University of South Bohemia
Faculty of Science
České Budějovice, Czech Republic

Analysis of repeat content in the genomes of green algae and early land plants

Bachelor thesis

Teodora Nazare

Supervisor: Abdoallah Sharaf Ph.D.

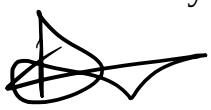
České Budějovice,

2022

I hereby declare that I have worked on my bachelor's thesis independently and used only the sources listed in the bibliography. I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full / in shortened form resulting from deletion of indicated parts to be kept in the Faculty of Science archive, in electronic form in a publicly accessible part of the IS STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

In České Budějovice 13.08.2022



Signature

Nazare, T., 2022: Analysis of repeat content in the genomes of green algae and early land plants. Bc. Thesis, in English. – 46., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Annotation

Genomes of 12 plant species ranging from uni-cellular green algae to land plants are analyzed using the Repeat Explorer 2 computational tool. The repetitive elements distribution and annotation is obtained and compared withing the 12 plant species.

Acknowledgements

I would like to thank my supervisors, Abdoallah Sharaf Ph.D., and Mgr. Iva Mozgová, Ph.D. for giving me the opportunity to study this interesting topic, and for their patience, implication, and professionalism. Their outstanding supervision was crucial towards the completion of this thesis, and the volume of information I have acquired from them will benefit me throughout my future career.

I would also like to thank my best friend, Gabriela Paul for constantly providing me with irreplaceable mental support and advice. Her sincere friendship is certainly one of the most valued possessions of mine.

This thesis is dedicated to my biological mother, Alina Nazare and my adoptive mother Simona Rednic, who dedicated themselves to see me succeed in life.

Special thanks go to Kateřina Bišová, Centre Algatech, Institute of Microbiology CAS, for providing me with the genome sequence for *Desmodesmus quadricauda*.

Computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

Contents

1. Introduction

Multiple definitions for plants have been introduced throughout history. Around 300 Before Christ (BC) the Greek philosopher Aristotle classified all living beings around him into non-motile plants and animals "with blood" and "without blood" (Manktelow, 2010). By the 1700's taxonomy evolved through Carl Linnaeus, who established the kingdoms of animals, vegetables and minerals (Podgorny & Irina, 2019). Current classification of living organisms (made by Thomas Cavalier-Smith) considers *Plantae* as a kingdom containing the biliphytes (*Glaucophyta* and *Rodophyta)* and *Viridiplantae* (Ruggiero et al., 2015). However, the most widely accepted definition of a plant is any organism belonging to the *Viridiplantae* subkingdom (Schreiber et al., 2022). *Viridiplantae* (translated "green plants") are a monophyletic group branched into three major lineages: *Prasinodermophyta*, *Chlorophyta* and *Streptophyta (Charophyta)*, which are part of the eukaryotic kingdom *Archaeplastida* (comprising *Glaucophyta*, *Rhodophyta* and *Viridiplantae*) (Leliaert, 2019; L. Li et al., 2020). *Archaeplastida* gained photosynthetic ability by engulfment of cyanobacteria in a process called primary symbiosis, resulting in chloroplasts. Throughout the text, I use the term "plants" or "green lineage" to describe *Viridiplantae* [footnote1]. By divergent evolution, green lineage species transitioned from the deep seas and oceans to fresh water, and later land (Zhang et al., 2022). This transition took place in three major steps: the emergence of plants through primary endosymbiosis (approximately 1.2 bya), the evolution of charophytes and embryophytes from a common green algal ancestor, and the angiosperms accelerated developmental domination over gymnosperms around 130 million years ago (Zhang et al., 2022).

Species of the green lineage are the principal organisms capable of oxygenic photosynthesis among the eukaryotic groups, and they have a substantial contribution to our quotidian life (Zhang et al., 2022). Plants use light energy to produce O_2, the essential gas for aerobic respiration. Furthermore, by assimilating CO_2 (one of the major greenhouse gases), green plants actively contribute to the global ecosystem and carbon recycling.

