination and replication-associated errors, rather than being solely attributable to methylated cytosine instability.

Conclusions and future outlook

Taken together, my projects converge on two intersecting themes: the incorporation of new DNA into plant genomes, and the variation spectrum at different time scales. The first encompasses processes such as the nuclear integration of organellar DNA and the presence of horizontally transferred genes in mitochondrial genomes, two natural mechanisms through which new sequences become embedded in the genome. The second theme explores how sequence changes arise and accumulate, either through substitutions that occur after integration, or through somatic mutations within individual organisms. In both cases, I observed consistent patterns in mutation spectra, particularly C>T transitions, pointing to shared mutational processes shaping genome evolution.

These findings also raise broader questions about how plant genomes respond to genetic novelty, whether that novelty arises spontaneously or is introduced artificially. This is particularly relevant in the context of genetically modified organisms (GMOs), where new DNA sequences are inserted intentionally for specific purposes. While tools like CRISPR-Cas9 now allow for precise and minimal edits, many applications still rely on knock-ins or transgenes, which involve the integration of foreign DNA. Understanding how such sequences behave over time, whether they mutate, persist, or are eventually silenced, will be essential for both fundamental biology and the future of responsible biotechnology. These are questions I hope to continue exploring in my future research.

References

1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65.

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

1001 Genomes Consortium. 2016. “1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis Thaliana.” *Cell* 166 (2): 481–91.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Alonso-Blanco, Carlos C., Haim Ashkenazy, Pierre Baduel, Zhigui Bao, Claude Becker, Erwann Caillieux, Vincent Colot, et al. 2024. “The 1001G+ Project: A Curated Collection of Arabidopsis Thaliana Long-Read Genome Assemblies to Advance Plant Research.” *Genomics*. bioRxiv. https://www.biorxiv.org/content/10.1101/2024.12.23.629943v1.

Anderson, Craig J., Lana Talmane, Juliet Luft, John Connelly, Michael D. Nicholson, Jan C. Verburg, Oriol Pich, et al. 2024. “Strand-Resolved Mutagenicity of DNA Damage and Repair.” *Nature* 630 (8017): 744–51.

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

Archibald, John M. 2015. “Genomic Perspectives on the Birth and Spread of Plastids.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (33): 10147–53.

Asri, Mobin, Pi-Chuan Chang, Juan Carlos Mier, Jouni Sirén, Parsa Eskandar, Alexey Kolesnikov, Daniel E. Cook, et al. 2025. “Pangenome-Aware DeepVariant.” *bioRxiv*. https://doi.org/10.1101/2025.06.05.657102.

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. 2010. “Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines.” *Nature* 465 (7298): 627–31.

Bankevich, Anton, Andrey V. Bzikadze, Mikhail Kolmogorov, Dmitry Antipov, and Pavel A. Pevzner. 2022. “Multiplex de Bruijn Graphs Enable Genome Assembly from Long, High-Fidelity Reads.” *Nature Biotechnology* 40 (7): 1075–81.

Beck, James B., Heike Schmuths, and Barbara A. Schaal. 2008. “Native Range Genetic Variation in Arabidopsis Thaliana Is Strongly Geographically Structured and Reflects Pleistocene Glacial Dynamics: ARABIDOPSIS THALIANA GENETIC STRUCTURE.” *Molecular Ecology* 17 (3): 902–15.

Beletskii, A., and A. S. Bhagwat. 1996. “Transcription-Induced Mutations: Increase in C to T Mutations in the Nontranscribed Strand during Transcription in Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (24): 13919–24.

Benjamin, David, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee

Lichtenstein. 2019. “Calling Somatic SNVs and Indels with Mutect2.” *bioRxiv*. bioRxiv. https://doi.org/10.1101/861054.
Bergthorsson, Ulfar, Aaron O. Richardson, Gregory J. Young, Leslie R. Goertzen, and Jeffrey D. Palmer. 2004. “Massive Horizontal Transfer of Mitochondrial Genes from Diverse Land Plant Donors to the Basal Angiosperm Amborella.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (51): 17747–52.
Bhagwat, Ashok S., Weilong Hao, Jesse P. Townes, Heewook Lee, Haixu Tang, and Patricia L. Foster. 2016. “Strand-Biased Cytosine Deamination at the Replication Fork Causes Cytosine to Thymine Mutations in Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (8): 2176–81.
Bi, Changwei, Fei Shen, Fuchuan Han, Yanshu Qu, Jing Hou, Kewang Xu, Li-An Xu, Wenchuang He, Zhiqiang Wu, and Tongming Yin. 2024. “PMAT: An Efficient Plant Mitogenome Assembly Toolkit Using Low-Coverage HiFi Sequencing Data.” *Horticulture Research* 11 (3): uhae023.
Bohra, Abhishek, Uday C. Jha, Premkumar Adhimoolam, Deepak Bisht, and Narendra P. Singh. 2016. “Cytoplasmic Male Sterility (CMS) in Hybrid Breeding in Field Crops.” *Plant Cell Reports* 35 (5): 967–93.
Bolognini, Davide, Alma Halgren, Runyang Nicolas Lou, Alessandro Raveane, Joana L. Rocha, Andrea Guarracino, Nicole Soranzo, Chen-Shan Chin, Erik Garrison, and Peter H. Sudmant. 2024. “Recurrent Evolution and Selection Shape Structural Diversity at the Amylase Locus.” *Nature* 634 (8034): 617–25.
Brownfield, Lynette. 2021. “Plant Breeding: Revealing the Secrets of Cytoplasmic Male Sterility in Wheat.” *Current Biology: CB* 31 (11): R724–26.
Budar, Françoise, and Fabrice Roux. 2011. “The Role of Organelle Genomes in Plant Adaptation: Time to Get to Work!” *Plant Signaling & Behavior* 6 (5): 635–39.
Butenko, Anzhelika, Julius Lukeš, Dave Speijer, and Jeremy G. Wideman. 2024. “Mitochondrial Genomes Revisited: Why Do Different Lineages Retain Different Genes?” *BMC Biology* 22 (1): 15.
Cagan, Alex, Adrian Baez-Ortega, Natalia Brzozowska, Federico Abascal, Tim H. H. Coorens, Mathijs A. Sanders, Andrew R. J. Lawson, et al. 2022. “Somatic Mutation Rates Scale with Lifespan across Mammals.” *Nature* 604 (7906): 517–24.
Calderón, Luciano, Pablo Carbonell-Bejerano, Claudio Muñoz, Laura Bree, Cristobal Sola, Daniel Bergamin, Walter Tulle, et al. 2024. “Diploid Genome Assembly of the Malbec Grapevine Cultivar Enables Haplotype-Aware Analysis of Transcriptomic Differences Underlying Clonal Phenotypic Variation.” *Horticulture Research* 11 (5): uhae080.
Chaudhry, Smita, and Gagan Preet Singh Sidhu. 2022. “Climate Change Regulated Abiotic Stress Mechanisms in Plants: A Comprehensive Review.” *Plant Cell Reports* 41 (1): 1–31.
Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. “Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm.” *Nature Methods* 18 (2): 170–75.
Chuong, Julie N., Nadav Ben Nun, Ina Suresh, Julia Cano Matthews, Titir De, Grace Avecilla, Farah Abdul-Rahman, Nathan Brandt, Yoav Ram, and David Gresham. 2024. “Template Switching during DNA Replication Is a Prevalent Source of Adaptive Gene Amplification.” *bioRxivorg*.

https://doi.org/10.1101/2024.05.03.589936.
Clark, Richard M., Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, et al. 2007. “Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis Thaliana.” *Science (New York, N.Y.)* 317 (5836): 338–42.
Colomé-Tatché, Maria, Sandra Cortijo, René Wardenaar, Lionel Morgado, Benoit Lahouze, Alexis Sarazin, Mathilde Etcheverry, et al. 2012. “Features of the Arabidopsis Recombination Landscape Resulting from the Combined Loss of Sequence Variation and DNA Methylation.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (40): 16240–45.
Dehaene, Noémie, Clément Boussardon, Philippe Andrey, Delphine Charif, Dennis Brandt, Clémence Gilouppe Taillefer, Thomas Nietzel, et al. 2024. “The Mitochondrial orf117Sha Gene Desynchronizes Pollen Development and Causes Pollen Abortion in Arabidopsis Sha Cytoplasmic Male Sterility.” *Journal of Experimental Botany* 75 (16): 4851–72.
Dierckxsens, Nicolas, Patrick Mardulyn, and Guillaume Smits. 2017. “NOVOPlasty: De Novo Assembly of Organelle Genomes from Whole Genome Data.” *Nucleic Acids Research* 45 (4): e18.
Dolzhenko, Egor, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, et al. 2019. “ExpansionHunter: A Sequence-Graph-Based Tool to Analyze Variation in Short Tandem Repeat Regions.” *Bioinformatics (Oxford, England)* 35 (22): 4754–56.
Duan, Yifan, Jiping Yan, Yue Zhu, Cheng Zhang, Xiuhua Tao, Hongli Ji, Min Zhang, Xianrong Wang, and Long Wang. 2022. “Limited Accumulation of High-Frequency Somatic Mutations in a 1700-Year-Old Osmanthus Fragrans Tree.” *Tree Physiology* 42 (10): 2040–49.
Durvasula, Arun, Andrea Fulgione, Rafal M. Gutaker, Selen Irez Alacakaptan, Pádraic J. Flood, Célia Neto, Takashi Tsuchimatsu, et al. 2017. “African Genomes Illuminate the Early History and Transition to Selfing in Arabidopsis Thaliana.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (20): 5213–18.
Exposito-Alonso, Moises, 500 Genomes Field Experiment Team, Hernán A. Burbano, Oliver Bossdorf, Rasmus Nielsen, and Detlef Weigel. 2019. “Natural Selection on the Arabidopsis Thaliana Genome in Present and Future Climates.” *Nature* 573 (7772): 126–29.
Exposito-Alonso, Moises, Claude Becker, Verena J. Schuenemann, Ella Reiter, Claudia Setzer, Radka Slovak, Benjamin Brachi, et al. 2018. “The Rate and Potential Relevance of New Mutations in a Colonizing Plant Lineage.” *PLoS Genetics* 14 (2): e1007155.
Exposito-Alonso, Moises, François Vasseur, Wei Ding, George Wang, Hernán A. Burbano, and Detlef Weigel. 2018. “Genomic Basis and Evolutionary Potential for Extreme Drought Adaptation in Arabidopsis Thaliana.” *Nature Ecology & Evolution* 2 (2): 352–58.
Fan, Hao, and Jia-You Chu. 2007. “A Brief Review of Short Tandem Repeat Mutation.” *Genomics, Proteomics & Bioinformatics* 5 (1): 7–14.
Fields, Peter D., Gus Waneka, Matthew Naish, Michael C. Schatz, Ian R. Henderson, and Daniel B. Sloan. 2022. “Complete Sequence of a 641-Kb Insertion of Mitochondrial DNA in the Arabidopsis Thaliana Nuclear Genome.” *Genome Biology and Evolution* 14 (5). https://doi.org/10.1093/gbe/evac059.
Flood, Pádraic J., Joost van Heerwaarden, Frank Becker, C. Bastiaan de Snoo,

Jeremy Harbinson, and Mark G. M. Aarts. 2016. “Whole-Genome Hitchhiking on an Organelle Mutation.” *Current Biology: CB* 26 (10): 1306–11.
Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. 2011. “A Map of Local Adaptation in Arabidopsis Thaliana.” *Science (New York, N.Y.)* 334 (6052): 86–89.
Fujii, S., M. Akiyama, K. Aoki, Y. Sugaya, K. Higuchi, M. Hiraoka, Y. Miki, et al. 1999. “DNA Replication Errors Produced by the Replicative Apparatus of Escherichia Coli.” *Journal of Molecular Biology* 289 (4): 835–50.
Garcia, Laura E., Alejandro A. Edera, Jeffrey D. Palmer, Hector Sato, and M. Virginia Sanchez-Puerta. 2021. “Horizontal Gene Transfers Dominate the Functional Mitochondrial Gene Space of a Holoparasitic Plant.” *The New Phytologist* 229 (3): 1701–14.
Garrison, Erik, Andrea Guarracino, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, et al. 2024. “Building Pangenome Graphs.” *Nature Methods* 21 (11): 2008–12.
Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, et al. 2018. “Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference.” *Nature Biotechnology* 36 (9): 875–79.
Gebow, D., N. Miselis, and H. L. Liber. 2000. “Homologous and Nonhomologous Recombination Resulting in Deletion: Effects of p53 Status, Microhomology, and Repetitive DNA Length and Orientation.” *Molecular and Cellular Biology* 20 (11): 4028–35.
Goel, Manish, José A. Campoy, Kristin Krause, Lisa C. Baus, Anshupa Sahu, Hequan Sun, Birgit Walkemeier, et al. 2024. “The Vast Majority of Somatic Mutations in Plants Are Layer-Specific.” *Genome Biology* 25 (1): 194.
Goel, Manish, Hequan Sun, Wen-Biao Jiao, and Korbinian Schneeberger. 2019. “SyRI: Finding Genomic Rearrangements and Local Sequence Differences from Whole-Genome Assemblies.” *Genome Biology* 20 (1): 277.
Go, Sangjin, Hyunjin Koo, Minah Jung, Seongmin Hong, Gibum Yi, and Yong-Min Kim. 2024. “Pan-Chloroplast Genomes for Accession-Specific Marker Development in Hibiscus Syriacus.” *Scientific Data* 11 (1): 246.
Guiblet, Wilfried M., Marzia A. Cremona, Monika Cechova, Robert S. Harris, Iva Kejnovská, Eduard Kejnovsky, Kristin Eckert, Francesca Chiaromonte, and Kateryna D. Makova. 2018. “Long-Read Sequencing Technology Indicates Genome-Wide Effects of Non-B DNA on Polymerization Speed and Error Rate.” *Genome Research* 28 (12): 1767–78.
Hancock, Angela M., Benjamin Brachi, Nathalie Faure, Matthew W. Horton, Lucien B. Jarymowycz, F. Gianluca Sperone, Chris Toomajian, Fabrice Roux, and Joy Bergelson. 2011. “Adaptation to Climate across the Arabidopsis Thaliana Genome.” *Science (New York, N.Y.)* 334 (6052): 83–86.
Hanlon, Vincent C. T., Sarah P. Otto, and Sally N. Aitken. 2019. “Somatic Mutations Substantially Increase the per-Generation Mutation Rate in the Conifer Picea Sitchensis.” *Evolution Letters* 3 (4): 348–58.
Hanson, Maureen R., and Stéphane Bentolila. 2004. “Interactions of Mitochondrial and Nuclear Genes That Affect Male Gametophyte Development.” *The Plant Cell* 16 Suppl (suppl_1): S154–69.
Hasenauer, Flavia C., Hugo C. Barreto, Chantal Lotton, and Ivan Matic. 2025. “Genome-Wide Mapping of Spontaneous DNA Replication Error-Hotspots Using Mismatch Repair Proteins in Rapidly Proliferating Escherichia Coli.” *Nucleic*

Acids Research 53 (2): gkae1196.
Hofmeister, Brigitte T., Johanna Denkena, Maria Colomé-Tatché, Yadollah Shahryary, Rashmi Hazarika, Jane Grimwood, Sujan Mamidi, et al. 2020. “A Genome Assembly and the Somatic Genetic and Epigenetic Mutation Rate in a Wild Long-Lived Perennial Populus Trichocarpa.” *Genome Biology* 21 (1): 259.
Hohmann, Nora, Eva M. Wolf, Martin A. Lysak, and Marcus A. Koch. 2015. “A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History.” *The Plant Cell* 27 (10): 2770–84.
Horton, Matthew W., Angela M. Hancock, Yu S. Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N. Wayan Muliyati, et al. 2012. “Genome-Wide Patterns of Genetic Variation in Worldwide Arabidopsis Thaliana Accessions from the RegMap Panel.” *Nature Genetics* 44 (2): 212–16.
Huang, C. Y., N. Grünheit, N. Ahmadinejad, J. N. Timmis, and W. Martin. 2005. “Mutational Decay and Age of Chloroplast and Mitochondrial Genomes Transferred Recently to Angiosperm Nuclear Chromosomes.” *Plant Physiology* 138 (3). https://doi.org/10.1104/pp.105.060327.
Hu, Tina T., Pedro Pattyn, Erica G. Bakker, Jun Cao, Jan-Fang Cheng, Richard M. Clark, Noah Fahlgren, et al. 2011. “The Arabidopsis Lyrata Genome Sequence and the Basis of Rapid Genome Size Change.” *Nature Genetics* 43 (5): 476–81.
Igolkina, Anna A., Sebastian Vorbrugg, Fernando A. Rabanal, Hai-Jun Liu, Haim Ashkenazy, Aleksandra E. Kornienko, Joffrey Fitz, et al. 2024a. “Towards an Unbiased Characterization of Genetic Polymorphism: A Comparison of 27 A. Thaliana Genomes.” *Genomics*. bioRxiv. https://www.biorxiv.org/content/10.1101/2024.05.30.596703v1.
———. 2024b. “Towards an Unbiased Characterization of Genetic Polymorphism: A Comparison of 27*A. Thaliana*genomes.” *bioRxiv*. https://doi.org/10.1101/2024.05.30.596703.
Itabashi, Etsuko, Natsuko Iwata, Sota Fujii, Tomohiko Kazama, and Kinya Toriyama. 2011. “The Fertility Restorer Gene, Rf2, for Lead Rice-Type Cytoplasmic Male Sterility of Rice Encodes a Mitochondrial Glycine-Rich Protein: Rf2 for CMS Rice Encodes a Glycine-Rich Protein.” *The Plant Journal: For Cell and Molecular Biology* 65 (3): 359–67.
Jaegle, Benjamin, Rahul Pisupati, Luz Mayela Soto-Jiménez, Robin Burns, Fernando A. Rabanal, and Magnus Nordborg. 2023. “Extensive Sequence Duplication in Arabidopsis Revealed by Pseudo-Heterozygosity.” *Genome Biology* 24 (1): 44.
Jiao, Wen-Biao, and Korbinian Schneeberger. 2020. “Chromosome-Level Assemblies of Multiple Arabidopsis Genomes Reveal Hotspots of Rearrangements with Altered Evolutionary Dynamics.” *Nature Communications* 11 (1): 989.
Jin, Jian-Jun, Wen-Bin Yu, Jun-Bo Yang, Yu Song, Claude W. dePamphilis, Ting-Shuang Yi, and De-Zhu Li. 2020. “GetOrganelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly of Organelle Genomes.” *Genome Biology* 21 (1): 241.
Kang, Minghui, Haolin Wu, Huanhuan Liu, Wenyu Liu, Mingjia Zhu, Yu Han, Wei Liu, et al. 2023. “The Pan-Genome and Local Adaptation of Arabidopsis Thaliana.” *Nature Communications* 14 (1): 6259.
Karlicki, Michał, Stanisław Antonowicz, and Anna Karnkowska. 2022. “Tiara: Deep Learning-Based Classification System for Eukaryotic Sequences.” *Bioinformatics (Oxford, England)* 38 (2): 344–50.

Karnkowska, Anna, Vojtěch Vacek, Zuzana Zubáčová, Sebastian C. Treitli, Romana Petrželková, Laura Eme, Lukáš Novák, et al. 2016. “A Eukaryote without a Mitochondrial Organelle.” *Current Biology: CB* 26 (10): 1274–84.
Keeling, Patrick J. 2010. “The Endosymbiotic Origin, Diversification and Fate of Plastids.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1541): 729–48.
Kennedy, Scott R., Michael W. Schmitt, Edward J. Fox, Brendan F. Kohrn, Jesse J. Salk, Eun Hyun Ahn, Marc J. Prindle, et al. 2014. “Detecting Ultralow-Frequency Mutations by Duplex Sequencing.” *Nature Protocols* 9 (11): 2586–2606.
Kim, Sangtae, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, et al. 2018. “Strelka2: Fast and Accurate Calling of Germline and Somatic Variants.” *Nature Methods* 15 (8): 591–94.
Kim, Sung, Vincent Plagnol, Tina T. Hu, Christopher Toomajian, Richard M. Clark, Stephan Ossowski, Joseph R. Ecker, Detlef Weigel, and Magnus Nordborg. 2007. “Recombination and Linkage Disequilibrium in Arabidopsis Thaliana.” *Nature Genetics* 39 (9): 1151–55.
Krämer, Carolin, Christian R. Boehm, Jinghan Liu, Michael Kien Yin Ting, Alexander P. Hertle, Joachim Forner, Stephanie Ruf, Mark A. Schöttler, Reimo Zoschke, and Ralph Bock. 2024. “Removal of the Large Inverted Repeat from the Plastid Genome Reveals Gene Dosage Effects and Leads to Increased Genome Copy Number.” *Nature Plants* 10 (6): 923–35.
Krämer, Ute. 2015. “Planting Molecular Functions in an Ecological Context with Arabidopsis Thaliana.” *eLife* 4 (March). https://doi.org/10.7554/eLife.06100.
Kreutzer, D. A., and J. M. Essigmann. 1998. “Oxidized, Deaminated Cytosines Are a Source of C --> T Transitions in Vivo.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (7): 3578–82.
Lian, Qichao, Bruno Huettel, Birgit Walkemeier, Baptiste Mayjonade, Céline Lopez-Roques, Lisa Gil, Fabrice Roux, Korbinian Schneeberger, and Raphael Mercier. 2024. “A Pan-Genome of 69 Arabidopsis Thaliana Accessions Reveals a Conserved Genome Structure throughout the Global Species Range.” *Nature Genetics* 56 (5): 982–91.
Lian, Qun, Shuai Li, Shenglong Kan, Xuezhu Liao, Sanwen Huang, Daniel B. Sloan, and Zhiqiang Wu. 2024. “Association Analysis Provides Insights into Plant Mitonuclear Interactions.” *Molecular Biology and Evolution* 41 (2): msae028.
Liu, Haoxuan, and Jianzhi Zhang. 2021. “The Rate and Molecular Spectrum of Mutation Are Selectively Maintained in Yeast.” *Nature Communications* 12 (1): 4044.
Liu, Hongfang, Wei Zhao, Wei Hua, and Jing Liu. 2022. “A Large-Scale Population Based Organelle Pan-Genomes Construction and Phylogeny Analysis Reveal the Genetic Diversity and the Evolutionary Origins of Chloroplast and Mitochondrion in Brassica Napus L.” *BMC Genomics* 23 (1): 339.
Liu, Mei Hong, Benjamin M. Costa, Emilia C. Bianchini, Una Choi, Rachel C. Bandler, Emilie Lassen, Marta Grońska-Pęski, et al. 2024. “DNA Mismatch and Damage Patterns Revealed by Single-Molecule Sequencing.” *Nature* 630 (8017): 752–61.
Lynch, Michael, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas, and Patricia L. Foster. 2016. “Genetic Drift, Selection and the Evolution of the Mutation Rate.” *Nature Reviews. Genetics* 17 (11): 704–14.
Martiniano, Rui, Erik Garrison, Eppie R. Jones, Andrea Manica, and Richard Durbin.

2020. “Removing Reference Bias and Improving Indel Calling in Ancient DNA Data Analysis by Mapping to a Sequence Variation Graph.” *Genome Biology* 21 (1): 250.
Meinke, D. W., J. M. Cherry, C. Dean, S. D. Rounsley, and M. Koornneef. 1998. “Arabidopsis Thaliana: A Model Plant for Genome Analysis.” *Science (New York, N.Y.)* 282 (5389): 662, 679–82.
Melonek, Joanna, Jorge Duarte, Jerome Martin, Laurent Beuf, Alain Murigneux, Pierrick Varenne, Jordi Comadran, et al. 2021. “The Genetic Basis of Cytoplasmic Male Sterility and Fertility Restoration in Wheat.” *Nature Communications* 12 (1): 1036.
Michalovova, M., B. Vyskot, and E. Kejnovsky. 2013. “Analysis of Plastid and Mitochondrial DNA Insertions in the Nucleus (NUPTs and NUMTs) of Six Plant Species: Size, Relative Age and Chromosomal Localization.” *Heredity* 111 (4): 314–20.
Monroe, J. Grey, Kevin D. Murray, Wenfei Xian, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, et al. 2023. “Reply to: Re-Evaluating Evidence for Adaptive Mutation Rate Variation.” *Nature* 619 (7971): E57–60.
Monroe, J. Grey, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink, Moises Exposito-Alonso, Marie Klein, et al. 2022. “Mutation Bias Reflects Natural Selection in Arabidopsis Thaliana.” *Nature* 602 (7895): 101–5.
Naish, Matthew, Michael Alonge, Piotr Wlodzimierz, Andrew J. Tock, Bradley W. Abramson, Anna Schmücker, Terezie Mandáková, et al. 2021. “The Genetic and Epigenetic Landscape of the Arabidopsis Centromeres.” *Science (New York, N.Y.)* 374 (6569): eabi7489.
Ng, Kah Ying, Guleycan Lutfullahoglu Bal, Uwe Richter, Omid Safronov, Lars Paulin, Cory D. Dunn, Ville O. Paavilainen, et al. 2022. “Nonstop mRNAs Generate a Ground State of Mitochondrial Gene Expression Noise.” *Science Advances* 8 (46): eabq5234.
Nordborg, Magnus, Justin O. Borevitz, Joy Bergelson, Charles C. Berry, Joanne Chory, Jenny Hagenblad, Martin Kreitman, et al. 2002. “The Extent of Linkage Disequilibrium in Arabidopsis Thaliana.” *Nature Genetics* 30 (2): 190–93.
Nordborg, Magnus, Tina T. Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, et al. 2005. “The Pattern of Polymorphism in Arabidopsis Thaliana.” *PLoS Biology* 3 (7): e196.
Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. “The Complete Sequence of a Human Genome.” *Science (New York, N.Y.)* 376 (6588): 44–53.
Osborne, Anne H., Derek Vance, Eelco J. Rohling, Nick Barton, Mike Rogerson, and Nuri Fello. 2008. “A Humid Corridor across the Sahara for the Migration of Early Modern Humans out of Africa 120,000 Years Ago.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (43): 16444–47.
Ossowski, Stephan, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M. Clark, Ruth G. Shaw, Detlef Weigel, and Michael Lynch. 2010. “The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis Thaliana.” *Science (New York, N.Y.)* 327 (5961): 92–94.
Parmesan, Camille, and Mick E. Hanley. 2015. “Plants and Climate Change: Complexities and Surprises.” *Annals of Botany* 116 (6): 849–64.
Platt, Alexander, Matthew Horton, Yu S. Huang, Yan Li, Alison E. Anastasio, Ni Wayan Mulyati, Jon Agren, et al. 2010. “The Scale of Population Structure in

Arabidopsis Thaliana." *PLoS Genetics* 6 (2): e1000843.
Plomion, Christophe, Jean-Marc Aury, Joëlle Amselem, Thibault Leroy, Florent Murat, Sébastien Duplessis, Sébastien Faye, et al. 2018. "Oak Genome Reveals Facets of Long Lifespan." *Nature Plants* 4 (7): 440–52.
Qing, Xue, Michał Karlicki, Fan Guo, Anna Karnkowska, and Hongmei Li. 2024. "Soil Nematode Community Profiling Using Reference-Free Mito-Metagenomics." *Soil Biology & Biochemistry* 199 (109613): 109613.
Quesneville, Hadi. 2020. "Twenty Years of Transposable Element Analysis in the Arabidopsis Thaliana Genome." *Mobile DNA* 11 (1): 28.
Quiroz, Daniela, Satoyo Oya, Diego Lopez-Mateos, Kehan Zhao, Alice Pierce, Lissandro Ortega, Alissza Ali, et al. 2024. "H3K4me1 Recruits DNA Repair Proteins in Plants." *The Plant Cell* 36 (6): 2410–26.
Rautiainen, Mikko, Sergey Nurk, Brian P. Walenz, Glennis A. Logsdon, David Porubsky, Arang Rhie, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. 2023. "Telomere-to-Telomere Assembly of Diploid Chromosomes with Verkko." *Nature Biotechnology* 41 (10): 1474–82.
Readman, Chiu, Rajan-Babu Indhu-Shree, Friedman Jan M, and Birol Inanc. 2021. "Straglr: Discovering and Genotyping Tandem Repeat Expansions Using Whole Genome Long-Read Sequences." *Genome Biology* 22 (1): 224.
Richardson, Aaron O., and Jeffrey D. Palmer. 2007. "Horizontal Gene Transfer in Plants." *Journal of Experimental Botany* 58 (1): 1–9.
Roulet, M. Emilia, Luis Federico Ceriotti, Leonardo Gatica-Soria, and M. Virginia Sanchez-Puerta. 2024. "Horizontally Transferred Mitochondrial DNA Tracts Become Circular by Microhomology-Mediated Repair Pathways." *The New Phytologist* 243 (6): 2442–56.
Satake, Akiko, Ryosuke Imai, Takeshi Fujino, Sou Tomimoto, Kayoko Ohta, Mohammad Na'iem, Sapto Indrioko, et al. 2024. "Somatic Mutation Rates Scale with Time Not Growth Rate in Long-Lived Tropical Trees." *eLife* 12 (October):RP88456.
Schmid-Siegert, Emanuel, Namrata Sarkar, Christian Iseli, Sandra Calderon, Caroline Gouhier-Darimont, Jacqueline Chrast, Pietro Cattaneo, et al. 2017. "Low Number of Fixed Somatic Mutations in a Long-Lived Oak Tree." *Nature Plants* 3 (12): 926–29.
Schmitt, Sylvain, Patrick Heuret, Valérie Troispoux, Mélanie Beraud, Jocelyn Cazal, Émilie Chancerel, Charlotte Cravero, et al. 2024. "Low-Frequency Somatic Mutations Are Heritable in Tropical Trees Dicorynia Guianensis and Sextonia Rubra." *Proceedings of the National Academy of Sciences of the United States of America* 121 (10): e2313312121.
Schneeberger, Korbinian, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. 2009. "Simultaneous Alignment of Short Reads against Multiple Genomes." *Genome Biology* 10 (9): R98.
Schultz, S. T., M. Lynch, and J. H. Willis. 1999. "Spontaneous Deleterious Mutation in Arabidopsis Thaliana." *Proceedings of the National Academy of Sciences of the United States of America* 96 (20): 11393–98.
Selby, Christopher P., Laura A. Lindsey-Boltz, Wentao Li, and Aziz Sancar. 2023. "Molecular Mechanisms of Transcription-Coupled Repair." *Annual Review of Biochemistry* 92 (1): 115–44.
Serero, Alexandre, Claire Jubin, Sophie Loeillet, Patricia Legoix-Né, and Alain G. Nicolas. 2014. "Mutational Landscape of Yeast Mutator Strains." *Proceedings of*

the National Academy of Sciences of the United States of America 111 (5): 1897–1902.
Sharbrough, Joel, Laura Bankers, Emily Cook, Peter D. Fields, Joseph Jalinsky, Kyle E. McElroy, Maurine Neiman, John M. Logsdon, and Jeffrey L. Boore. 2023. “Single-Molecule Sequencing of an Animal Mitochondrial Genome Reveals Chloroplast-like Architecture and Repeat-Mediated Recombination.” *Molecular Biology and Evolution* 40 (1). https://doi.org/10.1093/molbev/msad007.
Shimada, Atsushi, Jonathan Cahn, Evan Ernst, Jason Lynn, Daniel Grimanelli, Ian Henderson, Tetsuji Kakutani, and Robert A. Martienssen. 2024. “Retrotransposon Addiction Promotes Centromere Function via Epigenetically Activated Small RNAs.” *Nature Plants* 10 (9): 1304–16.
Sirén, Jouni, Jean Monlong, Xian Chang, Adam M. Novak, Jordan M. Eizenga, Charles Markello, Jonas A. Sibbesen, et al. 2021. “Pangenomics Enables Genotyping of Known Structural Variants in 5202 Diverse Genomes.” *Science (New York, N.Y.)* 374 (6574): abg8871.
Sloan, Daniel B., Andrew J. Alverson, John P. Chuckalovcak, Martin Wu, David E. McCauley, Jeffrey D. Palmer, and Douglas R. Taylor. 2012. “Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates.” *PLoS Biology* 10 (1): e1001241.
Song, Jia-Ming, Wen-Zhao Xie, Shuo Wang, Yi-Xiong Guo, Dal-Hoe Koo, Dave Kudrna, Chenbo Gong, et al. 2021. “Two Gap-Free Reference Genomes and a Global View of the Centromere Architecture in Rice.” *Molecular Plant* 14 (10): 1757–67.
Soorni, Aboozar, David Haak, David Zaitlin, and Aureliano Bombarely. 2017. “Organelle_PBA, a Pipeline for Assembling Chloroplast and Mitochondrial Genomes from PacBio DNA Sequencing Data.” *BMC Genomics* 18 (1): 49.
Steely, Cody J., W. Scott Watkins, Lisa Baird, and Lynn B. Jorde. 2022. “The Mutational Dynamics of Short Tandem Repeats in Large, Multigenerational Families.” *Genome Biology* 23 (1): 253.
Štorchová, Helena, and Manuela Krüger. 2024. “Methods for Assembling Complex Mitochondrial Genomes in Land Plants.” *Journal of Experimental Botany* 75 (17): 5169–74.
Sun, Hequan, Patrick Abeli, José Antonio Campoy, Thea Rütjes, Kristin Krause, Wen-Biao Jiao, Randy Beaudry, and Korbinian Schneeberger. 2024. “The Identification and Analysis of Meristematic Mutations within the Apple Tree That Developed the RubyMac Sport Mutation.” *BMC Plant Biology* 24 (1): 912.
Taanman, J. W. 1999. “The Mitochondrial Genome: Structure, Transcription, Translation and Replication.” *Biochimica et Biophysica Acta* 1410 (2): 103–23.
Tang, Huiwu, Xingmei Zheng, Chuliang Li, Xianrong Xie, Yuanling Chen, Letian Chen, Xiucai Zhao, et al. 2017. “Multi-Step Formation, Evolution, and Functionalization of New Cytoplasmic Male Sterility Genes in the Plant Mitochondrial Genomes.” *Cell Research* 27 (1): 130–46.
Tomkova, Marketa, Michael John McClellan, Gilles Crevel, Akbar Muhammed Shahid, Nandini Mozumdar, Jakub Tomek, Emelie Shepherd, Sue Cotterill, Benjamin Schuster-Böckler, and Skirmantas Kriaucionis. 2024. “Human DNA Polymerase ε Is a Source of C>T Mutations at CpG Dinucleotides.” *Nature Genetics* 56 (11): 2506–16.
Tubbs, Anthony, and André Nussenzweig. 2017. “Endogenous DNA Damage as a Source of Genomic Instability in Cancer.” *Cell* 168 (4): 644–56.
Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries,

Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.” *Nature Reviews. Methods Primers* 1 (1): 1–21.
Unseld, M., J. R. Marienfeld, P. Brandt, and A. Brennicke. 1997. “The Mitochondrial Genome of Arabidopsis Thaliana Contains 57 Genes in 366,924 Nucleotides.” *Nature Genetics* 15 (1): 57–61.
Vahrenholz, C., G. Riemen, E. Pratje, B. Dujon, and G. Michaelis. 1993. “Mitochondrial DNA of Chlamydomonas Reinhardtii: The Structure of the Ends of the Linear 15.8-Kb Genome Suggests Mechanisms for DNA Replication.” *Current Genetics* 24 (3): 241–47.
Valiente-Mullor, Carlos, Beatriz Beamud, Iván Ansari, Carlos Francés-Cuesta, Neris García-González, Lorena Mejía, Paula Ruiz-Hueso, and Fernando González-Candelas. 2021. “One Is Not Enough: On the Effects of Reference Genome for the Mapping and Subsequent Analyses of Short-Reads.” *PLoS Computational Biology* 17 (1): e1008678.
Varshney, Rajeev K., Andreas Graner, and Mark E. Sorrells. 2005. “Genic Microsatellite Markers in Plants: Features and Applications.” *Trends in Biotechnology* 23 (1): 48–55.
Voichek, Yoav, and Detlef Weigel. 2020. “Identifying Genetic Variants Underlying Phenotypic Variation in Plants without Complete Genomes.” *Nature Genetics* 52 (5): 534–40.
Vorbrugg, Sebastian, Ilja Bezrukov, Zhigui Bao, Wenfei Xian, and Detlef Weigel. 2024. “Gfa2bin Enables Graph-Based GWAS by Converting Genome Graphs to Pan-Genomic Genotypes.” *Bioinformatics*. bioRxiv. https://www.biorxiv.org/content/10.1101/2024.12.05.626966v1.
Waneka, Gus, Braden Pate, J. Grey Monroe, and Daniel B. Sloan. 2024. “Exploring the Relationship between Gene Expression and Low-Frequency Somatic Mutations in Arabidopsis with Duplex Sequencing.” *Genome Biology and Evolution* 16 (10): evae213.
Wang, D., D. A. Kreutzer, and J. M. Essigmann. 1998. “Mutagenicity and Repair of Oxidative DNA Damage: Insights from Studies Using Defined Lesions.” *Mutation Research* 400 (1-2): 99–115.
Wang, Jie, Shenglong Kan, Xuezhu Liao, Jiawei Zhou, Luke R. Tembrock, Henry Daniell, Shuangxia Jin, and Zhiqiang Wu. 2024. “Plant Organellar Genomes: Much Done, Much More to Do.” *Trends in Plant Science* 29 (7): 754–69.
Wang, Nan, Chaochao Li, Lihua Kuang, Xiaomeng Wu, Kaidong Xie, Andan Zhu, Qiang Xu, et al. 2022. “Pan-Mitogenomics Reveals the Genetic Basis of Cytonuclear Conflicts in Citrus Hybridization, Domestication, and Diversification.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (43): e2206076119.
Wang, Xi, Detlef Weigel, and Lisa M. Smith. 2013. “Transposon Variants and Their Effects on Gene Expression in Arabidopsis.” *PLoS Genetics* 9 (2): e1003255.
Weber, J. L., and P. E. May. 1989. “Abundant Class of Human DNA Polymorphisms Which Can Be Typed Using the Polymerase Chain Reaction.” *The American Journal of Human Genetics* 44 (3): 388–96.
Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. “Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome.” *Nature Biotechnology* 37 (10): 1155–62.
Weng, Mao-Lun, Claude Becker, Julia Hildebrandt, Manuela Neumann, Matthew T.

Rutter, Ruth G. Shaw, Detlef Weigel, and Charles B. Fenster. 2019. “Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis Thaliana.” *Genetics* 211 (2): 703–14.
Willems, Thomas, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. 2017. “Genome-Wide Profiling of Heritable and de Novo STR Variations.” *Nature Methods* 14 (6): 590–92.
Wlodzimierz, Piotr, Fernando A. Rabanal, Robin Burns, Matthew Naish, Elias Primetis, Alison Scott, Terezie Mandáková, et al. 2023. “Cycles of Satellite and Transposon Evolution in Arabidopsis Centromeres.” *Nature* 618 (7965): 557–65.
Woodward, Andrew W., and Bonnie Bartel. 2018. “Biology in Bloom: A Primer on the Arabidopsis Thaliana Model System.” *Genetics* 208 (4): 1337–49.
Wu, Guohong Albert, Javier Terol, Victoria Ibanez, Antonio López-García, Estela Pérez-Román, Carles Borredá, Concha Domingo, et al. 2018. “Genomics of the Origin and Evolution of Citrus.” *Nature* 554 (7692): 311–16.
Wu, Hsin-Yen Larry, Qiaoyun Ai, Rita Teresa Teixeira, Phong H. T. Nguyen, Gaoyuan Song, Christian Montes, J. Mitch Elmore, Justin W. Walley, and Polly Yingshan Hsu. 2024. “Improved Super-Resolution Ribosome Profiling Reveals Prevalent Translation of Upstream ORFs and Small ORFs in Arabidopsis.” *The Plant Cell* 36 (3): 510–39.
Wulfridge, Phillip, and Kavitha Sarma. 2024. “Intertwining Roles of R-Loops and G-Quadruplexes in DNA Repair, Transcription and Genome Organization.” *Nature Cell Biology* 26 (7): 1025–36.
Xiao, Senlin, Jingfeng Xing, Tiange Nie, Aiguo Su, Ruyang Zhang, Yanxin Zhao, Wei Song, and Jiuran Zhao. 2022. “Comparative Analysis of Mitochondrial Genomes of Maize CMS-S Subtypes Provides New Insights into Male Sterility Stability.” *BMC Plant Biology* 22 (1): 469.
Zhang, Guo-Jun, Ran Dong, Li-Na Lan, Shu-Fen Li, Wu-Jun Gao, and Hong-Xing Niu. 2020. “Nuclear Integrants of Organellar DNA Contribute to Genome Structure and Evolution in Plants.” *International Journal of Molecular Sciences* 21 (3): 707.
Zhivagui, Maria, Areebah Hoda, Noelia Valenzuela, Yi-Yu Yeh, Jason Dai, Yudou He, Shuvro P. Nandi, Burcak Otlu, Bennett Van Houten, and Ludmil B. Alexandrov. 2023. “DNA Damage and Somatic Mutations in Mammalian Cells after Irradiation with a Nail Polish Dryer.” *Nature Communications* 14 (1): 276.
Zhong, Yuyang, Miki Okuno, Nobuhiro Tsutsumi, and Shin-Ichi Arimura. 2025. “Mitochondrial DNA and the Largest Nuclear-Mitochondrial DNA in Arabidopsis Can Be Separated by Their Methylation Levels.” *Plant Physiology* 197 (3). https://doi.org/10.1093/plphys/kiaf069.
Zhou, Chenxi, Max Brown, Mark Blaxter, Shane A. McCarthy, and Richard Durbin. 2024. “Oatk: A de Novo Assembly Tool for Complex Plant Organelle Genomes.” *Bioinformatics*. bioRxiv. https://www.biorxiv.org/content/10.1101/2024.10.23.619857v1.
Zimorski, Verena, Chuan Ku, William F. Martin, and Sven B. Gould. 2014. “Endosymbiotic Theory for Organelle Origins.” *Current Opinion in Microbiology* 22 (December):38–48.

Thesis Appendix

Thesis Appendix I

TIPPo: A User-Friendly Tool for De Novo Assembly of Organellar Genomes with High-Fidelity Data

Wenfei Xian [1], Ilja Bezrukov [1], Zhigui Bao [1], Sebastian Vorbrugg [1], Anupam Gautam [2,3], Detlef Weigel [*,1,4]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[2]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

[3]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[4]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

***Corresponding author:** E-mail: weigel@tue.mpg.de.

Associate editor: Michael Purugganan

Abstract

Plant cells have two major organelles with their own genomes: chloroplasts and mitochondria. While chloroplast genomes tend to be structurally conserved, the mitochondrial genomes of plants, which are much larger than those of animals, are characterized by complex structural variation. We introduce TIPPo, a user-friendly, reference-free assembly tool that uses PacBio high-fidelity long-read data and that does not rely on genomes from related species or nuclear genome information for the assembly of organellar genomes. TIPPo employs a deep learning model for initial read classification and leverages *k*-mer counting for further refinement, significantly reducing the impact of nuclear insertions of organellar DNA on the assembly process. We used TIPPo to completely assemble a set of 54 complete chloroplast genomes. No other tool was able to completely assemble this set. TIPPo is comparable with PMAT in assembling mitochondrial genomes from most species but does achieve even higher completeness for several species. We also used the assembled organelle genomes to identify instances of nuclear plastid DNA (NUPTs) and nuclear mitochondrial DNA (NUMTs) insertions. The cumulative length of NUPTs/NUMTs positively correlates with the size of the nuclear genome, suggesting that insertions occur stochastically. NUPTs/NUMTs show predominantly C:G to T:A changes, with the mutated cytosines typically found in CG and CHG contexts, suggesting that degradation of NUPT and NUMT sequences is driven by the known elevated mutation rate of methylated cytosines. Small interfering RNA loci are enriched in NUPTs and NUMTs, consistent with the RdDM pathway mediating DNA methylation in these sequences.

Key words: **chloroplast genome, mitochondrial genome, genome assembly, nuclear insertions of organellar genomes, PacBio HiFi reads.**

Introduction

In the cells of green plants, DNA is found in three main locations: chloroplasts or chloroplast-related plastids, mitochondria, and the nucleus. The chloroplast is the primary site of photosynthesis, converting solar energy into chemical energy, while mitochondria are crucial for cellular energy metabolism. Chloroplasts and mitochondria are thought to have originated from ancient endosymbiosis events (Zimorski et al. 2014). Due to secondary and tertiary endosymbiosis, chloroplasts or plastids are present across various kingdoms, collectively referred to as photosynthetic eukaryotes (Yoon et al. 2004).

Chloroplast genomes are structurally conserved across species, and they typically comprise four distinct fragments: one large single copy (LSC), one small single copy (SSC), and two inverted repeats (IRs). In contrast, the genomes of mitochondria, present in all eukaryotic organisms except for the microorganism *Monocercomonoides* sp. (Karnkowska et al. 2016), vary significantly across kingdoms. The structure of bilaterian mitochondrial genomes is conserved, presenting as a single small circular DNA with sizes around 17 kb (Ladoukakis and Zouros 2017). The situation is very different in plants, which have structurally complex mitochondrial genomes with large variation in size, with the largest known mitochondrial genomes reaching up to 11 Mb (Sloan et al. 2012; Putintseva et al. 2020).

Compared with nuclear genomes, much less attention has been paid to the high-quality assembly of organellar genomes. Short-read data are useful, with some caveats, for the assembly of the relatively small and conserved mitochondrial genomes of animals and chloroplast genomes of plants (Dierckxsens et al. 2017; Jin et al. 2020), but their

Received: July 26, 2024. **Revised:** November 15, 2024. **Accepted:** November 18, 2024

Open Access

Methods

utility is limited for the larger and more complex mitochondrial genomes of plants (Štorchová and Krüger 2024). Long and highly accurate read data have substantially enhanced our ability to assemble nuclear genomes (Wenger et al. 2019; Sereika et al. 2022). With the help of long reads, even highly repetitive regions such as centromeres and telomeres can be assembled (Naish et al. 2021; Nurk et al. 2022; Wlodzimierz et al. 2023), although challenges persist with the assembly of rDNA clusters. Moreover, the typically very high coverage of organellar genomes in data sets of genomic DNA interferes with productive assembly using standard tools, which are optimized for the nuclear genomes (Cheng et al. 2021). In addition, chloroplast and mitochondrial DNA fragments are often transferred to the nucleus, which also interferes with assembly of the true organellar genomes (Uliano-Silva et al. 2023).