As a source of food, plants benefit not only humans (through agriculture), but also other organisms such as bacteria and heterotrophic eukaryotes (such as amoebae and ciliates) including animals, which use their sugar contents as an energy and carbon source. According

[footnote 1] Throughout the text, I use italics for all taxa levels in Latin and regular font with small letters for familiarized versions of the names (e.g. *Chlorophyta* vs. chlorophytes, *Angiospermae* vs. angiosperms).

to the United Nations (UN) world population prospects, the world population has rapidly increased over the past 200 years. The number of inhabitants on our planet has reached 7.9 billion in November 2021, more than double the population number of approximately 4 billion in 1975 (United Nations, DESA, Population Division, 2022). As the continuously growing population increases, so do the dietary needs to be met. The United Nations Food and Agriculture Organization (FAO) estimated in 2011 that in order for the human population to be fed, crop production should increase by 70% by the year 2050 (FAO, 2009). Medicinal applications are also widely employed within the green lineage, especially throughout the green algae (*Chlorophyta*) and vascular plants (*Tracheophyta*) taxa, for the anti-inflammatory, antimicrobial, and antioxidant properties they possess (Cowan, 1999; Haq et al., 2019; Wiart, 2006). Also mentionable are the herbs and condiments used in worldwide cuisine and the trees and shrubs utilized in the latex, paper and construction industry.

The most ancient phylum of *Viridiplantae* is *Prasinodermophyta* (or "prasinodermophytes"), a group of green algae which is a sister branch to the clade diverging into *Chlorophyta* and *Streptophyta* (L. Li et al., 2020). Adapted to low-light and low-nutrient conditions, prasinodermophytes are marine planktonic microalgae (Piganeau, 2020; Schreiber et al., 2022). Their ancestor is now believed to have been flagellated, which is a novel discovery for this group (Piganeau, 2020).

Chlorophyta (or "chlorophytes") forms a group of unicellular and multicellular eukaryotic photosynthetic organisms with simply organized thalli. Chlorophytes are aquatic, motile green algae. From unicellular, to colonial, filamentous and multicellular, chlorophytes exhibit a substantial morphological variation (Schreiber et al., 2022). They are a well-known protein rich food source commercially cultivated in Asia, also used in the pigment and cosmetic industry (Abu-Ghannam & Gupta, 2010; J. K. Kim et al., n.d.; Shah et al., 2020).

Streptophyta (or streptophytes) are the first green organisms adapted to life on land. They include the streptophytic green algae and land plants (*Embryophyta*) (Figure 1). Streptophyte algae can further be classified into the *Klebsomidiophyceae*, *Chlorokybophyceae*, and *Mesostigmatophyceae* (KCM) order and the more advanced *Zygnematophyceae*, *Coleochaetophyceae*, and *Charophyceae* (ZCC) order (Schreiber et al., 2022).