Several tools have been developed to enable the specific use of long-read data for organellar genome assembly, primarily focusing on chloroplast genomes, such as Organelle_PBA (Soorni et al. 2017), ptGAUL (Zhou et al. 2023), and CLAW (Phillips et al. 2024). The general approach begins with extracting chloroplast reads from the data set by aligning long reads to the chloroplast genomes of closely related species. This is straightforward and effective for the chloroplast genome, as there are now over 12,000 published chloroplast genomes available, making it almost always possible to find a sufficiently closely related species for successful extraction of chloroplast reads. However, this approach has limitations for mitochondrial genomes, given the much smaller number of available plant mitochondrial genomes (~500 as of July 2023) and the much lower conservation of mitochondrial genomes, even between closely related species. There have been ongoing efforts to assemble complex mitochondrial genomes. GSAT (He et al. 2022) begins by using short reads to construct the assembly graph and then simplifies it using long reads. However, the assembly graph created from short reads struggles to handle highly repetitive regions, making it challenging to assemble complete genomes. Recently, an alternative approach has been proposed—PMAT (Bi et al. 2024). It begins with downsampling the initial read data set to an estimated coverage of the organellar genomes that is suitable for standard assembly tools. Next, a normal assembly is performed, and then the contigs that appear to belong to organellar genomes are identified based on the presence of conserved protein-coding genes. While useful, this approach may result in incomplete assemblies, especially for species with multichromosomal mitochondrial genomes where some chromosomes lack coding genes (Sanchez-Puerta et al. 2017). Clearly, the preferred approach would be a (largely) reference free and tool for organellar genome assembly that has similar power for both chloroplast and mitochondrial genomes.

As stated above, organellar DNA can be transferred to the nucleus, and it is common to find organellar sequences in the nuclear genome (Richly and Leister 2004a,b; Hazkani-Covo et al. 2010; Michalovova et al. 2013; Zhang et al. 2020). These sequences are known as nuclear mitochondrial DNA (NUMTs) and nuclear chloroplast DNA (NUPTs). The nuclear genome evolves much faster than mitochondrial genome, typically by an order of magnitude (Wolfe et al. 1987; Drouin et al. 2008). Accordingly, NUPTs and NUMTs tend to diverge from the ancestral organellar genomes quite rapidly. By aligning NUPTs and NUMTs, which should not carry any function, to the corresponding organellar genomes, one can explore presumably neutral processes of sequence change in the integrated organellar DNA (Huang et al. 2005; Rousseau-Gueutin et al. 2011; Yoshida et al. 2014; Fields et al. 2022). Questions of interest are whether NUPTs and NUMTs behave in a similar manner, and how their evolutionary fate compares with that of other large insertions, such as transposons (Wang et al. 2013; Maumus and Quesneville 2014).

We have developed TIPPo, a user-friendly, reference-free assembly tool for plant organellar genomes that integrates TIARA, a deep learning–based approach for organellar DNA classification (Karlicki et al. 2022), eliminating the need for knowledge of organellar genomes from closely related species genomes or nuclear genome information of the target species. We use *k*-mer information to optimize TIARA's output, distinguishing NUPTs, NUMTs, and misclassifications caused by repetitive sequences. Using TIPPo, we not only successfully assembled 54 complete chloroplast genomes but also demonstrated superior performance in mitochondrial assembly compared with PMAT, revealing the complex structure of mitochondrial genomes. Additionally, we detailed the insertion patterns of NUPTs and NUMTs and analyzed nucleotide substitutions in NUPTs and NUMTs.

Approach

We designed and implemented a reference-free, user-friendly tool for the assembly of plant organellar genomes called TIPPo from highly accurate PacBio high-fidelity (HiFi) long reads. It begins with a deep learning model to identify candidate organelle reads, followed by the use of a *k*-mer count approach to filter out the remaining nuclear reads and finishing with the assembly of the organellar genomes. Figure 1 illustrates the entire workflow.

TIPPo uses TIARA (Karlicki et al. 2022) to classify the reads, a deep learning–based approach that follows a two-step process: first, it classifies reads as nuclear or organellar, and then further categorizes the organellar reads into plastid or mitochondrial. We evaluated the accuracy of TIARA (Karlicki et al. 2022) using *Arabidopsis thaliana* and *Oryza sativa* (supplementary figs. S1 and S20, Supplementary Material online). As described in the original paper, TIARA will classify the NUMTs/NUPTs as organelle reads, and there is also an increased proportion of misclassification in highly repeated regions, such as centromeres and rDNA clusters. Hence, further filtering is necessary.

The assumption for subsequent filtering is that true organellar reads are the largest class in the TIARA output, and that misclassifications are relatively rare. We use KMC3 (Kokot et al. 2017) to generate a *k*-mer ($k = 31$) count database from the reads identified by TIARA. Next, we perform filtering based on *k*-mer counts separately for chloroplast

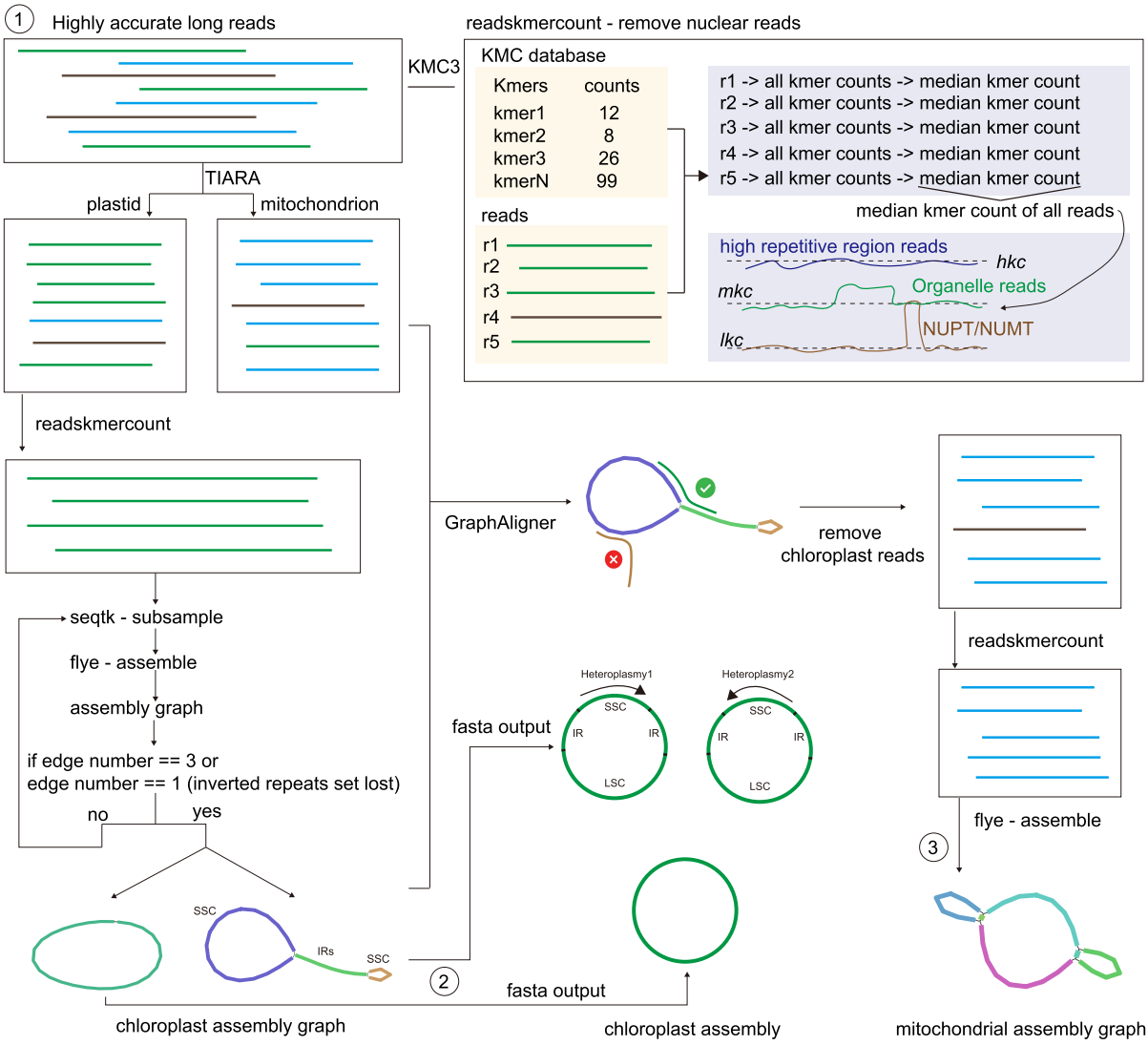


Fig. 1. Workflow of TIPPo.

and mitochondrial reads. We use readskmercount to obtain the read median *k*-mer count *rmkc*, which is used as a representative for each read. Reads labeled as plastid are processed first because chloroplast genomes are more conserved than mitochondrial genomes.

After calculating *rmkc* for all input reads, the median *k*-mer count *mkc* of all chloroplast reads, and of all mitochondrial reads after chloroplast assembly, will be used for filtering. To this end, we set the low *k*-mer count threshold *lkc* to 0.3 × *mkc*, and the high *k*-mer count threshold *hkc* to 5 × *mkc*. A read is removed if more than one-fifth of its *k*-mer counts are either lower than *lkc* or higher than *hkc*. Reads with many *k*-mer counts below the *lkc* threshold likely originate from the nucleus, and possibly correspond to NUPTs or NUMTs. Reads with many *k*-mer counts above the *hkc* threshold are likely from highly repetitive nuclear regions such as centromeres, and rDNA clusters. After filtering, flye (Kolmogorov et al. 2019) is used to assemble the chloroplast genome in the first assembly step. The assembly is performed iteratively with a random selection of reads, until the assembly graph matches the typical chloroplast structure. In each assembly round, only 800 reads are used, which is around 100× coverage, since excessive coverage might negatively affect flye results. Following the assembly with flye, the assembly graph is checked for a typical chloroplast structure or a circular DNA when IRs were set as lost. The structural check is aiming to match two isomeric chloroplast genomes that coexist equimolarly, differing only in the orientation of the LSC and SSC, as is the case in most land plants and algae (Palmer 1983; Aldrich et al. 1985; Wang and Lanfear 2019). Once this is achieved, the cycle ends with output of two typical heteroplasmic fasta sequences or one circular sequence.

The next step is the assembly of the mitochondrial genome. Considering that some chloroplast reads might be misclassified as mitochondrion by TIARA, GraphAligner

(Rautiainen and Marschall 2020) is used to align all reads labeled as mitochondrion to the chloroplast assembly graph as a further refinement step. If the read alignment is almost end-to-end (left clip length ≤ 100 bp, right clip length ≤ 100 bp, and identity > 95%), reads are considered as likely originating from the chloroplast and are removed. It is worth noting that mapping reads directly to an organelle assembly graph is the optimal solution for the organellar genome alignment, since linearized circular DNA combined with heteroplasmy will lead to clipped alignments. CLAW (Phillips et al. 2024) also addresses the alignment issues caused by a linearized circular DNA target by joining the two linear DNA sequences. Although this approach avoids clipping alignment, it introduces the issue of mapping quality of zero.

As a final step in TIPPo, the reads remaining after alignment to the chloroplast assembly graph will be processed by readskmercount to exclude reads originating from the nucleus, as described above. Given that the coverage of mitochondrion is generally lower than that of chloroplast, and the genome sizes are usually larger, all finally remaining reads serve directly as input to flye for generating the assembly graph.

Results and Discussion

Chloroplast Genome Assembly

Given the conserved structure of chloroplast genomes, we categorized the assemblies based on the structure on the assembly graph into three classes: (i) containing only the typical chloroplast genome or one circular DNA (complete genome); (ii) consisting of the complete genome and other sequences; and (iii) incomplete assembly (Fig. 2).

To test the performance of TIPPo, we selected 54 phylogenetically diverse planta and compared the performance with that of ptGAUL and CLAW. Using TIPPo, we successfully assembled all 54 complete chloroplast genomes without any extraneous sequences (supplementary fig. S2, Supplementary Material online) and assembled 48 complete chloroplast genomes at 0.5× nuclear genome coverage (supplementary fig. S23, Supplementary Material online). We obtained two chloroplast genomes (supplementary fig. S2, Supplementary Material online) for *Acorus gramineus*, suggesting that the sample might contain reads from two genotypes. It was therefore excluded from downstream analysis. ptGAUL assembled 46 complete genomes, produced 6 assemblies containing complete chloroplast genomes along with other sequences, and was unable to assemble 1 species (supplementary fig. S3, Supplementary Material online). CLAW successfully assembled 35 complete chloroplast genomes, 14 assemblies included complete chloroplast genomes, as well as other sequences, and it did not assemble 4 species (supplementary fig. S4, Supplementary Material online).

Out of 54 species, 4 species have lost IRs, resulting in a single circular structure in the assembly graph (supplementary fig. S2, Supplementary Material online). For the remaining 50 species with IRs, the assembly graph typically consists of 3 nodes: 1 representing the LSC, 1 representing the SSC, and 1 representing the IR, as shown in supplementary fig. S2, Supplementary Material online. The heteroplasmy in chloroplasts is mainly mediated by an IR, so for outputs with typical chloroplast structures in TIPPo, we provide two separate configurations of the chloroplast genome.

Whole-genome alignments against published chloroplast genomes indicated high consistency between the published and TIPPo assemblies (supplementary fig. S6, Supplementary Material online). Typical chloroplast genomes have three distinct regions: SSC, LSC, and IR. Public chloroplast genomes are typically presented as linear circular DNA sequences. Thus, in whole-genome alignments, the single IR from the assembly graph aligned to two regions of the linear representations, with one forward and one reverse orientation (supplementary fig. S6, Supplementary Material online). We also assembled two previously unpublished chloroplast genomes, *Adenosma buchneroides* (153,640 bp; supplementary fig. S7, Supplementary Material online) and *Helichrysum umbraculigerum* (154,011 bp; supplementary fig. S8, Supplementary Material online). Comparing the chloroplast genome lengths across 53 species, we observed that those from green algae are larger than those from terrestrial plants, with terrestrial plant chloroplast genomes generally around 150 kb (Fig. 2; supplementary table S4, Supplementary Material online). The base consensus approach was also applied to *A. thaliana* and *Silene conica*, resulting in only a 1 and 2 bp insertion, respectively, compared with the published reference genome, both occurring in homopolymer regions (supplementary fig. S21, Supplementary Material online). These could be true minor differences between the exact germplasm used, or due to assembly errors in either the published genomes or in our assemblies.

Mitochondrial Genome Assembly

Only PMAT assembled also mitochondrial genomes, and we therefore compared the ability of TIPPo to assemble mitochondrial genomes with PMAT. Given that PMAT assembled genomes often contain sequences from both organelles, we aligned distinct parts of the mitochondrial assembly graph from both PMAT and TIPPo against the chloroplast genomes assembled by TIPPo. For PMAT, mitochondrial genome assemblies from 33 out of 53 species contained also chloroplast sequences (supplementary fig. S9, Supplementary Material online). For *Musa acuminata*, *Ad. buchneroides*, *Trapa bicornis* (master), and *Fragaria vesca* (master), all parts aligned fully to the chloroplast genome graph, indicating that the assembly of mitochondrial genomes had failed. Since TIPPo removes chloroplast reads first, none of the assemblies contain chloroplast sequences (supplementary fig. S9, Supplementary Material online). Thus, in subsequent analyses, we removed the chloroplast sequences from PMAT mitochondrial genome assemblies.

Given the structural diversity of plant mitochondrial genomes, it is challenging to assess the completeness of results from the assembly graph structure as we did with chloroplasts (Wang 2024). Inspired by BUSCO (Seppey

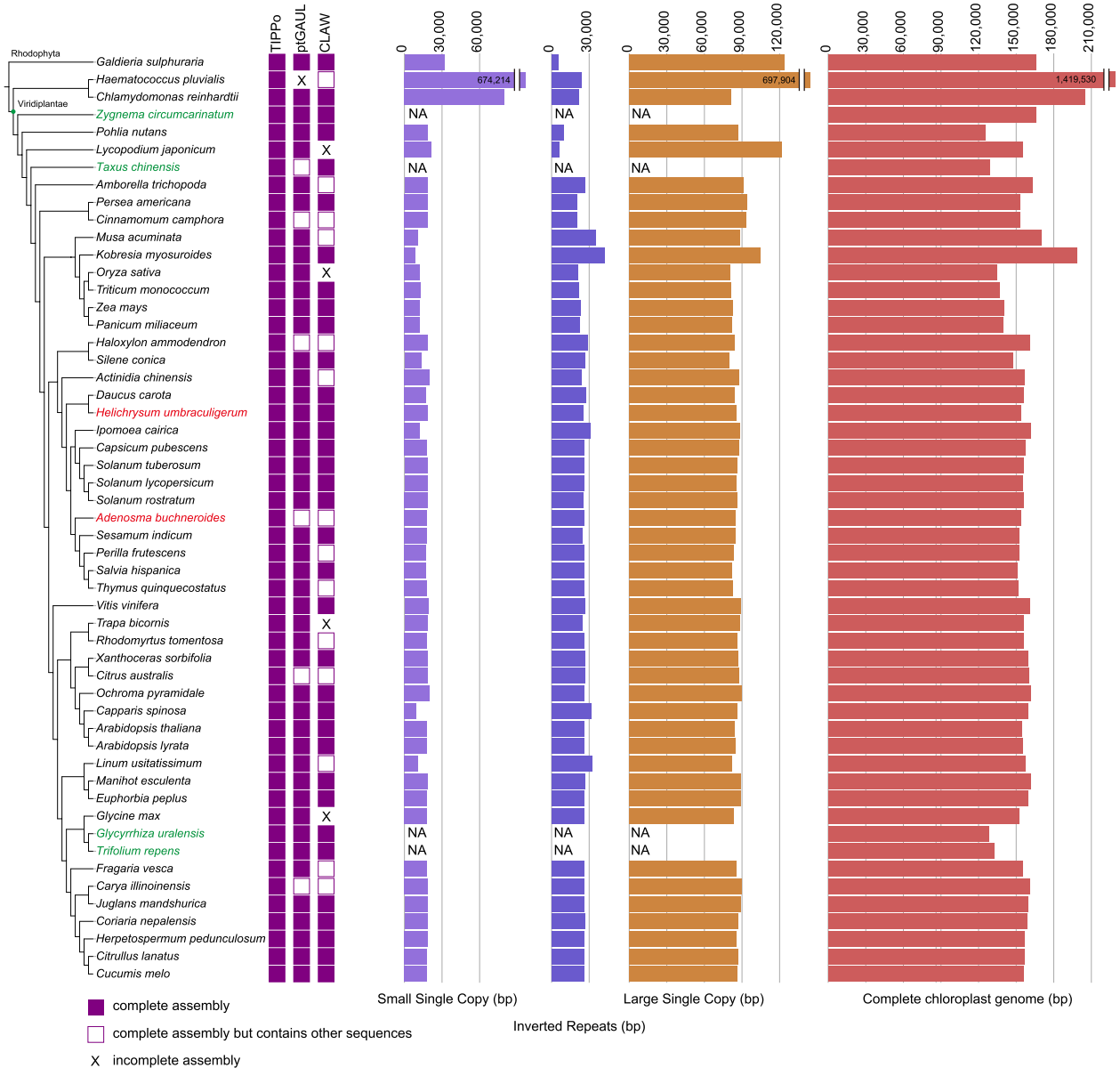


Fig. 2. Benchmarking of four chloroplast genome assembly tools and genome statistics. See Materials and Methods for phylogenetic tree. The assemblies for *Adenosma buchneroides* and *Helichrysum umbraculigerum* are presented here for the first time. *Zygnema circumcarinatum*, *Taxus chinensis*, *Glycyrrhiza uralensis*, and *Trifolium repens* have lost IRs, and the three topologically defined regions are therefore not measured.

et al. 2019) for assessing the completeness for nuclear genomes, we use 41 protein-coding genes collected by mitopy (Alverson et al. 2010) to evaluate the completeness of mitochondrial assemblies. Out of the 53 species, 35 mitochondrial genomes had previously been published, which we also included in our evaluation (supplementary tables S5 to S8, Supplementary Material online). Considering that the output of mitochondrial genomes from PMAT and TIPPo is in the form of assembly graphs, where large repetitive fragments are represented only once, we focused on the presence or absence of genes, and did not consider orientation or copy number.

As shown in Fig. 3a, TIPPo and PMAT are in agreement regarding the completeness of protein-coding genes in the mitochondrial assemblies of 43 species. The results based on protein-coding genes are consistent with alignments to the published mitochondrial genomes (supplementary table S16, Supplementary Material online). In eight species, the mitochondrial genomes assembled by TIPPo had higher protein-coding genes completeness, while for two species, PMAT outperformed TIPPo. For *M. acuminata*, both TIPPo and PMAT failed to assemble the mitochondrial genome.

The seven species in which TIPPo was superior include one red alga and two Chlorophyta for which PMAT failed to output mitochondrial genomes, with the TIPPo assemblies matching the published assemblies for these three species. Although *Haematococcus lacustris* and *Haematococcus pluvialis* belong to the same genus, their mitochondrial genomes exhibit poor synteny (supplementary fig. S10, Supplementary Material online).

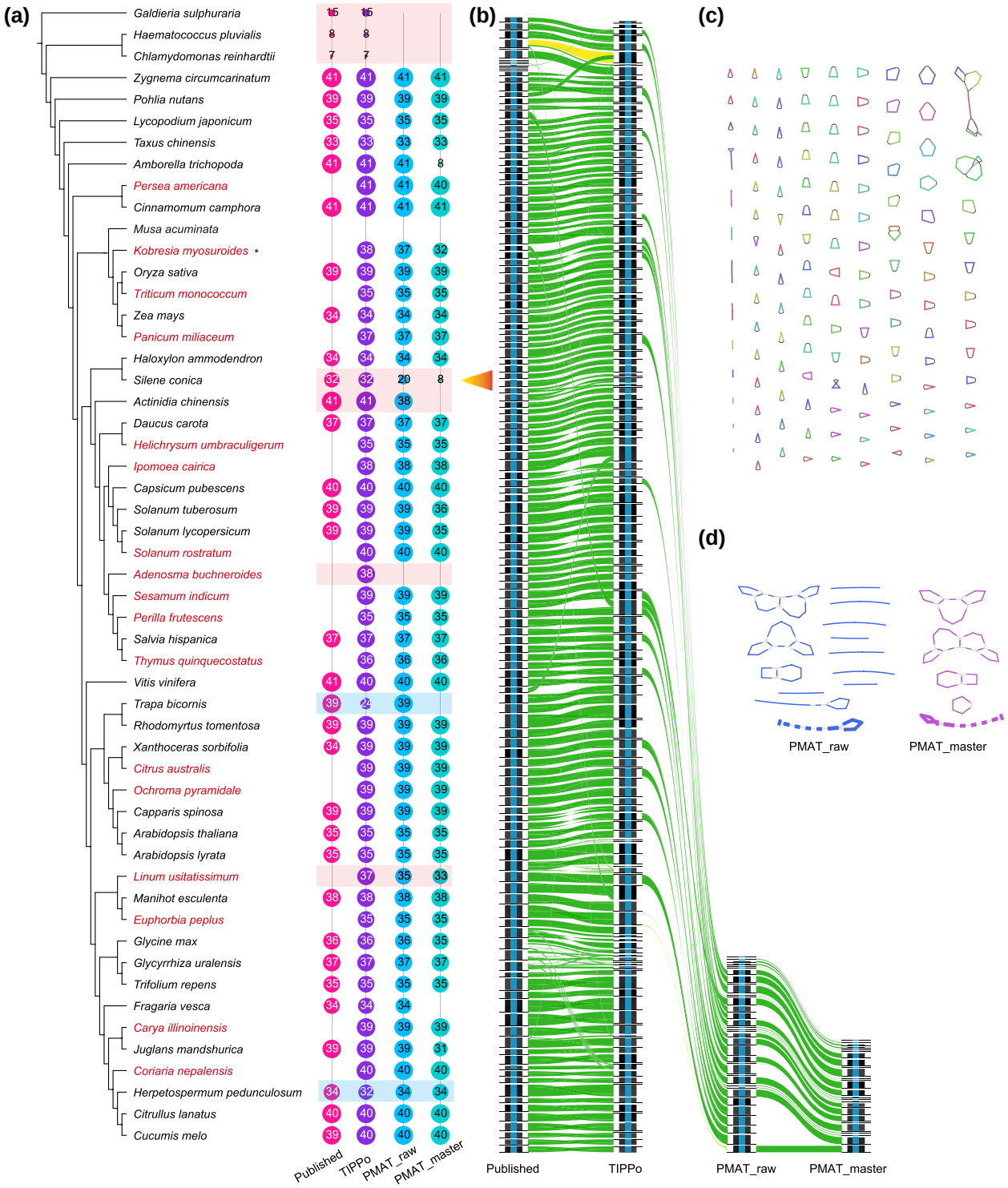


Fig. 3. Benchmarking of mitochondrial genome assembly. a) See Materials and Methods for phylogenetic tree. The assemblies for *Persea americana*, *Kobresia myosuroides*, *Triticum monococcum*, *Panicum miliaceum*, *Helichrysum umbraculigerum*, *Ipomoea cairica*, *Solanum rostratum*, *Adenosma buchneroides*, *Sesamum indicum*, *Perilla frutescens*, *Thymus quinquecostatus*, *Citrus australis*, *Ochroma pyramidale*, *Linum usitatissimum*, *Euphorbia peplus*, *Carya illinoinensis*, and *Coriaria nepalensis* are presented here for the first time. The numbers inside the circles indicate the number of nonredundant protein-coding genes in the assembly. Light shading indicates superior results with TIPPo or PMAT. b) Whole-genome alignment, including the published, TIPPo and PMAT assemblies (both raw and master), of the *S. conica* mitochondrial genome, visualized with Alitv (v1.0.6). c) TIPPo assembly graph of *S. conica* visualized with Bandage (v0.9.0). d) PMAT assembly graph of *S. conica* visualized with Bandage (v0.9.0).

In *S. conica*, which has one of the largest mitochondrial genomes (11 Mb), TIPPo assembled a mitochondrial genome that was highly consistent with the published genome (Fig. 3b, supplementary fig. S17, Supplementary Material online). PMAT, in contrast, only assembled parts of the mitochondrial genome. The mitochondrial assembly

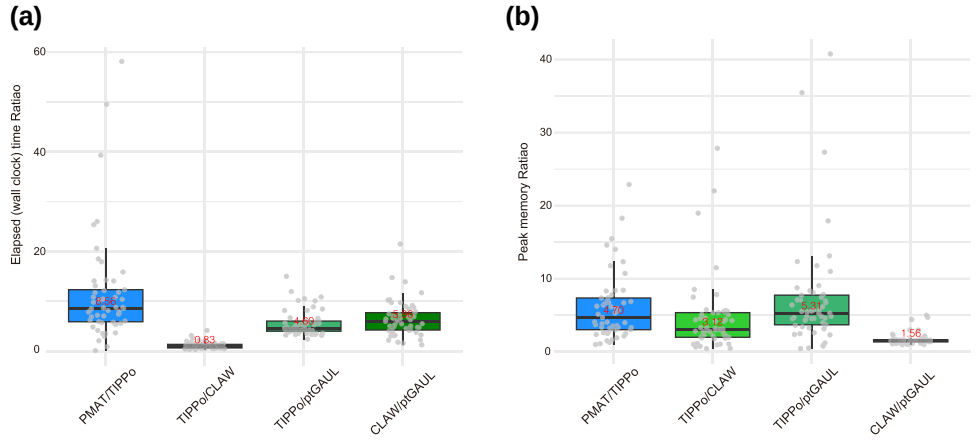


Fig. 4. Computational cost. a) Ratio of elapsed times between each pair of the four tools. b) Ratio of peak memory usage between each pair of the four tools. Gray dots indicate different species. The means are shown as horizontal lines, with the upper and lower box indicating the interquartile range (IQR), and the whiskers extending to the most extreme values within 1.5 times the IQR from the first and third quartiles.

graph from TIPPo had numerous small circular DNAs (Fig. 3c), which PMAT failed to identify (Fig. 3d). A similar issue with missing small circular DNAs in PMAT occurred in *Actinidia chinensis* and *Linum usitatissimum*. TIPPo assembly of *Ac. chinensis* matched the published genome, which includes a large circular DNA of 724 kb and a smaller circular DNA of 200 kb, whereas PMAT only generates a linearized sequence of the large circle (supplementary fig. S12, Supplementary Material online). In *L. usitatissimum*, the PMAT assembly had lost two protein-coding genes, *rpl5* and *rps14*, which are present in a circular DNA sequence assembled by TIPPo. Whole-genome alignment again indicated that PMAT the assembly had lost the circular DNA with these two genes (supplementary fig. S13, Supplementary Material online). In *Ad. buchneroides*, PMAT failed to assemble the mitochondrial genome, whereas TIPPo assembled a 346 kb linear DNA sequence containing 38 protein-coding genes (supplementary fig. S14, Supplementary Material online). Given the number of protein-coding genes in related species—39 in *Sesamum*, 35 in *Perilla*, 37 in *Salvia*, and 36 in *Thymus*—this suggests that the linear DNA sequence from TIPPo is largely complete. When using low sequence depth data as inputs for the assemblies (1× and 0.5× nuclear genome coverage), both PMAT and TIPPo showed varying degrees of incomplete assembly. Therefore, for assembling mitochondrial genomes, we do not recommend using ultra-low coverage (supplementary fig. S22, Supplementary Material online).

As mentioned, PMAT outperformed TIPPo for two species. For *T. bicornis*, the TIPPo assembly graph comprised only linear DNA fragments, indicating the erroneous identification of a large number of nonmitochondrial reads. Using verkko to construct a whole-genome assembly graph revealed that *T. bicornis* possesses a large rDNA cluster that is misidentified by TIARA (supplementary fig. S15, Supplementary Material online). For *Herpetospermum pedunculosum*, the TIPPo assembly lacked two genes, *nad3* and *atp6*, due to overfiltering by the *k*-mer approach (supplementary fig. S19, Supplementary Material online). However, the PMAT raw assembly included nonmitochondrial fragments (supplementary fig. S16, Supplementary Material online).

Computational Cost

Using data from 53 species, we performed chloroplast genome assembly with 3 different tools: TIPPo (chloroplast mode), ptGAUL, and CLAW, all with default parameters. Our results show that both TIPPo and CLAW are approximately five times slower than ptGAUL (Fig. 4a). Regarding peak memory usage, TIPPo required the most memory, consuming three times more than CLAW and five times more than ptGAUL (Fig. 4b). For mitochondrial genome assembly, we utilized PMAT in mt mode and TIPPo in organelle mode. PMAT was approximately eight times slower than TIPPo and consumed four times more memory (Fig. 4a and b). For detailed time and memory usage at the different coverages used, please refer to supplementary tables S9 and S10, Supplementary Material online.

Identification of NUPTs/NUMTs

Next, we wanted to know whether we could improve on the accurate identification of NUPTs and NUMTs and the elimination of potential contamination of nuclear assemblies with pieces of organellar genomes. High-quality nuclear genomes assembled from PacBio HiFi data are available for all of the species used in this study except *Lycopodium japonicum*, *Ochroma pyramidale*, and *Perilla frutescens*, with the assemblies of the latter two being highly fragmented. Because algal genomes are small and have very few NUPTs and NUMTs (Zhang et al. 2020), we excluded them from further analysis. *Musa acuminata* was not included either, because we had not been able to assemble the mitochondrial genome. For all other 45 nuclear genome assemblies, we retrieved all contigs/scaffolds over 500 kb.

The species with the longest cumulative lengths of NUMTs were *S. conica*, *Amborella trichopoda*, *Triticum monococcum*, *Capsicum pubescens*, and *Taxus chinensis*. This might be attributed to *S. conica* and *Am. trichopoda*

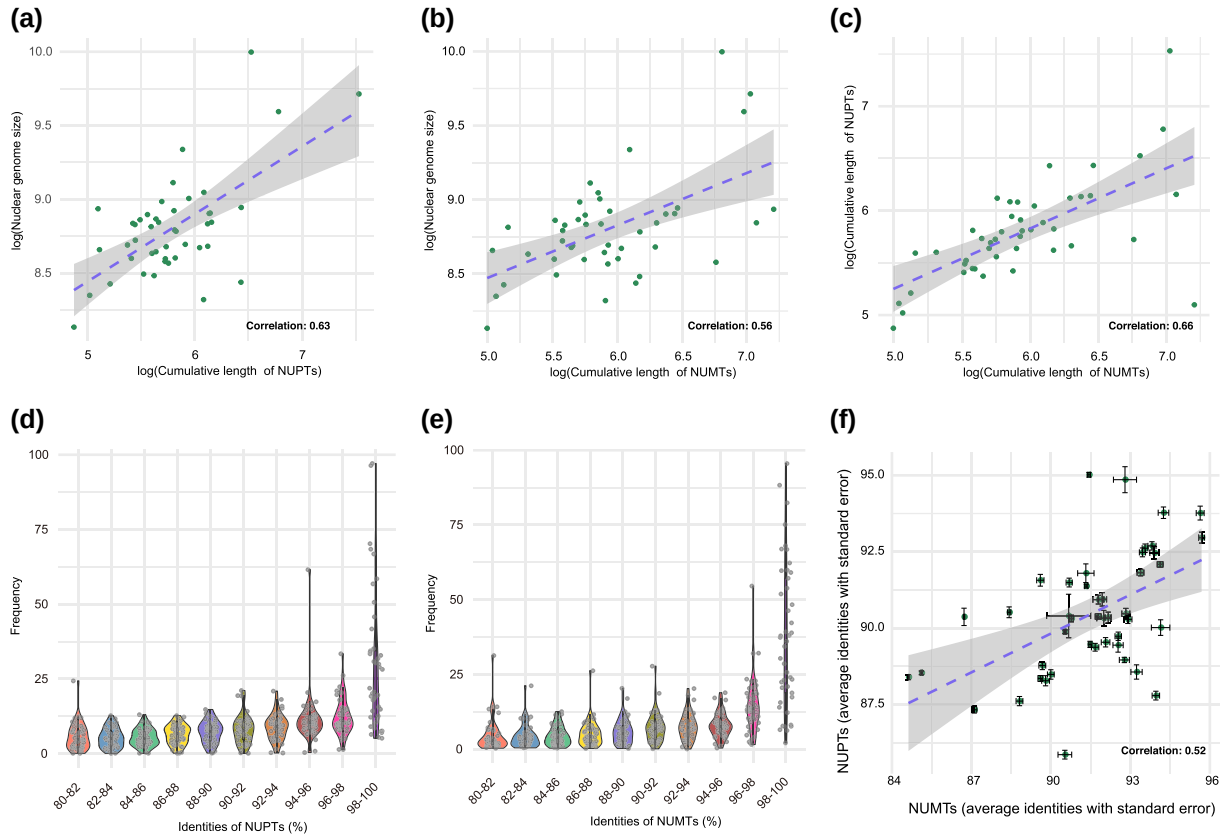


Fig. 5. Comparison of NUPT and NUMT sequences and the corresponding organellar genomes. a) Comparison of cumulative lengths of NUPTs and of nuclear genome size. b) Comparison of cumulative lengths of NUMTs and of nuclear genome size. c) Comparison of cumulative lengths of NUPTs and of NUMTs. d) Cumulative length distribution of NUPTs across different identities. e) Cumulative length distribution of NUMTs as a function of sequence identity with the corresponding mitochondrial genome. f) Correlation between NUPT/chloroplast genome identity and NUMT/mitochondrial genome identity. Bars indicate standard errors.

having large mitochondrial genomes (11 and 3.9 Mb) and *Tri. monococcum*, *C. pubescens*, and *Ta. chinensis* having large nuclear genomes (5, 3.9, and 10 Gb). The latter three species also had the highest cumulative lengths of NUPTs (supplementary table S11, Supplementary Material online). As observed before (Zhang et al. 2020), both NUPT and NUMT lengths are positively correlated with nuclear genome size in plants (Pearson's correlation coefficients of 0.63 and 0.56; Fig. 5a and b). However, no significant association was found when comparing species across different kingdoms (Richly and Leister 2004a,b). Since NUPTs and NUMTs are part of the nuclear genome, their lengths are also positively correlated (Fig. 5c).

NUPTs and NUMTs appear to evolve mostly neutrally, as evidenced by the gradual accumulation of mutations (Huang et al. 2005; Noutsos et al. 2005). Because the substitution rates of plant organellar genomes is typically an order of magnitude lower than that of nuclear genomes (Wolfe et al. 1987; Drouin et al. 2008), the number of differences between NUPT and NUMT sequences and the corresponding organellar genomes reflect the age of nuclear insertions (Richly and Leister 2004a,b; Michalovova et al. 2013; Yoshida et al. 2019). We found that recent insertion events, with sequence identities of 98% to 100%, are most frequent (Fig. 5d and e, supplementary table S12, Supplementary Material online), which is also reflected by the correlation of average sequence identities between NUPTs and NUMTs and their organellar genomes being well correlated (Pearson's correlation coefficient = 0.52; Fig. 5f). We conclude that NUPTs and NUMTs tend to degrade rapidly, which is consistent with individual NUPTs and NUMTs in *A. thaliana* genomes having low allele frequencies (Igolkina et al. 2024).

Substitution Spectra of NUPTs/NUMTs

C:G > T:A substitutions dominate the substitution spectrum in *A. thaliana* mutation accumulation lines, both in the greenhouse and in the wild, although not in older natural populations (Ossowski et al. 2010; Cao et al. 2011; Exposito-Alonso et al. 2018; Weng et al. 2019). The excess of C:G > T:A substitutions has been attributed to spontaneous deamination of methylated cytosines (Ossowski et al. 2010), which is found in plants in three contexts, CG, CHG, and CHH, with most of it in the CG context (Law and Jacobsen 2010). Previous studies have found that C:G > T:A substitutions to be the most common substitutions in NUPTs and NUMTs (Huang et al. 2005; Rousseau-Gueutin et al. 2011; Fields et al. 2022). We confirm this phenomenon in our set of 45 species, with the highest substitution rates

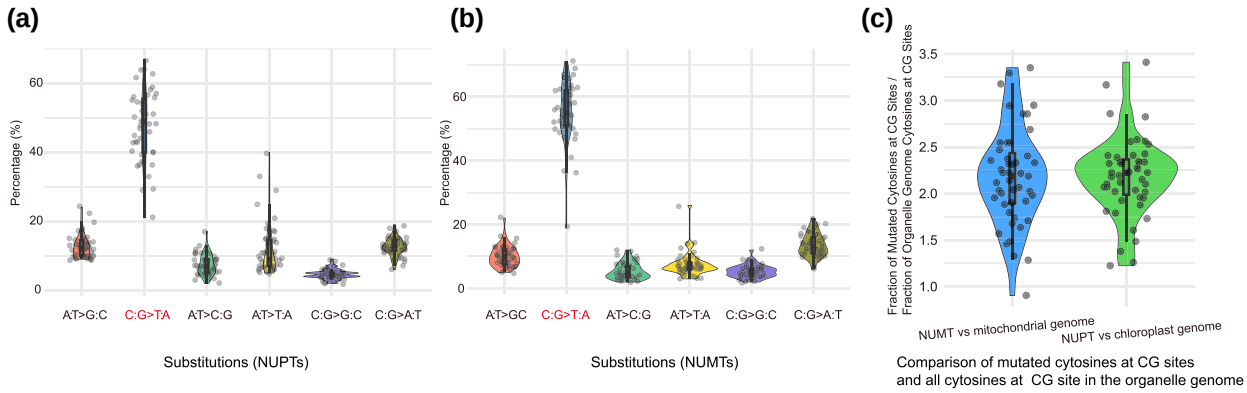


Fig. 6. The landscape of substitutions in NUPTs and NUMTs. a) Distribution of nucleotide substitutions in NUPTs, inferred from sequence comparison with the corresponding chloroplast genome. b) Distribution of nucleotide substitutions in NUMTs, inferred from sequence comparison with the corresponding mitochondrial genome. c) Enrichment of cytosine substitutions in NUPTs and NUMTs at CG sites.

at CG sites (Fig. 6, supplementary tables S13 and S14, Supplementary Material online).

Small interfering RNA Targeting NUPTs and NUMTs

The increased substitution rate at CG sites in NUPTs and NUMTs suggested that these are often methylated, which has been directly confirmed in several instances (Yoshida et al. 2014; Fields et al. 2022). The most common type of DNA methylation in plants, RNA-directed DNA methylation, is associated with small interfering RNAs (siRNAs) (Sigman and Slotkin 2016), and we therefore tested the hypothesis that NUPTs and NUMTs are enriched for siRNAs. In a previous study, siRNA data were generated for 11 of the 45 species that we investigated (Lunardon et al. 2020), and we annotated siRNA loci by mapping siRNA reads (Axtell 2013).

For all 11 species, the overlap of siRNA loci with NUPT/NUMTs was significantly higher than expected by chance (Fig. 7, supplementary table S15, Supplementary Material online), demonstrating that siRNAs are indeed enriched in NUPTs and NUMTs.

Conclusions

We introduce TIPPo, a user-friendly, reference-free approach for assembling plant organellar genomes. TIPPo provides a streamlined and universal assembly process without the need for external reference genomes. For both chloroplast and mitochondrial genomes, we provide assembly graphs. For chloroplast genomes, we provide in addition information on heteroplasmy. A limitation of our approach is that it can only use high-quality long reads, but we feel this is justified given that this technology underpins many of the ongoing large-scale genome sequencing and assembly projects (Rhie et al. 2021; Darwin Tree of Life Project Consortium 2022; Lewin et al. 2022). We also note that another newly released assembler for plant organellar genomes that comes from some of the colleagues leading these large-scale efforts is also restricted to the use of high-quality long reads (Zhou et al. 2024).

TIPPo outperforms all other tested assemblers for chloroplast genomes. Compared with chloroplast genomes, assessing the performance for mitochondrial genomes is more difficult due to the diversity of plant mitochondrial genomes. Based on the completeness of protein-coding genes, TIPPo outperforms the second-best tool PMAT (Bi et al. 2024) in eight species, while PMAT was superior for two species, *T. bicornis* and *H. pedunculosum*. A significant factor appears to be the presence of a large rDNA cluster in the nuclear genome of *T. bicornis*, which results in poor classification by Tiara (Karlicki et al. 2022), the initial tool used by TIPPo for selecting input reads for the assembly. The incomplete mitochondrial assembly of *H. pedunculosum* is likely the result of excessive filtering using a *k*-mer-based approach.

Materials and Methods

Data Sources

HiFi datasets were downloaded from publicly available databases, details refer to supplementary table S1, Supplementary Material online. The accession numbers for chloroplast and mitochondrial genomes are provided in supplementary tables S2 and S3, Supplementary Material online. A phylogenetic tree of the 53 species was constructed with rtrees (https://github.com/daijiang/rtrees; Li 2023).

Evaluation of Tiara for Read Classification

First, minimap2 (2.24-r1122) with the parameter map-hifi was used to align all HiFi reads to the *A. thaliana* (Rabanal et al. 2022) and *O. sativa* (Shang et al. 2023) reference genomes, retaining only the primary alignments. Next, Tiara (1.0.3; Karlicki et al. 2022) was used to classify HiFi reads as organellar. A 100 kb sliding window was applied to calculate the proportion of reads classified as organellar by Tiara compared with minimap2 in each window. The results were visualized using ggplot2 (3.5.1).

Parameter Selection for TIARA and Flye

For evaluating the impact of parameters on Flye, we tested: (i) default parameters; (ii) default parameters with –meta;

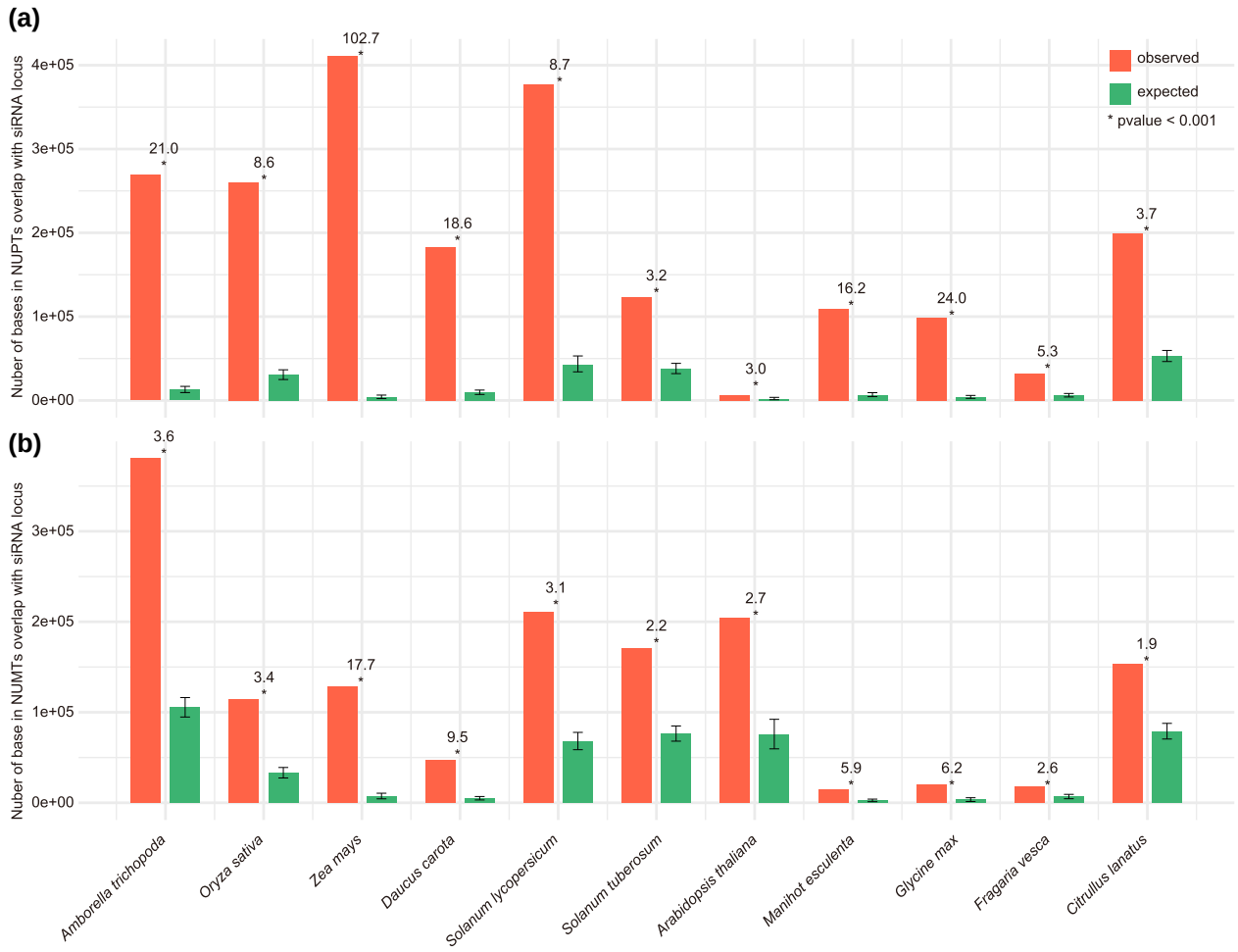


Fig. 7. Enrichment of siRNAs in NUPTs and NUMTs. a) Overlap of siRNA loci with NUPTs. b) Overlaps of siRNA loci with NUMTs. Species in (a) and (b) annotated at the bottom. The numbers on top of each bar represent the enrichment, and the error bars represent the 95% CI from random sampling of the genome.

(iii) default parameters with –keep-haplotypes; and (iv) default parameters with both –meta and –keep-haplotypes (supplementary fig. S18, Supplementary Material online). For selecting the best parameter for TIARA, we used different parameter combinations: *k*1 with 3 values (4 to 6), *k*2 with 4 values (4 to 7), and *p* with 15 values (0.3 to 1), resulting in a total of 180 combinations for reads classification.

Assembly of Organellar Genomes

We used fxTools (v0.1.0; https://github.com/moold/fxTools) for subsampling PacBio HiFi reads to approximate 4× nuclear genome coverage for each species, except for 2× for *Ta. chinensis*, which has a particularly large nuclear genome (Xiong et al. 2021). For *S. conica*, with its large mitochondrial genome (Sloan et al. 2012), we used 10× nuclear genome coverage. For *Ly. japonicum*, the sequenced data coverage is only 0.59× (Bi et al. 2024). We used identical datasets for assembly with the different tools. TIPPo (v2.1) with default parameters was used to assemble chloroplast and mitochondrial genomes simultaneously. PMAT (v1.5.3; Bi et al. 2024) is optimized for the assembly of plant mitochondrial genomes and has not been optimized for chloroplast assembly (supplementary fig. S5, Supplementary Material online). For PMAT, the auto mode was first used with the parameters -tp mt and -tp all, applied separately. Subsequently, the buildgraph mode was applied using the output from the auto mode. For ptGAUL (v1.0.5; Zhou et al. 2023) and CLAW (https://github.com/aaronphillips7493/CLAW; Phillips et al. 2024), which only assemble chloroplast genomes, the chloroplast genome sequences of closely related species were provided and run with default parameters.

Whole-Genome Alignment and Visualization

To compare genomes assembled from different sources, whole-genome alignments were performed with MiniTV (https://github.com/weigelworld/minitv), which uses minimap2 (v2.24-r1122; Li 2018) for alignment, followed by visualization with AliTV (v1.0.6) (https://alitvteam.github.io/AliTV/d3/AliTV.html; Ankenbrand et al. 2017).

Removal of Chloroplast Sequences From Mitochondrial Assemblies

First, we converted the mitochondrial assembly graphs into fragments. Given that the TIPPo chloroplast assembly

results are the cleanest and the most complete, we aligned the mitochondrial contigs from PMAT (v1.5.3; Bi et al. 2024) to the TIPPo chloroplast genome using minimap2 (2.24-r1122; Li 2018). Contigs that were covered over >90% of their length by the chloroplast genome and had >95% similarity to it were labeled as "chloroplast." Using Bandage (v0.9.0; Wick et al. 2015), we colored the nodes identified as chloroplast sequences in green and confirmed their identity after visual inspection. We removed the chloroplast sequences from the mitochondrial assemblies.

Assessing Assembly Completeness

We obtained amino acid sequence files for 41 conserved mitochondrial genes from mitopy (https://github.com/dsenalik/mitofy; Alverson et al. 2010). We used BLASTX (2.9.0+; McGinnis and Madden 2004) to align mitochondrial genome assemblies to each of the 41 genes, using a threshold of 1e−3. Considering that the current mitochondrial assembly results are presented in the format of an assembly graph, where long repeats will be collapsed into a single node, we evaluate gene completeness based on the presence or absence of genes, without accounting for their copy number.

Performance Benchmarking

All organellar genomes were assembled on an AMD EPYC 7742 processor with 64 cores and 1 TB of RAM. Runtime and peak memory usage were calculated using the/usr/bin/time -v command. All the assembly tools were set to run with 40 threads.

NUMT and NUPT Analysis

To identify NUPTs and NUMTs in the nuclear genome, we used BLASTN (2.9.0+; McGinnis and Madden 2004) with the parameters -evalue 1e−5, -dust no, -penalty -2, -word_size 9, and -outfmt 6. We aligned the chloroplast and mitochondrial genomes to their respective nuclear genomes and retained hits with an identity of >80% and a length >100 bp. Considering the redundancy in the BLASTN output, we removed all high-scoring segment pairs (HSPs) completely embedded in longer HSPs. We merged overlapping HSPs with bedtools (v2.31.1; Quinlan and Hall 2010). The identity of the merged interval in the nuclear genome to the organellar genome was calculated as the average of the identities before merging.

To identify substitutions in NUPTs and NUMTs relative to the chloroplast and mitochondrial genomes, we used minimap2 (version 2.24-r1122; Li 2018) with the parameters −paf-no-hit -ax asm5 −cs -r2k to generate alignment files. Finally, we used htsbox (version r345) (https://github.com/lh3/htsbox) with the parameters pileup -q5 -evcf to call variants.

Annotation of siRNA Loci and Overlap With NUPTs/NUMTs

For each of the selected 11 species, we downloaded data from 2 libraries. We used ShortStack (v4.0.4; Axtell 2013) with default parameters to annotate siRNA loci. In short, reads with one or no mismatch were retained, and multimapping reads were assigned to a single location with the U model. GAT (v1.3.5; Heger et al. 2013) was used to test whether the siRNA locus overlaps were greater than expected by chance with the parameter -num-samples = 1,000.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Andrea Movilli, Adrian Contreras, Yueqi Tao, Svitlana Sushko, Li He, and Haim Ashkenazy for the discussions. A.G. acknowledges support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no. INST 37/935-1 FUGG, and the de.NBI Cloud within the German Network for Bioinformatics Infrastructure and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, and W-de.NBI-022) for providing computational resources to carry out software testing in this work. This work was supported by the Max Planck Society and the Novozymes Prize of the Novo Nordisk Foundation (D.W.).

Author Contributions

W.X. designed the project, conducted the analyses, and wrote the first draft of the manuscript. I.B. set up the computational environment. I.B., Z.B., S.V., and A.G. tested the tool. D.W. supervised the project. W.X. and D.W. prepared the final manuscript with inputs from all authors.

Conflict of Interest

D.W. holds equity in Computomics, which advises plant breeders. D.W. also consults for KWS SE, a plant breeder and seed producer with activities throughout the world. All other authors declare no conflicts.

Data Availability

Chloroplast and mitochondrial assembly graphs are available on Figshare at https://doi.org/10.6084/m9.figshare.26362141.v1. TIPPo is available at Github (https://github.com/Wenfei-Xian/TIPP). Code to reproduce results from this paper can be found at Github (https://github.com/Wenfei-Xian/Reproducible_for_TIPP_paper).

References

Aldrich J, Cherney B, Merlin E, Williams C, Mets L. Recombination within the inverted repeat sequences of the *Chlamydomonas*

reinhardii chloroplast genome produces two orientation isomers. Curr Genet. 1985:**9**(3):233–238. https://doi.org/10.1007/BF00420317.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol. 2010:**27**(6):1436–1448. https://doi.org/10.1093/molbev/msq029.

Ankenbrand MJ, Hohlfeld S, Hackl T, Förster F. AliTV—interactive visualization of whole genome comparisons. PeerJ Comput Sci. 2017:**3**:e116. https://doi.org/10.7717/peerj-cs.116.

Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. RNA. 2013:**19**(6):740–751. https://doi.org/10.1261/rna.035279.112.

Bi C, Shen F, Han F, Qu Y, Hou J, Xu K, Xu L-A, He W, Wu Z, Yin T. PMAT: an efficient plant mitogenome assembly toolkit using low-coverage HiFi sequencing data. Hortic Res. 2024:**11**(3):uhae023. https://doi.org/10.1093/hr/uhae023.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 2011:**43**(10):956–963. https://doi.org/10.1038/ng.911.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021:**18**(2):170–175. https://doi.org/10.1038/s41592-020-01056-5.

Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin Tree of Life Project. Proc Natl Acad Sci U S A. 2022:**119**(4):e2115642118. https://doi.org/10.1073/pnas.2115642118.

Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017:**45**(4):e18. https://doi.org/10.1093/nar/gkw955.

Drouin G, Daoud H, Xia J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol. 2008:**49**(3):827–831. https://doi.org/10.1016/j.ympev.2008.09.009.

Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, et al. The rate and potential relevance of new mutations in a colonizing plant lineage. PLoS Genet. 2018:**14**(2):e1007155. https://doi.org/10.1371/journal.pgen.1007155.

Fields PD, Waneka G, Naish M, Schatz MC, Henderson IR, Sloan DB. Complete sequence of a 641-kb insertion of mitochondrial DNA in the *Arabidopsis thaliana* nuclear genome. Genome Biol Evol. 2022:**14**(5):evac059. https://doi.org/10.1093/gbe/evac059.

Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 2010:**6**(2):e1000834. https://doi.org/10.1371/journal.pgen.1000834.

He W, Xiang K, Chen C, Wang J, Wu Z. Master graph: an essential integrated assembly model for the plant mitogenome based on a graph-based framework. Brief Bioinform. 2022:**24**(1):bbac522. https://doi.org/10.1093/bib/bbac522.

Heger A, Webber C, Goodson M, Ponting CP, Lunter G. GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics. 2013:**29**(16):2046–2048. https://doi.org/10.1093/bioinformatics/btt343.

Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 2005:**138**(3):1723–1733. https://doi.org/10.1104/pp.105.060327.

Igolkina A, Vorbrugg S, Rabanal F, Liu H-J, Ashkenazy H, Kornienko A, Fitz J, Collenberg M, Kubica C, Morales AM, et al. Towards an unbiased characterization of genetic polymorphism. bioRxiv 596703. https://doi.org/10.1101/2024.05.30.596703, 30 May 2024, preprint: not peer reviewed.

Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020:**21**(1):241. https://doi.org/10.1186/s13059-020-02154-5.

Karlicki M, Antonowicz S, Karnkowska A. Tiara: deep learning-based classification system for eukaryotic sequences. Bioinformatics. 2022:**38**(2):344–350. https://doi.org/10.1093/bioinformatics/btab672.

Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, et al. A eukaryote without a mitochondrial organelle. Curr Biol. 2016:**26**(10):1274–1284. https://doi.org/10.1016/j.cub.2016.03.053.

Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 2017:**33**(17):2759–2761. https://doi.org/10.1093/bioinformatics/btx304.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019:**37**(5):540–546. https://doi.org/10.1038/s41587-019-0072-8.

Ladoukakis ED, Zouros E. Evolution and inheritance of animal mitochondrial DNA: rules and exceptions. J Biol Res (Thessalon). 2017:**24**(1):1–7. https://doi.org/10.1186/s40709-017-0060-4.

Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010:**11**(3):204–220. https://doi.org/10.1038/nrg2719.

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. The Earth BioGenome Project 2020: starting the clock. Proc Natl Acad Sci U S A. 2022:**119**(4):e2115635118. https://doi.org/10.1073/pnas.2115635118.

Li D. rtrees: an R package to assemble phylogenetic trees from megatrees. Ecography. 2023:**2023**(7):06643. https://doi.org/10.1111/ecog.06643.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:**34**(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Lunardon A, Johnson NR, Hagerott E, Phifer T, Polydore S, Coruh C, Axtell MJ. Integrated annotations and analyses of small RNA-producing loci from 47 diverse plants. Genome Res. 2020:**30**(3):497–513. https://doi.org/10.1101/gr.256750.119.

Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. Nat Commun. 2014:**5**(1):4104. https://doi.org/10.1038/ncomms5104.

McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004:**32**(Web Server):W20–W25. https://doi.org/10.1093/nar/gkh435.

Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity (Edinb). 2013:**111**(4):314–320. https://doi.org/10.1038/hdy.2013.51.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, et al. The genetic and epigenetic landscape of the Arabidopsis centromeres. Science. 2021:**374**(6569):eabi7489. https://doi.org/10.1126/science.abi7489.

Noutsos C, Richly E, Leister D. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Res. 2005:**15**(5):616–628. https://doi.org/10.1101/gr.3788705.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science. 2022:**376**(6588):44–53. https://doi.org/10.1126/science.abj6987.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, ClarkRM, Shaw RG, Weigel D, Lynch M. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 2010:**327**(5961):92–94. https://doi.org/10.1126/science.1180677.

Palmer JD. Chloroplast DNA exists in two orientations. Nature. 1983:**301**(5895):92–93. https://doi.org/10.1038/301092a0.

Phillips AL, Ferguson S, Burton RA, Watson-Haigh NS. CLAW: an automated Snakemake workflow for the assembly of chloroplast genomes from long-read data. PLoS Comput Biol. 2024:**20**(2): e1011870. https://doi.org/10.1371/journal.pcbi.1011870.

Putintseva YA, Bondar EI, Simonov EP, Sharov VV, Oreshkova NV, Kuzmin DA, Konstantinov YM, Shmakov VN, Belkov VI, Sadovsky MG, et al. Siberian larch (*Larix sibirica* Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome. BMC Genomics. 2020:**21**(1):654. https://doi.org/10.1186/s12864-020-07061-4.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010:**26**(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rabanal FA, Gräff M, Lanz C, Fritschi K, Llaca V, Lang M, Carbonell-Bejerano P, Henderson I, Weigel D. Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes. Nucleic Acids Res. 2022:**50**(21): 12309–12327. https://doi.org/10.1093/nar/gkac1115.

Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 2020:**21**(1):253. https://doi.org/10.1186/s13059-020-02157-2.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:**592**(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.

Richly E, Leister D. NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. Mol Biol Evol. 2004a:**21**(10): 1972–1980. https://doi.org/10.1093/molbev/msh210.

Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 2004b:**21**(6):1081–1084. https://doi.org/10.1093/molbev/msh110.

Rousseau-Gueutin M, Ayliffe MA, Timmis JN. Conservation of plastid sequences in the plant nuclear genome for millions of years facilitates endosymbiotic evolution. Plant Physiol. 2011:**157**(4): 2181–2193. https://doi.org/10.1104/pp.111.185074.

Sanchez-Puerta MV, García LE, Wohlfeiler J, Ceriotti LF. Unparalleled replacement of native mitochondrial genes by foreign homologs in a holoparasitic plant. New Phytol. 2017:**214**(1):376–387. https://doi.org/10.1111/nph.14361.

Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol. 2019:**1962**: 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14.

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022:**19**(7):823–826. https://doi.org/10.1038/s41592-022-01539-7.

Shang L, He W, Wang T, Yang Y, Xu Q, Zhao X, Yang L, Zhang H, Li X, Lv Y, et al. A complete assembly of the rice Nipponbare reference genome. Mol Plant. 2023:**16**(8):1232–1236. https://doi.org/10.1016/j.molp.2023.08.003.

Sigman MJ, Slotkin RK. The first rule of plant transposable element silencing: location, location, location. Plant Cell. 2016:**28**(2): 304–313. https://doi.org/10.1105/tpc.15.00869.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 2012:**10**(1):e1001241. https://doi.org/10.1371/journal.pbio.1001241.

Soorni A, Haak D, Zaitlin D, Bombarely A. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics. 2017:**18**(1):49. https://doi.org/10.1186/s12864-016-3412-9.

Štorchová H, Krüger M. The overview of methods for assembling complex mitochondrial genomes in land plants. J Exp Bot. 2024:**75**(17):5169–5174. https://doi.org/10.1093/jxb/erae034.

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K; Darwin Tree of Life Consortium, Formenti G, Abueg L, Torrance J, Myers EW, Durbin R, Blaxter M, et al. Mitohifi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 2023:**24**(1):288. https://doi.org/10.1186/s12859-023-05385-y.

Wang J. Plant organellar genomes: much done, much more to do. Trends Plant Sci. 2024:**29**(7):754–769. https://doi.org/10.1016/j.tplants.2023.12.014.

Wang W, Lanfear R. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. Genome Biol Evol. 2019:**11**(12):3372–3381. https://doi.org/10.1093/gbe/evz256.

Wang X, Weigel D, Smith LM. Transposon variants and their effects on gene expression in Arabidopsis. PLoS Genet. 2013:**9**(2): e1003255. https://doi.org/10.1371/journal.pgen.1003255.

Weng M-L, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. Genetics. 2019:**211**(2):703–714. https://doi.org/10.1534/genetics.118.301721.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019:**37**(10): 1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015:**31**(20): 3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Wlodzimierz P, Rabanal FA, Burns R, Naish M, Primetis E, Scott A, Mandáková T, Gorringe N, Tock AJ, Holland D, et al. Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature. 2023:**618**(7965):557–565. https://doi.org/10.1038/s41586-023-06062-z.

Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 1987:**84**(24):9054–9058. https://doi.org/10.1073/pnas.84.24.9054.

Xiong X, Gou J, Liao Q, Li Y, Zhou Q, Bi G, Li C, Du R, Wang X, Sun T, et al. The Taxus genome provides insights into paclitaxel biosynthesis. Nat Plants. 2021:**7**(8):1026–1036. https://doi.org/10.1038/s41477-021-00963-5.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 2004:**21**(5):809–818. https://doi.org/10.1093/molbev/msh075.

Yoshida T, Furihata HY, Kawabe A. Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. DNA Res. 2014:**21**(2):127–140. https://doi.org/10.1093/dnares/dst045.

Yoshida T, Furihata HY, To TK, Kakutani T, Kawabe A. Genome defense against integrated organellar DNA fragments from plastids into plant nuclear genomes through DNA methylation. Sci Rep. 2019:**9**(1):2060. https://doi.org/10.1038/s41598-019-38607-6.

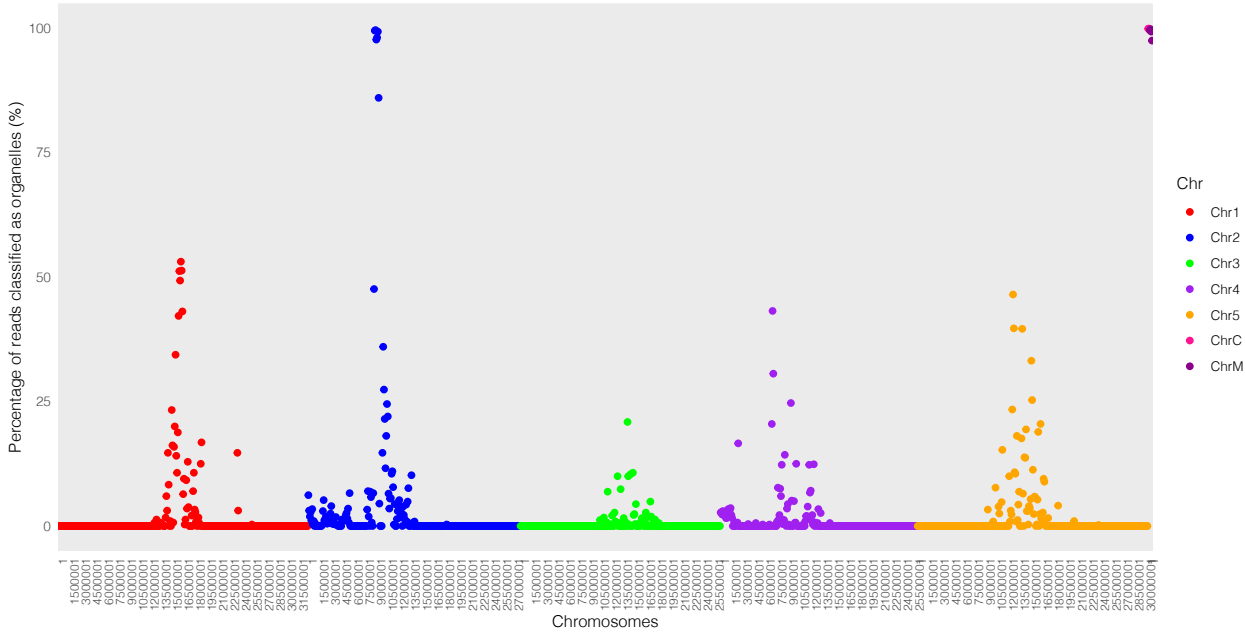
Zhang G-J, Dong R, Lan L-N, Li S-F, Gao W-J, Niu H-X. Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. Int J Mol Sci. 2020:**21**(3):707. https://doi.org/10.3390/ijms21030707.

Zhou C, Brown M, Blaxter M; The Darwin Tree of Life Project Consortium, McCarthy SA, Durbin R. Oatk: a de novo assembly tool for complex plant organelle genomes. bioRxiv 619857. https://doi.org/10.1101/2024.10.23.619857, 28 October 2024, preprint: not peer reviewed.

Zhou W, Armijos CE, Lee C, Lu R, Wang J, Ruhlman TA, Jansen RK, Jones AM, Jones CD. Plastid genome assembly using long-read data. Mol Ecol Resour. 2023:**23**(6):1442–1457. https://doi.org/10.1111/1755-0998.13787.

Zimorski V, Ku C, Martin WF, Gould SB. Endosymbiotic theory for organelle origins. Curr Opin Microbiol. 2014:**22**:38–48. https://doi.org/10.1016/j.mib.2014.09.008.

Arabidopsis thaliana - Col-0



Oryza sativa ssp. japonica cv. Nipponbare

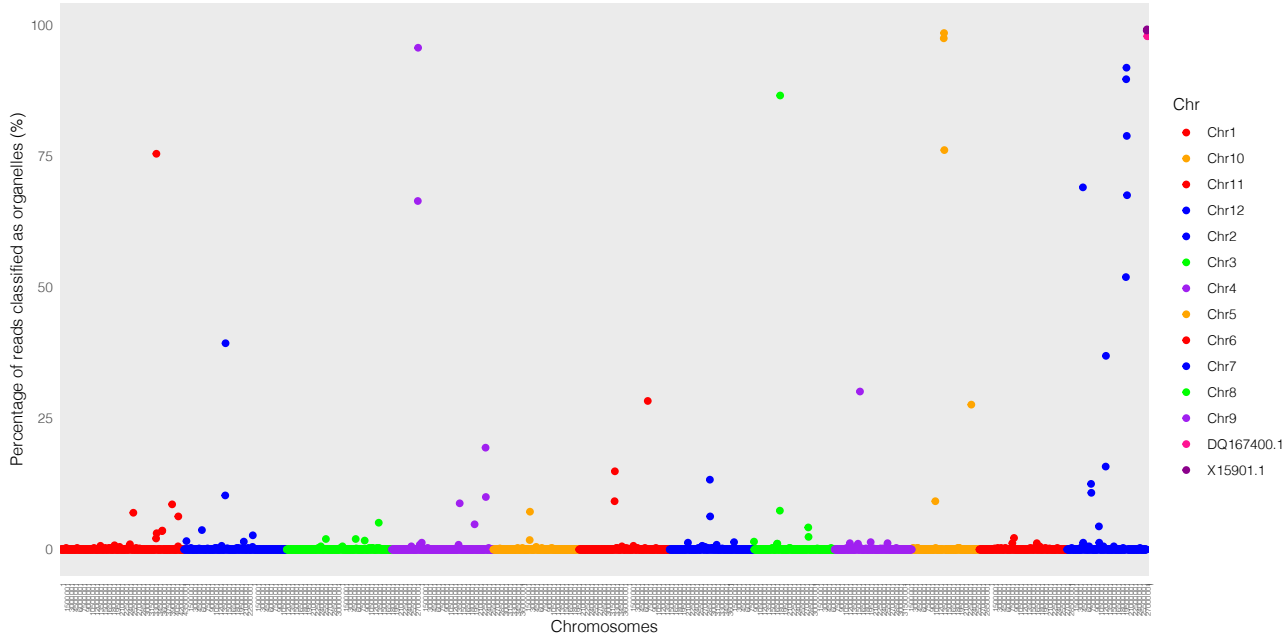


Figure S1. Testing the performance of TIARA in two plants.

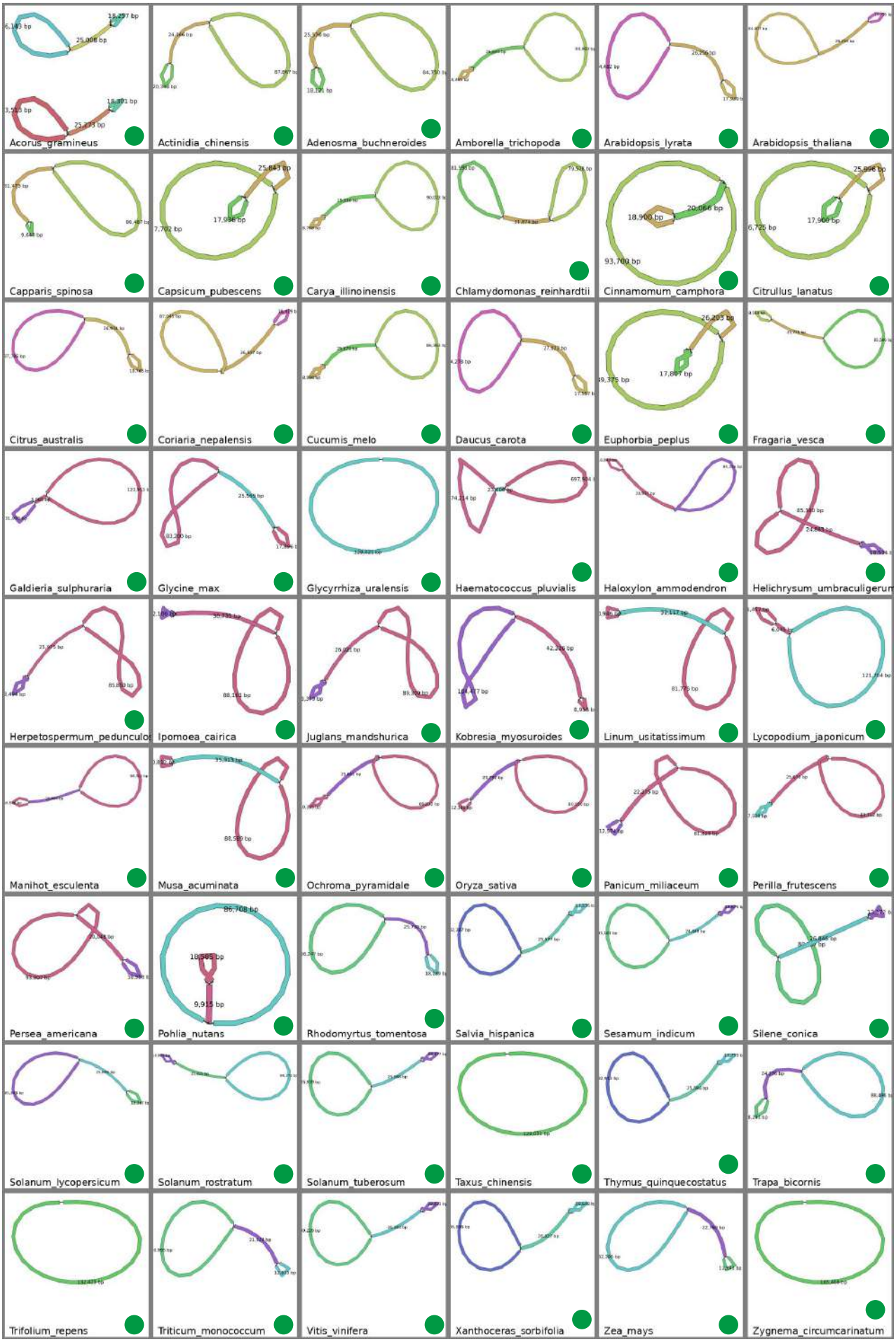


Figure S2. Chloroplast genomes of 54 species assembled using TIPPo. Green solid dots represent class 1 complete genomes. Blue solid dots represent class 2 complete genomes and other sequences. Brown solid dots represent class 3 incomplete assemblies.

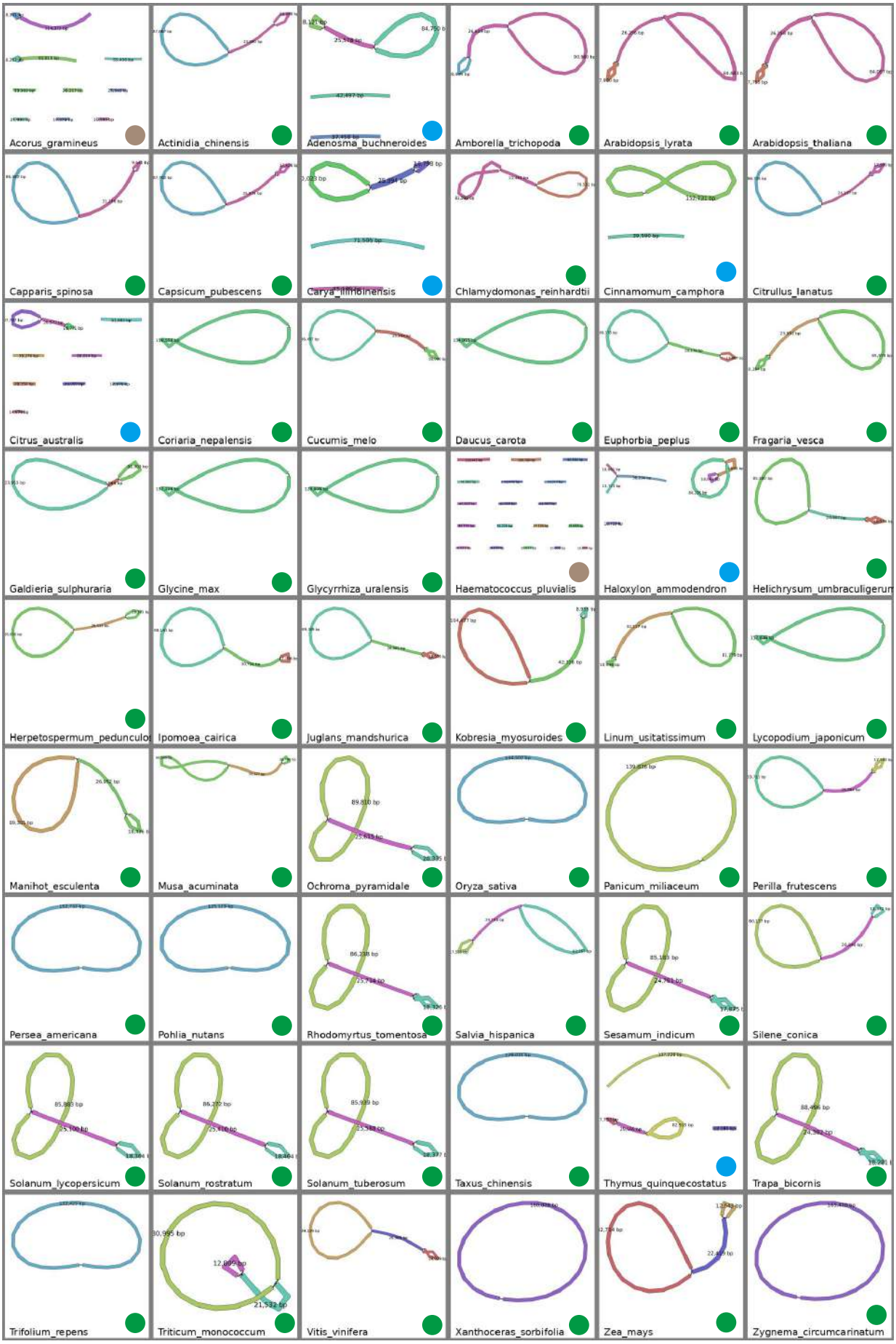


Figure S3. Chloroplast genomes of 54 species assembled using ptgaul. Green solid dots represent class 1 complete genomes. Blue solid dots represent class 2 complete genomes and other sequences. Brown solid dots represent class 3 incomplete assemblies.

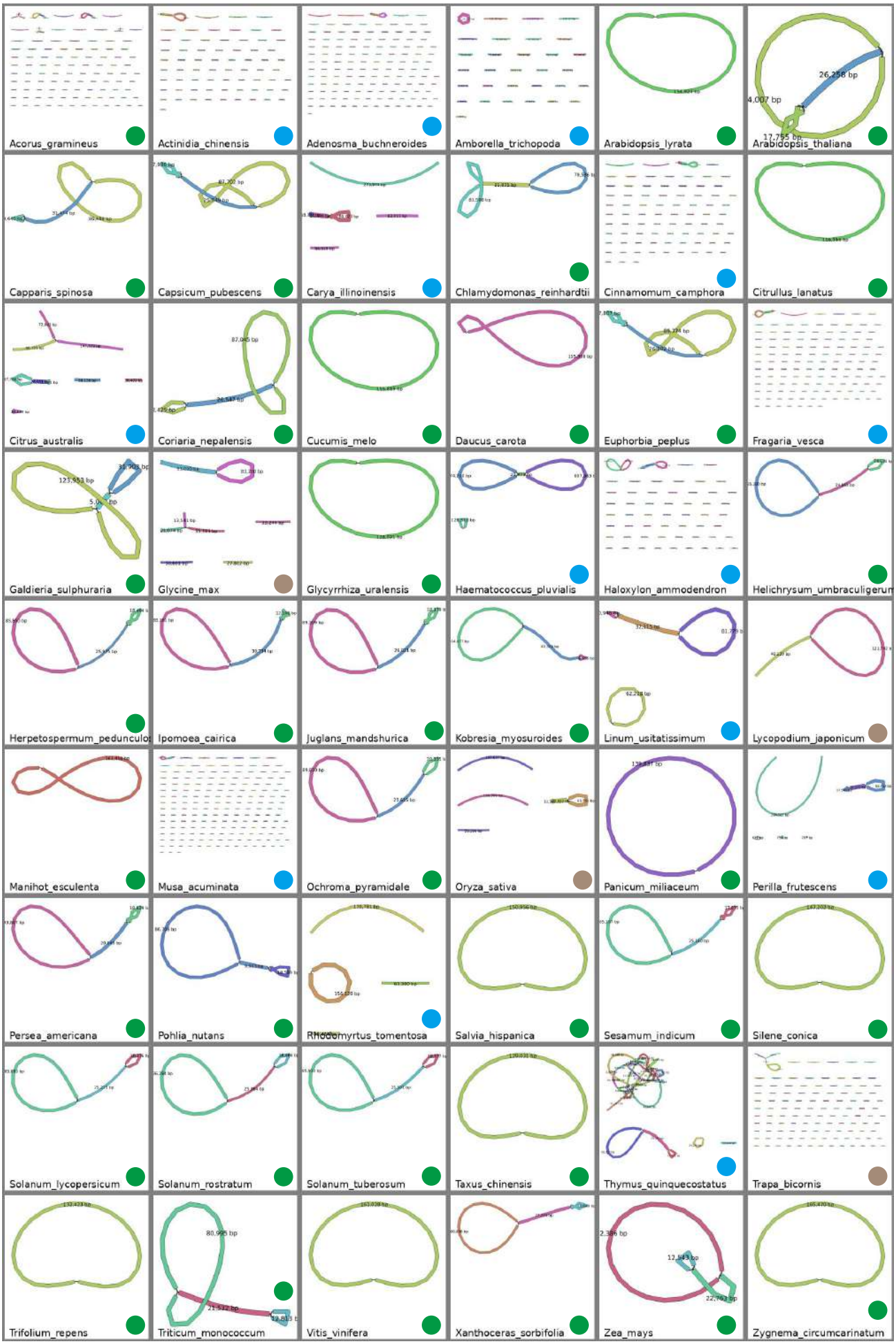


Figure S4. Chloroplast genomes of 54 species assembled using CLAW. Green solid dots represent class 1 complete genomes. Blue solid dots represent class 2 complete genomes and other sequences. Brown solid dots represent class 3 incomplete assemblies.

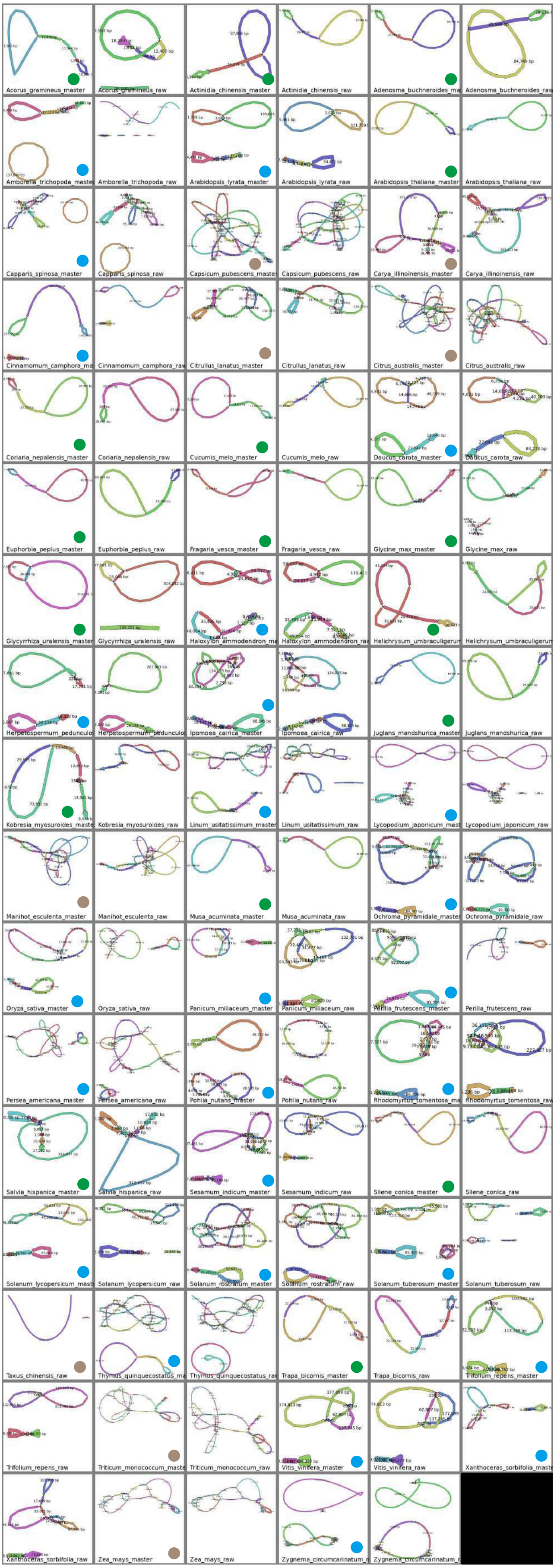


Figure S5. Chloroplast genomes of 54 species assembled using PMAT pt. Green solid dots represent class 1 complete genomes. Blue solid dots represent class 2 complete genomes and other sequences. Brown solid dots represent class 3 incomplete assemblies.

Figure S6. Whole genome alignment of chloroplast genomes. Species names ending with "TIPPo" were assembled using the TIPPo tool, while those ending with an accession ID were downloaded from NCBI.

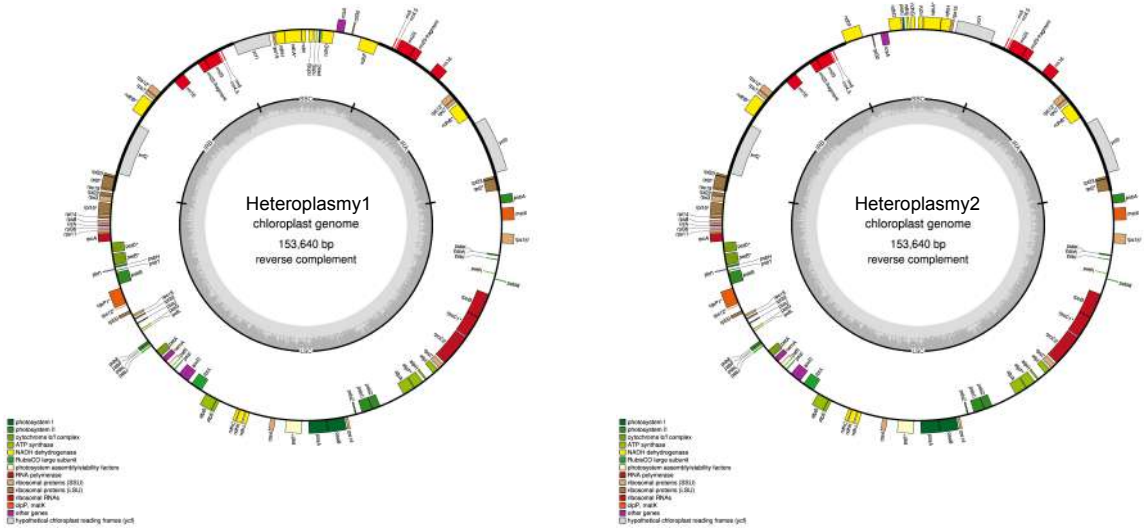


Figure S7. Chloroplast genome of *Adenosma buchneroides.*

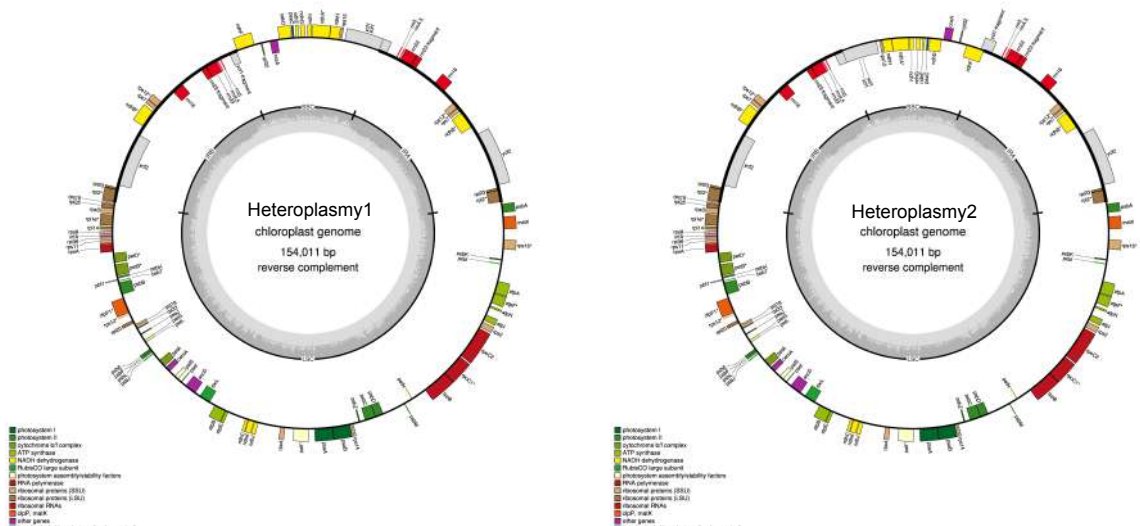


Figure S8. Chloroplast genome of *Helichrysum umbraculigerum.*

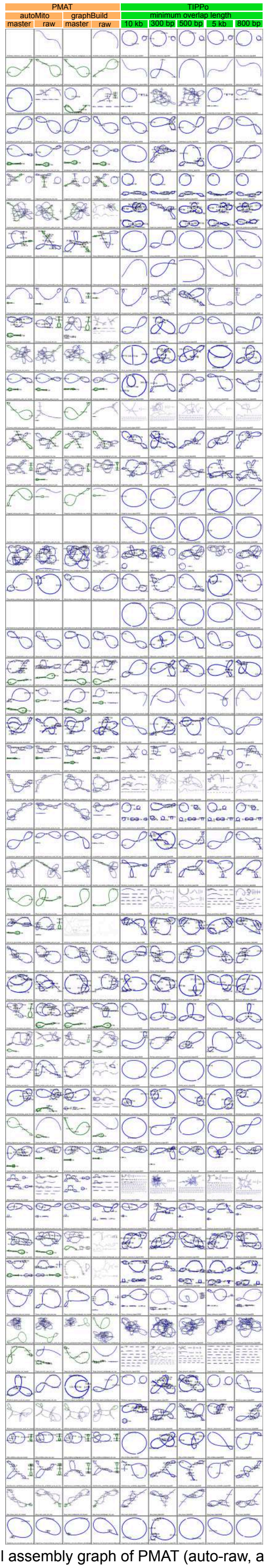


Figure S9: Mitochondrial assembly graph of PMAT (auto-raw, auto-master, graphbuild-raw, graphbuild-master) and TIPPo (minimum overlap length: 300, 500, 800, 5000 and 1000 bp). Green nodes represent chloroplast sequences, and blue nodes represent non-chloroplast sequences.

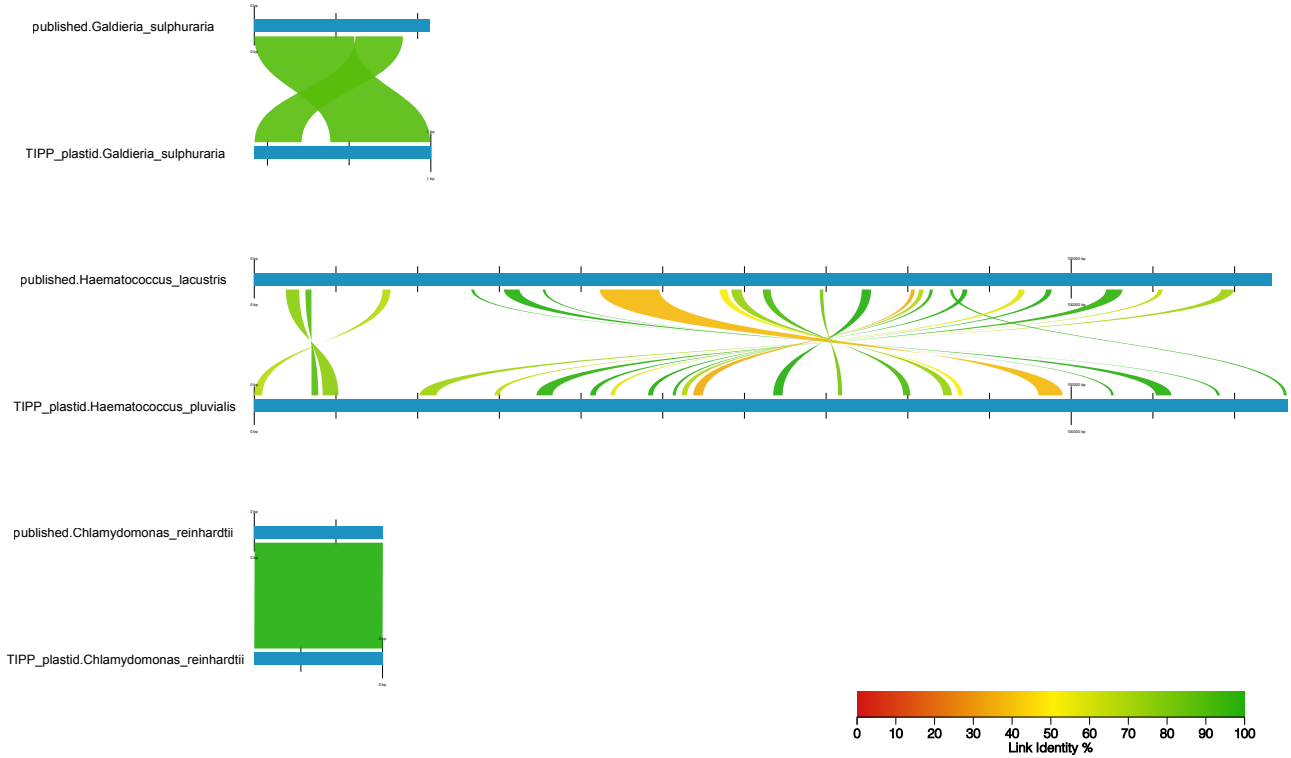


Figure S10. Whole genome alignment of mitochondrial genomes.