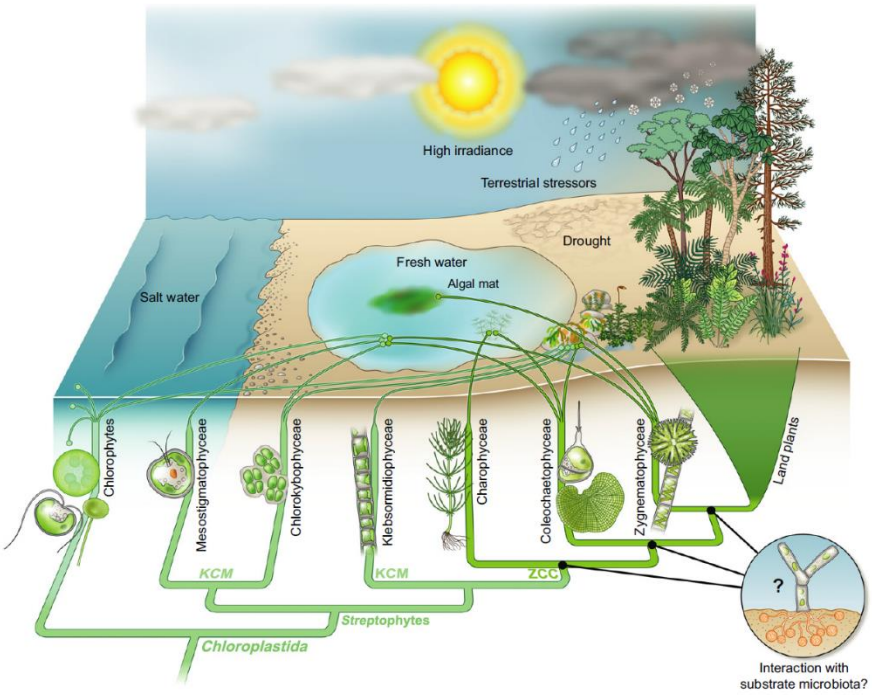


Figure 1 *Streptophyta cladogram*, as previously described in (de Vries & Archibald, 2018) showing *Chlorophyta* as outgroup.

Among ZCC clade, the most ancient are the *Charophyceae* dating back to the Devonian era (>380 mya). The extant order of this group, Charales, are called stoneworts due to their calcified surface (Lambert, 2009). Streptophyte algal species range in morphology from unicellular to filamentous and multicellular (Turmel & Lemieux, 2018). The latest diverged groups are the *Coleochaetophyceae*, which developed the thallus alike extant land plants, and *Zygnematophyceae* (Fig. 1), the largest ZCC group considered to be the precursor of land plants, despite its morphological reduction (Zhang et al., 2022).

Embryophyta ("embryophytes" or "land plants") are the most specialized multicellular representatives of plants. They originate from the Ordovician era (450 mya), when they gave rise to a novel biosphere. *Embryophytes* currently dominate the global terrestrial flora, with angiosperms accounting for more than 95% of the approximately 300 000 extant species (classified in Figure 2). Their name is derived from their ability to produce a diploid embryo via mitotic division of the zygote (Broutin, 2019).

The transition from aquatic charophytic algae (stoneworts) to land plants has been discussed in multiple phylogenetic studies. The ancestor of the embryophytes is thought to be a member of *Zygnematophyceae* which molecularly adapted to stress factors such as drought and initiated apical cell growth, developing the primordial 3-D growth and branching (Rensing, 2020; Schreiber et al., 2022). Fossil records document the existence of land plants around 450 million years ago (mya). A 455 million-year-old fossil suggests an unknown moss species as a possible candidate for the first avascular land plant (Cardona-Correa et al., 2016). Another study from 2018 considers *Cooksonia barrandei* (an amphibious species living around 432 mya), as the earliest vascular land plant. *Cooksonia* is the oldest known plant to have had a stem with vascular tissue, and had a sporophytic life stage with trumpet-shaped sporangia (Libertín et al., 2018).