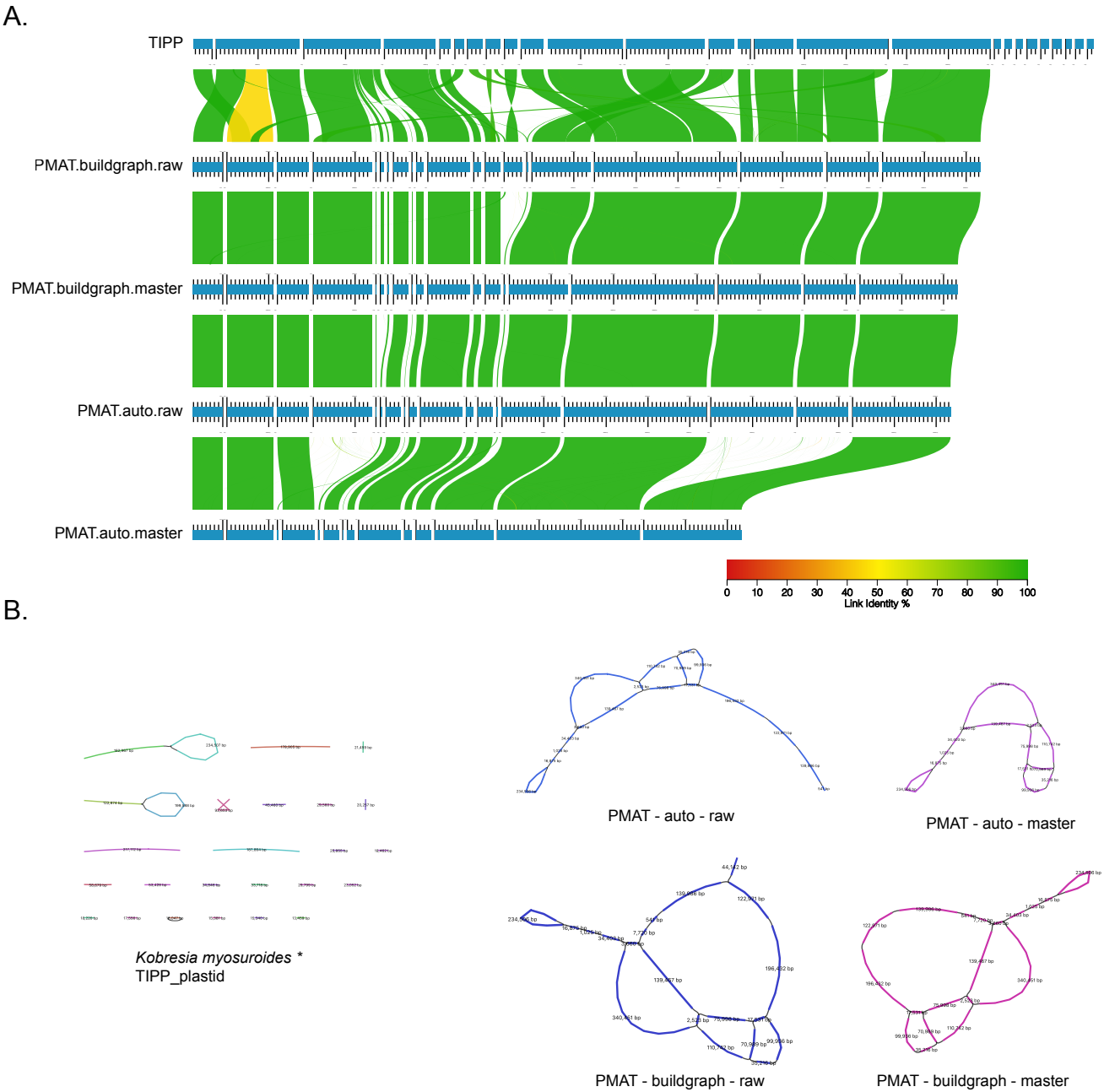
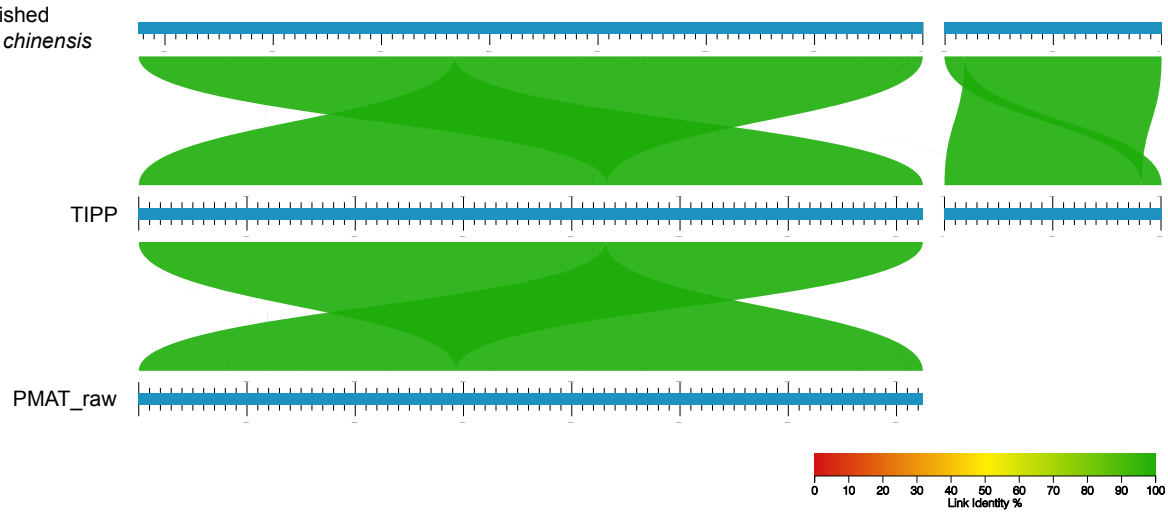
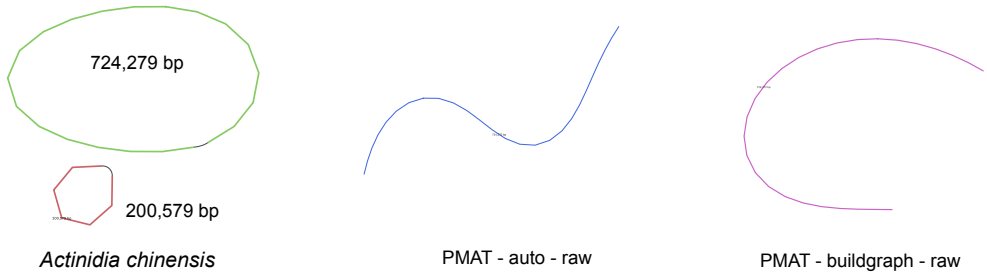


Figure S11. Mitchondrial assemblies of *Kobresia myosuroides*. A. whole genome alignment. B. visualization of assembly graphs.

A.



B.



No difference between auto mode and buildgraph mode.

C.

```
/tmp/global2/wxian/software/PMAT/bin/PMAT graphBuild -rs Actinidia_chinensis.4X.fastq.gz -o Actinidia_chinensis.PMAT_buildGraph  -gs 616M -c Actinidia_chinensis.PMAT/assembly_result/PMATContigGraph.txt
 -a Actinidia_chinensis.PMAT/assembly_result/PMATAllContigs.fna
Reading files ...
2024-09-18 11:02:21
[ INFO 2024-09-18 11:02:25 ] Contig number : 65087
[ INFO 2024-09-18 11:02:25 ] Longest Contig : contig1 724272bp
[ INFO 2024-09-18 11:02:25 ] ----------------------------------------------------------------
Candidate seeds search start ...
2024-09-18 11:02:25

BLASTn encountered an error:
Warning: [blastn] Examining 5 or more matches is recommended

[ INFO 2024-09-18 11:04:24 ] 5 contigs are used as candidate seeds
Contigs         Length          Depth
-------------------------------------
contig00001     724272bp        27.8X
contig19935     22465bp         4.0X
contig03346     137721bp        3.5X
contig37452     24411bp         2.7X
contig49476     10081bp         1.7X
[ INFO 2024-09-18 11:04:25 ] ----------------------------------------------------------------
Seeds extension start ...
2024-09-18 11:04:25
Mt Extension No.1: 100% |#################################################################################################################################################
##########################|
Seeds extension end ...
2024-09-18 11:04:25
[ INFO 2024-09-18 11:04:25 ] ----------------------------------------------------------------
[ INFO 2024-09-18 11:04:32 ] Start generating the gfa file ...
[ INFO 2024-09-18 11:04:57 ] save gfa for 23.73s
[ INFO 2024-09-18 11:04:57 ] 1 contigs are added to a raw graph
[ ERROR 2024-09-18 11:04:57 ] There is no master structure for this seeds extension result.
[ INFO 2024-09-18 11:04:57 ] Generate gfa task end.
[ INFO 2024-09-18 11:04:57 ] Task over, bye!
```

Figure S12. Mitchondrial genome assemblies of *Actinidia chinensis*.
A. whole genome alignment. B. visualization of assembly graphs. C. log of PMAT buildGraph.

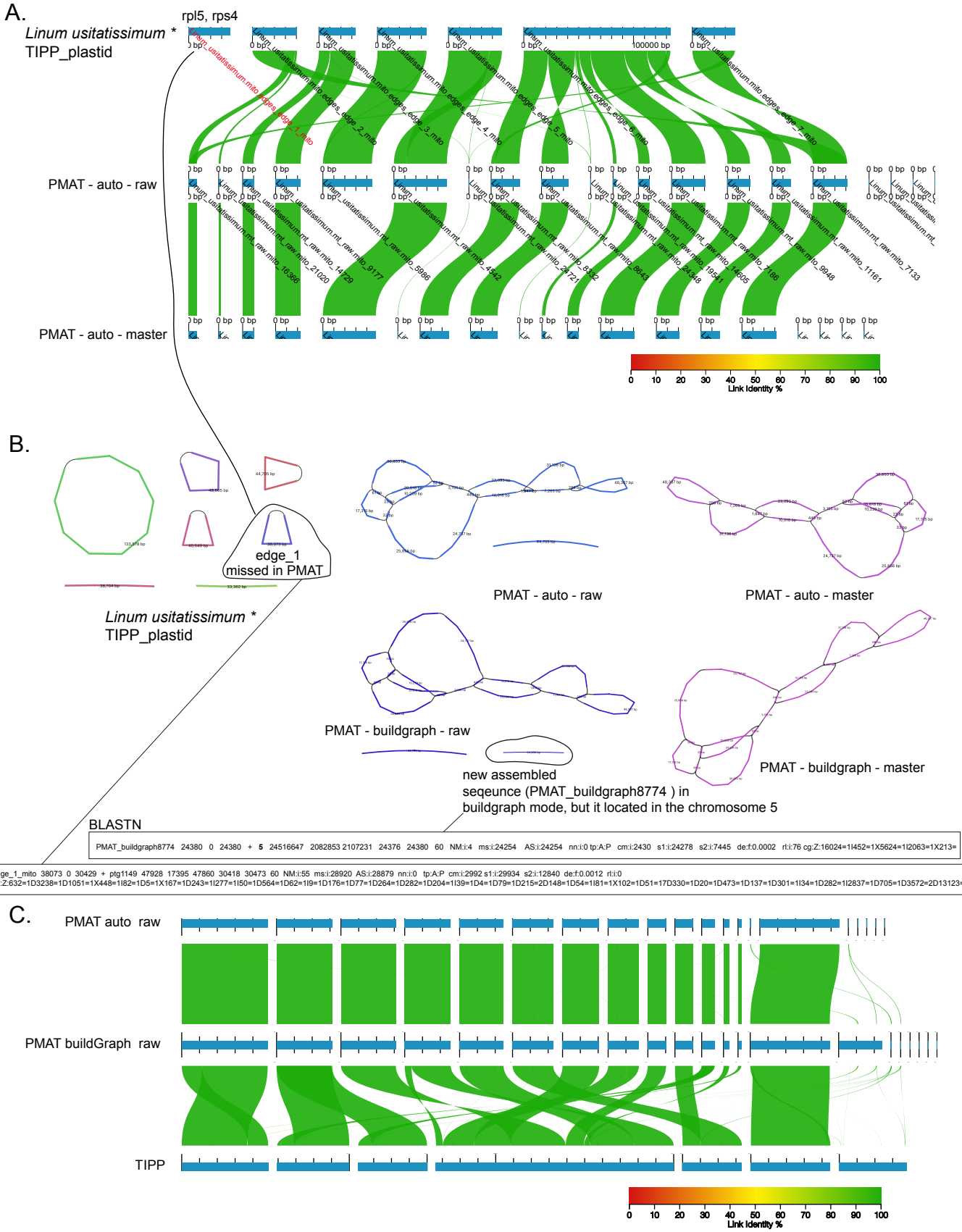
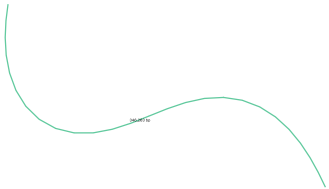
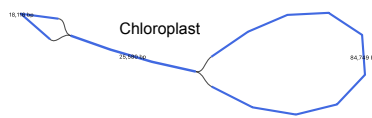
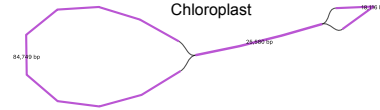


Figure S13. Mitchondrial genome assemblies of *Linum usitatissimum*.
A. whole genome alignment of TIPPo, PMAT-auto_raw and PMAT-auto-master assemblies.
B. visualization of assembly graphs. C. Whole genome alignment of TIPPo, PMAT-auto-raw and PMAT-buildgraph-raw.

A.







Adenosma buchneroides *
TIPP_plastid

PMAT - raw

PMAT - master

In buildGraph model, only chloroplast genome is presented.
No difference between auto model and buildgraph model.

B.

```
/tmp/global2/wxian/software/PMAT/bin/PMAT graphBuild -rs Adenosma_buchneroides.4X.fastq.gz -o Adenosma_buchneroides.PMAT_buildGraph  -gs 442M -c Adenosma_buchneroides.PMAT/assembly_result/
PMATContigGraph.txt -a Adenosma_buchneroides.PMAT/assembly_result/PMATAllContigs.fna
Reading files ...
2024-09-18 11:33:05
[ INFO 2024-09-18 11:33:05 ] Contig number : 5691
[ INFO 2024-09-18 11:33:05 ] Longest Contig : contig1 1637434bp
[ INFO 2024-09-18 11:33:05 ] ----------------------------------------------------------------
Candidate seeds search start ...
2024-09-18 11:33:05

BLASTn encountered an error:
Warning: [blastn] Examining 5 or more matches is recommended

[ INFO 2024-09-18 11:34:12 ] 5 contigs are used as candidate seeds
Contigs         Length          Depth
-------------------------------------
contig02570     25580bp         200.5X
contig00118     512004bp        4.4X
contig02179     19298bp         2.8X
contig03105     14421bp         1.0X
contig03809     6545bp          1.0X
[ INFO 2024-09-18 11:34:13 ] ----------------------------------------------------------------
Seeds extension start ...
2024-09-18 11:34:13
Mt Extension No.1: 100% |##########################################################################################################################################################
##########################|
Mt Extension No.2: 100% |##########################################################################################################################################################
##########################|
Seeds extension end ...
2024-09-18 11:34:13
[ INFO 2024-09-18 11:34:13 ] ----------------------------------------------------------------
[ INFO 2024-09-18 11:34:16 ] Start generating the gfa file ...
[ INFO 2024-09-18 11:34:32 ] save gfa for 14.83s
[ INFO 2024-09-18 11:34:32 ] 3 contigs are added to a raw graph
[ INFO 2024-09-18 11:34:32 ] 3 contigs are added to a master graph
[ INFO 2024-09-18 11:34:32 ] Generate gfa task end.
[ INFO 2024-09-18 11:34:32 ] Task over, bye!
```

Figure S14. Mitchondrial genome assemblies of *Adenosma buchneroides*.
A. Visualization of assembly graphs. B. log of PMAT buildGraph.

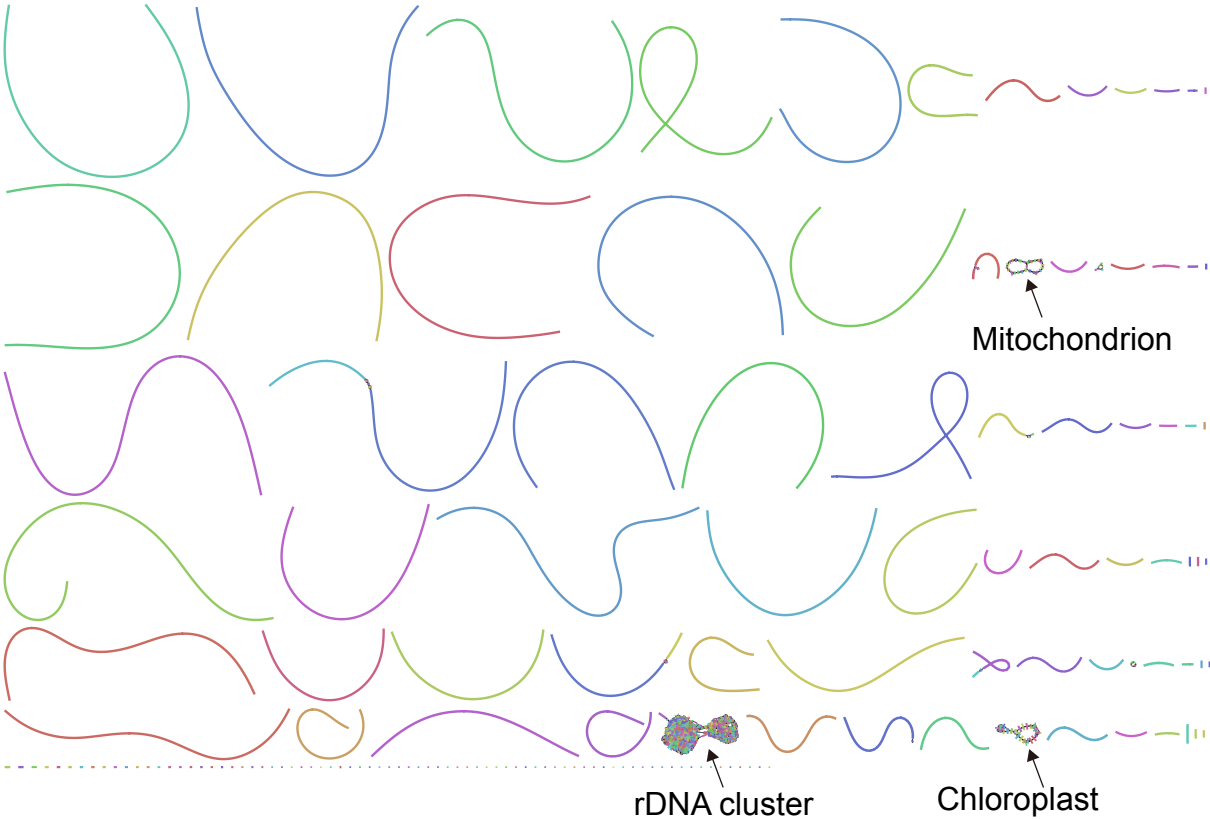


Figure S15. visualization of whole genome assembly graph of *Trapa bicornis*.

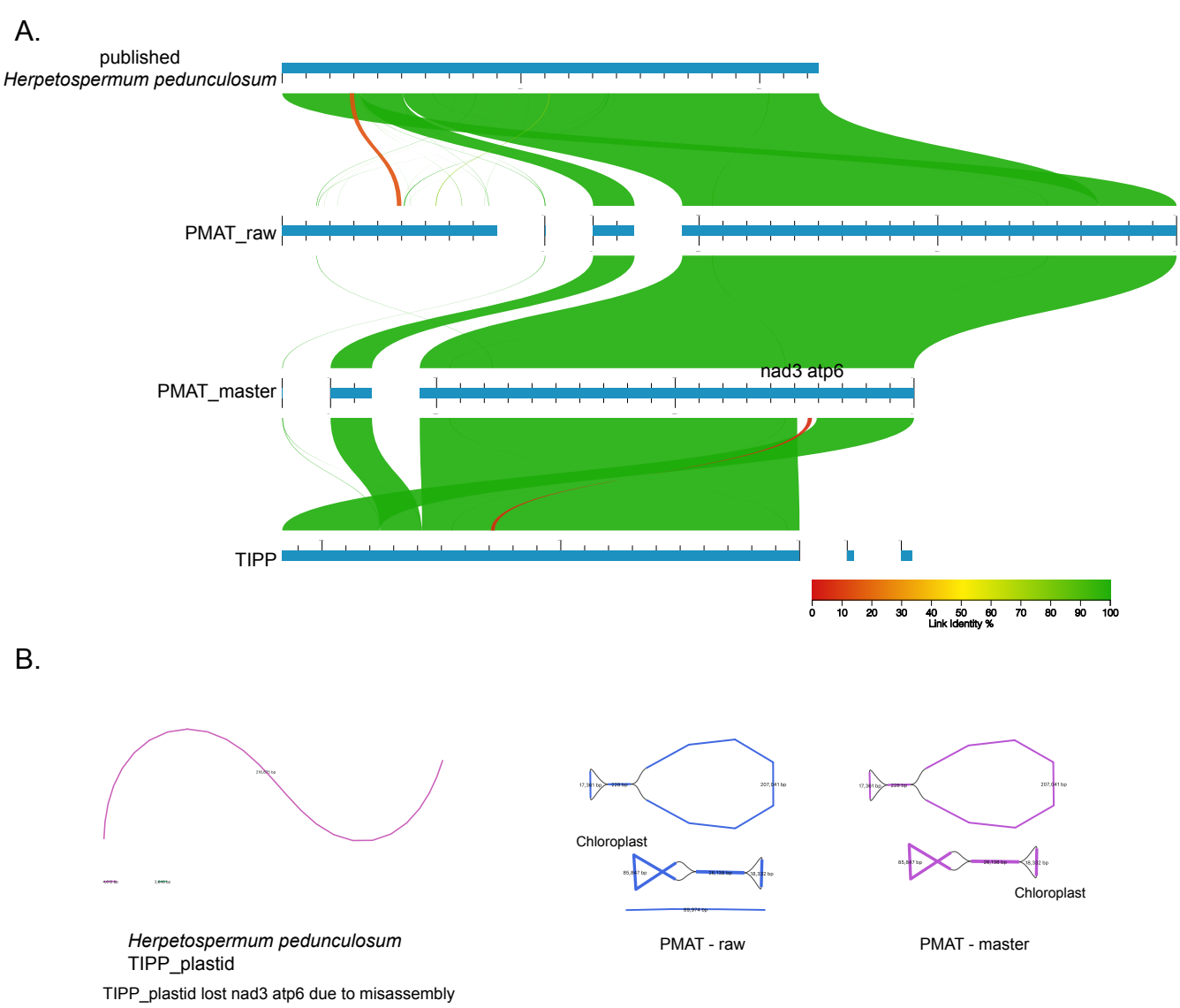


Figure S16. Mitchondrial genome assemblies of *Herpetospermum pedunculosum*. A. whole genome alignment. B. visualization of assembly graphs.

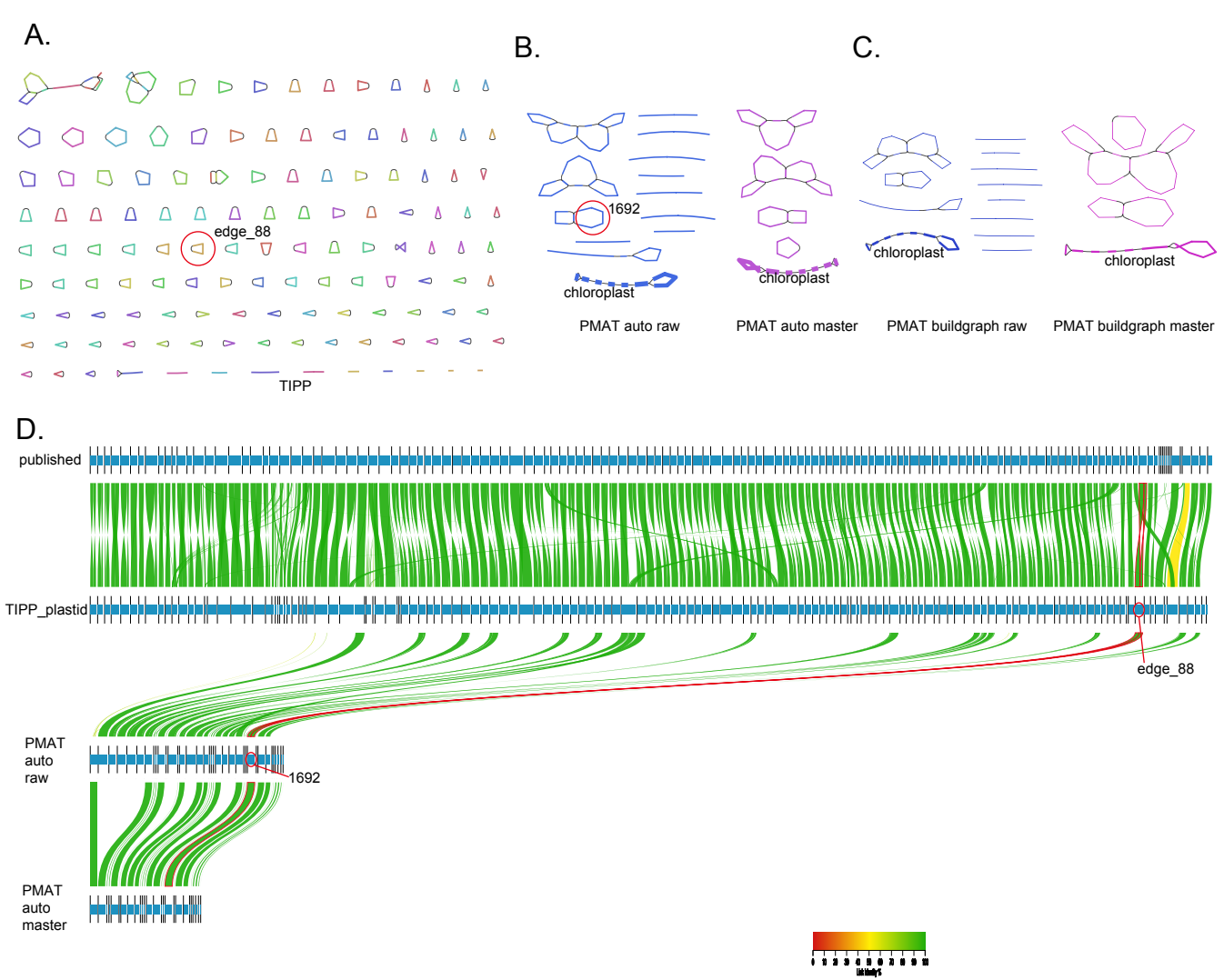


Figure S17. Assembly graph of *Silene conica.* A. assembly graph of TIPPo. B, assembly graph of PMAT auto model. C, assembly graph of PMAT buildgraph model. D. Whole genome alignment of published, TIPPo assembly, PMAT auto raw and PMAT auto master assembly. Blue bars represent the nodes in the assembly graph. Such as the node1692 in PMAT auto raw assembly show high similarity with the node edge_88 in TIPPo.

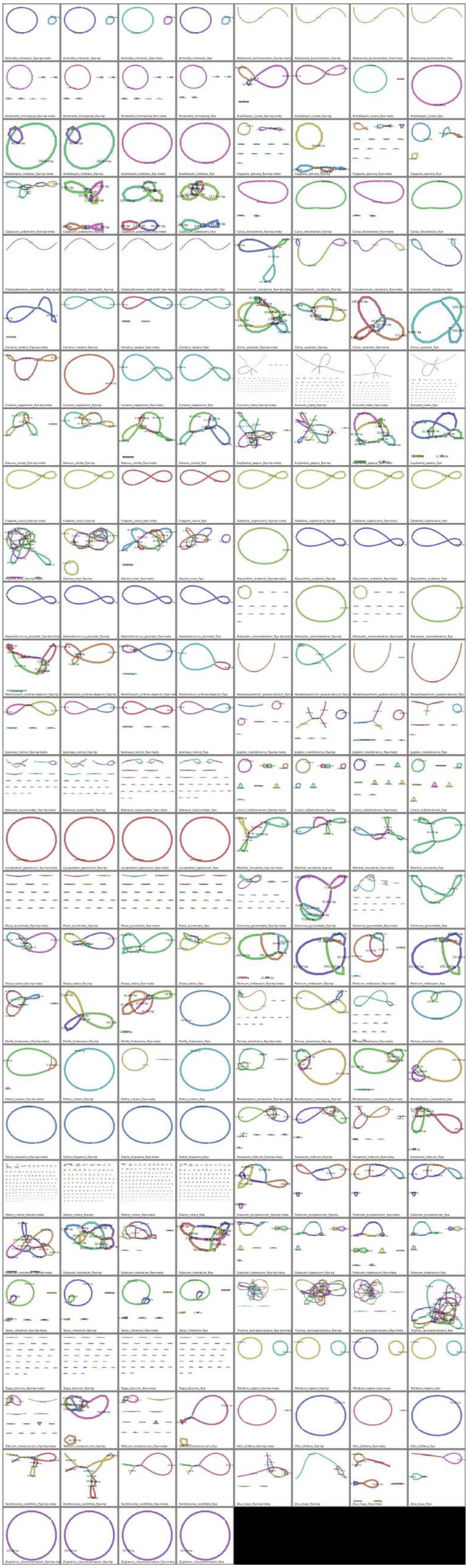


Figure S18: Mitochondrial assembly graph of TIPPo with different Flye parameters.
Flye default (species_flye), Flye --meta (species_flye-meta),
Flye --keep-haplotype (species_flye-kp), Flye --meta --keep-haplotype (species_flye-kp-meta).

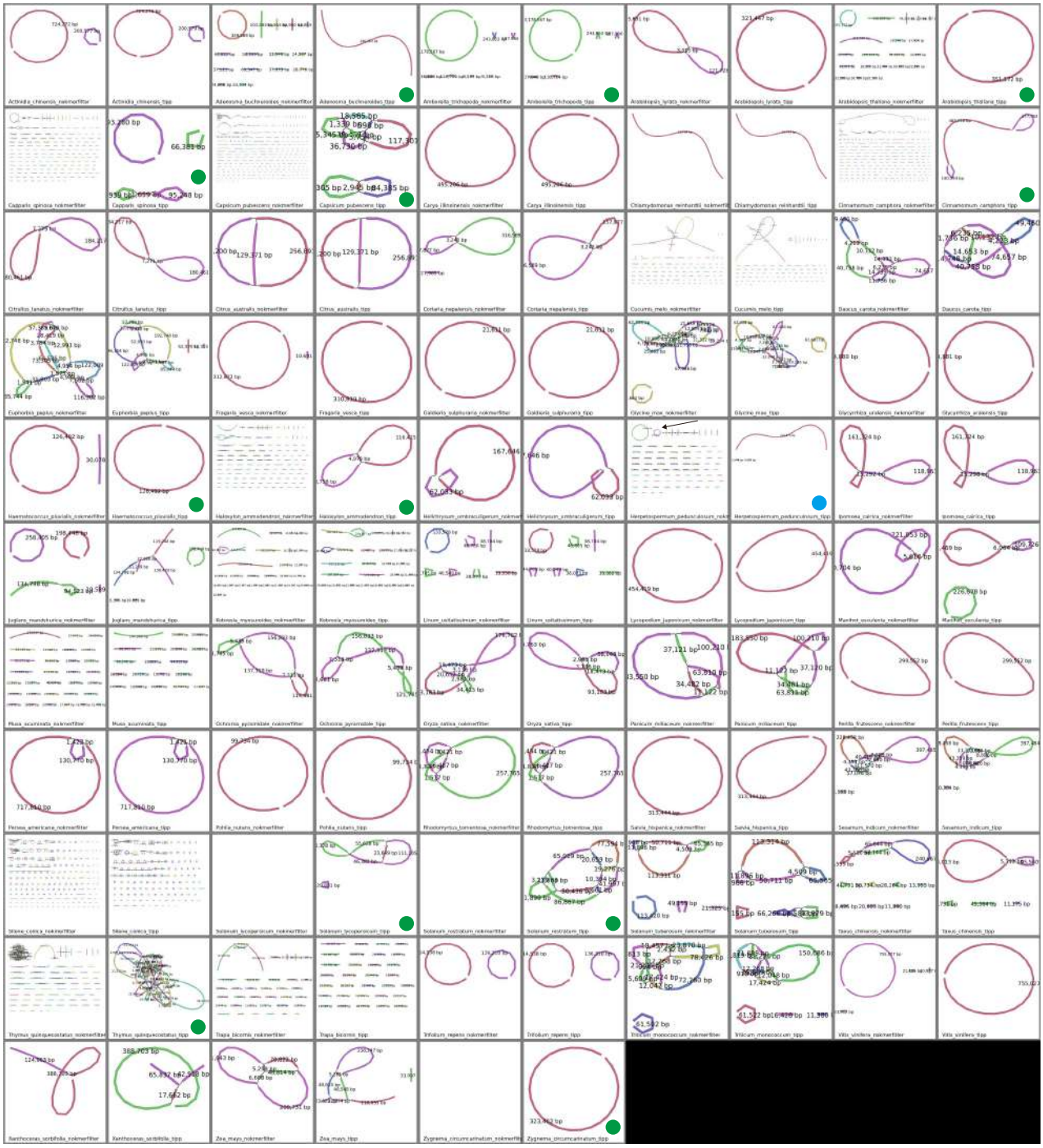
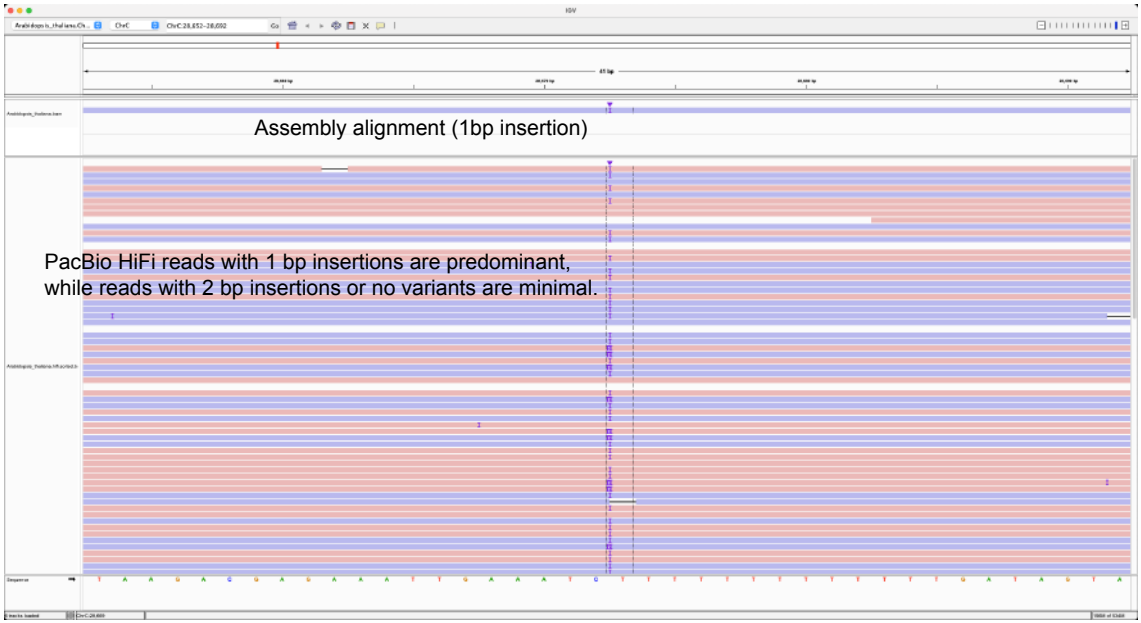


Figure S19: Mitochondrial assembly graph of TIPPo (species_tipp) and TIARA + Flye (species_nokmerfilter). Green solid dots represent that assemblies benefited from k-mer filtering. Blue solid dot represent that overfiltering by k-mer filtering.

Figure S20. *A. thaliana* mitochondrial assembly graph of Flye assembly after claffification with TIARA with different parameter combinations: k1 with 3 values (4, 5, 6), k2 with 4 values (4 to 7), and p with 15 values (0.3 to 1), resulting in a total of 180 combinations.

A.



B.

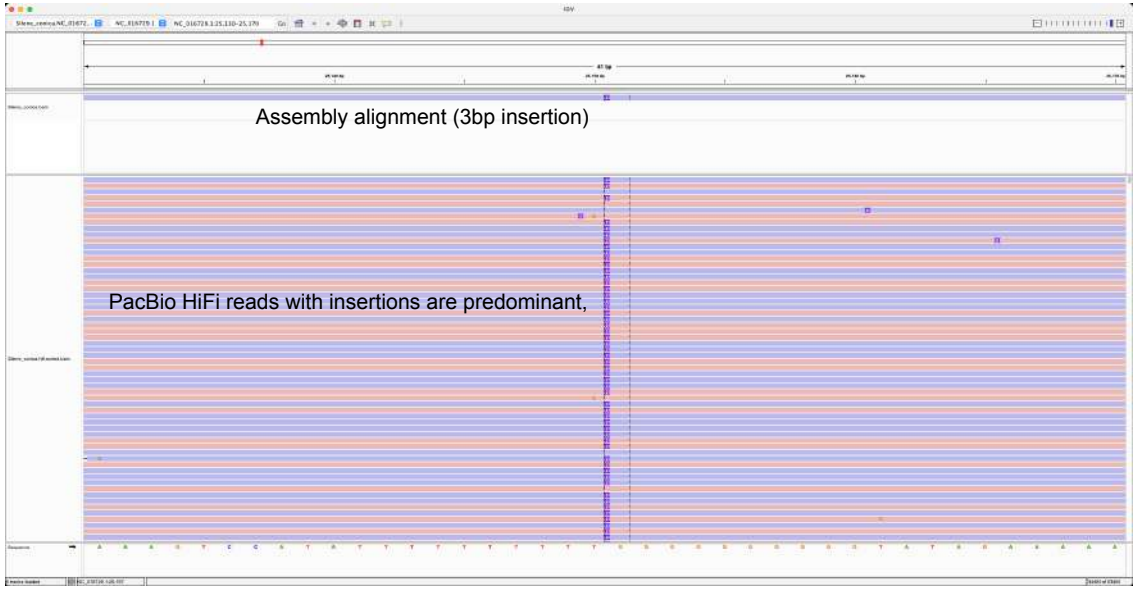
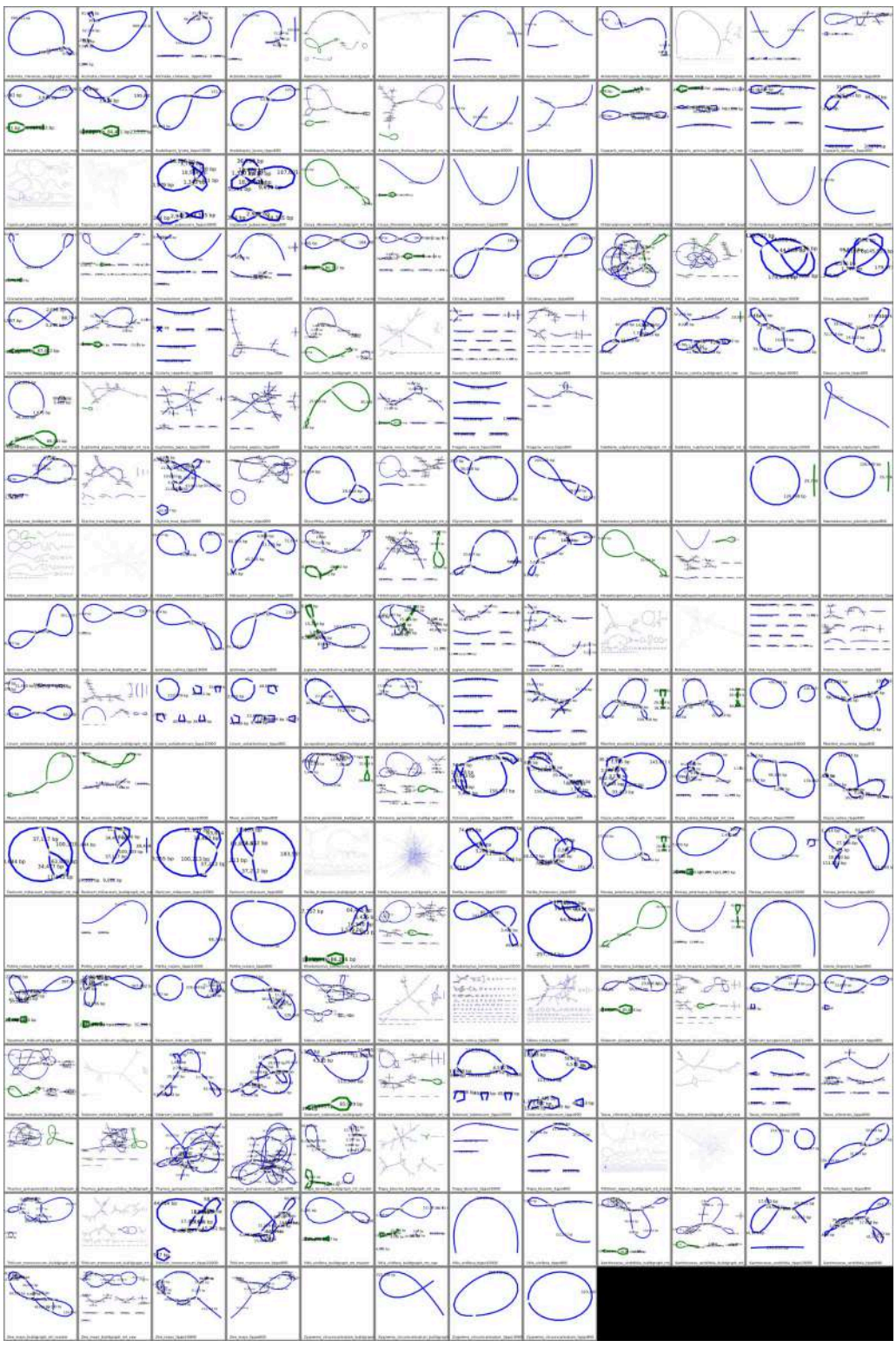


Figure S21. Base consensus compared with the published genomes. A. *Arabidopsis thaliana* Col-0 chloroplast. B. *Silene conica* chloroplast.

A.



B.

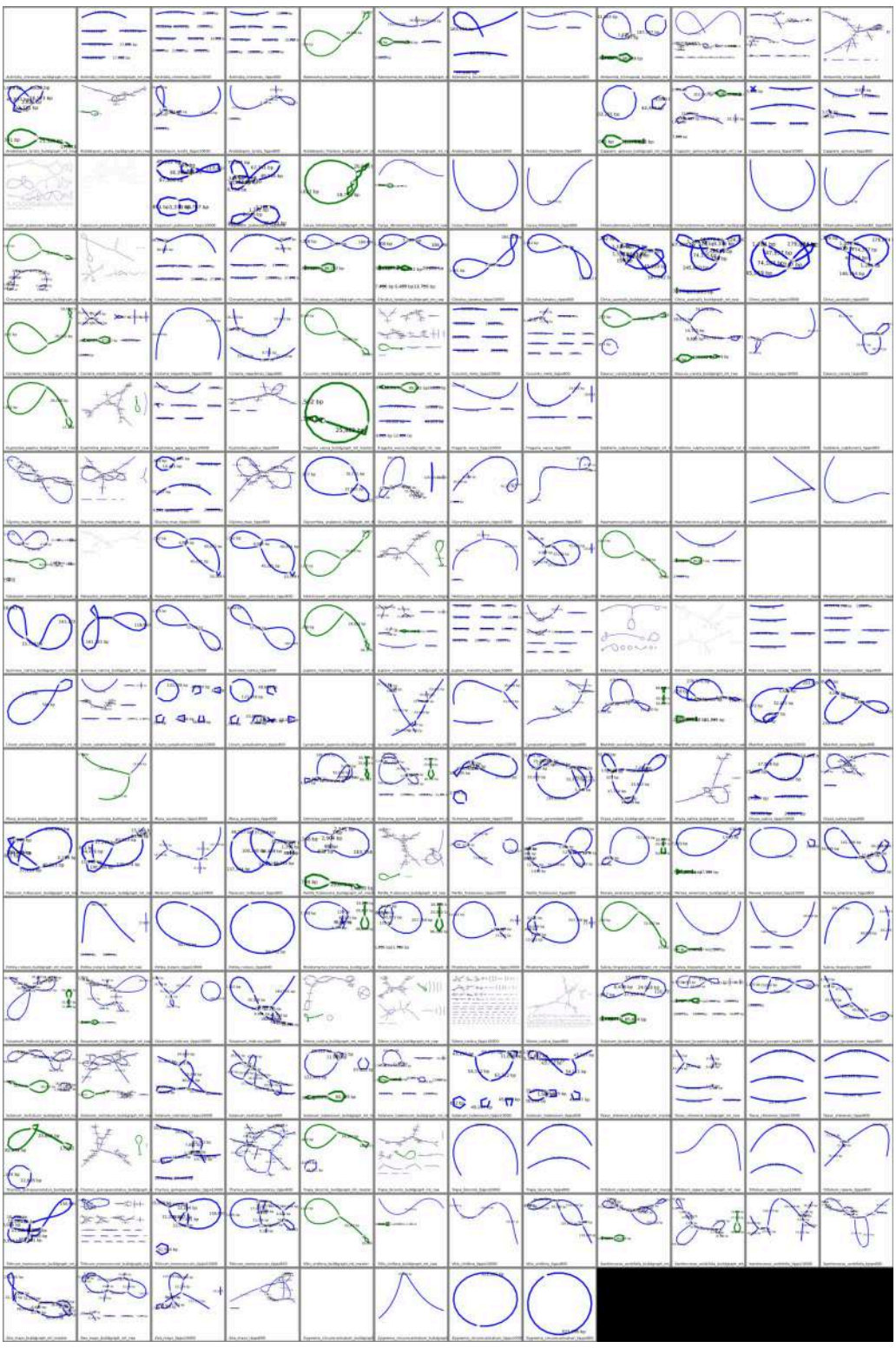


Figure S22: Mitochondrial assembly graph of PMAT (buildgraph-raw, buildgraph-master) and TIPPo (minimum overlap length: 800 and 10000 bp). A. mitochondrial assembly using 1x coverage for each species, except for 0.5x for *T. chinensis*, 2.5x for *S. conica* and 0.15x for *L. japonicum*. B. mitochondrial assembly using 0.5x for each species, except for 0.25x for *T. chinensis*, 1.25x for *S. conica* and 0.075x for *L. japonicum*. Green nodes represent chloroplast sequences, and blue nodes represent non-chloroplast sequences.