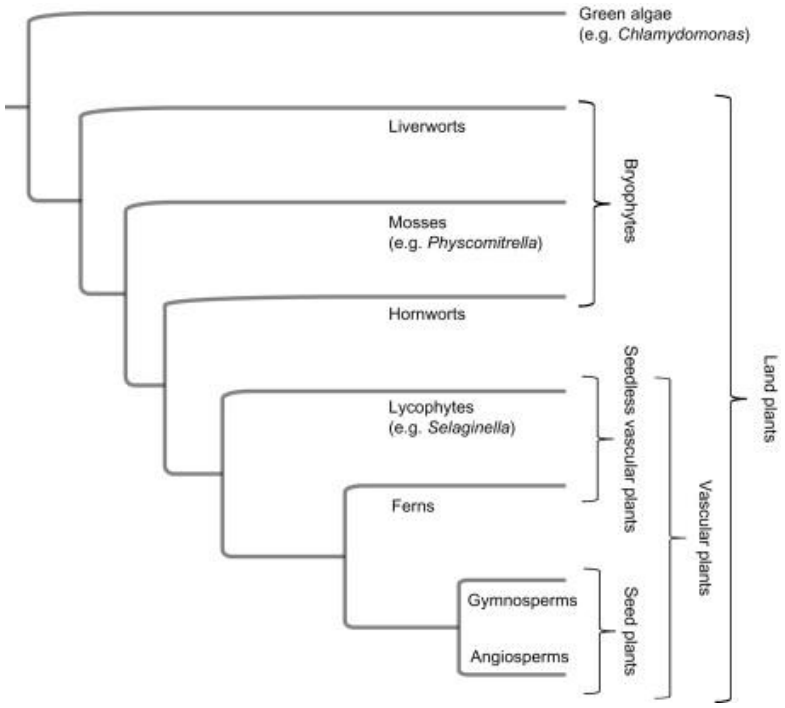


Figure 2 *Embryophyta cladogram* as presented in (Jiao & Guo, 2014)

1.1 The plant genome

Mechanisms driving genome evolution are still unclear, in spite of decades of intense study. Genome evolution is driven by accumulation of changes, including DNA mutations, duplications and rearrangements (Soltis & Soltis, 1999). Understanding the processes and pathways which underlie the evolution of the green lineage genomes provides a considerable insight into plant evolution, body form development and basic mechanisms of genome function and control, with implications for plant genetic manipulation or breeding. As a result,

studying the plant genomes is important both for fundamental science as well as for agricultural applications. The agriculture field would benefit by acquiring genomic data applicable for genetically modified (GM) plants. GM plants may be engineered for resistance to drought or other environmental stresses, reducing the amount of fertilizer or water required, inducing resistance to pathogens or pesticides, and even increasing nutritional levels and flavor (Halagarda & Popek, 2018; Kumar, 2021). From the scientific community point of view, analyzing the genomes of the green lineage, especially repetitive elements which contribute to genome evolution and gene expression, will deliver valuable information about the key drivers of evolution and epigenetic mechanisms which dictate how plants adapt to their surroundings and shape environmental diversity.

Plants have a larger number of genes than vertebrates, most of them conserved during evolution (Armisén et al., 2008; Claros et al., 2012). However, in contrast to vertebrates, their genomes exhibit extensive variations in size. Land plant species have a 2340-fold variation of genome sizes, chlorophytes and streptophyte alga 4680 (Leebens-Mack et al., 2019; Leitch & Leitch, 2013). The main responsible causes for this complexity are whole genome duplications followed by rearrangements and reductions, and repetitive element copy number expansion (Kress et al., 2022).

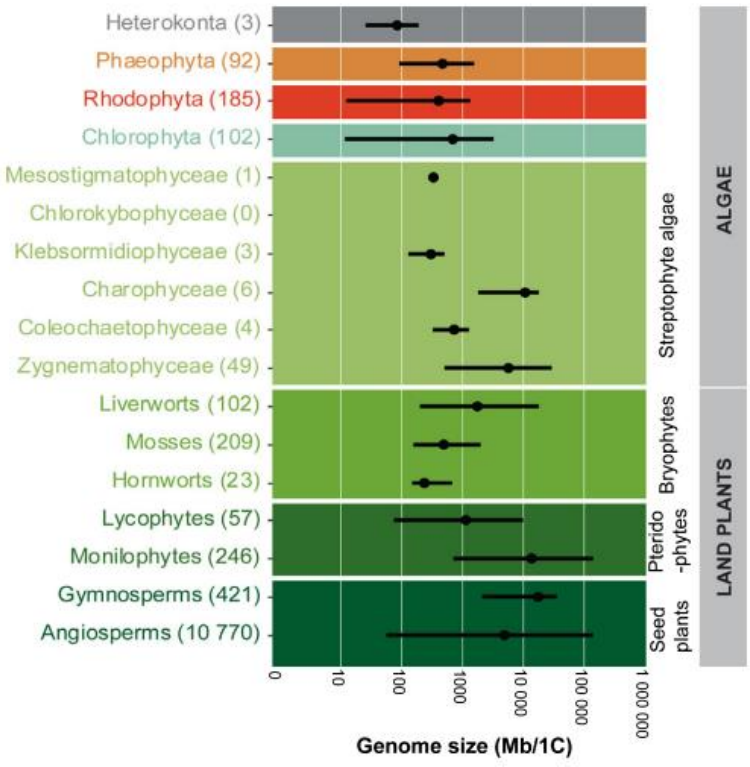


Figure 3 *Genome sizes across algae and land plant species* from (Pellicer & Leitch, 2020)

The C-value is the amount of DNA enclosed in a haploid cell nucleus given in picograms, regardless of the level of ploidy (Fleury et al., 2012). C-values among land plants show a dynamic distribution, alternating between genome reductions and expansions over approximately 4 orders of magnitude (Michael, 2014). The "C-value paradox" states that the developmental and body plan complexity of an organism is not connected to the size of its genome (Ågren & Clark, 2018). For example, the unicellular protozoan *Polychaos dubium* (earlier *Amoeba dubia*) has the largest known C-value of all organisms (approximately 685 pg), whilst humans have a 200 times smaller C-value (approximately 3 pg) (Doležel et al., 2003; Karami, 2013).

Although multiple plant genomes have been sequenced and annotated, much remains to be done and discovered. Only 2202 plant genome assemblies are currently deposited on the National Center for Biotechnology Information (NCBI) website, of which 819 are completed at a chromosomal level, and only 7 complete. This number is extremely low in comparison to the sequence data availabe for animals (9083). Multiple explanations are possible for this discrepancy. Possibly the most relevant one would be the general lack of interest in sequencing non-agricultural plants (Kersey, 2019). Another explanation could be the complications caused by the abundance of repetitive elements in the plant genomes. Repetitive elements in plants may account for 15% (*Genlisea nigrocaulis Steyerm*) up to 85% (*Zea mays*) of their genome size (Schnable et al., 2009; Vu et al., 2015). Also, the extensive number of introns developed in most of the genomes of green plants coupled with the high number of repetitive elements complicate gene annotation (Kress et al., 2022).

However, the recent advancement of bioinformatic and sequencing tools in the past 20 years has considerably improved the plant genomics field (Marks et al., 2021). With the optimization of long read sequencing technologies (further described in chapter 1.4.2.3), the number of sequenced plant species has increased from 383 in 2018 to 812 in 2020 and continues to rise exponentially (Kersey, 2019; Kress et al., 2022). A more in-depth study and interpretation of green plants genomic identities will broaden our understanding of speciation, evolution, and physiological development (Kress et al., 2022).

1.2 DNA sequencing

In 1944, the molecule carrying genetic information was identified by three members of the Rockefeller Institute for Medical Research in New York. Avery, MacLeod and McCarty proved that deoxyribonucleic acid (DNA), a long α-helical structured polymer is the genetic

material responsible for cell identity (Avery et al., 1944). This and other discoveries prompted scientists to decipher the structure of DNA, which was officially announced in 1953 by J. Watson and F. Crick (Watson & Crick, 1953). DNA is a double helix consisting of a (deoxyribose-phosphate backbone to which four types of nitrogenous bases (purines and pyrimidines) are attached. Adenine (A) and thymine (T), and cytosine (C) and guanine (G) respectively pair with each other through hydrogen bonds, forming two antiparallel nucleotide strands.

According to the central dogma of molecular biology, the expression of genetic information in a cell starts with DNA-encoded "instructions", transcribes into RNA intermediates which determine the synthesis (translation) of functional proteins (Crick, 1958). Sequencing methods enabled researchers to "read" genomes and opened the era of genomics: an area of genetics that concerns the sequencing and analysis of genetic information from an organism (Mandal, 2020).