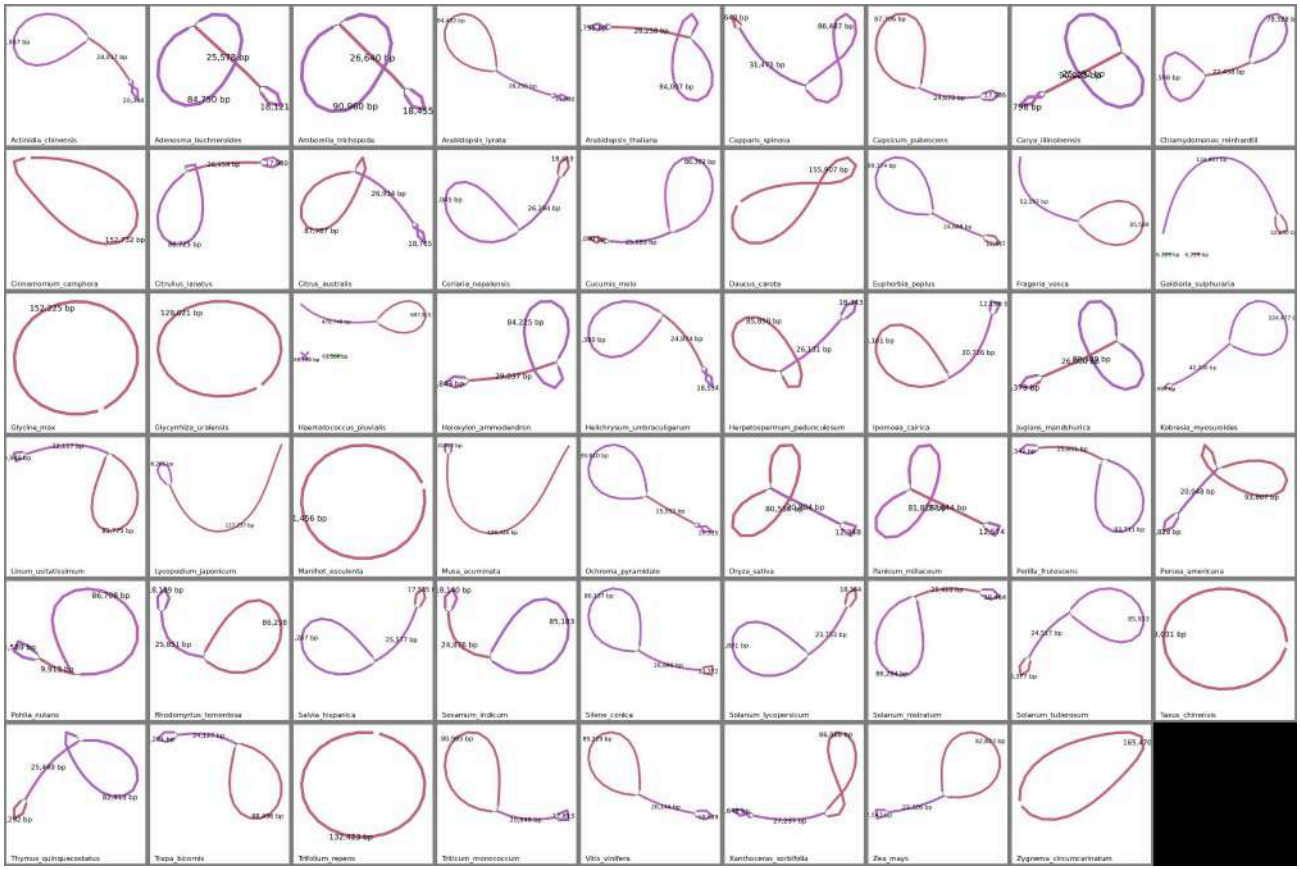


Figure S23: Chloroplast assembly graph of TIPPo. The coverage is set to 0.5x for each species, except for 0.25x for *T. chinensis,* 1.25x for *S. conica* and 0.075x for *L. japonicum.*

Thesis Appendix II

The structure of mitochondrial genomes is associated with geography in *Arabidopsis thaliana*

Wenfei Xian[1], Zhigui Bao[1], Sebastian Vorbrugg[1], Yueqi Tao[1], Andrea Movilli[1], Ilja Bezrukov[1], and Detlef Weigel[1,2†]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany

[†]For correspondence: weigel@tue.mpg.de

Abstract

Chloroplasts and mitochondria are the primary sites for photosynthesis and respiration, each harboring its own unique genome. Although the organellar genomes are considerably smaller compared to the nuclear genome, they are nonetheless essential for survival of the organism. A common feature of many chloroplast and mitochondrial genomes is the presence of large repeated sequences longer than 1 kb. These can be either in inverted or direct orientation, and recombination between them leads to structural heteroplasmy. To understand the intraspecific evolution of organellar genomes, we assembled chloroplast and mitochondrial genomes of 143 *A. thaliana* accessions from PacBio HiFi sequencing data. We find large repeats to be associated with heteroplasmy and structural variation. Our extensive genome annotation identifies novel open reading frames (ORFs) in those accessions that lost large repeats, potentially introduced via horizontal gene transfer, illuminating additional paths for diversification of plant organelles. The loss of large repeats correlates with geography and phenotypes, pointing to their adaptive importance. The assembled and annotated organellar genomes constitute a rich source for future functional studies of the interaction between the three genomes of a plant.

Introduction

Climate change-induced increases in temperature and drought affect plant photosynthesis and respiration (Ribas-Carbo et al. 2005; Hu et al. 2020; Trösch et al. 2022), threatening food security (Hasegawa et al. 2018). Photosynthesis occurs in chloroplasts, while respiration takes place in mitochondria. Both chloroplasts and mitochondria have their own independent genomes that have co-evolved closely with the nuclear genome (Sloan et al. 2018a; Lian et al. 2024b). While population-level genomic studies have deepened our understanding of species evolution and helped pinpoint alleles relevant to adaptation of wild and crop plants (1001 Genomes Consortium 2016; Wang et al. 2018; Liu et al. 2020; Zhou et al. 2022; Wilson et al. 2019; Evans et al. 2014; Jayakodi et al. 2020; Cheng et al. 2024), these studies have typically focused on the nuclear genome, with variation in the genomes of chloroplasts and mitochondria remaining less well understood. Of the efforts that have analyzed variation in organellar genomes at the population scale, these have mostly

concentrated on chloroplast genomes (Go et al. 2024; Magdy et al. 2019), with analyses of mitochondrial genomes typically being based on a small number of assemblies (Sun et al. 2022; Wang et al. 2022; Fan et al. 2022).

In vascular plants, most chloroplast and mitochondrial genomes feature large repeated sequences longer than 1 kb (Wynn and Christensen 2019; Sloan 2013). Pairs of large repeats in either inverted or direct orientation have been found to trigger recombination, which leads in turn to structural heteroplasmy, defined as the presence of multiple chloroplast or mitochondrial genomes in a single individual (Palmer 1983; Kozik et al. 2019; Klein et al. 1994; Wang and Lanfear 2019). Heteroplasmy complicates genome representation, as a single linear sequence cannot encapsulate all configurations (Palmer 1983).

Frequent recombination-driven rearrangements hinder comparisons both between and within species, as whole-genome alignment software is primarily designed for conserved synteny (Li 2018). Rearrangements are biologically relevant, as they can impact fitness (Juszczuk et al. 2007; Shedge et al. 2010). Chimeric open reading frames (ORFs) resulting from rearrangements may cause cytoplasmic male sterility (Sandhu et al. 2007).

The chloroplast genome of the *Arabidopsis thaliana* reference accession Col-0 has a pair of inverted repeats that are 26 kb long (Sato et al. 1999). The mitochondrial genome of this accession contains two large repeats, the direct 4.2 kb long repeat I and the inverted 6.5 kb long repeat II (Unseld et al. 1997). Recombination between the two copies of the direct repeat I leads to two subgenomic DNAs of 134 and 234 kb (Klein et al. 1994; Unseld et al. 1997). The assembly of a small number of *A. thaliana* mitochondrial genomes from short or long reads has also shown that not all of them contain two copies of repeat II (Davila et al. 2011; Zou et al. 2022).

For this study, we assembled organellar genomes of 143 *A. thaliana* accessions from Pacbio HiFi data. We describe the structural diversity of both chloroplast and mitochondrial genomes, report a likely instance of horizontal gene transfer from another member of the Brassicaceae to *A. thaliana* mitochondrial genomes, and uncover an association between the loss of large repeats in mitochondrial genomes with geography and phenotypes.

Results

Structural Diversity of Organellar Genomes

We obtained PacBio HiFi reads for 143 *A. thaliana* accessions, for which nuclear genome assemblies have been recently reported (Lian et al. 2024a; Wlodzimierz et al. 2023; Kang et al. 2023; Rabanal et al. 2022). We used the TIPPo tool for assembly of chloroplast genomes (Xian et al. 2024). We used both TIPPo (Xian et al. 2024) and PMAT (Bi et al. 2024) to assemble mitochondrial genomes, manually selecting the completeness assembly for each accession. For three accessions (Geg-14, HR-10, Nz-1), assembly of the mitochondrial genome failed, and we used Verkko (Rautiainen et al. 2023) to perform whole-genome assemblies, visualized the assembly graph in Bandage (Wick et al. 2015). We then manually selected the appropriate nodes based on the structure of the subgraph, provided these as input to Ribotin (Rautiainen 2024) to extract the constituent HiFi reads, and finally reassembled these with flye (Kolmogorov et al. 2019).

Chloroplast genomes are found in the canonical form in all accessions, with one large single copy (LSC), one small single copy (SSC) and a pair of inverted repeats (IRs) (**Figure 1A** and **Figure S1**). The average size of the chloroplast genomes across all accessions is 154,297 bp, with a range of 152,495 bp to 154,909 bp (**Figure 1B** and **Table S1**).

The mitochondrial genomes can be categorized into four primary types, distinguished by the copy number and orientation of large repeats within the assembly graph: 51 accessions of type 1a contain a pair of inverted repeat I and a single copy of repeat II; 16 accessions of type 1b have a pair of direct repeat I and a single copy of repeat II; 6 accessions of type 2a feature a pair of repeat I in direct orientation and a pair of repeat II in inverted orientation; and 70 accessions of type 2b have a pair of the two repeats, with both repeat I and II in inverted orientations (**Figure 1A**, **Figure S2** and **Table S2**). For accessions with two pairs of large repeats (types 2), the average mitochondrial genome size is 368,714 bp, ranging from 367,334 bp to 368,974 bp (**Figure 1B** and **Table S1**). The average mitochondrial genome size of accessions with only one pair of large repeats (type 1) is 350,579 bp and has a larger range, from 337,339 bp to 365,358 bp (**Figure 1B** and **Table S1**). The 67 mitochondrial genomes of type 1 are much more variable in size, with a standard deviation of 5,106 bp compared to 374 bp for the 76 type 2 genomes (**Figure 1B** and **Table S1**).

To validate the single copy of repeat II sequence in 67 accessions (type 1), we used read coverage to infer the copy number by mapping HiFi reads to repeat I and II sequences (**Table S3**). If the mitochondrial genome contains two copies of both repeat I and II (type 2), the coverage ratio of repeat I and repeat II should be close to 1. With two copies of repeat I and one copy of repeat II (types 1), the ratio should be closer to 2.

As expected, the mean coverage ratio for two copies of repeat I and II accessions (type 2) is 0.98 with a standard deviation of 0.06. However, the average coverage ratio for two copies of repeat I and one copy of repeat II accessions (type 2) is 1.71, with a standard deviation of 0.11 (**Figure 1C**). This did not change much when we considered only the first 4.2 kb of repeat II, which is over 50% longer than repeat I (**Table S4**). The variance to the expected ratio of 2 between repeat I and II in type 1 mitochondrial genomes cases suggests that these genomes have fewer copies of repeat I or more copies of repeat II than expected.

One possible explanation is the coexistence of type 2 mitochondrial genomes in these accessions, increasing the abundance of repeat II. A pair of large repeats has high recombination activity, resulting in heteroplasmy (Ramsey and Mandel 2019; Palmer 1983; Wang and Lanfear 2019; Sloan 2013). Assuming that repeat II has two copies, recombination will result in four combinations of repeat II and adjacent sequences, but only one combination when repeat II is present in only one copy (**Figure 1D**).

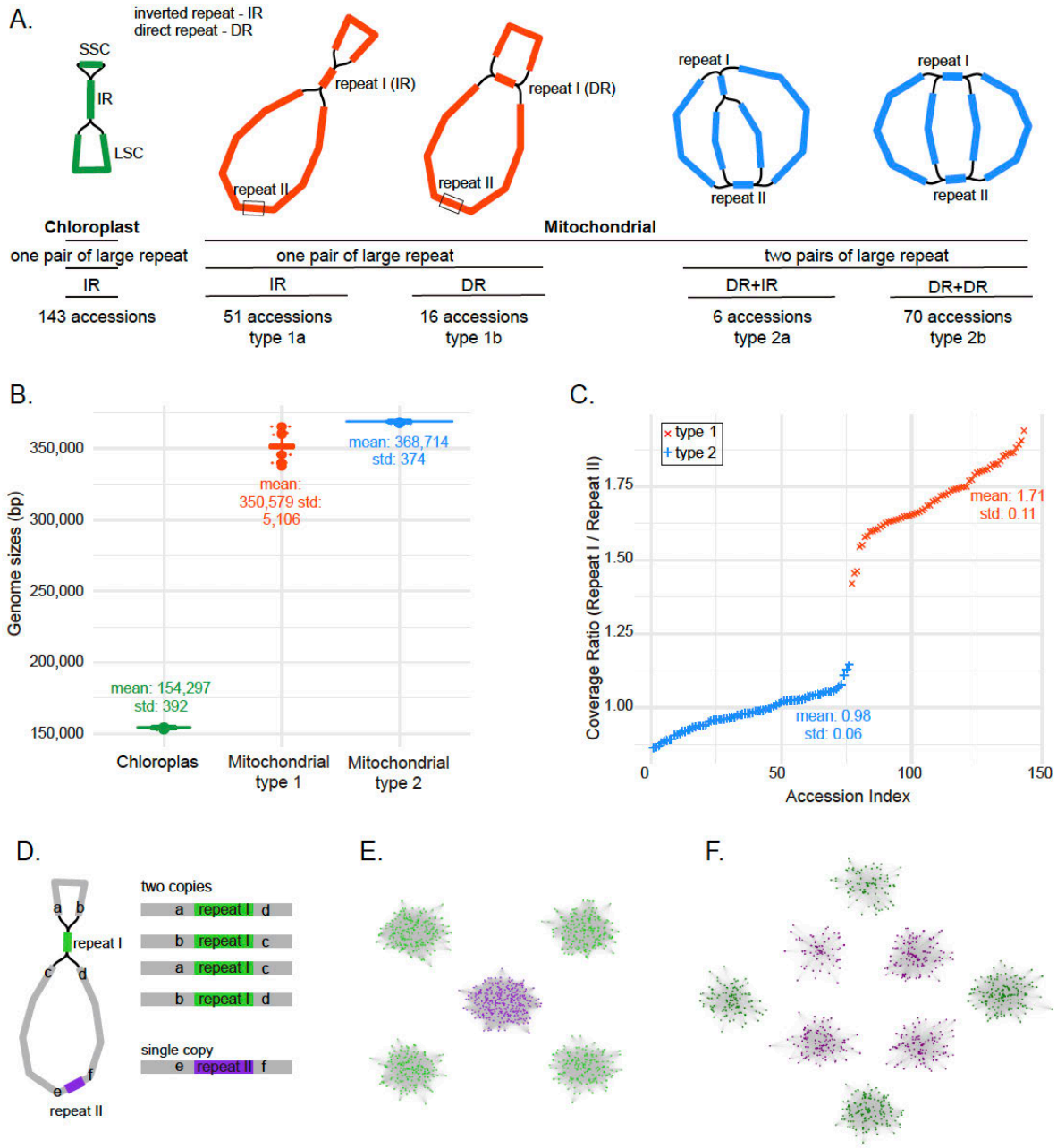


Figure 1. Structure diversity of organellar genomes. **A)** Major types of assembly graphs of chloroplast and mitochondrial genomes in *A. thaliana*. **B)** Distribution of genome sizes of chloroplast and mitochondrial genomes. **C)** Coverage ratio between repeat I and repeat II for inferring their relative copy number in mitochondrial genomes. **D)** Diagram illustrating the outcome of recombination with a pair of large repeats in a mitochondrial genome. **E)** Clusters of reads containing repeat I (green) and II (purple) sequences in accession Ler-0 (type 1). **F)** Clusters of reads containing repeat I (green) and II (purple) sequences in the reference accession Col-0 (type 2).

We searched for HiFi reads completely encompassing repeat II, and then extracted 1 kb sequences upstream and downstream of repeat II, and subjected these to an all-versus-all alignment using minimap2. Pairs of reads with > 95% mutual coverage were then provided as input for MCL clustering. As a control, the same analysis was conducted for repeat I. We found that in type 1 mitochondrial genome, repeat II reads formed only one

cluster in each accession, indicating a lack of recombination (**Figure 1E, Table S5 and S6**). However, both repeat I of type 1 mitochondrial genome and repeats I and II of type 2 mitochondrial genomes formed four clusters in an accession, indicative of recombination activity (**Figure 1F, Table S5 and S6**). Thus, there is no evidence for the co-existence of the two different structural types of mitochondrial genomes in the type 1 accessions.

Rearrangements in organellar genomes

Considering that recombination mediated by repeats can lead to rearrangements of organellar genomes (Davila et al. 2011; Cole et al. 2018). If we assume that each organellar genome contains only one pair of large repeats, two linear DNA genomes could result akin to haplotypes for individuals with a diploid nuclear genome (Wang and Lanfear 2019). Previous studies have often erroneously interpreted the chloroplast SSC as a hotspot for inversions that distinguish species because they generally ignored heteroplasmy caused by large repeats within individuals (Walker et al. 2015). This has happened as a consequence of the fact that prevailing alignment methods are built for comparison of linear DNA sequences. To avoid the impact of intragenomic recombination-induced rearrangements on alignments across individuals, we focus on rearrangements between single-copy regions and we do not use representations that connect the single-copy fragments into larger linear molecules via large repeats.

In this way, we identified rearrangement events through pairwise genome comparisons. If no rearrangement was detected, the genomes were classified as belonging to the same cluster. For each cluster, we randomly selected one accession as a representative.

We did not detect obvious rearrangement in chloroplast genomes (**Figure S3**). In mitochondrial genomes, however, rearrangements are frequent. In total, we identified 25 clusters (**Figure 2A, Figure S4**). The largest cluster, C1, contains 37 accessions (**Figure 2B**). When we projected these 25 clusters onto a nuclear phylogenetic tree (**Figure 2C**), we found that accessions belonging to the same cluster were not necessarily closely related, especially for larger clusters.

Evidence for horizontal gene transfer in mitochondrial genomes

Initially, we identified all ORFs of at least 50 amino acids across organellar genomes for 143 accessions and derived orthologous groups (OGs) with OrthoFinder (Emms and Kelly 2019). In the mitochondrial genome, rearrangements are considered as one of the sources for new ORFs (Wang et al. 2024). Since we found no rearrangements between chloroplast genomes, we did not expect to find many new ORFs in chloroplast genomes. Consistent with expectations, we identified 92 chloroplast OGs, of which 89 were conserved across all genomes. The remaining three, which were not similar to known proteins, were found in only one or two accessions, and we cannot exclude that these are due to assembly or annotation errors.

In the much larger mitochondrial genomes, we identified 625 OGs, with 516 missing in no more than one accession and 109 missing in at least two accessions (**Table S7**). Notably these 109 non-conserved ORFs were typically quite common. If the variation in ORFs is related to genome rearrangements, the distribution of OGs should correspond to the rearrangement-based clusters identified above (**Figure 2A**). We found that

Presence/Absence Variation (PAV) of non-conserved OGs largely follow cluster assignments (**Figure 3A**). As absences are rarer than presences, we conclude that rearrangements are more likely to break existing ORFs than leading to the formation of new ORFs.

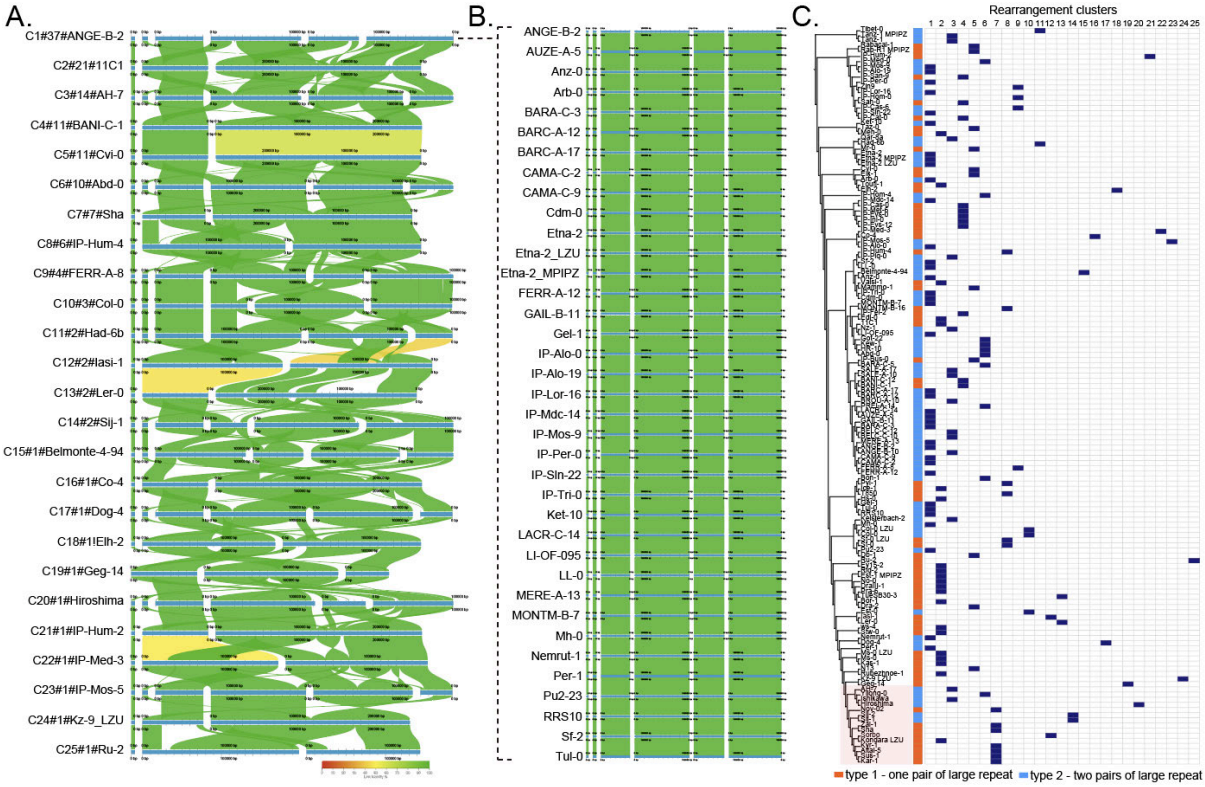


Figure 2. Rearrangements in mitochondrial Genomes. **A)** Whole genome alignments among the 25 clusters of mitochondrial genomes. Using "#" as a separator, the first string represents the cluster name, the second string indicates the number of accessions in the cluster, and the third string denotes the representative accession. **B)** Whole genome alignments of cluster 1. **C)** The distribution of cluster members across the nuclear phylogenetic tree.

We then focused on those OGs that are cluster-specific (cluster size ≥2). We found that a block with the three OGs OG0000593-595 is specific to cluster 7, and that a block with the 11 OGs OG0000602-612 is specific to cluster 12. For each OG, we selected a representative sequence and aligned it using DIAMOND BLASTP (Buchfink et al. 2021) to the UniRef90 database, where 320 OGs matched homologous sequences under thresholds of ≥75% coverage and ≥75% identity (**Figure S5**). OG000608, unique to cluster 12, aligned with a *Coffea canephora* protein from unassembled WGS sequence at close to 100% coverage but only about 30% amino acid identity (**Figure S5**). Specifically, the homologous sequences for OG000594, which is unique to cluster 7, showed 100% coverage and identity in all seven accessions of this cluster. This ORF was previously reported as orf117Sha in *A. thaliana* accession Sha, where it is associated with cytoplasmic male sterility (CMS) (Gobron et al. 2013). The original report also indicated that this ORF was detectable in accessions from Central Asia, which agrees with our finding that cluster 7 is restricted to this geographic region (Lian et al. 2024a).

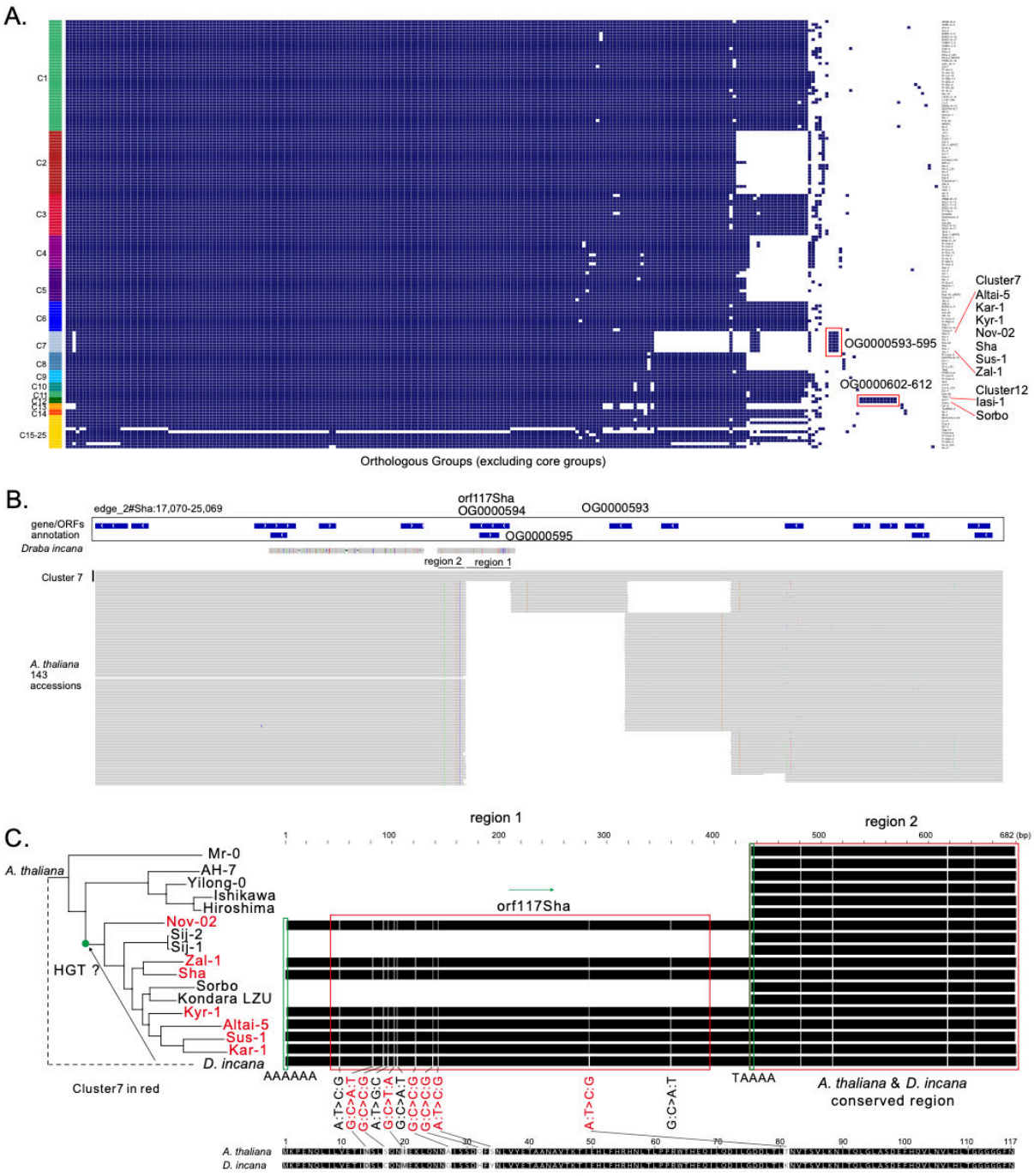


Figure 3. New ORFs in Mitochondrial genomes. A) Orthologous groups (OGs) of mitochondrial genes and ORFs. Accessions were ordered by their structures. **B)** IGV screenshot of the coverage of orf117Sha-related sequences in other accessions and annotation in the *D. incana* mitochondrial genome. Only one representative was chosen for each *A. thaliana* cluster. **C)** Close-up of the alignment of orf117Sha and adjacent sequences in several *A. thaliana* accessions.

It had been suggested that orf117Sha originates from an unknown source (Gobron et al. 2013). Given that OGs OG000593-595 are specific to cluster 7, it is likely that the sequences containing the ORFs in these OGs are unique to cluster 7 at the genomic level – indeed, our dataset shows the orf117Sha region to be unique to cluster 7 (**Figure 3A and 3B**). Searching the NT database, we found a remarkably close hit, 96.8% at the DNA level, in

the mitochondrial genome of another Brassicaceae species, *Draba incana* (GenBank: OY755218.1, data from Darwin Tree of Life project), as well as more distant hits in the mitochondria of other plants, including a member of the Fabaceae (**Figure S6**).

With a larger region from *D. incana* around the orf117Sha related sequence as a query, we identified homologous sequences in five closely related *A. thaliana* accessions in addition to the OG000594 region in cluster 7 accessions (**Figure 3C**). Based on a multiple sequence alignment, the *D. incana* orf117Sha related sequence block can be divided into two regions: Region 1, which includes orf117Sha plus about 30 bp on either side and which is present only in all cluster 7 accessions; region 2 is conserved between *D. incana* and *A. thaliana* (**Figure 3B & 3C**).

Comparing Region 1/orf117Sha between *A. thaliana* and *D. incana*, we find 11 single-nucleotide differences, seven of which are non-synonymous mutations. The adjacent Region 2, which is only about half as long as Region 1, contains only four single-nucleotide differences. Additionally, we detected adenine (A)-rich micro-homologies enriched at both ends of Region 1/orf117Sha (**Figure 3C**). The seven accessions of cluster 7 and four additional accessions Sij-2, Sij-1, Sorbo and Kondara_LZU form a monophyletic clade, as can be inferred from Figure 3C, consistent with a potential HGT event when this clade arose.

For the 11 cluster 12-specific OGs (3,427 bp), we also observed a potential origin from HGT. The best match in the NT database belongs to the mitochondrial genome of *Arabidopsis lyrata* (NC_081483.1, coverage 99%, similarity 99.33%), for which no further provenance information is available (**Figure S7**). We assembled the complete mitochondrial genomes of *A. lyrata* accessions MN47 and NT1 (Wlodzimierz et al. 2023) using PacBio HiFi reads, but did not find any homologous sequences, suggesting that this cluster also segregates in *A. lyrata*.

Loss of repeat II is associated with geography

We analyzed the distribution of repeat I and II in mitochondrial genomes on a larger scale using published Illumina short read sequencing data of 995 accessions (**Table S8**) (1001 Genomes Consortium 2016; Zou et al. 2017). We first confirmed that short read coverage can be used to infer the relative copy number with 38 accessions that had sequenced with both Pacbio HiFi long reads and Illumina short reads, ensuring that the read sets were indeed from the same accessions (**Figure S8**). Coverage ratios were highly consistent between long and short reads (**Figure S9**).

We observed that HPG1 accessions, members of a nearly clonal lineage that was introduced to North America about 400 years ago (Exposito-Alonso et al. 2018), contain two copies of repeat I and repeat II, indicating that the copy number of repeat II has not changed over approximately 400 years of evolution **(Figure S10**). In Central Asia, accessions with only one pair of large repeats (type 1) predominate, but in the Yangtze River Basin (Zou et al. 2017) are mostly accessions with two pairs of large repeats (types 2) (**Figure 4A**).

Turning to geography more broadly, there was a significant positive correlation between latitude and coverage ratio ($\rho = 0.33$, p-value < 2.2e-16), indicating that accessions with two pairs of repeats are more common at lower latitudes, while those with one pair of repeats are more common at higher latitudes (**Figure 4B**). There was no significant correlation with longitude (p-value = 0.099) (**Figure S11**).

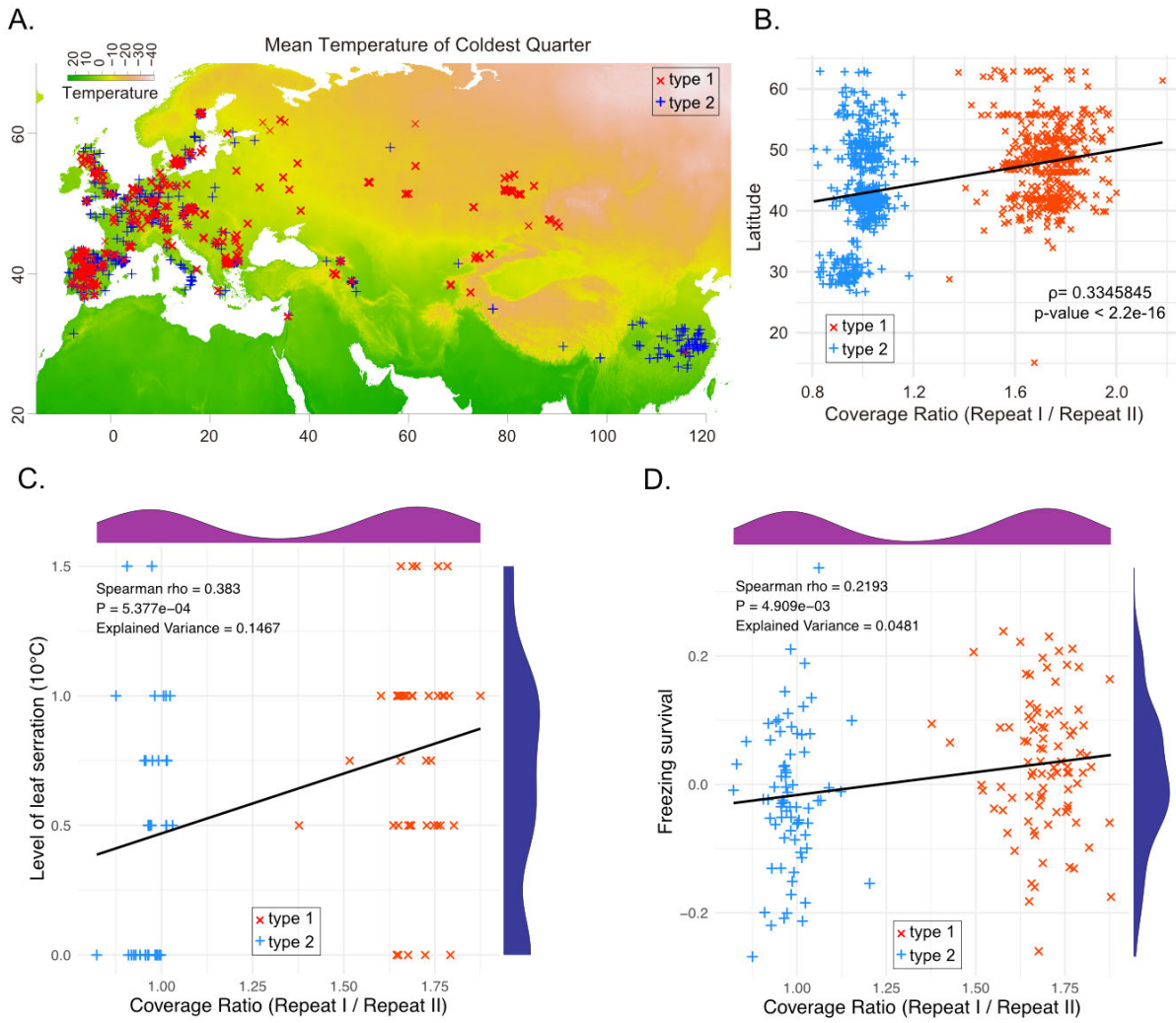


Figure 4. Loss of large repeats is associated with local geography and potentially local adaptation. A) Geographic distribution of 995 accessions in which repeat structure of mitochondrial genomes was estimated with short reads. **B)** Correlation between coverage ratio and latitude. Spearman's ρ and P-values are shown. **C)** Correlation between coverage ratio and level of leaf serration in 10°C. **D)** Correlation between coverage ratio and freezing survival.

We examined whether the loss of repeat II is associated with specific phenotypes. We performed a correlation analysis with 1,695 phenotypes (Voichek and Weigel 2020); only 13 phenotypes showed a correlation that was significant at a p-value threshold of 0.01 (**Table S9**). The strongest correlation was with leaf serration at 10°C (p-value = 0.0005, rho = 0.383) (**Figure 4C**), with higher phenotypic values corresponding to sharper and more jagged serrations (Atwell et al. 2010). The leaves of type 1 accessions are more highly serrated than those of type 2 accessions in 10°C. However, differences in leaf serration at 16°C and 22°C were not statistically significant, suggesting a connection to cold adaptation. One other significant phenotype was also related to cold adaptation: freezing survival (p-value = 0.005, rho = 0.219) (Horton et al. 2016), with type I accessions having higher survival rates upon freezing.

Discussion

Pioneering research conducted 40 years ago in seven angiosperm species revealed extensive rearrangements in the chloroplast genomes of two legume species, pea and broad bean, both of which are characterized by the loss of a large inverted repeat sequence (Palmer and Thompson 1982). In contrast, virtual collinearity is observed among the chloroplast genomes of spinach, petunia, and cucumber, all of which retain both copies of the inverted repeat (Palmer and Thompson 1982). This finding suggested that genomes containing inverted repeats tend to be relatively more stable, whereas genomes that have lost the large repeats are more dynamic. Similar patterns of increased rearrangement rates were observed in the genus *Cephalotaxus*, which had also lost its inverted repeats (Ji et al. 2021).

Previous studies of inter- and intraspecific organellar genome variation in plants have primarily focused on plastid genomes. Here, we extend these studies through a large-scale investigation of large repeats in mitochondrial genomes of *A. thaliana* accessions. We confirm that genomes with two copies of repeat I and II are more stable, whereas those that have lost one copy of repeat II are more diverse. Specifically, mitochondrial genome size is stable in accessions with two copies, but it becomes more variable in those with only one copy of repeat II (**Figure 1B**). Furthermore, based on the PAV distribution of ORFs (OGs), the mitochondrial genomes of accessions in clusters 1, 3, and 6, which contain two copies of the repeats, are similar, whereas mitochondrial genomes in clusters 2, 4, 5, 7, and 8, which contain only one copy of repeat II, show significant differences (**Figure 3A**).

Heteroplasmy in plant organellar genomes is an intriguing phenomenon that poses several analytical challenges. Heteroplasmy complicates the representation of organellar genomes, which cannot be simply depicted as a single linear DNA sequence. Although the sequence content may remain unchanged, the sequence order can vary, making it difficult to identify structural variations across different species or even within different individuals of the same species, potentially leading to misinterpretation (Walker et al. 2015).

We wanted to minimize the impact of large repeat-mediated heteroplasmy—which can result in multiple genome sequences with identical content but different arrangements. We therefore did not connect non-repetitive fragments through large repeats into a single master ring. Instead, we directly compared the rearrangements between each fragment (**Figure 3A**). This approach is similar to a recent method used for identifying structural variations (SVs) in the *A. thaliana* nuclear genome, where the genome is divided into chunks before alignment to a reference genome, in order to avoid alignment errors due to rearrangements (Igolkina et al. 2024). The same principle can be applied to transposon alignments (Anderson et al. 2019), as orthologous transposons typically do not reside within large orthologous blocks in whole-genome alignments.

With the sequencing of more and more genomes, reports of horizontal gene transfer are increasing (Haimlich et al. 2024; Cheng et al. 2019). In this study, the *orf117Sha* gene, which was thought to have come from an "unknown source" (Gobron et al. 2013) due to the limited number of sequenced genomes at the time, can now be linked to a very similar sequence found in another Brassicaceae, *D. incana* (Bright 2022), with the majority of other Brassicaceae lacking orf117Sha related sequences. Given that the *A. thaliana* accessions with *orf117Sha* form a monophyletic clade in a nuclear phylogenetic tree and that the orf117Sha sequence is identical, if it was transferred by HGT, it likely took place in the

ancestor of these accessions. A recent study revealed that the parasitic species *Lophophytum mirabile* acquires up to 74% of its mitochondrial DNA through recurrent horizontal gene transfer and the authors proposed that horizontally transferred mitochondrial DNA segments become circularized via microhomology-mediated repair pathways, subsequently integrating into the mitochondrial genome through recombination (Emilia Roulet et al. 2024). Similarly, in our findings, we observed A-rich repeats flanking the insertion fragments, consistent with the formation of circular DNA being an important ingredient of HGT of organellar DNA.

Although recent experimental work demonstrated that the removal of the inverted repeats from the tobacco plastid genome by genome editing had little impact on gross morphology or stress responses (Krämer et al., 2024), we found that in *A. thaliana*, the loss of large repeats in the mitochondrial genome is associated with geography and temperature-related phenotypes, in support of the evolutionary relevance of variation in organellar genomes.

Materials and Methods

Assembling organellar genomes

For each accession, we randomly selected PacBio HiFi reads corresponding to 1 and 2 Gb. For the chloroplast genomes, we used the assembly graph generated by TIPPo (Xian et al. 2024). For the mitochondrial genomes, we used both TIPPo v1.0 (Xian et al. 2024) and PMAT v1.5.3 (Bi et al. 2024) for primary assembly. We visualized each assembly graph using Bandage v0.9.0 (Wick et al. 2015) and manually selected the complete assembly graph. In three accessions (Geg-14, HR-10, Nz-1), we were unable to obtain a complete mitochondrial genome using TIPPo or PMAT. For these accessions, we first performed whole-genome assembly with Verkko v1.4.1 (Rautiainen et al. 2023) followed by visual inspection of the graph to identify mitochondrial nodes. We used Ribotin v1.2 (Rautiainen 2024) to extract the corresponding HiFi reads and finally assembled the mitochondrial genome using Flye 0.7.17-r1188 (Kolmogorov et al. 2019).

Estimating the copy number of large repeats using sequencing depth

Considering that repeat I and repeat II are highly conserved in *A. thaliana* mitochondrial genomes, we aligned the PacBio HiFi data from each accession to the published *A. thaliana* Col-0 reference sequences of repeat I and repeat II (Sloan et al. 2018b) using minimap2 0.7.17-r1188 (Li 2018), retaining alignments with coverage of at least 90% of the length of the reference sequence. Next, we used mosdepth v0.3.3 (Pedersen and Quinlan 2018) to calculate the sequencing depth, and derived the ratio of sequencing depths of repeat I to repeat II. The same method was applied to Illumina 100 bp paired-end short-read sequencing data (1001 Genomes Consortium 2016; Zou et al. 2017), except that we used BWA 0.7.17-r1188 (Li and Durbin 2009) as aligner.

Estimating the extent of heteroplasmy

To estimate the abundance of each heteroplasmic genome, we used GraphAligner v1.0.17 (Rautiainen and Marschall 2020) to align PacBio HiFi reads to the organellar assembly graph, ensuring that the reads fully spanned the large repeats. The relative orientation of the

nodes adjacent to the large repeats was used to define different heteroplasmy states. Finally, we counted the number of reads corresponding to each state to determine their respective abundances.

Detecting rearrangements

First, we performed pairwise whole genome alignments of mitochondrial genomes using minimap2 0.7.17-r1188 (Li 2018). In the absence of rearrangements, the alignment target coverage should be close to 100%, and the number of alignment blocks should equal the number of fragments in the target genome. Based on this principle, we first identified pairs of accessions without evidence of rearrangements. We then clustered mitochondrial genomes without apparent rearrangements with Markov Cluster Algorithm (MCL) (14-137) (Enright et al. 2002). We performed multiple sequence alignment of all genomes in each cluster using minitv ([CSL STYLE ERROR: reference with no printed form.]) and visualized the results with AliTV (Ankenbrand et al. 2017). Finally, we visually inspected the alignments to validate the absence of obvious errors in clustering.

Construction of nuclear phylogenetic tree

The nuclear phylogenetic tree was constructed using Mashtree v1.4.6 (Katz et al. 2019), with the assembly of each accession as input.

Annotation of Organellar Genomes

For each organellar genome, we identified all ORFs longer than 149 bp beginning with a start and ending with a stop codon with Getorf (Rice et al. 2000). We annotated the core genes using miniprot 0.12-r237 (Li 2023) based on the Col-0 mitochondrial genome annotation (Sloan et al. 2018b). We removed ORFs that overlapped with the core genes to generate the final annotation.

We used OrthoFinder v2.5.4 (Emms and Kelly 2019) to identify orthologous gene families. With the OrthoGroup (OG) results in hand, we selected a representative sequence for each OG based on median length of the OG and used DIAMOND v2.0.13 (Buchfink et al. 2015) BLASTp to align these sequences against the UniRef90 database. If a match was found under the default threshold, we selected the best alignment as the representative match.

Data availability

HiFi reads used in this study were downloaded from ENA/NCBI/CNCB: PRJEB62038, PRJEB55353, PRJEB55632, PRJEB50694 and PRJCA012695 (Lian et al. 2024a; Wlodzimierz et al. 2023; Kang et al. 2023; Rabanal et al. 2022).

Acknowledgments

We thank Wei Yuan, Li He and Haim Ashkenazy for the discussions. This study was supported by the Max Planck Society and the Novozymes Prize of the Novo Nordisk Foundation (D.W.).

Competing Interests

DW holds equity in Computomics, which advises plant breeders. DW also consults for KWS SE, a globally active plant breeder and seed producer. All other authors declare no competing interests.

References

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

Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, O'Connor CH, Hirsch CD, Ross-Ibarra J, Hirsch CN, Springer NM. 2019. Transposable elements contribute to dynamic genome content in maize. *The Plant Journal* **100**: 1052–1065.

Ankenbrand MJ, Hohlfeld S, Hackl T, Förster F. 2017. AliTV—interactive visualization of whole genome comparisons. *PeerJ Comput Sci* **3**: e116.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**: 627–631.

Bi C, Shen F, Han F, Qu Y, Hou J, Xu K, Xu L-A, He W, Wu Z, Yin T. 2024. PMAT: an efficient plant mitogenome assembly toolkit using low coverage HiFi sequencing data. *Hortic Res* uhae023.

Bright M. 2022. *Darwin's tree of life*. Wayland, London, England.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**: 366–368.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Cheng S, Feng C, Wingen LU, Cheng H, Riche AB, Jiang M, Leverington-Waite M, Huang Z, Collier S, Orford S, et al. 2024. Harnessing landrace diversity empowers wheat breeding. *Nature*. http://dx.doi.org/10.1038/s41586-024-07682-9.

Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, Sun W, Li X, Xu Y, Zhang Y, et al. 2019. Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179**: 1057–1067.e14.

Cole LW, Guo W, Mower JP, Palmer JD. 2018. High and Variable Rates of Repeat-Mediated Mitochondrial Genome Rearrangement in a Genus of Plants. *Mol Biol Evol* **35**: 2773–2785.

Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu Y-Z, Weigel D, Mackenzie SA. 2011. Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. *BMC Biol* **9**: 64.

Emilia Roulet M, Ceriotti LF, Gatica-Soria L, Virginia Sanchez-Puerta M. 2024. Horizontally transferred mitochondrial DNA tracts become circular by microhomology-mediated repair pathways. *New Phytologist* **243**: 2442–2456.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale

detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, et al. 2014. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat Genet* **46**: 1089–1096.

Fan W, Liu F, Jia Q, Du H, Chen W, Ruan J, Lei J, Li D-Z, Mower JP, Zhu A. 2022. Fragaria mitogenomes evolve rapidly in structure but slowly in sequence and incur frequent multinucleotide mutations mediated by microinversions. *New Phytol* **236**: 745–759.

Gobron N, Waszczak C, Simon M, Hiard S, Boivin S, Charif D, Ducamp A, Wenes E, Budar F. 2013. A cryptic cytoplasmic male sterility unveils a possible gynodioecious past for Arabidopsis thaliana. *PLoS One* **8**: e62450.

Go S, Koo H, Jung M, Hong S, Yi G, Kim Y-M. 2024. Pan-chloroplast genomes for accession-specific marker development in Hibiscus syriacus. *Sci Data* **11**: 246.

Haimlich S, Fridman Y, Khandal H, Savaldi-Goldstein S, Levy A. 2024. Widespread horizontal gene transfer between plants and bacteria. *ISME Commun* **4**: ycae073.

Hasegawa T, Fujimori S, Havlík P, Valin H, Bodirsky BL, Doelman JC, Fellmann T, Kyle P, Koopman JFL, Lotze-Campen H, et al. 2018. Risk of increased food insecurity under stringent global climate change mitigation policy. *Nat Clim Chang* **8**: 699–703.

Horton MW, Willems G, Sasaki E, Koornneef M, Nordborg M. 2016. The genetic architecture of freezing tolerance varies across the range of Arabidopsis thaliana. *Plant, cell & environment* **39**. https://pubmed.ncbi.nlm.nih.gov/27487257/ (Accessed November 24, 2024).

Hu S, Ding Y, Zhu C. 2020. Sensitivity and Responses of Chloroplasts to Heat Stress in Plants. *Front Plant Sci* **11**: 375.

Igolkina AA, Vorbrugg S, Rabanal FA, Liu H-J, Ashkenazy H, Kornienko AE, Fitz J, Collenberg M, Kubica C, Morales AM, et al. 2024. Towards an unbiased characterization of genetic polymorphism. *bioRxiv* 2024.05.30.596703. https://www.biorxiv.org/content/10.1101/2024.05.30.596703v1.full (Accessed August 9, 2024).

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*. https://doi.org/10.1038/s41586-020-2947-8.

Ji Y, Liu C, Landis JB, Deng M, Chen J. 2021. Plastome phylogenomics of Cephalotaxus (Cephalotaxaceae) and allied genera. *Ann Bot* **127**: 697–708.

Juszczuk IM, Flexas J, Szal B, Dabrowska Z, Ribas-Carbo M, Rychter AM. 2007. Effect of mitochondrial genome rearrangement on respiratory activity, photosynthesis, photorespiration and energy status of MSC16 cucumber (Cucumis sativus) mutant. *Physiol Plant* **131**: 527–541.

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.

Katz LS, Griswold T, Morrison SS, Caravas JA, Zhang S, den Bakker HC, Deng X, Carleton HA. 2019. Mashtree: a rapid comparison of whole genome sequence files. *Journal of open source software* **4**: 10.21105/joss.01762.

Klein M, Eckert-Ossenkopp U, Schmiedeberg I, Brandt P, Unseld M, Brennicke A, Schuster W. 1994. Physical mapping of the mitochondrial genome of Arabidopsis thaliana by

cosmid and YAC clones. *Plant J* **6**: 447–455.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540–546.

Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC. 2019. The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet* **15**: e1008373.

Krämer C, Boehm CR, Liu J, Ting MKY, Hertle AP, Forner J, Ruf S, Schöttler MA, Zoschke R, Bock R. 2024. Removal of the large inverted repeat from the plastid genome reveals gene dosage effects and leads to increased genome copy number. *Nat Plants* **10**: 923–935.

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

Lian Q, Li S, Kan S, Liao X, Huang S, Sloan DB, Wu Z. 2024b. Association Analysis Provides Insights into Plant Mitonuclear Interactions. *Mol Biol Evol* **41**. http://dx.doi.org/10.1093/molbev/msae028.

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

Li H. 2023. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**. http://dx.doi.org/10.1093/bioinformatics/btad014.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, et al. 2020. Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**: 162–176.e13.

Magdy M, Ou L, Yu H, Chen R, Zhou Y, Hassan H, Feng B, Taitano N, van der Knaap E, Zou X, et al. 2019. Pan-plastome approach empowers the assessment of genetic variation in cultivated Capsicum species. *Hortic Res* **6**: 108.

Palmer JD. 1983. Chloroplast DNA exists in two orientations. *Nature* **301**: 92–93.

Palmer JD, Thompson WF. 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* **29**: 537–550.

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

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.

Ramsey AJ, Mandel JR. 2019. When one genome is not enough: Organellar heteroplasmy in plants. *Annual Plant Reviews online* 619–658. https://onlinelibrary.wiley.com/doi/10.1002/9781119312994.apr0616.

Rautiainen M. 2024. Ribotin: automated assembly and phasing of rDNA morphs. *Bioinformatics* **40**. http://dx.doi.org/10.1093/bioinformatics/btae124.

Rautiainen M, Marschall T. 2020. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol* **21**: 253.

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.

Ribas-Carbo M, Taylor NL, Giles L, Busquets S, Finnegan PM, Day DA, Lambers H, Medrano H, Berry JA, Flexas J. 2005. Effects of water stress on respiration in soybean leaves. *Plant Physiol* **139**: 466–473.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**. https://pubmed.ncbi.nlm.nih.gov/10827456/ (Accessed November 24, 2024).

Sandhu APS, Abdelnoor RV, Mackenzie SA. 2007. Transgenic induction of mitochondrial rearrangements for cytoplasmic male sterility in crop plants. *Proc Natl Acad Sci U S A* **104**: 1766–1770.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete structure of the chloroplast genome of Arabidopsis thaliana. *DNA Res* **6**: 283–290.

Shedge V, Davila J, Arrieta-Montiel MP, Mohammed S, Mackenzie SA. 2010. Extensive rearrangement of the Arabidopsis mitochondrial genome elicits cellular conditions for thermotolerance. *Plant Physiol* **152**: 1960–1970.

Sloan DB. 2013. One ring to rule them all? Genome sequencing provides new insights into the “master circle” model of plant mitochondrial DNA structure. *New Phytol* **200**: 978–985.

Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, Havird JC. 2018a. Cytonuclear integration and co-evolution. *Nat Rev Genet* **19**: 635–648.

Sloan DB, Wu Z, Sharbrough J. 2018b. Correction of persistent errors in Arabidopsis reference mitochondrial genomes. *Plant Cell* **30**: 525–527.

Sun M, Zhang M, Chen X, Liu Y, Liu B, Li J, Wang R, Zhao K, Wu J. 2022. Rearrangement and domestication as drivers of Rosaceae mitogenome plasticity. *BMC Biol* **20**: 181.

Trösch R, Ries F, Westrich LD, Gao Y, Herkt C, Hoppstädter J, Heck-Roth J, Mustas M, Scheuring D, Choquet Y, et al. 2022. Fast and global reorganization of the chloroplast protein biogenesis network during heat acclimation. *Plant Cell* **34**: 1075–1099.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. *Nat Genet* **15**: 57–61.

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

Walker JF, Jansen RK, Zanis MJ, Emery NC. 2015. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. *Am J Bot* **102**: 1751–1752.

Wang J, Kan S, Liao X, Zhou J, Tembrock LR, Daniell H, Jin S, Wu Z. 2024. Plant organellar genomes: much done, much more to do. *Trends Plant Sci* **29**: 754–769.

Wang N, Li C, Kuang L, Wu X, Xie K, Zhu A, Xu Q, Larkin RM, Zhou Y, Deng X, et al. 2022. Pan-mitogenomics reveals the genetic basis of cytonuclear conflicts in citrus hybridization, domestication, and diversification. *Proc Natl Acad Sci U S A* **119**: e2206076119.

Wang W, Lanfear R. 2019. Long-Reads Reveal That the Chloroplast Genome Exists in Two Distinct Versions in Most Plants. *Genome Biol Evol* **11**: 3372–3381.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**: 43–49.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**: 3350–3352.

Wilson PB, Streich JC, Murray KD, Eichten SR, Cheng R, Aitken NC, Spokas K, Warthmann N, Gordon SP, Accession Contributors, et al. 2019. Global Diversity of the Brachypodium Species Complex as a Resource for Genome-Wide Association Studies Demonstrated for Agronomic Traits in Response to Climate. *Genetics* **211**: 317–331.

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.

Wynn EL, Christensen AC. 2019. Repeats of Unusual Size in Plant Mitochondrial Genomes: Identification, Incidence and Evolution. *G3* **9**: 549–559.

Xian W, Bezrukov I, Bao Z, Vorbrugg S, Gautam A, Weigel D. 2024. TIPP_plastid: A User-Friendly Tool for De Novo Assembly of Organellar Genomes with HiFi Data. *bioRxiv* 2024.01.29.577798. https://www.biorxiv.org/content/10.1101/2024.01.29.577798v2 (Accessed August 12, 2024).

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K, et al. 2022. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**: 527–534.

Zou Y-P, Hou X-H, Wu Q, Chen J-F, Li Z-W, Han T-S, Niu X-M, Yang L, Xu Y-C, Zhang J, et al. 2017. Adaptation of Arabidopsis thaliana to the Yangtze River basin. *Genome Biol* **18**: 239.

Zou Y, Zhu W, Sloan DB, Wu Z. 2022. Long-read sequencing characterizes mitochondrial and plastid genome variants in Arabidopsis msh1 mutants. *Plant J* **112**: 738–755.

minitv: Alignment frontend for AliTV. Github https://github.com/weigelworld/minitv (Accessed August 9, 2024).

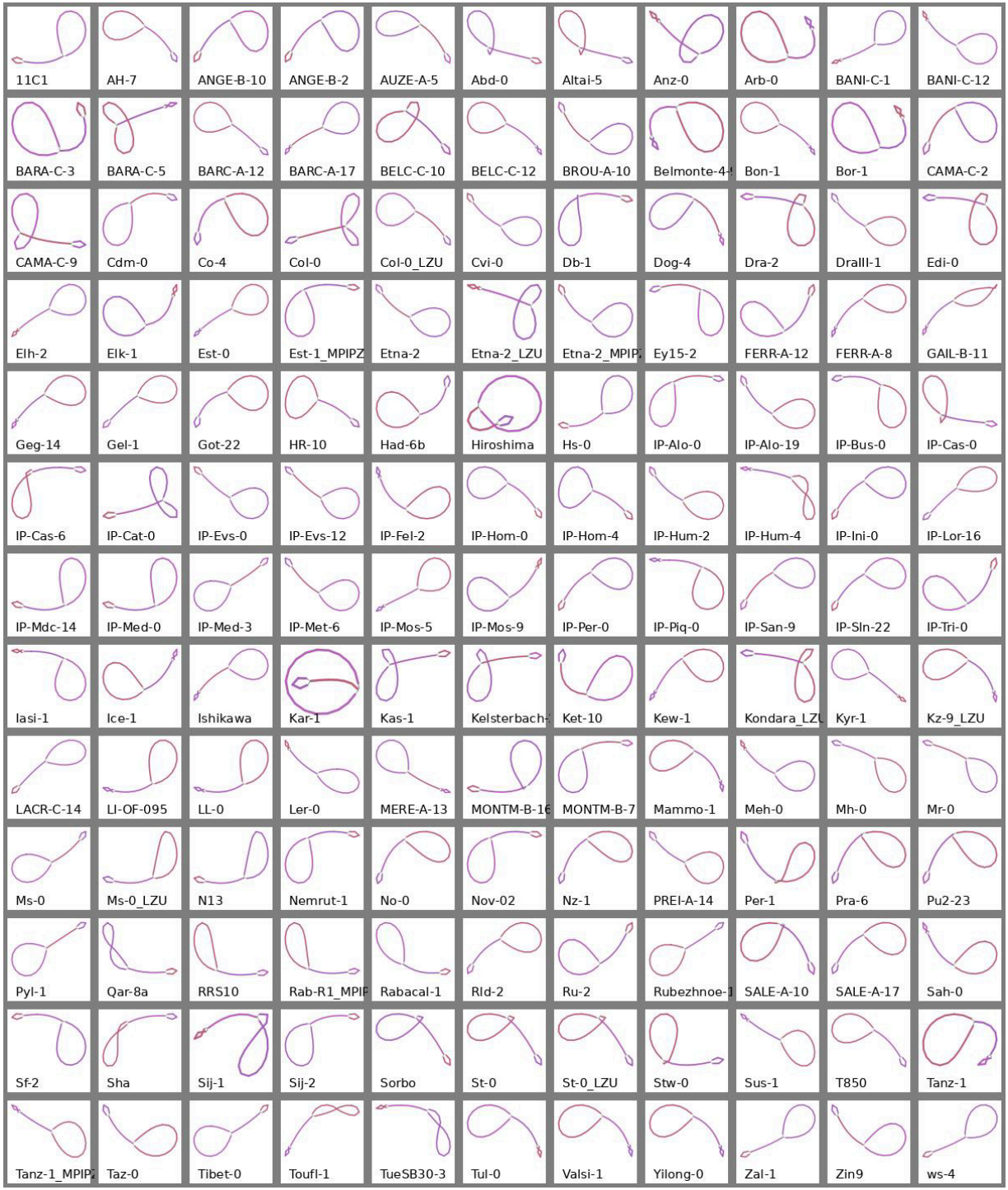


Figure S1. Visualization of chloroplast assembly graphs for 143 *Arabidopsis thaliana* accessions.

Figure S2. Visualization of mitochondrial assembly graphs for 143 *Arabidopsis thaliana* accessions.

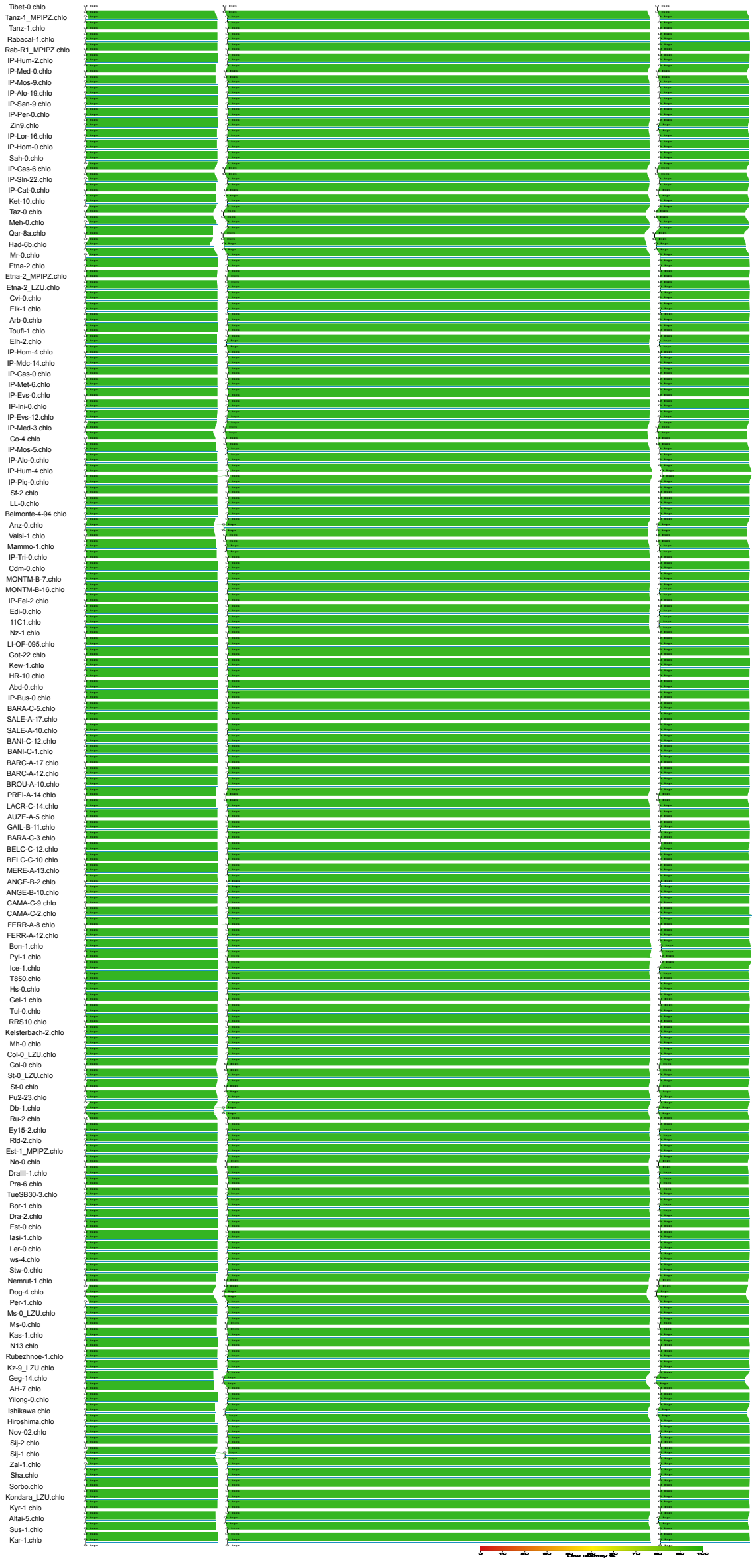


Figure S3. Multiple whole genome alignment of chloroplast genome among 143 accessions.

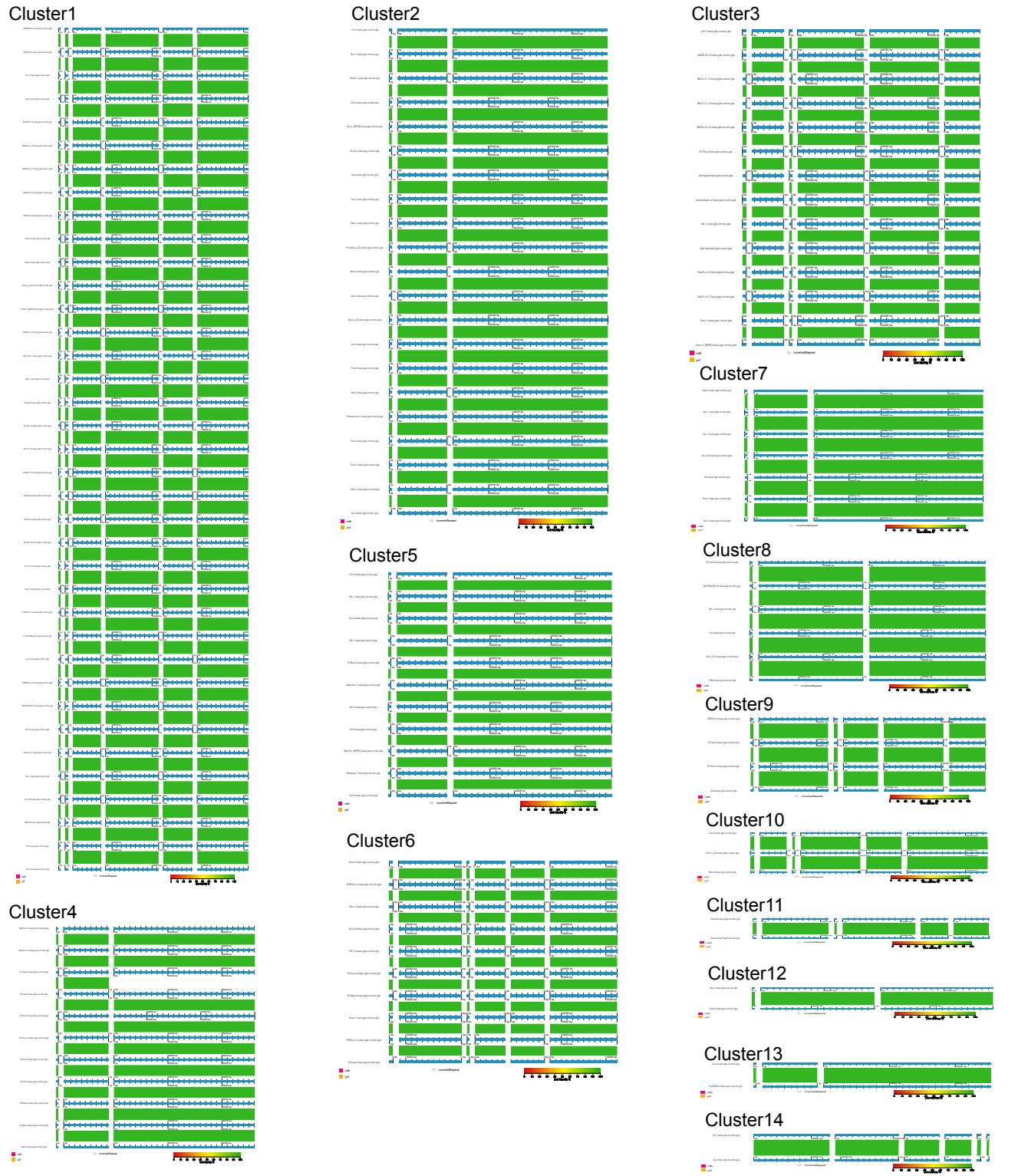


Figure S4. Mulitple / pair whole genome alignment for each mitochondrial genome cluster.

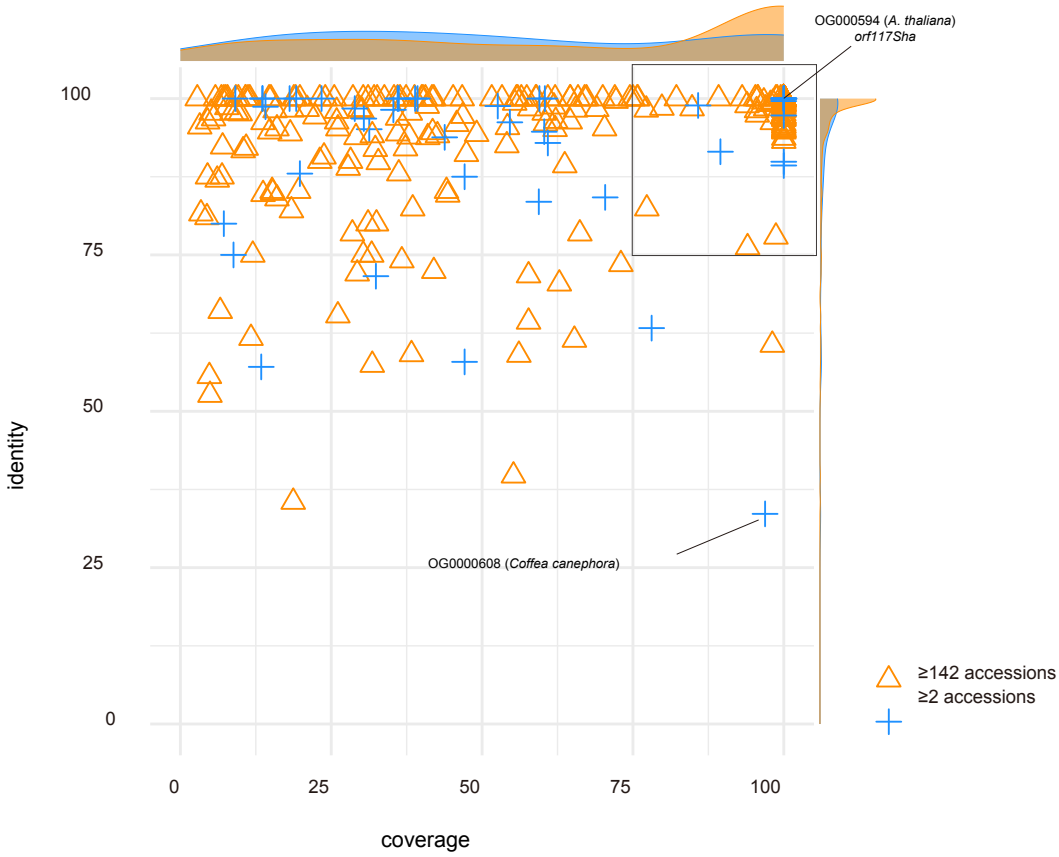


Figure S5. The alignment of representative sequences from mitochondrial genes orthologous groups to uniref90 database.

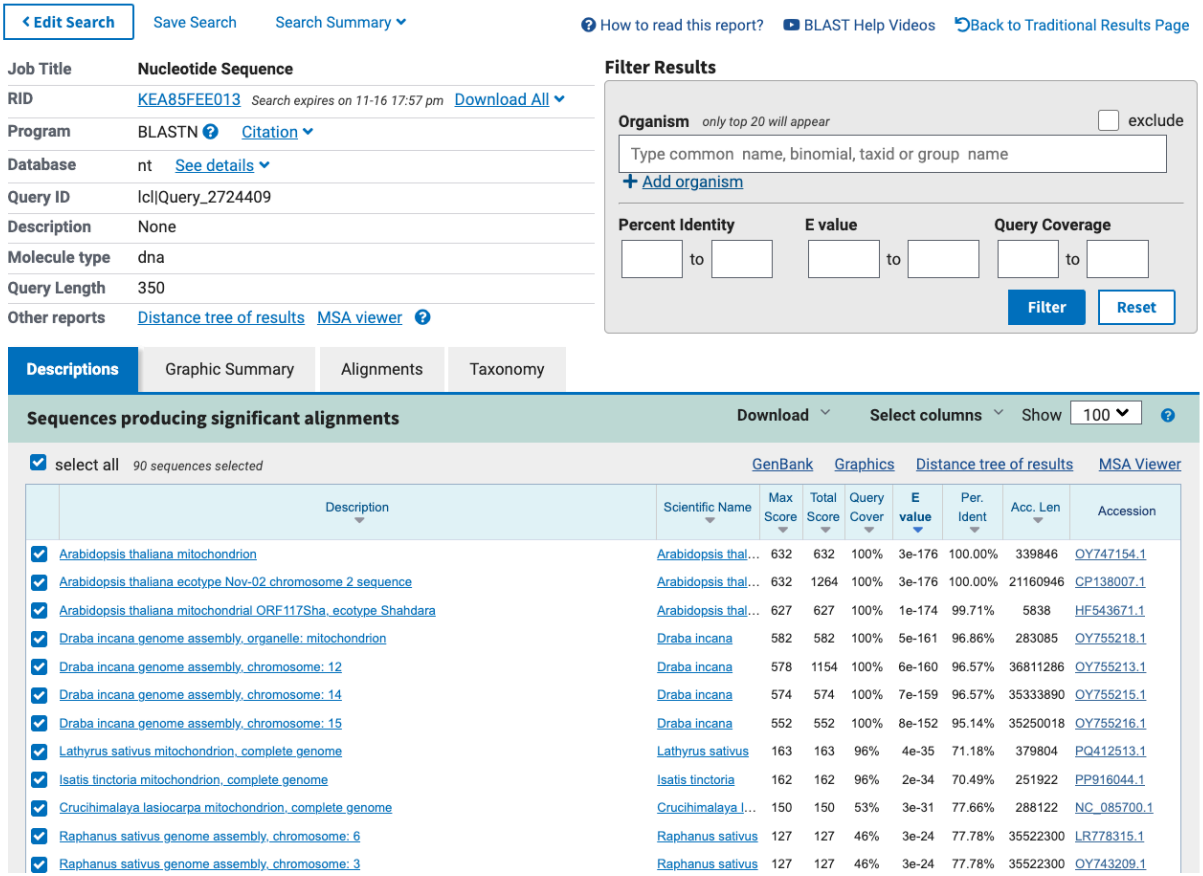


Figure S6. The Sha orf117 sequence was aligned against the NCBI NT database using BLAST.

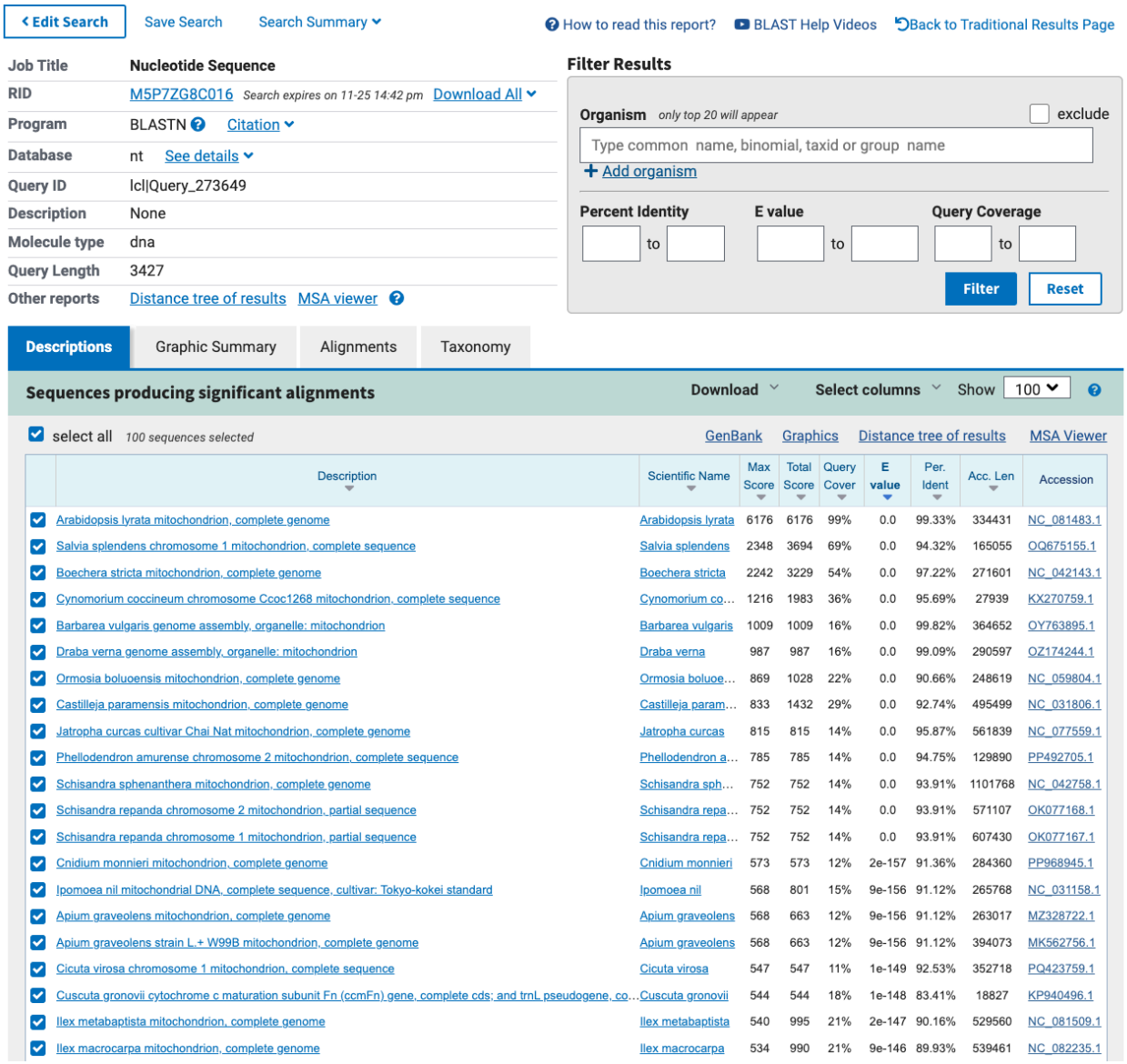


Figure S7. The sequence of OG0000602-612 was aligned against the NCBI NT database using BLAST in NCBI.

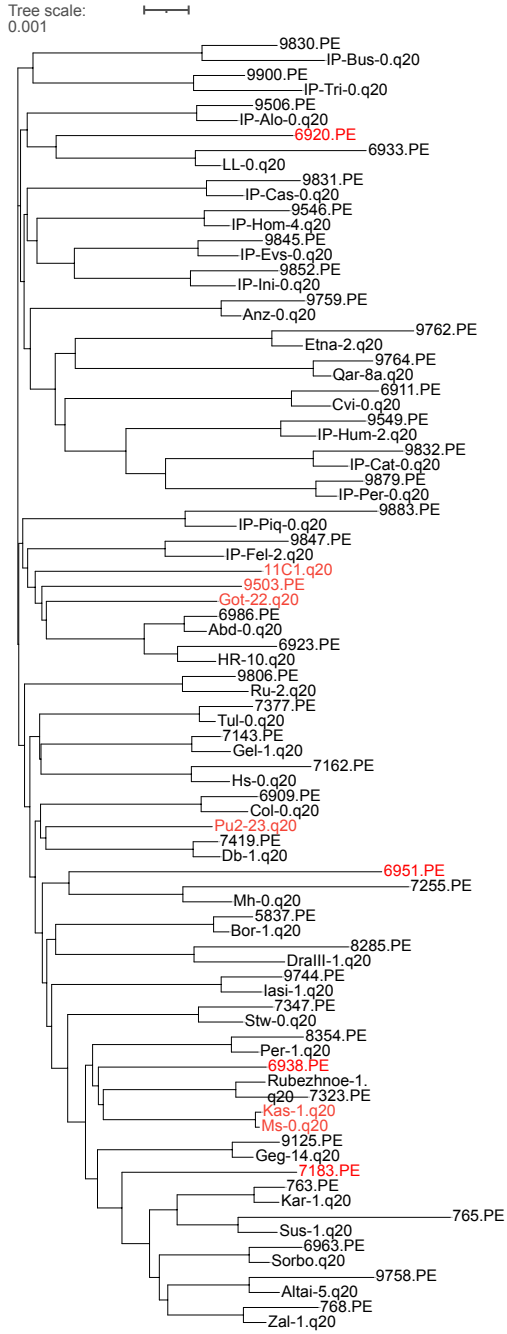


Figure S8. Phylogenetic Tree of Short Read and HiFi Read Data. IDs with the "q20" suffix represent HiFi data, while IDs with the "PE" suffix represent short-read data. Accessions with inconsistencies are marked in red and were excluded from the analysis.

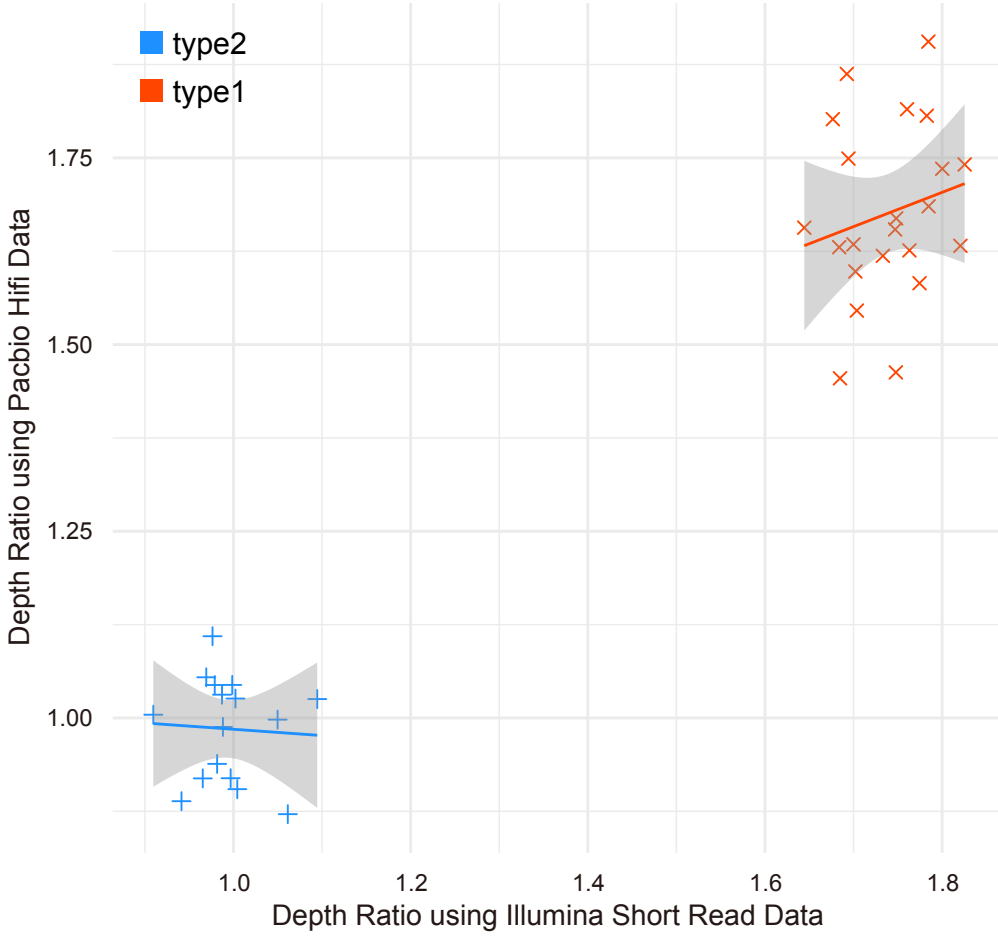


Figure S9. Distribution of Depth Ratios Inferred from Short Reads and HiFi Reads. Each dot represents one accession.

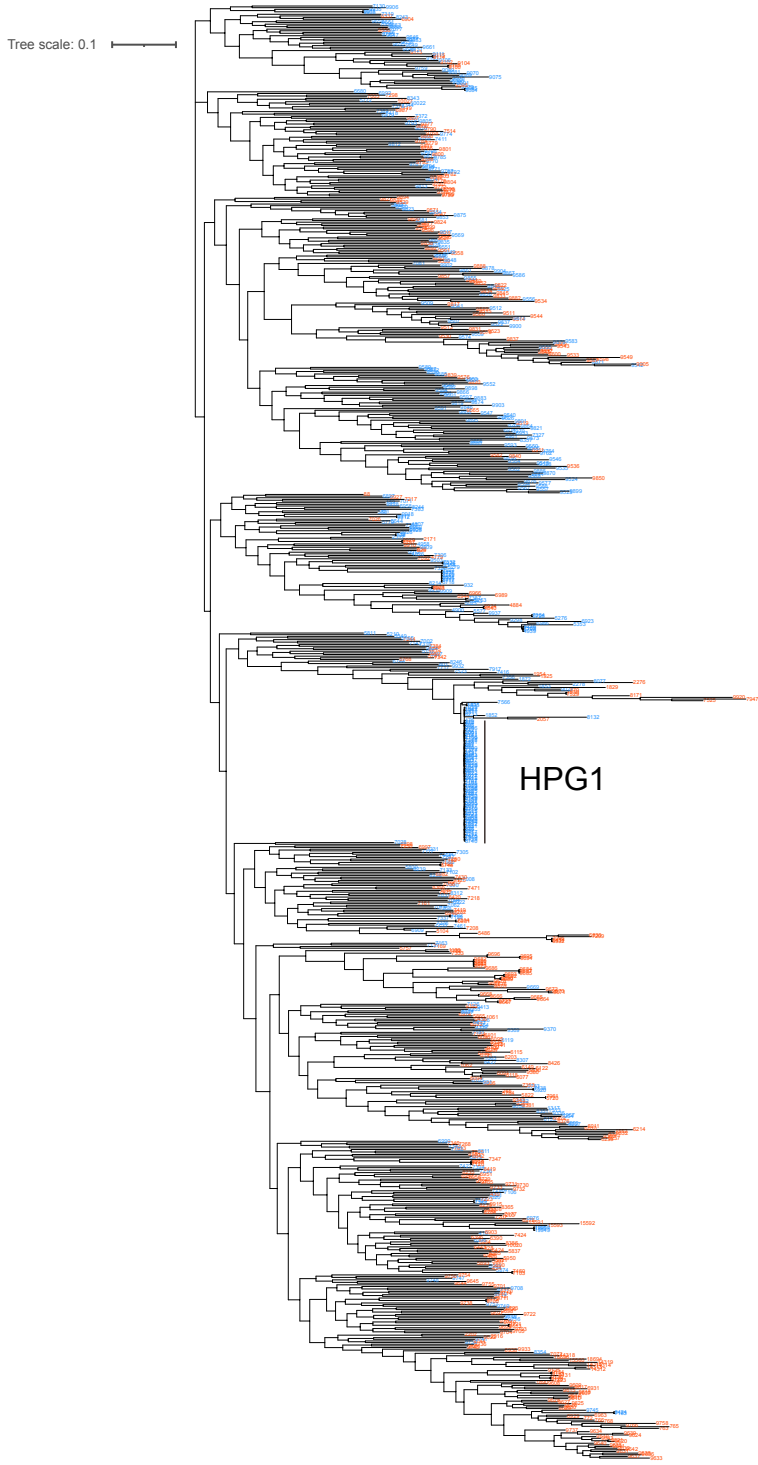


Figure S10. Distribution of Type 1 and Type 2 Mitochondria genomes projected onto the phylogenetic tree. Red IDs represent Type 1, while blue IDs represent Type 2.

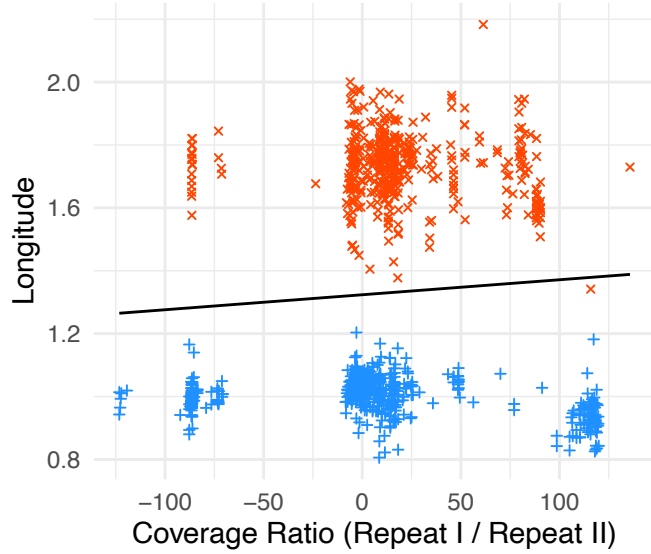


Figure S11. Correlation between coverage ratio and longitude.

Thesis Appendix III

G3, 2025, **15(8)**, jkaf143
https://doi.org/10.1093/g3journal/jkaf143
Advance Access Publication Date: 21 June 2025
Plant Genetics and Genomics

Minimizing detection bias of somatic mutations in a highly heterozygous oak genome

Wenfei Xian,[1] Pablo Carbonell-Bejerano ,[1,2] Fernando A. Rabanal ,[1] Ilja Bezrukov,[1,*] Philippe Reymond ,[3] Detlef Weigel [1,4,*]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen 72076, Germany
[2]Instituto de Ciencias de la Vid y del Vino, ICVV, CSIC, Universidad de La Rioja, Gobierno de La Rioja, Logroño, La Rioja 26007, Spain
[3]Department of Plant Molecular Biology, University of Lausanne, Lausanne 1015, Switzerland
[4]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen 72076, Germany

*Corresponding author: Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen 72076, Germany. Email: ilja.bezrukov@tuebingen.mpg.de;
*Corresponding author: Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Tübingen 72076, Germany. Email: weigel@tue.mpg.de

Somatic mutations are particularly relevant for long-lived organisms. Sources of somatic mutations include imperfect DNA repair, replication errors, and exogenous damage such as ultraviolet radiation. A previous study estimated a surprisingly low number of somatic mutations in a 234-year-old individual of the pedunculate oak (*Quercus robur*), known as the Napoleon Oak. It has been suggested that the true number of somatic mutations was underestimated due to gaps in the reference genome and too conservative filtering of potential mutations. We therefore generated new high-fidelity long-read data for the Napoleon Oak ($n = 12$) to produce both a pseudo-haploid genome assembly and a partially phased diploid assembly. The high heterozygosity allowed for complete reconstruction of phased and gapless centromeres for 22 of the 24 chromosomes. On the other hand, the high heterozygosity posed challenges for short-read alignments. Use of only the pseudo-haploid assembly as a reference led to potential misalignments, while use of only the diploid assembly reduced variant detection sensitivity. Since most somatic mutations are layer-specific, the fraction of reads covering a specific somatic mutation is expected to be relatively low, even where all cells in a single layer contain a specific mutation. To address this challenge, we employed a read assignment strategy, selecting the appropriate reference sequence (pseudo-haploid or diploid) based on alignment score and mapping quality. Ultimately, we identified 198 high-confidence somatic mutations, compared with 17 somatic mutations identified before with the same set of short reads. Our approach thus increased the total estimated annual mutation rate by a factor of 5.

Keywords: Somatic mutations; *Quercus robur*; HiFi assembly

Introduction

Fifteen years ago, a seminal study by Sally Otto and colleagues showed that mutations reducing male fertility accumulate in a long-lived deciduous tree in a clock-like manner, suggesting that there are natural limits to how long trees could produce viable offspring (Ally *et al.* 2010). Since that early paper, a series of studies has attempted to directly detect the accumulation of mutations in trees using a range of sequencing strategies, resulting in a remarkably wide range of mutation rate estimates (summarized in Johannes 2024).

Oaks, widely distributed in Northern Hemisphere forests, are renowned for their longevity, with many individuals living for centuries (Leroy *et al.* 2020). One noteworthy, over 200-year-old specimen, known as the Napoleon Oak, which grows on the campus of the University of Lausanne, was the subject of a recent somatic mutation study (Schmid-Siegert *et al.* 2017), which used a reference genome assembled from long reads of this individual in combination with short reads of material from a lower and an upper branch. Ten single nucleotide variants (SNVs) were identified in the upper branch and 7 SNVs in the lower branch. Taking into account potential false negatives, the fixed somatic mutation rate was estimated to be $4.2–5.2 \times 10^{-8}$ substitutions per site per generation (Schmid-Siegert *et al.* 2017). Subsequent work with another oak tree suggested that the number of somatic mutations identified in the Napoleon Oak was a likely underestimate because the minimization of false positives (specificity) had been prioritized over the identification of all true positives (sensitivity) (Plomion *et al.* 2018). Apart from the reference genome having been generated with an early long-read technology, the threshold for variant calling might have been overly stringent. In addition, the high heterozygosity of oak genomes could have impacted variant calling, as only a single phase of the Napoleon Oak genome had been assembled.

When reads are aligned to a haploid assembly, regions that are highly divergent in sequence between the 2 haplotypes might produce misalignments, as a haploid reference cannot fully represent both haplotypes. This has been quantitatively confirmed in highly heterozygous sweet oranges, where a phased assembly as reference yielded more than twice the number of somatic mutations compared to a haploid assembly (Wang *et al.* 2024). Using a diploid reference is, however, not without its own challenges. In regions that are identical in the 2 haplotypes, aligners will randomly assign reads to 1 of the 2 phases, reducing the power of variant

Received on 13 February 2025; accepted on 29 May 2025

identification, particularly when there are only very few variant reads covering a specific position. Conversely, high-frequency variants may appear on both haplotypes, leading to an overestimate of true variants.

A recent study of 2 tropical trees found that almost all somatic mutations are present at low frequency in the sampled tissue (Schmitt et al. 2024). This finding highlights the difficulty of distinguishing somatic mutations that are not present in all cells of the sampled tissue from sequencing errors. This is a particular issue for plants, where multiple tissue layers are established early on in development, with these layers staying largely segregated throughout the life of the plant (Steeves and Sussex 1989). Thus, even somatic mutations that are present in all cells of a layer in the sampled tissue may go undetected if allele frequency thresholds are applied too strictly. Several studies have confirmed that most somatic mutations are layer-specific, as has long been known from the study of mutations with phenotypic impact, such as grape color (Dermen 1960; Amundson et al. 2023; Goel et al. 2024). Since the samples sequenced in the Napoleon Oak study were the entire leaves (Schmid-Siegert et al. 2017), this might have further reduced the observed frequency of detected somatic mutations. These considerations underscore the need for more sensitive alignment and variant-calling strategies that can accurately capture true mutations, even when present at low frequencies.

We produced a new genome assembly of the Napoleon Oak and optimized a somatic mutation-calling pipeline for highly heterozygous genomes, using both haploid and diploid reference genomes. Utilizing this improved approach, we identify a larger set of high-confidence somatic mutations in the Napoleon Oak, which lead to a substantial upward revision of the annual somatic mutation rate.

Methods

Biological material and PacBio HiFi sequencing

Six grams of leaves were sampled on August 2021 from the bottom branch of the Napoleon Oak (*Quercus robur*) on the campus of the University of Lausanne (Switzerland, 46° 31′ 18.9″ N, 6° 34′ 44.5″ E). Leaves were stored in liquid nitrogen. Prior to DNA extraction, to avoid contamination from pathogens, we removed any visibly spotted areas of the leaves. To improve DNA extraction efficiency, we also trimmed away the leaf veins. High molecular weight (HMW) DNA was obtained from frozen leaf powder as described by Calderón et al. (2024). In brief, the Nanobind Plant Nuclei Big DNA Kit (Circulomics) was used to extract HMW DNA from nuclei isolated according to Workman et al. (2018). We sheared the HMW DNA into appropriately sized fragments using a g-TUBE and then prepared a PacBio library using the SMRTbell Express Template Prep Kit 2.0 from 10 µg of sheared DNA. The library was selected for >10 kb fragments using a BluePippin instrument (Sage Science). Sequencing was performed using two 8 M SMRTCells on a Sequel II.

Genome assembly and annotation

PacBio CCS (https://github.com/PacificBiosciences/ccs) (v6.4.0) was used to generate CCS reads from the raw subreads with the parameter --min-rq=0.88. Subreads were aligned to CCS reads with ACTC (https://github.com/PacificBiosciences/actc) (v0.3.1). CCS reads were further polished by DeepConsensus (Baid et al. 2023) (v1.2.0) resulting in 63.65 Gb high-quality HiFi reads. Short and long reads of sample 0 were used to estimate the heterozygosity from kmer frequencies using KMC3 and GenomeScope v2.0.

HiFi reads were assembled into a haploid assembly using Hifiasm v0.24.0 (Cheng et al. 2021), producing 16 long contigs. Based on length ranking, the 16th contig was 3.5 Mb long, while the 17th contig was much smaller, only 0.5 Mb. Therefore, we retained only the 16 longer contigs. Among these, 8 contained telomeric repeats at both ends, and the remaining 8 contigs had telomeric repeats at only 1 end. A diploid assembly was produced with Verkko v1.4.1 (Rautiainen et al. 2023) with default parameters.

Commonly used scaffolding methods rely on reference genomes (Alonge et al. 2022), but this approach may introduce reference bias. Moreover, only 4 gaps remained in our haploid assembly, 3 of which likely contained rDNA clusters. Previous studies indicated that rDNA clusters form tangles in the Verkko assembly graph, and we visualized the Verkko diploid assembly graph (Rautiainen et al. 2023) assembly with Bandage v0.9 (Wick et al. 2015) to identify tangles caused by rDNA clusters. We manually selected the node IDs corresponding to each tangle in Bandage —representing 4 haplotype-phased 45S rDNA clusters and 1 haplotype-collapsed 5S rDNA cluster—and used them as input for Ribotin to generate consensus sequences. The 18S, 5.8S, and 25S rDNA units from *Arabidopsis thaliana* were aligned to each 45S rDNA consensus sequence, and the 5S rDNA unit from *A. thaliana* was aligned to the oak 5S rDNA consensus sequence using both BWA 0.7.17-r1188 (Li and Durbin 2009) and BLASTN 2.12.0+ (Camacho et al. 2009). A flow diagram of the described assembly steps is shown in Supplementary Fig. 14.

Using Minimap2 2.24-r1122 (Li 2018), we aligned the 8 contigs with telomeric repeats at only 1 end to the Verkko assembly. From the unitig-popped.layout.scfmap file, we identified the corresponding nodes in the assembly graph. Based on the connectivity information in the graph, we manually linked 2 contigs with 100 Ns, ultimately obtaining 12 pseudo-chromosomes. Chromosomes were assigned to these 12 scaffolds based on alignment to the 3P Oak genome assembly (Plomion et al. 2018). Therefore, the haploid assembly refers to the scaffolded sequences, while the diploid assembly corresponds to the Verkko contigs. We assembled the oak organellar genome using TIPPo v2.1 (Xian et al. 2025).

Quality evaluation of our haploid assembly with Merqury v1.3 indicated a QV of 53 (Rhie et al. 2020) with short reads only, and Compleasm assessment showed 99% completeness of the embryophyta_odb10 conserved gene set (Huang and Li 2023). The same evaluations were also performed on the previous assembly of the Napoleon Oak (Schmid-Siegert et al. 2017) for comparison.

Liftoff (Shumate and Salzberg 2021) was used to project protein-coding genes from the 3P Oak assembly (Plomion et al. 2018) onto the Napoleon Oak assembly. Transposable elements were annotated with EDTA v2.0.0 (Ou et al. 2019) with parameters --sensitive 1 --anno 1. CpG methylation profiles were identified from HiFi reads by ccsmeth (Ni et al. 2023). Satellite repeats were annotated with Tandem Repeat Annotation and Structural Hierarchy (TRASH) (Wlodzimierz et al. 2023).

To verify the accuracy of the CEN146 array, HiFi reads were aligned to the diploid assembly with Minimap2 (Li 2018), Winnowmap2 (Jain et al. 2022), and VerityMap (Bzikadze et al. 2022). StainedGlass (Vollger et al. 2022) and pyGenomeTracks (Lopez-Delisle et al. 2021) were used to visualize HiFi read coverage, TE annotation, and the similarity of the CEN146 arrays.

CEN146 arrays from homologous chromosomes were aligned separately using UniAligner (Bzikadze and Pevzner 2023) and Minimap2 (Li 2018). Alignment length and mismatch count were analyzed using UniAligner's built-in script, cigar_histogram.py. Dot plots of rare kmers were generated using smart_dotplot.py, also from UniAligner. For Minimap2 alignments, dot plot

visualization was performed using paf2dot (https://github.com/pangenome/paf2dot). To select unique homologous chromosomes, we manually identified 2 contigs forming a bubble in the assembly graph and aligned them using Minimap2. The specific sequences can be found in Supplementary Table 7.

Short-read processing and alignment

The Illumina short reads for the upper and lower branches have been published (Schmid-Siegert *et al.* 2017) and were downloaded from NCBI (accession PRJNA327502). Reads were trimmed using fastp v0.23.1 (Chen *et al.* 2018) with the parameters -q 20 -l 75 -w 12 --cut_tail --cut_mean_quality 20.

Trimmed reads were separately aligned to the diploid and haploid assemblies using BWA 0.7.17-r1188 (Li and Durbin 2009). Potential PCR duplicates were marked using Picard v2.26.7 ("Picard toolkit" 2019).

Alignment comparison between haploid and diploid assemblies as reference

We compared insert size distribution, pairing orientation, and the proportion of properly paired reads between haploid and diploid assemblies as reference using samtools v1.10 stat.

To identify the potential misalignment of short reads, we first aligned the short reads to the diploid assembly to serve as a reference framework and then extracted the flanking 2 kb sequences for every read and aligned them to the haploid assembly with Minimap2 2.24-r1122. Using the flanking sequences from the diploid assembly as an intermediary, we inferred the expected positions of short reads in the haploid assembly and compared these to the actual alignment positions when short reads were mapped directly against the haploid genome. Reads that did not align within the range of expected positions were identified (Supplementary Fig. 10).

We used DeepVariant v1.8.0 (Poplin *et al.* 2018) to call germline SNVs on alignments of short reads from either the upper or lower branch to the haploid assembly. To ensure the reliability of variants identified in the short-read dataset, we retained only variants with a depth between 0.5× and 2× sample coverage, a minimum variant allele frequency (VAF) of 0.3, and those marked as PASS. Because DeepVariant may produce false negatives, we extracted the variants identified in the short-read data and used Bam-readcount v1.0.1 (Khanna *et al.* 2022) to verify whether these variants were present in the HiFi data. We applied a very lenient threshold: if a variant was supported by at least 5 HiFi reads (the genome-wide coverage was 77×), we considered it reproduced in HiFi data.

Identification of somatic mutations

We extracted the MAPQ and alignment scores for each read from the BAM files of short reads aligned to either the haploid or diploid reference genome. If the alignment scores differed, a read was assigned to the reference genome with the higher alignment score. If the alignment scores were the same, the read was assigned to the reference genome with the higher MAPQ value. If the MAPQ values were also identical, the read was assigned to the diploid assembly. Clipped reads were removed. Reads from the upper and lower branches were used as the mutant and control sample, respectively.

For the diploid assembly as the reference genome, we used bam-readcount v1.0.1 (Khanna *et al.* 2022) with MAPQ of ≥0 and base quality of ≥0 for the control sample. These lenient thresholds were used to reduce false negatives from the control. Only references called sites with coverage of ≥20×, and no indels, were collected as the "set of control positions".

For the mutant sample, to reduce false positives, we applied stringent parameters using bam-readcount v1.0.1 (Khanna *et al.* 2022): MAPQ of ≥10, base quality of ≥30, and site coverage of ≤70×. Sites had to have biallelic bases without indels. The alternate allele had to have at least 3 supporting reads, with support on both strands. Reads with the reference allele had to have an average mismatch of ≤0.01. The average mismatch of reads carrying the alternative allele minus the mismatch of reads carrying the reference allele had to be ≤0.01. Only variant sites in the mutant sample overlapping with the "set of control positions" identified in the control sample were retained as likely somatic mutations.

For the haploid assembly as the reference genome, we used a similar workflow for the diploid assembly as reference, but with the following adjustments: control sample, coverage of >40×, mutant sample, and coverage of ≤120×. The somatic mutations identified against both the haploid and diploid assemblies were then combined. For comparison, Strelka2 v2.9.10 (Kim *et al.* 2018) was used to detect the somatic mutations using separately either the haploid or diploid assembly.

In the diploid assembly, we require a minimum depth of 20× for the control sample. We then quantified regions in the diploid assembly where the sequencing depth is above 20×, resulting in an effective site size of 1.3 Gb.

To identify the position of previously validated SNVs in the new assemblies, we extracted the flanking 100 bp from the previous Napoleon Oak assembly (Schmid-Siegert *et al.* 2017) and aligned the sequences to both our haploid and diploid assemblies with Minimap2 (Li 2018). To calculate the allele frequency of the 9 somatic SNVs shared between leaves and acorns in the 3P Oak (Plomion *et al.* 2018), we downloaded short reads from NCBI (accession number PRJEB8388). We used the same approach as for the Napoleon Oak to process and align the reads to the assembly. We extracted the allele frequency of the 9 SNVs from the bam file.

Results

Genome architecture of *Quercus robur*

To identify high-quality somatic mutations in the diploid pedunculate oak (*Q. robur*, $n = 12$), individually known as the Napoleon Oak, we first assembled a high-quality genome of this individual. In August 2021, fresh leaves were sampled from the bottom branch, at the same location where the material used for genome assembly and mutation calling had been sampled before (sample 0 of Schmid-Siegert *et al.* 2017). Two PacBio SMRTcells were used to generate a total of 63.7 Gb of HiFi data (Supplementary Table 1). Using Hifiasm (Cheng *et al.* 2021), we assembled a highly contiguous haploid genome comprising only 16 contigs (Fig. 1a). Eight of the contigs contained telomeric repeats at both ends and thus corresponded to 8 complete chromosomes.

We estimated a high level of heterozygosity, about 1.6%, for the Napoleon Oak by kmer-based analysis of Illumina short and PacBio long reads (Supplementary Fig. 1). Because the haploid Hifiasm assembly only partially represents all of the sequences present in the genome, we assembled a diploid genome from the same HiFi data using Verkko (Rautiainen *et al.* 2023; Fig. 1b).

We used information from the Verkko assembly graph to scaffold the 8 contigs of the Hifiasm assembly that did not have telomeres on both ends into additional chromosomes (Supplementary Table 2). One of the major challenges in assembling complete plant chromosomes comes from rDNA clusters

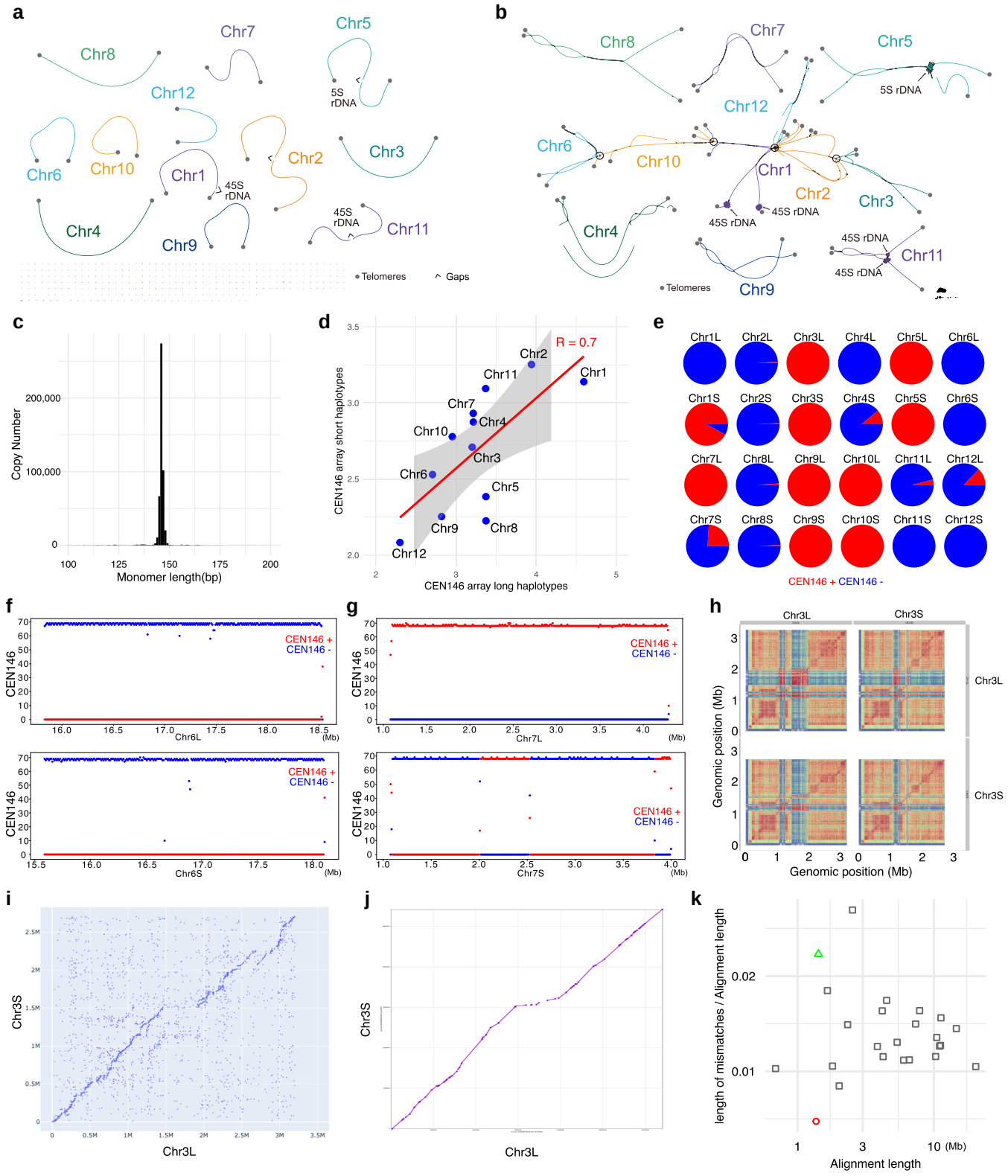


Fig. 1. Genome architecture of the Napoleon Oak. a) Scaffolded haploid assembly graph. b) Contig-level diploid assembly graph. c) Copy number of centromeric satellite repeats in the diploid assembly. d) Length of CEN146 arrays on different chromosomes. e) Orientation distribution of CEN146: forward (red, +) and reverse (blue, –) proportion for each chromosome and haplotype (homolog). L, longer haplotype; S, shorter haplotype. f) Strand-specific density of CEN146 in 10-kb windows along the CEN146 arrays on the 2 chromosome 6 haplotypes. Forward strand density in red and reverse strand density in blue. g) Strand-specific density of CEN146 in 10-kb windows along the CEN146 arrays on the 2 chromosome 7 haplotypes. Forward strand density in red and reverse strand density in blue. h) Heat maps depicting sequence identity within and between the 2 haplotypes of the CEN146 array on chromosome 3. i) Dot plot of rare kmers showing that the 2 haplotypes of the CEN146 array on chromosome 3 can be aligned with UniAligner. j) Minimap2 alignment of the CEN146 array on chromosome 3. k) Apparent substitution rates across various regions: green triangle represents chromosome 3 CEN146 arrays aligned using Minimap2, red circle indicates alignments using UniAligner, and gray rectangles indicate unique non-CEN146 sequences aligned with Minimap2. For details of the unique regions, see Supplementary Table 7.

(Rabanal *et al.* 2022; Fultz *et al.* 2023), and the genome of the pedunculate oak contains two 45S rDNA clusters and one 5S rDNA cluster (Bočkor *et al.* 2014). In the Verkko assembly graph, there were prominent tangles that connected sequences at both ends. Based on published descriptions (Rautiainen *et al.* 2023), we expected these tangles to correspond to rDNA clusters. We derived consensus sequences of each rDNA clusters and found these to have high similarity (>95%) to those of *A. thaliana*, although the lengths of the 45S rDNA units differed between chromosome 1 (9.8 kb) and chromosome 11 (7.7 kb) (Supplementary Table 15 and Supplementary Fig. 13).

We aligned the 8 contigs of the Hifiasm assembly without telomeres on both ends to the diploid assembly and projected them onto the diploid assembly graph (Supplementary Fig. 2). Based on the connections in the graph, we manually scaffolded the 8 contigs, resulting in 4 additional chromosome-level pseudomolecules. Each had 1 end capped by telomere repeats and large arrays of tandem repeats (either rDNA or another, unknown repeat type) at the other end (Supplementary Table 2). The final haploid genome had high accuracy (QV 53) and high completeness (BUSCO 99.8%), significantly surpassing the quality of the previous version generated from an earlier generation of long reads (Supplementary Fig. 3 and Supplementary Table 3) (Schmid-Siegert *et al.* 2017). The assembled haploid genome spans 810 Mb, in agreement with estimates for the 1C genome size for this species, which range from 759 to 1,068 Mb (Pellicer and Leitch 2020).

The long and highly accurate PacBio HiFi reads enabled the assembly of repetitive regions, such as centromeres (Naish *et al.* 2021). In each of the 12 chromosomes of the haploid assembly, we identified large arrays of a shared 146 bp satellite repeat unit, with a genome-wide total of 486,051 copies in the diploid assembly (Fig. 1c; Supplementary Fig. 4 and Supplementary Table 4) and with high levels of CpG methylation (Supplementary Fig. 5). These arrays, on average about 3 Mb long, are likely centromeres, with 22 of 24 being gapless and phased. In analogy with the *A. thaliana* nomenclature, we call these repeats CEN146 (Naish *et al.* 2021). The CEN146 arrays on chromosome 2 were manually scaffolded (Supplementary Table 5). Uniform coverage of HiFi read alignments suggest that there are no significant assembly errors (Supplementary Fig. 6). Within the CEN146 arrays, transposons were rare, but transposons were present in the flanking regions (Supplementary Fig. 6).

For each set of homologs, we categorized the CEN146 arrays as belonging to a long haplotype (L) or short haplotype (S) based on their lengths. The size of CEN146 arrays differed more between chromosomes than between homologs (Fig. 1d). The relative orientation of CEN146 repeats within the arrays was homolog (haplotype) rather than chromosome-specific (Fig. 1e–g). For instance, on Chr1S, Chr4S, Chr7S, and Chr12L, more than 8% of repeat units were in an orientation opposite to the majority of repeats, whereas in their counterparts (Chr1L, Chr4L, Chr7L, and Chr12S), over 99% of arrays were in the same orientation.

Due to the challenges of aligning long highly repeated sequences, previous studies have often greatly differed in their estimates for evolutionary turnover of centromere sequences (Logsdon *et al.* 2021; Bzikadze and Pevzner 2023). Here, we attempted to use our phased CEN146 arrays to examine the performance of 2 different alignment approaches for the discovery of centromeric variants: Minimap2 (Li 2018), based on standard concepts of molecular evolution, and UniAligner (Bzikadze and Pevzner 2023), which is designed for aligning long tandem repeats. First, we used UniAligner to align the 2 haplotypes of the CEN146 arrays of each chromosome. Dot plots of rare kmers revealed that only chromosome 3 contained a substantial number of rare kmers, indicating that alignment was feasible only for chromosome 3 (Fig. 1i and Supplementary Fig. 7). Because only chromosome 3 produced a Minimap2 alignment consistent with the UniAligner dot plot (Fig. 1j and Supplementary Fig. 8), we focused on comparing the CEN146 arrays of this chromosome (Fig. 1h). As UniAligner uses rare kmers as anchors to extend alignments, we used the alignment blocks of Minimap2 with the maximally possible mapping quality (MAPQ) of 60. The total alignment lengths were similar for the 2 tools, with UniAligner aligning 1.36 Mb and Minimap2 aligning 1.41 Mb. However, the substitution rate inferred by UniAligner is 0.004, which is 5 times lower than the one inferred by Minimap2, 0.02. Notably, the substitution rate from Minimap2 was higher than that of most non-CEN146 uniquely aligned regions using Minimap2 (Fig. 1k and Supplementary Tables 6 and 7). Because the 2 tools produced such diverging results, we excluded the centromeres from our analyses of somatic mutation rates.

Variant allele frequency estimates of layer-specific mutations

Our goal is to identify somatic mutations that had become fixed in stem cells that gave rise to the sampled tissue (Schmid-Siegert *et al.* 2017). We take the cell layer where the mutations occur into account, since the layer specificity of somatic mutations has not only long been known from phenotypic examination (Dermen 1960), subsequently confirmed by molecular analysis of specific mutations (Kobayashi *et al.* 2004), but has also recently shown to be the prevailing type by high-throughput sequencing in potato, apricot, and apple (Amundson *et al.* 2023; Goel *et al.* 2024 ; Sun *et al.* 2024).

The plant body comprises multiple, segregated layers of cells that arise from multiple layers of stem cells in the meristems (Steeves and Sussex 1989), where the outermost layer, L1, makes up the epidermis, L2 the photosynthetic tissue, and L3 the ground tissue. The frequency with which fixed somatic mutations are observed in samples such as leaves, which include cells derived from all meristem layers, will depend on the relative contribution of each layer to the total genomic DNA in that sample. Although we lack layer-specific data to evaluate the DNA content of each layer in oak leaves, one can roughly approximate the DNA content of L2 by analyzing shared mutations detected in leaves and acorns, since embryonic tissues originate from L2 (Amundson *et al.* 2023).

In the published oak accession 3P, 9 somatic SNVs were shared between leaves and the embryonic tissue in acorns (Plomion *et al.* 2018). Thus, these 9 SNVs are most likely derived from L2 of the meristem and can be considered as L2 markers. The median frequency of reads covering these 9 somatic mutations in the leaves was approximately 0.3 (Supplementary Table 8). Since a fixed mutation will appear in only 1 haplotype, the DNA of the mutant cells would have constituted 60% of the total DNA of the sample (Fig. 2a). This provides an upper bound for the contribution of the other 2 layers, with L1 contributing less than L3. Given that L1 and L3 together make up about 40% of all reads, this would suggest that L1 contributes <20% and L3 >20%. The expected observed frequency of L1-fixed SNVs in the haploid assembly, which serves as the reference, would thus be below 0.1 (below half of 20%). Therefore, the use of a VAF threshold substantially above 0.1 may prevent the identification of layer-specific somatic mutations even when they are fixed in the investigated sample.

Limitations of using either a single haploid or diploid assembly as reference

In highly heterozygous genomes, a single haploid assembly is an imperfect representation of the entire genome, and reads from regions that are highly diverged in the other haplophase cannot be

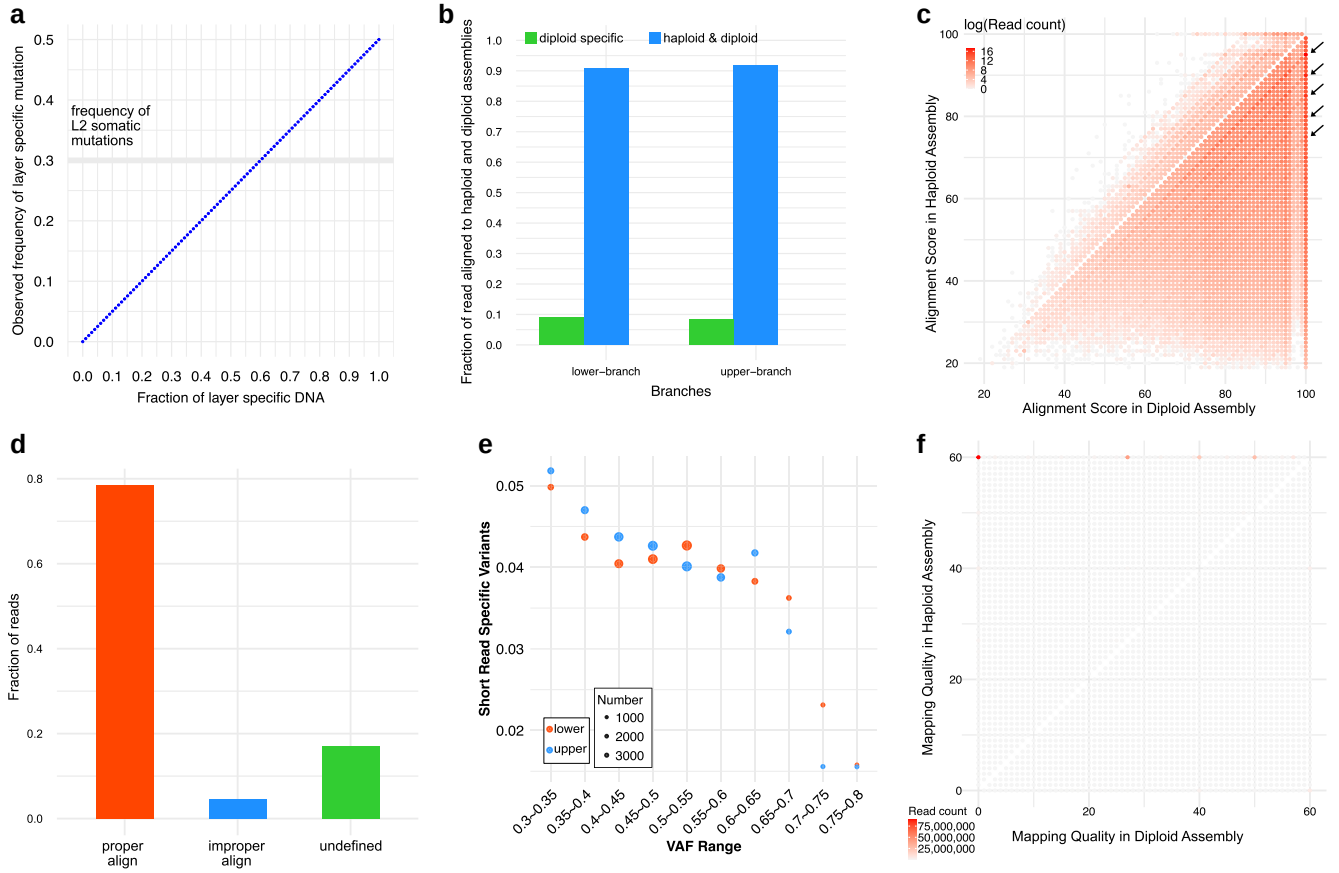


Fig. 2. Using haploid or diploid assemblies for short-read alignment. a) Frequencies of somatic mutations in sequencing reads relating to the contribution of a layer to the total DNA of the sample. b) Fractions of reads mapping only to the diploid assembly. c) Distribution of mapping (alignment) scores in 2 assemblies as reference. The dots indicated by arrows represent germline single nucleotide variants. d) Reads whose mapping (alignment) position matched the expected position (as inferred from alignments of 4.1 kb diploid assembly fragments centered on each read to the haploid assembly) were classified as having a proper alignment. Reads that aligned outside the expected position were classified as having an improper alignment. Undefined are reads where the 4.1 kb diploid assembly region centered on the read aligned to a region in the haploid assembly that was smaller than 4 kb. e) Fraction of variants identified with short reads that were not identified from HiFi reads. The sizes of the dots indicate the number of short-read-specific variants. f) Distribution of mapping qualities when using either the haploid or diploid assembly as reference.

aligned. We compared the short-read IDs that mapped to the haploid assembly vs those mapped to the diploid assembly and found that approximately 9% of the reads could only be aligned to the diploid assembly (Fig. 2b; Supplementary Table 11). Therefore, using only the haploid assembly as a reference genome will increase the false-negative rate in mutation identification.

Even though more reads can be mapped to the diploid than to the haploid assembly, the number of paired reads with discordant orientation or mapping to different chromosomes is reduced (Supplementary Table 12). Additionally, when reads are mapped to the diploid assembly, the average size of mapped fragments is smaller, with reduced standard deviation. This suggests that mapping to the diploid assembly can increase read mapping accuracy and reduce false positives.

We extracted reads that mapped to both assemblies and compared their alignment scores in the haploid and diploid assemblies (Fig. 2c and Supplementary Fig. 9). Although a correct alignment generally results in a higher alignment score, when aligning short reads to a haploid genome, heterozygous sites appear as mismatches, with BWA assigning +1 for a match and −4 for a mismatch by default (Li and Durbin 2010). For reads with an alignment score of 100 in the diploid assembly, alignment scores in haploid assembly values are predominantly 95, 90, 85, 80, or 75. Overall, aligning reads to the diploid genome increases the alignment score, indicating a reduction in alignment errors.

Because short reads are more prone to mismapping than long reads, we used long flanking regions from the diploid assembly to identify mismapped short reads in the haploid assembly. This analysis revealed that approximately 5% of reads aligned to the haploid genome were mismapped relative to their expected positions (Fig. 2d and Supplementary Table 14).

To further evaluate potential mapping errors introduced by using the haploid assembly as a reference, we examined whether germline SNVs identified in short-read data could be confirmed in the HiFi reads. As expected, the fraction of short-read-specific variants decreased as the VAF of short-read variants increased. Notably, when the VAF of short-read variants was between 0.3 and 0.35, the proportion of short-read-specific variants remained around 5% (Fig. 2e and Supplementary Table 13).

While aligning reads to a diploid genome can mitigate mapping errors caused by differences between the 2 haplotypes, using only the diploid genome also has limitations. For instance, in homozygous regions, MAPQ is often 0 (due to multiple mapping), and coverage is reduced by half, making low-frequency mutations more difficult to detect. Among reads with different MAPQ values when mapped against either the haploid or diploid reference

assembly, most had a MAPQ of 60 in the haploid assembly, but a MAPQ of 0 in the diploid assembly (Fig. 2f).

Optimizing short-read alignments in highly heterozygous genomes via dynamic reference genome selection

To overcome these issues, we utilized both haploid and diploid assemblies as reference genomes. The haploid assembly was employed to identify variants in regions identical in both haplotypes, while the diploid assembly was leveraged to resolve heterozygous regions unique to each haplotype. We followed a hierarchical decision-making strategy, in which reads were first aligned to both haploid and diploid assemblies and each read was then assigned to 1 of the 2 reference assemblies based on a simple prioritization rule: if a read had a higher alignment score in 1 assembly, it was assigned to that assembly. If alignment scores were identical, the read was assigned to the assembly with the higher MAPQ value. Finally, if both alignment scores and MAPQ values were equal, the read was assigned to the diploid assembly (Fig. 3a). The details of the variant-calling pipeline and subsequent filtering criteria are described in the Methods.

Somatic mutations in the Napoleon Oak

Initially, with reads assigned to the diploid assembly as reference, 137 mutations were called in the upper branch and 44 mutations in the lower branch. From reads assigned to the haploid assembly as reference, 53 mutations were called in the upper branch and 21 mutations in the lower branch. We then visualized the results in the Integrative Genomics Viewer (IGV) for inspection (Robinson *et al.* 2017), including the separate BAM files for short reads from both branches, the original BAM file of short reads from both branches, and the BAM file of HiFi reads used to generate the assemblies. IGV screenshots of both retained and SNVs removed with our manual filters are shown in Supplementary Fig. 11 (available on Figshare).

After IGV inspection, we retained 123 mutations in the upper branch and 14 in the lower branch with the diploid assembly as reference, and 51 mutations in the upper branch and 11 in the lower branch with the haploid assembly as reference. Combining the 2 datasets resulted in 173 mutations in the upper branch and 25 in the lower branch. Because we had called mutations in the upper branch using the lower branch as control and vice versa, there was by definition no overlap between the 2 branches, resulting in a total of 198 SNVs that were absent from the germline (Supplementary Table 9).

All of the 10 upper-branch variants that had been previously validated by amplicon sequencing (Schmid-Siegert *et al.* 2017) were identified with our approach: 2 were detected with the diploid assembly as reference, and 8 were detected with the haploid assembly as reference (Supplementary Table 10). As a control, we called SNVs using Strelka2 (Kim *et al.* 2018), a tool specifically developed for rare frequency somatic mutation. Of the 10 validated upper-branch SNVs (Schmid-Siegert *et al.* 2017), Strelka2 identified 9 SNVs with the haploid assembly as reference. One SNV could not be detected because its region of origin is absent from the haploid assembly. With the diploid assembly as reference, Strelka2 identified 7 of the 10 validated SNVs. Inspecting the alignments of the 3 missing SNVs in IGV indicates that failure to detect them with Streka2 is likely due to the MAPQ values of the corresponding reads being 0 (Supplementary Fig. 12).

For the 7 lower-branch variants confirmed by amplicon sequencing, we did find rare variant reads for these sites also in the upper branch, both in the original and filtered BAM files. Therefore, we do not consider these 7 variants to be true lower-branch-specific mutations compared to the upper branch.

Consistent with mutation spectra observed in other systems (Hofmeister *et al.* 2017; Exposito-Alonso *et al.* 2018; Weng *et al.* 2019; Belfield *et al.* 2021; Satake *et al.* 2023), we found that G:C > A:T mutations were predominant (Fig. 3b).

Discussion

What does our work say about germline mutation rates? To approximate the per-site germline mutation rates in the Napoleon Oak, we considered only mutations that could be transmitted to the next generation, meaning they should have occurred in L2.

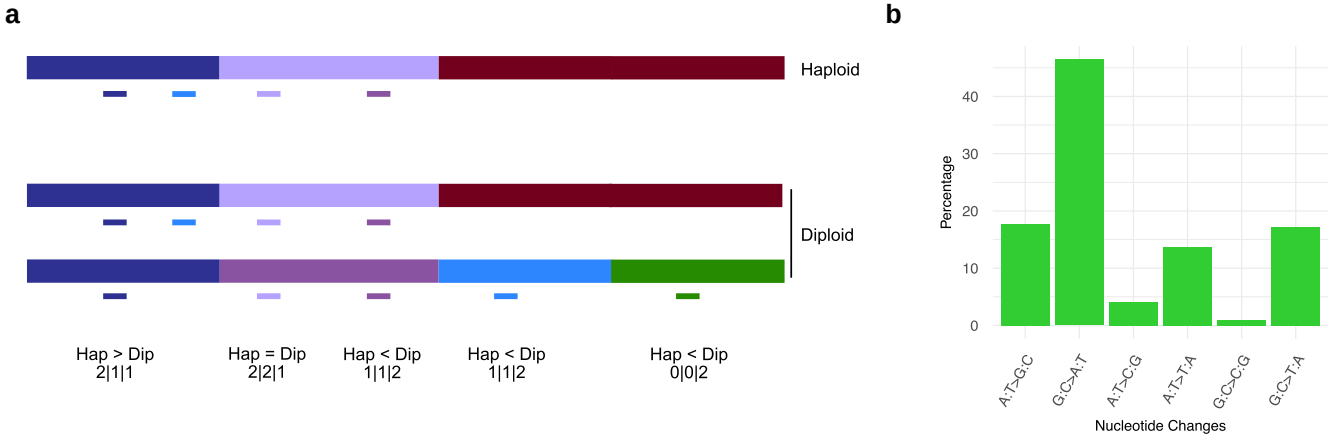


Fig. 3. Somatic mutations in the Napoleon Oak. a) Diagram of how inappropriately aligned short reads in the haploid and diploid assembly can be identified. Scores indicate alignment in haploid and in the diploid genome composed of 2 haplophases. 2, high alignment and MAPQ scores; 1, high alignment but low MAPQ scores, or low alignment but high MAPQ scores; 0, no alignment. Dark blue read: homozygous region, reads in both assemblies have the same alignment scores, but the MAPQ value is higher in haploid than in diploid assembly. Light purple read: heterozygous region, read has higher alignment score and MAPQ value in haploid vs diploid assembly, and alignment score differs between haplophases. Dark purple read: heterozygous region, read has higher alignment score in diploid vs haploid assembly, and alignment score differs between haplophases. Light blue read: heterozygous region, read has higher alignment score in diploid vs haploid assembly, and alignment score differs between haplophases. Green read: highly divergent region, read only aligned to 1 haplophase of the diploid assembly. b) Mutation spectrum for all 198 somatic SNVs.

As laid out in the Results, we reasoned that L2 accounts for around 60% of the total DNA in leaves. Each haploid set of chromosomes would thus contribute around 30% of the leaf DNA. To identify likely L2 mutations, we therefore used a VAF threshold of at least 0.2 for the haploid assembly as reference and 0.4 for the diploid assembly as reference. We found 77 such potential L2 mutations. For the Napoleon Oak, this would yield a raw mutation rate of approximately 6×10^{-8} bp^{-1}. Scaled to the age of the Napoleon Oak, 234 years, the annual mutation rate would be approximately 2.5×10^{-10} bp^{-1}, within the range of rate estimates for other tree species (see the summary in Johannes 2024). We do not address false-negative rates, but even if we accepted all of our 198 originally called mutations to have been fixed in L2 of the sampled leaves, the mutation rate estimate would still be within the range published for other trees.

For acorns produced by the Napoleon Oak today, which was 234 years old at the time of sampling (Schmid-Siegert *et al.* 2017), this would also be close to the generational mutation rate, since we are mostly ignoring mutations that occur only after the entire branch was formed. Assuming that most oak trees have, however, only a generation time closer to 50 years, the generational mutation rate would have about 1.2×10^{-8} bp^{-1} as a lower bound. This is not very different from the per-generation mutation rate of *A. thaliana* in nature, estimated to be about 3×10^{-9} bp^{-1} (Exposito-Alonso *et al.* 2018).

To assess the per-generation mutation rate, it is also useful to consider the number of cell divisions separating each generation. In the annual species *A. thaliana*, it has been estimated that there are around 30 such cell divisions (Hoffman *et al.* 2004; Watson *et al.* 2016). In trees, there are perhaps twice to thrice as many cell divisions separating each generation (Burian *et al.* 2016). Thus, the per-site per-cell division mutation rate in oak and *A. thaliana* would be very similar. Note that we focus on mutations that are likely to be transmitted to the next generation, and that our study therefore does not speak to whether or not an allometric scaling law needs to be invoked for somatic mutation rates in plants (Johannes 2024).

Recent studies have suggested that mutation rate is affected by a series of biological and environmental factors (Belfield *et al.* 2021). Notably, effective population size has been reported in 2 independent studies to correlate negatively with mutation rate (Sung *et al.* 2012; Lynch *et al.* 2016). In contrast, findings on the effects of generation time vary between taxa: while 1 meta-analysis across eukaryotes reports a positive correlation between generation time and mutation rate, a plant-specific study found a negative correlation (Wang and Obbard 2023; Johannes 2024). This discrepancy may stem from the absence of a segregated germline in plants. *Q. robur* has a long generation time, which may increase the effective germline mutation rate. On the other hand, the large effective population size of oaks may have an opposite effect (Kremer and Hipp 2020). How these factors together influence mutation rate in oaks and other long-lived plants warrants further investigation, using different species and individuals of different ages.

Data availability

PacBio HiFi sequencing data have been submitted to the ENA under the project accession PRJEB85730. Assemblies in fasta and gfa format and Supplementary Fig. 11 are available through https://doi.org/10.6084/m9.figshare.28397981. Scripts are available at https://github.com/Wenfei-Xian/Somatic_mutation_in_high_heterogyzous_genome.

Supplemental material available at G3 online.

Acknowledgments

We thank Christian Hardtke for encouragement, intellectual support, and discussion. We are especially grateful for the thoughtful feedback from Korbinian Schneeberger. We thank Christa Lanz for the help with PacBio sequencing. We are grateful to Adrián Contreras, Yueqi Tao, Zhigui Bao, Andrea Movilli, and Sebastian Vorbrugg for discussion.

Funding

This study was supported by the Max Planck Society and the Novozymes Prize of the Novo Nordisk Foundation (DW).

Conflicts of interest

DW holds equity in Computomics, which advises plant breeders. DW also consults for KWS SE, a globally active plant breeder and seed producer. All other authors declare no competing interests.

Author contributions

DW designed and supervised the project. PC-B, FAR, and PR prepared the leaf sample. PC-B and WX prepared the DNA library. WX, PC-B, and FAR analyzed the data. WX, IB, and DW prepared the final manuscript with inputs from all authors.

Literature cited

Ally D, Ritland K, Otto SP. 2010. Aging in a long-lived clonal tree. PLoS Biol. 8(8):e1000454. https://doi.org/10.1371/journal.pbio.1000454.

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(1):258. https://doi.org/10.1186/s13059-022-02823-7.

Amundson KR, Marimuthu MPA, Nguyen O, Sarika K, DeMarco IJ, Phan A, Henry IM, Comai L. 2023. Differential mutation accumulation in plant meristematic layers. bioRxiv. https://doi.org/10.1101/2023.09.25.559363.

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(2): 232–238. https://doi.org/10.1038/s41587-022-01435-7.

Belfield EJ, Brown C, Ding ZJ, Chapman L, Luo M, Hinde E, van Es SW, Johnson S, Ning Y, Zheng SJ, *et al.* 2021. Thermal stress accelerates *Arabidopsis thaliana* mutation rate. Genome Res. 31(1): 40–50. https://doi.org/10.1101/gr.259853.119.

Bočkor VV, Barišić D, Horvat T, Maglica Ž, Vojta A, Zoldoš V. 2014. Inhibition of DNA methylation alters chromatin organization, nuclear positioning and activity of 45S rDNA loci in cycling cells of *Q. robur*. PLoS One. 9(8):e103954. https://doi.org/10.1371/journal.pone.0103954.

Burian A, Barbier de Reuille P, Kuhlemeier C. 2016. Patterns of stem cell divisions contribute to plant longevity. Curr Biol. 26(11): 1385–1394. https://doi.org/10.1016/j.cub.2016.03.067.

Bzikadze AV, Mikheenko A, Pevzner PA. 2022. Fast and accurate mapping of long reads to complete genome assemblies with VerityMap. Genome Res. 32(11–12):2107–2118. https://doi.org/10.1101/gr.276871.122.

Bzikadze AV, Pevzner PA. 2023. UniAligner: a parameter-free framework for fast sequence alignment. Nat Methods. 20(9):1346–1354. https://doi.org/10.1038/s41592-023-01970-4.

Calderón L, Carbonell-Bejerano P, Muñoz C, Bree L, Sola C, Bergamin D, Tulle W, Gomez-Talquenca S, Lanz C, Royo C, *et al.* 2024. Diploid genome assembly of the Malbec grapevine cultivar enables haplotype-aware analysis of transcriptomic differences underlying clonal phenotypic variation. Hortic Res. 11(5): uhae080. https://doi.org/10.1093/hr/uhae080.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10(1):421. https://doi.org/10.1186/1471-2105-10-421.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

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(2):170–175. https://doi.org/10.1038/s41592-020-01056-5.

Dermen H. 1960. Nature of plant sports. Am Hortic Mag. 39:123–173.

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(2):e1007155. https://doi.org/10.1371/journal.pgen.1007155.

Fultz D, McKinlay A, Enganti R, Pikaard CS. 2023. Sequence and epigenetic landscapes of active and silent nucleolus organizer regions in Arabidopsis. Sci Adv. 9(44):eadj4509. https://doi.org/10.1126/sciadv.adj4509.

Goel M, Campoy JA, Krause K, Baus LC, Sahu A, Sun H, Walkemeier B, Marek M, Beaudry R, Ruiz D. 2024. The vast majority of somatic mutations in plants are layer-specific. Genome Biol. 25(1):1–18. https://doi.org/10.1186/s13059-024-03337-0.

Hoffman PD, Leonard JM, Lindberg GE, Bollmann SR, Hays JB. 2004. Rapid accumulation of mutations during seed-to-seed propagation of mismatch-repair-defective Arabidopsis. Genes Dev. 18(21):2676–2685. https://doi.org/10.1101/gad.1217204.

Hofmeister BT, Lee K, Rohr NA, Hall DW, Schmitz RJ. 2017. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. Genome Biol. 18(1):155. https://doi.org/10.1186/s13059-017-1288-x.

Huang N, Li H. 2023. compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics. 39(10):btad595. https://doi.org/10.1093/bioinformatics/btad595.

Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. 2022. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods. 19(6):705–710. https://doi.org/10.1038/s41592-022-01457-8.

Johannes F. 2024. Allometric scaling of somatic mutation and epimutation rates in trees. Evolution. 79(1):1–5. https://doi.org/10.1093/evolut/qpae150.

Khanna A, Larson DE, Srivatsan SN, Mosior M, Abbott TE, Kiwala S, Ley TJ, Duncavage EJ, Walter MJ, Walker JR, *et al.* 2022 Bam-readcount—rapid generation of basepair-resolution sequence metrics. J Open Source Softw., 7(69), 3722 https://doi.org/10.21105/joss.03722.

Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, *et al.* 2018. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 15(8): 591–594. https://doi.org/10.1038/s41592-018-0051-x.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. Science. 304(5673):982. https://doi.org/10.1126/science.1095011.

Kremer A, Hipp AL. 2020. Oaks: an evolutionary success story. New Phytol. 226(4):987–1011. https://doi.org/10.1111/nph.16274.

Leroy T, Plomion C, Kremer A. 2020. Oak symbolism in the light of genomics. New Phytol. 226(4):1012–1017. https://doi.org/10.1111/nph.15987.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics (Oxford, England). 26(5):589–595. https://doi.org/10.1093/bioinformatics/btp698.

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, *et al.* 2021. The structure, function and evolution of a complete human chromosome 8. Nature. 593(7857):101–107. https://doi.org/10.1038/s41586-021-03420-7.

Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Grüning B, Ramírez F, Manke T. 2021. pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics. 37(3): 422–423. https://doi.org/10.1093/bioinformatics/btaa692.

Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 17(11):704–714. https://doi.org/10.1038/nrg.2016.104.

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 *Arabidopsis* centromeres. Science. 374(6569):eabi7489. https://doi.org/10.1126/science.abi7489.

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(1):4054. https://doi.org/10.1038/s41467-023-39784-9.

Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Lugo CS, Elliott TA, Ware D, Peterson T, *et al.* 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20(1):275. https://doi.org/10.1186/s13059-019-1905-y.

Pellicer J, Leitch IJ. 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytol. 226(2):301–305. https://doi.org/10.1111/nph.16261.

Picard toolkit. 2019. Broad Institute. GitHub repository. https://broadinstitute.github.io/picard/.

Plomion C, Aury J-M, Amselem J, Leroy T, Murat F, Duplessis S, Faye S, Francillonne N, Labadie K, Le Provost G, *et al.* 2018. Oak genome reveals facets of long lifespan. Nat Plants. 4(7):440–452. https://doi.org/10.1038/s41477-018-0172-3.

Poplin R, Chang PC, 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(10):983–987. https://doi.org/10.1038/nbt.4235.

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(21): 12309–12327. https://doi.org/10.1093/nar/gkac1115.

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(10):1474–1482. https://doi.org/10.1038/s41587-023-01662-6.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21(1):1–27. https://doi.org/10.1186/s13059-020-02134-9.

Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. 2017. Variant review with the Integrative Genomics Viewer. Cancer Res. 77(21):e31–e34. https://doi.org/10.1158/0008-5472.CAN-17-0337.

Satake A, Imai R, Fujino T, Tomimoto S, Ohta K, Na'iem M, Indrioko S, Widiyatno W, Purnomo S, Morales AM, *et al.* 2023. Somatic mutation rates scale with time not growth rate in long-lived tropical trees. eLife. 12:RP88456. https://doi.org/10.7554/eLife.88456.3.

Schmid-Siegert E, Sarkar N, Iseli C, Calderon S, Gouhier-Darimont C, Chrast J, Cattaneo P, Schütz F, Farinelli L, Pagni M. 2017. Low number of fixed somatic mutations in a long-lived oak tree. Nat Plants. 3(12):926–929. https://doi.org/10.1038/s41477-017-0066-9.

Schmitt S, Heuret P, Troispoux V, Beraud M, Cazal J, Chancerel É, Cravero C, Guichoux E, Lepais O, Loureiro J, *et al.* 2024. Low-frequency somatic mutations are heritable in tropical trees *Dicorynia guianensis* and *Sextonia rubra*. Proc Natl Acad Sci U S A. 121(10):e2313312121. https://doi.org/10.1073/pnas.2313312121.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. Bioinformatics. 37(12):1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016.

Steeves TA, Sussex IM. 1989. Patterns in Plant Development. Cambridge, New York, Melbourne: Cambridge University Press.

Sun H, Abeli P, Campoy JA, Rütjes T, Krause K, Jiao WB, Beaudry R, Schneeberger K. 2024. The identification and analysis of meristematic mutations within the apple tree that developed the RubyMac sport mutation. BMC Plant Biol. 24(1):912. https://doi.org/10.1186/s12870-024-05628-x.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. 109(45):18488–18492. https://doi.org/10.1073/pnas.1216223109.

Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE. 2022. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. Bioinformatics. 38(7):2049–2051. https://doi.org/10.1093/bioinformatics/btac018.

Wang N, Chen P, Xu Y, Guo L, Li X, Yi H, Larkin RM, Zhou Y, Deng X, Xu Q. 2024. Phased genomics reveals hidden somatic mutations and provides insight into fruit development in sweet orange. Hortic Res. 11(2):uhad268. https://doi.org/10.1093/hr/uhad268.

Wang Y, Obbard DJ. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. Evol Lett. 7(4):216–226. https://doi.org/10.1093/evlett/qrad027.

Watson JM, Platzer A, Kazda A, Akimcheva S, Valuchova S, Nizhynska V, Nordborg M, Riha K. 2016. Germline replications and somatic mutation accumulation are independent of vegetative life span in Arabidopsis. Proc Natl Acad Sci U S A. 113(43):12226–12231. https://doi.org/10.1073/pnas.1609686113.

Weng M-L, Becker C, Hildebrandt J, Rutter MT, Shaw RG, Weigel D, Fenster CB. 2019. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. Genetics. 211(2):703–714. https://doi.org/10.1534/genetics.118.301721.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 31(20):3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Wlodzimierz P, Hong M, Henderson IR. 2023. TRASH: Tandem Repeat Annotation and Structural Hierarchy. Bioinformatics. 39(5):btad308. https://doi.org/10.1093/bioinformatics/btad308.

Workman R, Timp W, Fedak R, Kilburn D, Hao S, Liu K. 2018. High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. Protoc Exch. 2018:059. doi:10.17504/protocols.io.4vbgw2n.

Xian W, Bezrukov I, Bao Z, Vorbrugg S, Gautam A, Weigel D. 2025. TIPPo: a user-friendly tool for de novo assembly of organellar genomes with high-fidelity data. Mol Biol Evol. 42(1):msae247. https://doi.org/10.1093/molbev/msae247.

Editor: D. Schrider

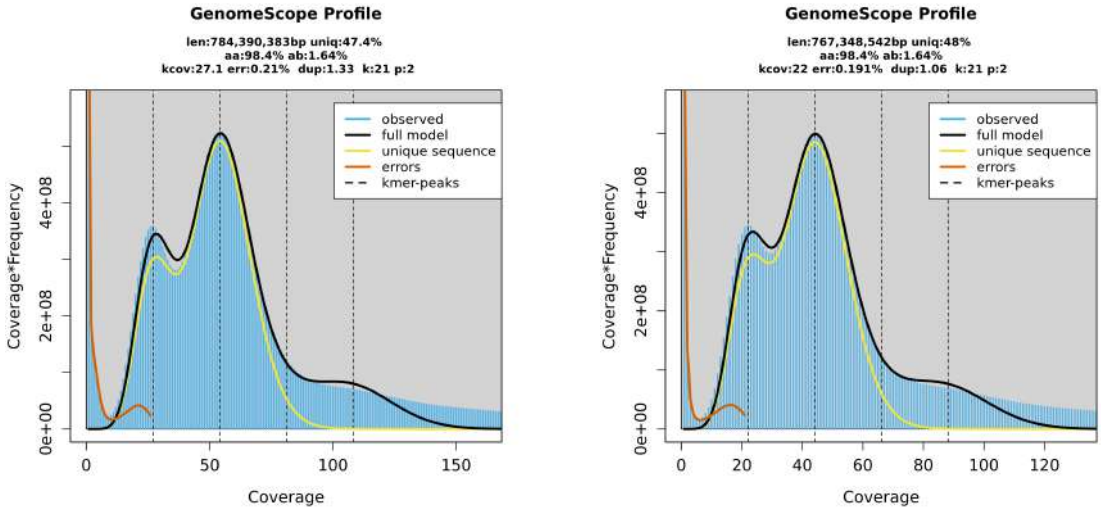


Figure S1. Genome survey using GenomeScope2. A. using illumina reads from lower branch. B. using illumina reads from upper branch.

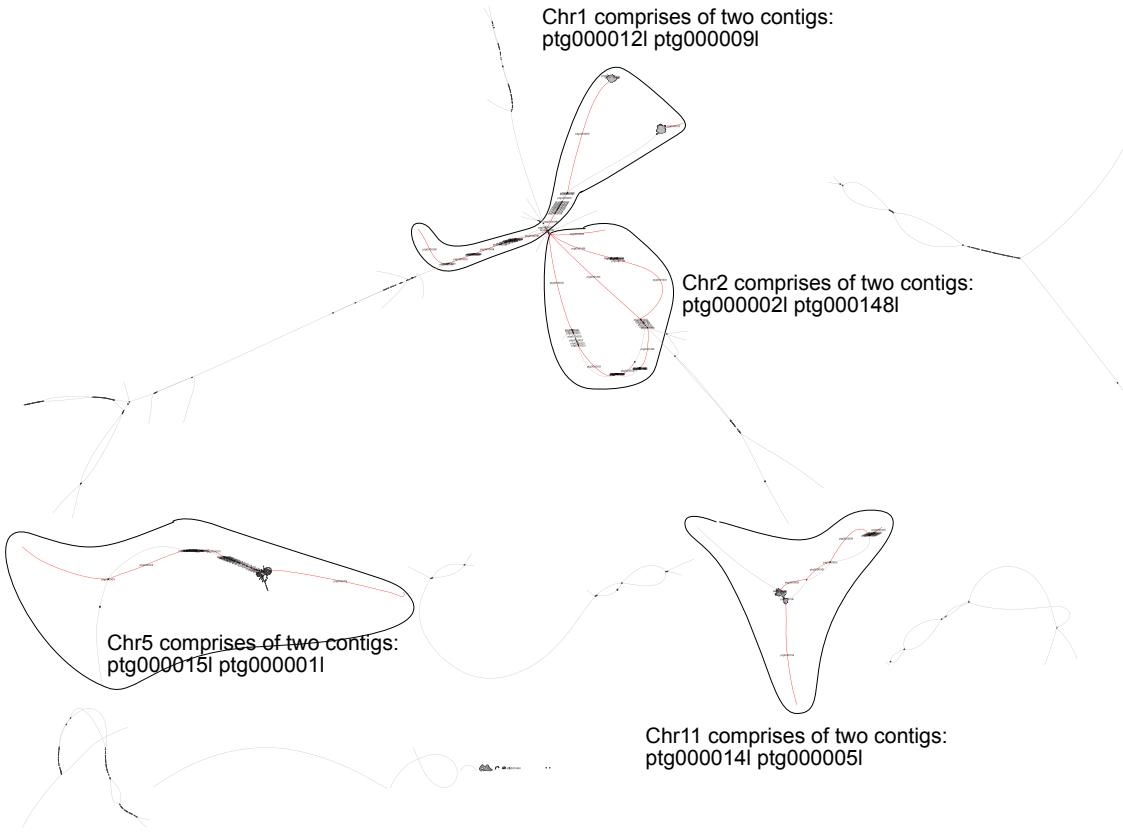


Figure S2. Eight long contigs (in red) lacking two telomeric repeats in the diploid assembly graph for manually scaffolding.

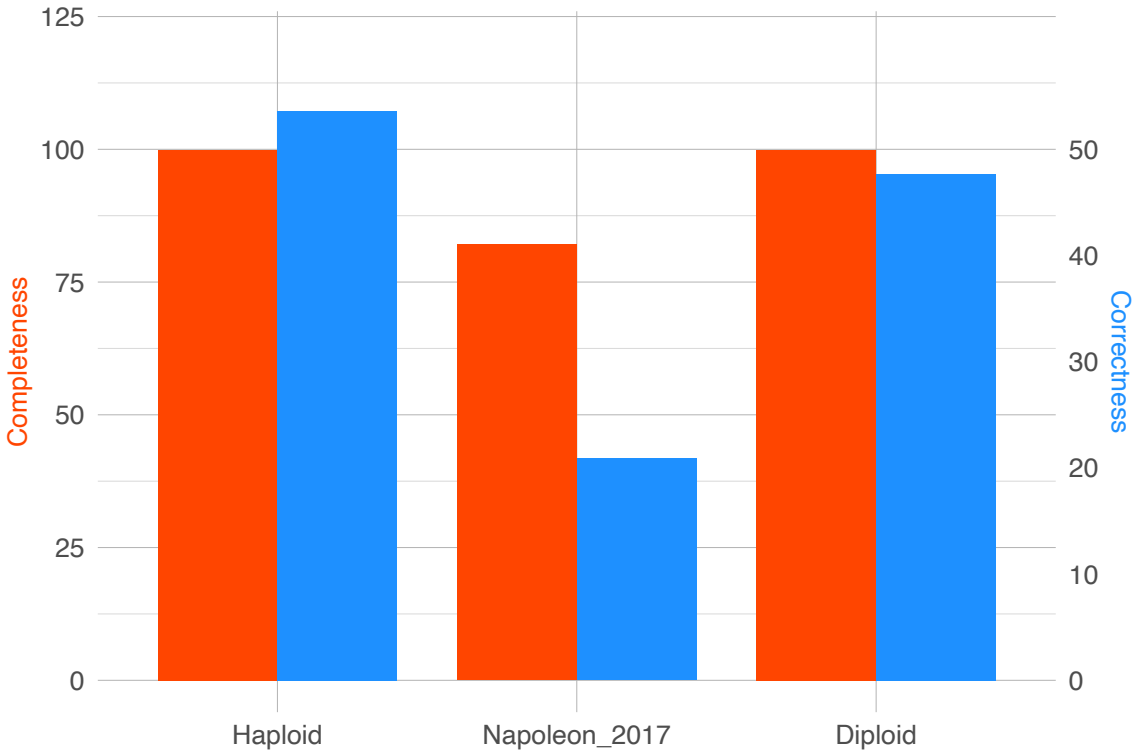


Figure S3. Quality evaluation of two assemblies.

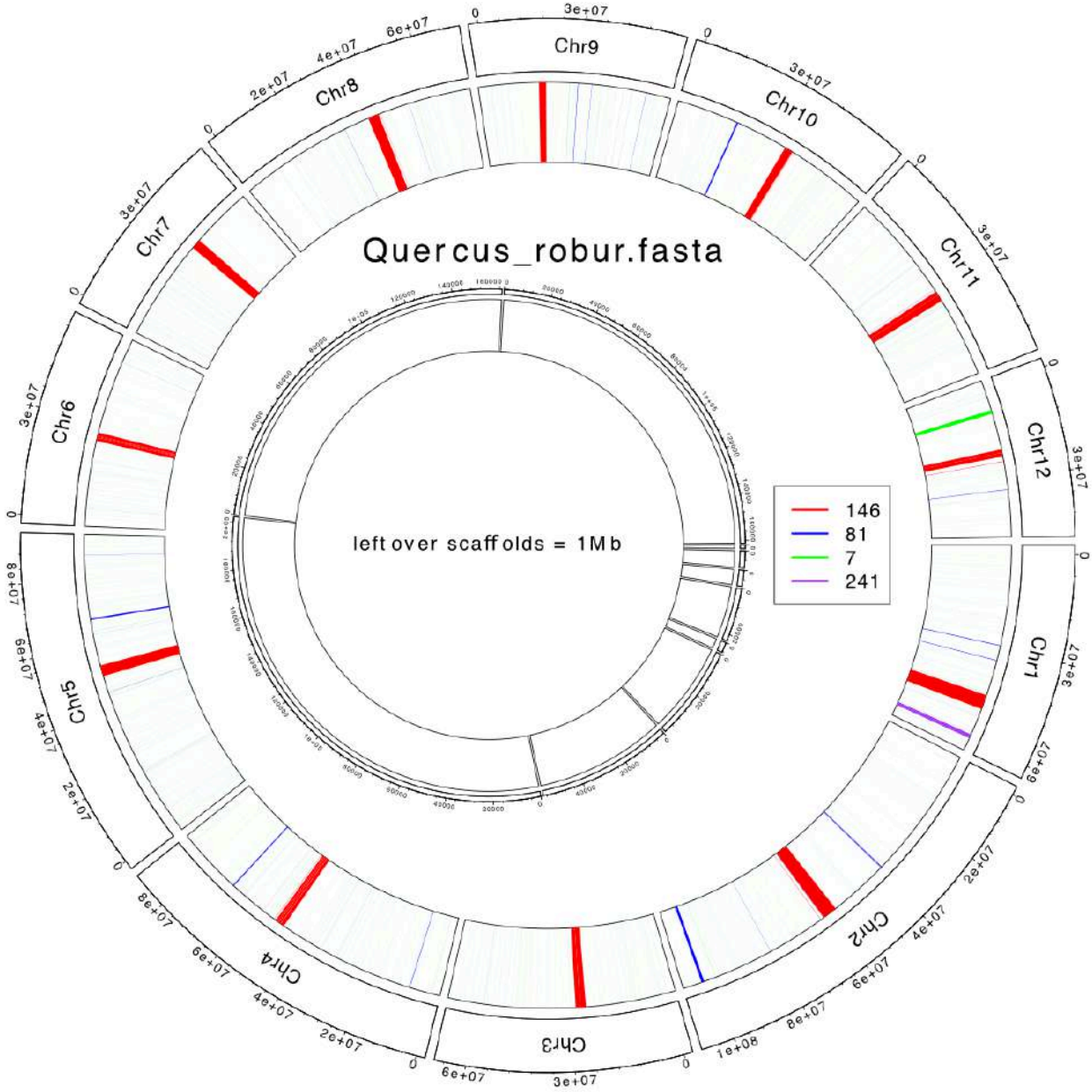


Figure S4. Satellite repeats in haploid assembly.

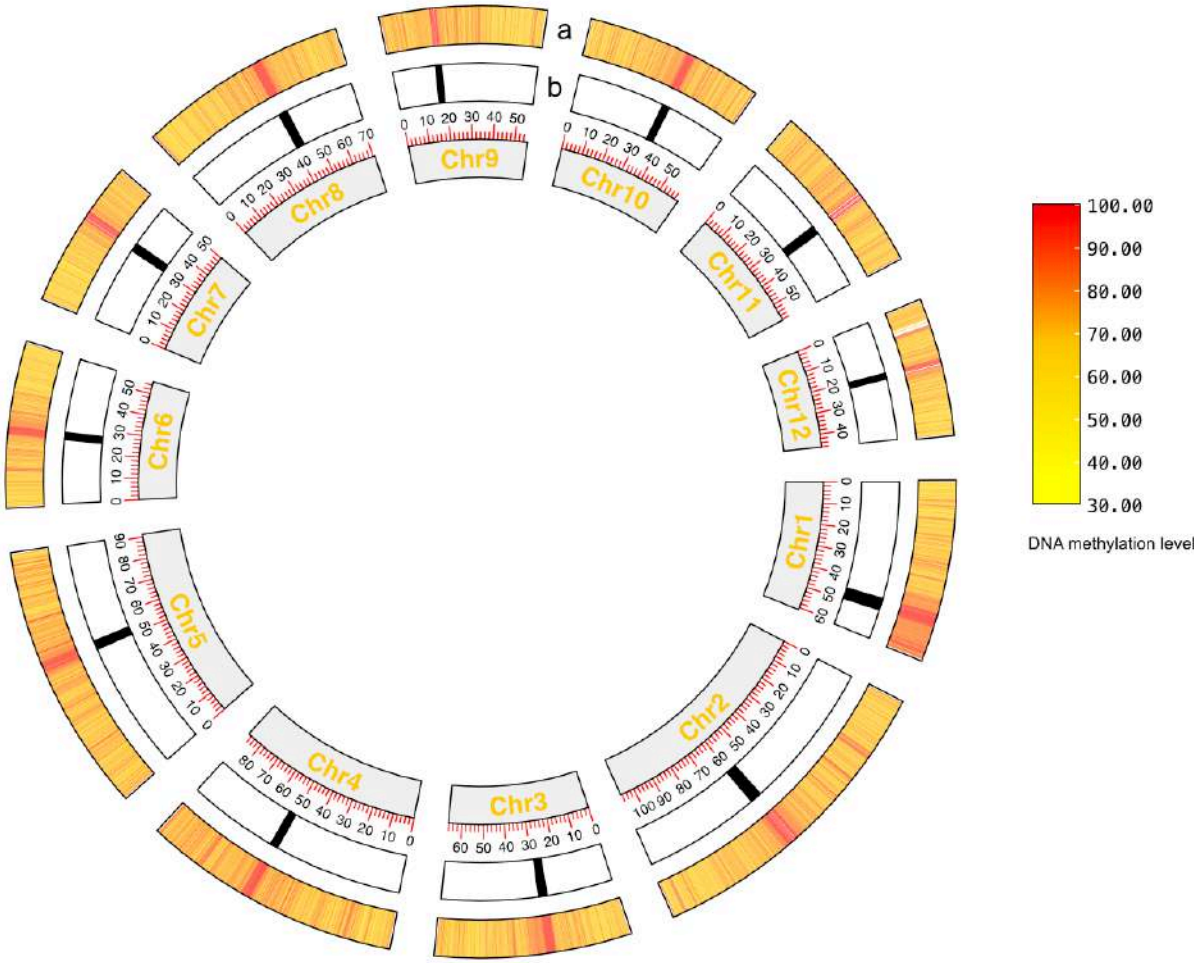


Figure S5. Circos plot of CEN146 and DNA methylation. a track indicates the DNA methylation level and b track indicates the location of CEN146.

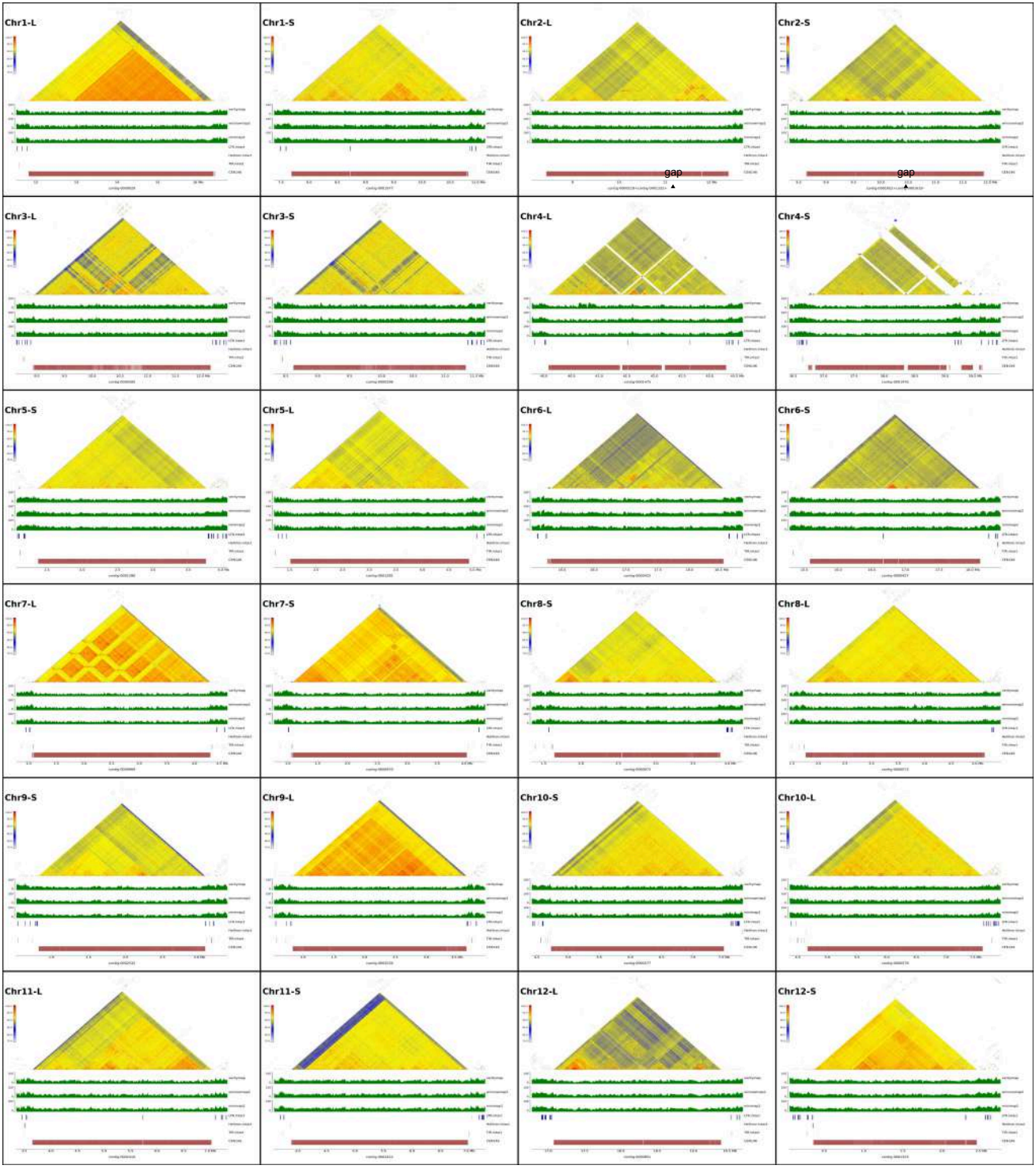


Figure S6. HiFi coverage, the present of intact TE and similarity of CEN146 arrays.

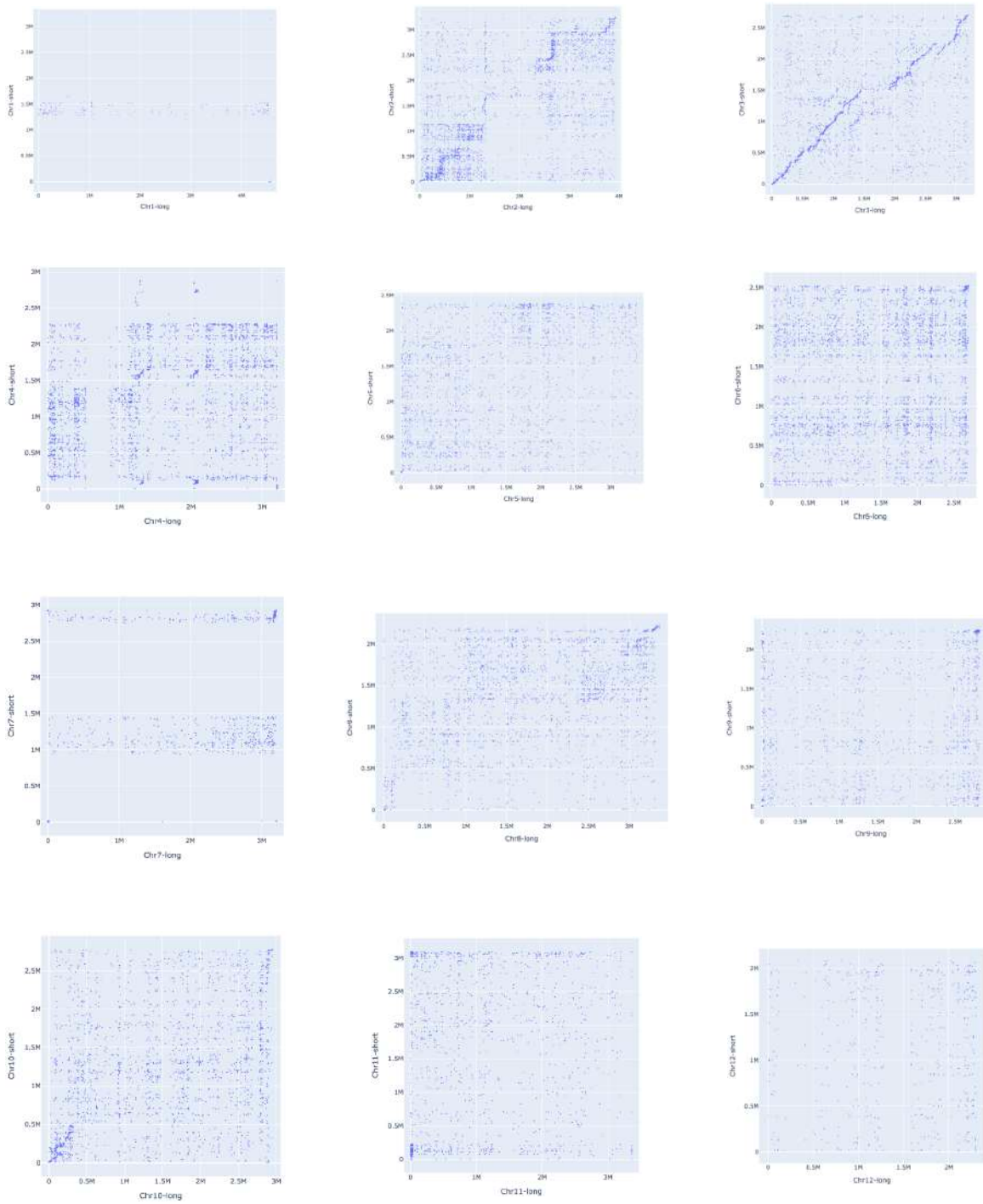


Figure S7. Rare dot-plot of CEN146 arrays between homologous chromosome generated by Unialigner.

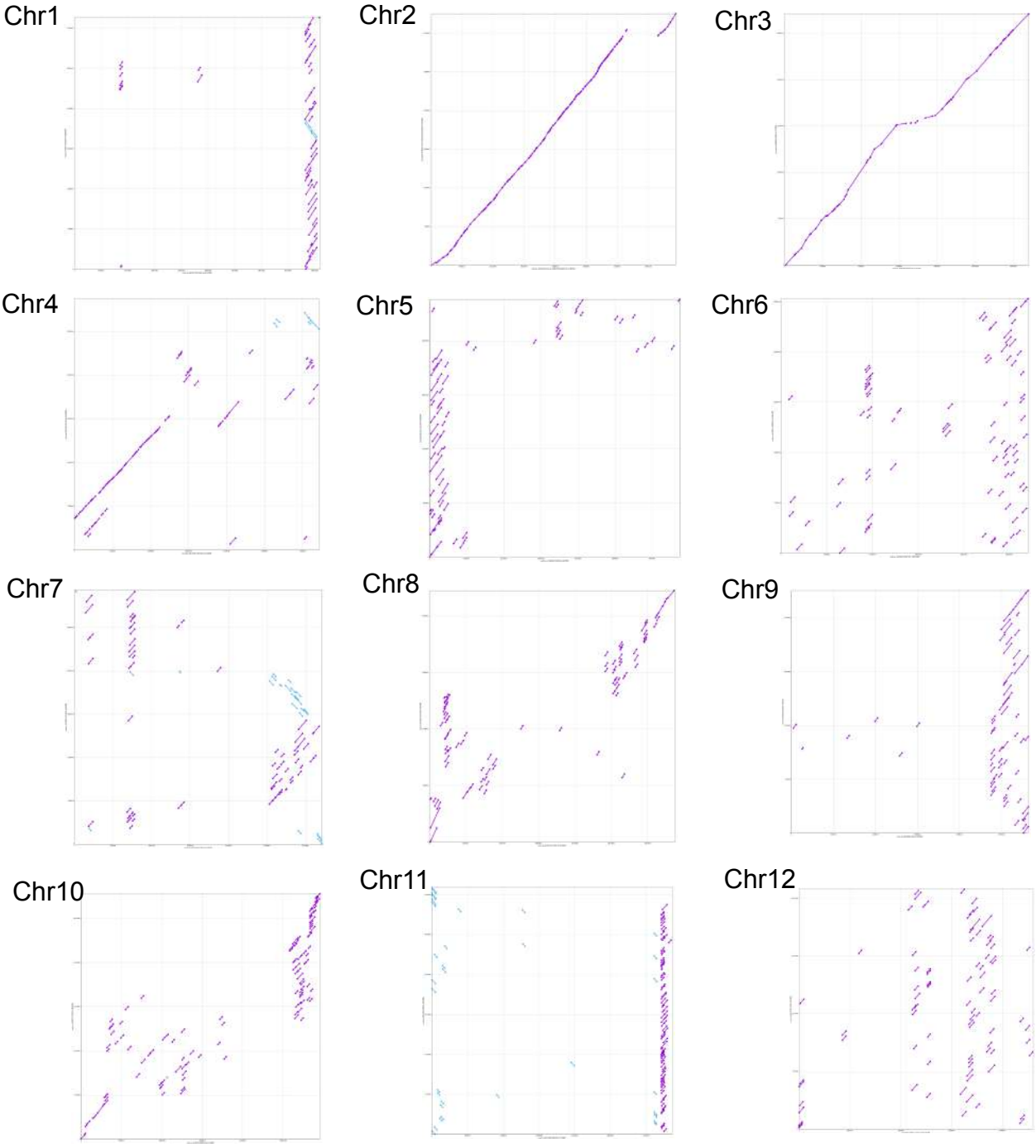


Figure S8. Dot-plot of CEN146 arrays between homologous chromosome by Minimap2.

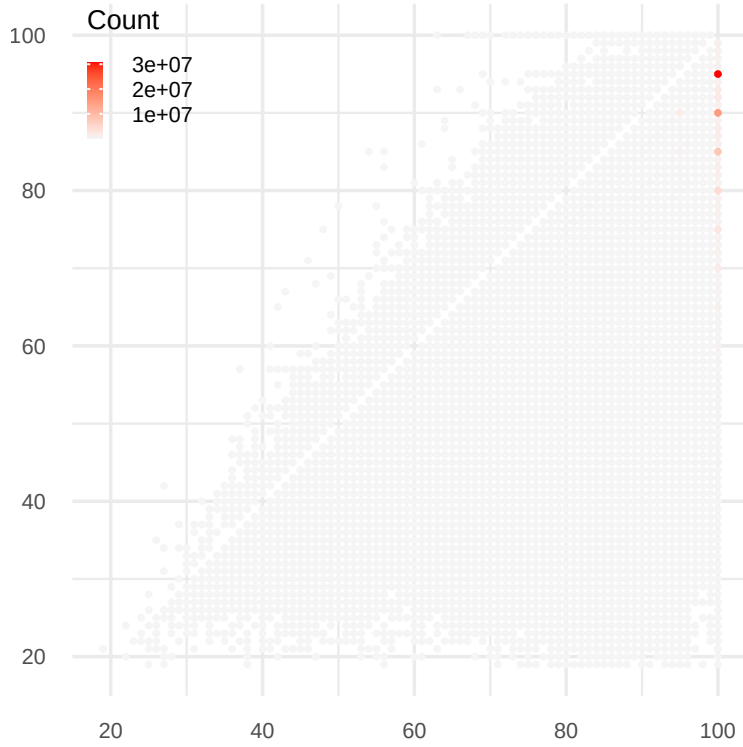


Figure S9. Distribution of alignment score in two assemblies as reference. The color intensity of the dots represent the actual number of reads.

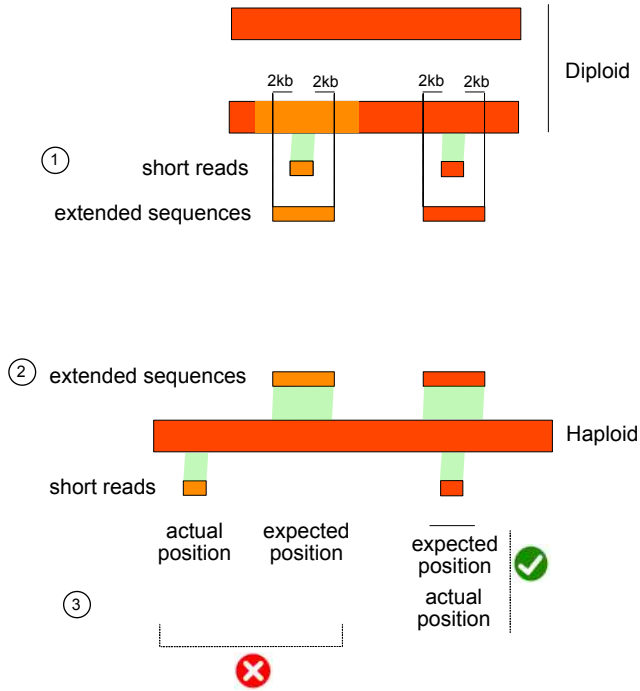
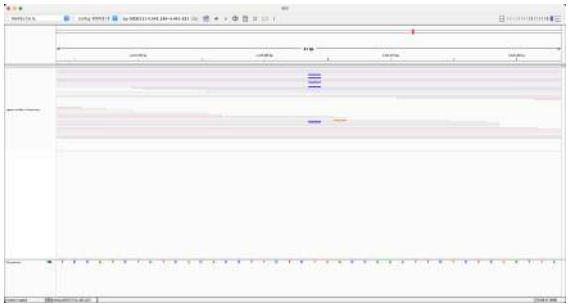
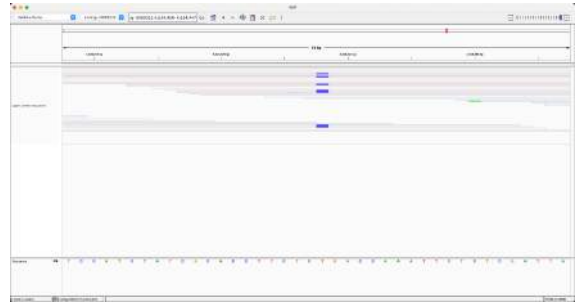


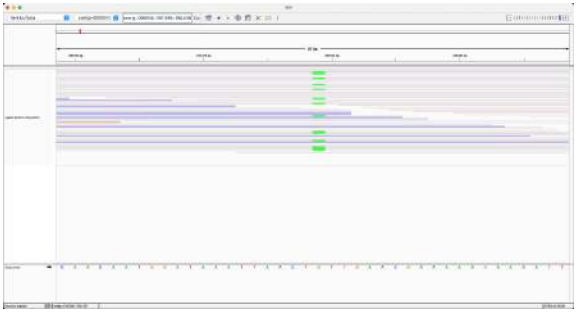
Figure S10. Identification of misaligned short reads in the haploid assembly. Step 1: Align short reads to the diploid assembly and extract the flanking 2 kb sequences. Step 2: Align the short reads and the corresponding extended sequences to the haploid assembly. Step 3: Compare the positions of the short reads and the extended sequences.

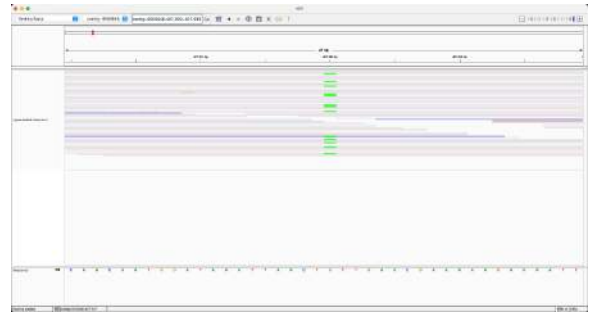
SNV1



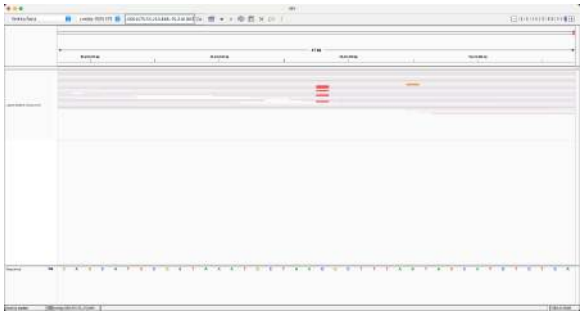


SNV2





SNV3



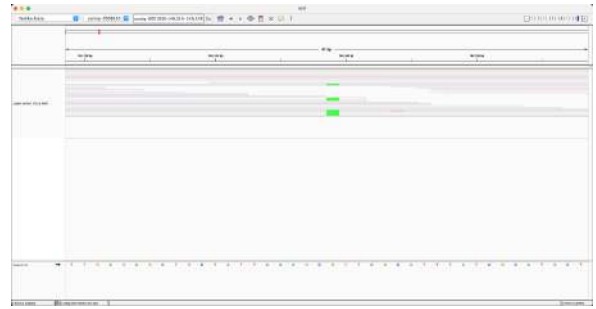


Figure S12. Read alignments for the three undetected SNVs using the diploid assembly as the reference.

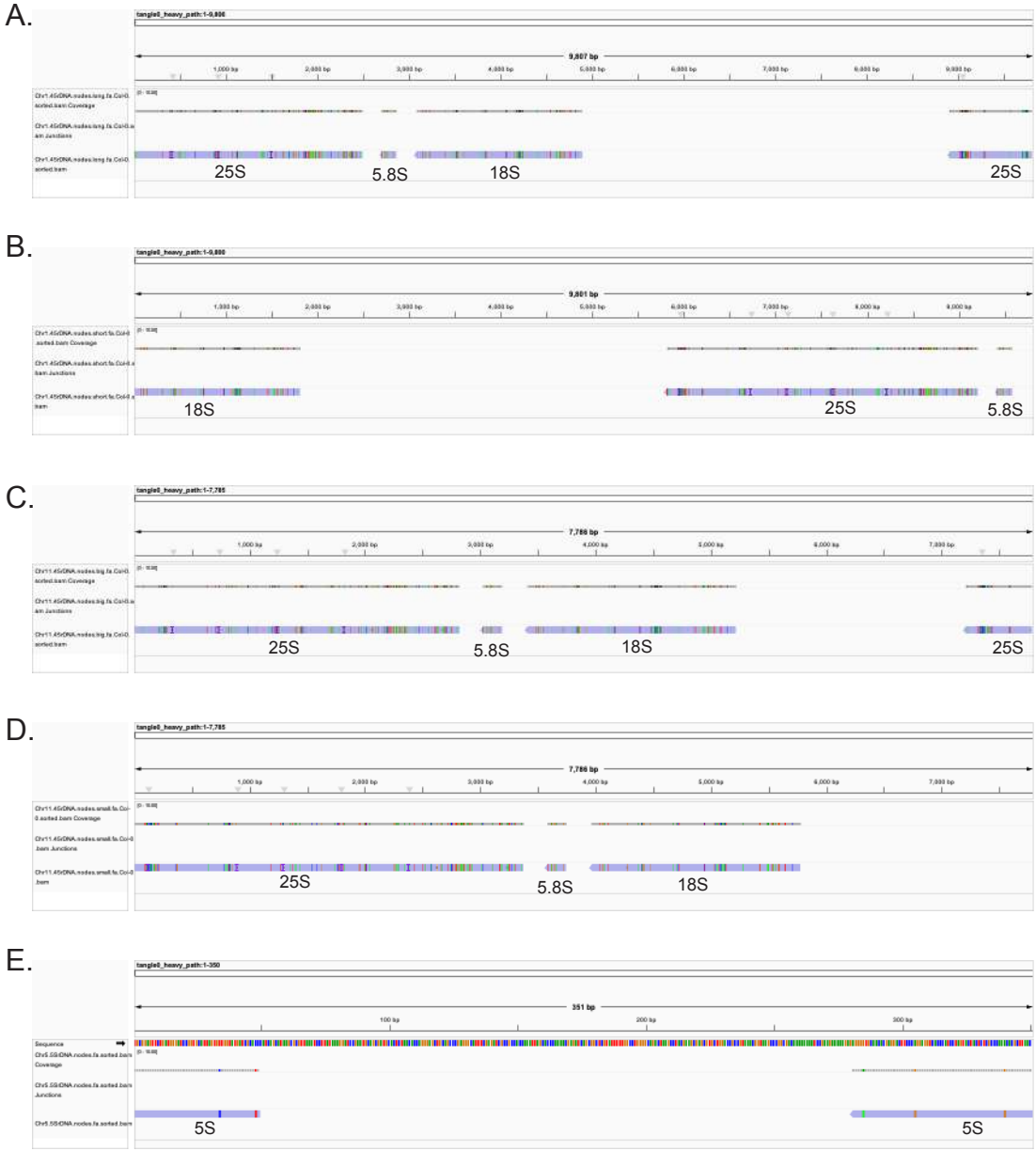


Figure S13. Alignment of *Arabidopsis thaliana* rDNA unit to the oak consensus sequence of each rDNA cluster with BWA. A. 45S rDNA on Chr1 (haplotype 1). B. 45S rDNA on Chr1 (haplotype 2). C. 45S rDNA on Chr11 (haplotype 1). D. 45S rDNA on Chr11 (haplotype 2). E. 5S rDNA on Chr5.

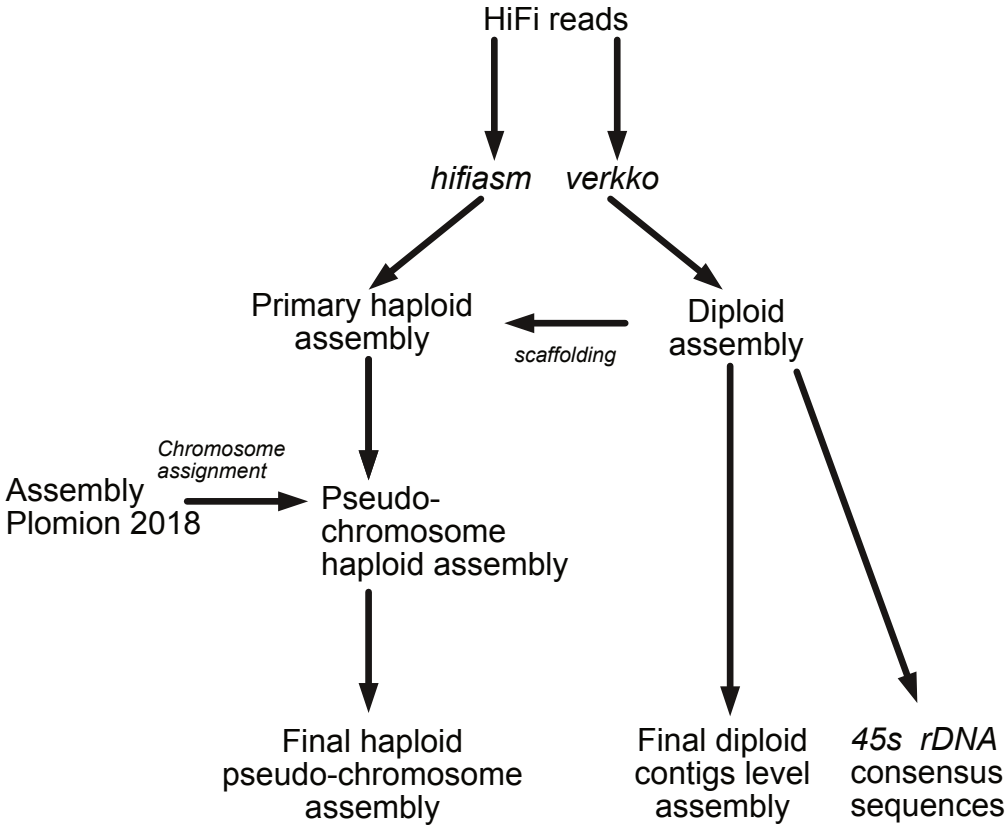


Figure S14. A flow diagram of the nuclear genome assembly steps.