1.2.1 Sanger sequencing

The dideoxy DNA sequencing method was invented in 1977 by Frederick Sanger. Sanger discovered that incorporation of a chemically modified dideoxynucleotide (ddNTP) which lacked the OH group on the 3' end in a newly synthesized DNA strand triggers DNA synthesis termination. This led to the development of the "Dideoxynucleotide trisphosohate chain terminator" also called "Sanger sequencing", which implied four reactions running in four separate tubes, each containing all deoxynucleotides (dNTPs), and one dideoxynucleotide (ddNTP): ddATP, ddTTP, ddCTP and ddGTP, respectively. In each reaction, it was a matter of chance when the ddNTP was integrated, and the synthesis is terminated. This resulted in multiple fragments of DNA, each terminated at a different length. These fragments could be separated on the basis of their size by running each reaction tube on an electrophoresis gel apparatus. The sequence could then be read one nucleotide at a time from bottom to top in a 5' to 3' direction (Valencia et al., 2013).

Despite its high efficiency, this method was very laborious and time consuming, as it could only sequence a maximum of 1000 base pairs at a time. In addition, the use of radioactivity made library preparation complicated. Automatic Sanger sequencing machines which used fluorescently labelled ddNTP´s and capillary electrophoresis were produced in 1987 by Applied Biosystems, and became the primary means for DNA sequencing (Liu et al., 2012; Penn State, 2017).

In 1990, the Human Genome Project planned the complete sequencing of the human genome using the Sanger method over a time span of 15 years. Laboratories from all over the world joined forces and each started sequencing short fractions of human genes using Bacterial Artificial Chromosome Cloning (BAC) in a hierarchical shotgun strategy. Hierarchical because each fragment of nucleotides with a known position in the genome belonged to a larger fragment. Shotgun because the BAC clones were eventually truncated and randomly sequenced. After sequencing, the fragments were reassembled by studying the sequence overlap (Waterston et al., 2002).

By 1998 and about halfway to accomplish the Human Genome Project, Celera Genomics was founded by Craig Venter (Shampo & Kyle, 2011). The revolutionary idea behind this investment was to use the random shotgun method for the entire genome without employing BAC clones and to use computers to map all the obtained fragments sequences by sequence similarity (Waterston et al., 2002). This meant that portions of the genome were deciphered based on chance, so each one of them could be selected multiple times, meaning that throughput would increase and the required time would considerably decrease. With hundreds of sequencing machines running constantly, Celera Genomics announced the plan to complete the human genome from scratch in less than two years (Marshall, 2000). This initiated the historical race between the publicly founded Human Genome Project and Celera Genomics to publish the human genome. The final outcome was the publishing of 90 % of the human DNA by both opponents simultaneously in 2001 (Shampo & Kyle, 2011). Ultimately, the complete sequence of a human genome was just published by the Telomere-to-Telomere (T2T) Consortium presenting a gapless assembly for all chromosomes except the Y chromosome (Nurk et al., 2022).

1.2.2 Next generation sequencing

After the scientific breakthrough of obtaining the human genome sequence brought about by Sanger DNA sequencing, a new era of genome sequencing emerged. Next Generation Sequencing (NGS) entails innovative sequencing techniques with improved throughput and lower costs (Liu et al., 2012). Perhaps the most notable improvement offered by NGS was the novel ability to sequence multiple fragments at the same time, greatly increasing the speed. The steps it employs are library preparation and sequencing. Subsequently, many bioinformatic tools and algorithms have been improved to analyze NGS data (Van Dijk et al., 2014).

On the basis of the starting material, NGS can be used on RNA, DNA, or proteins. Classified by their principle of operation, NGS technologies include sequencing by synthesis and sequencing by oligo ligation detection (DeWitt, 2020). Sequencing by synthesis functions by amplification, denaturation, and subsequent DNA replication taking place in cycles until the desired amount of DNA is reached. The sequencing by synthesis protocols are diverse from one company to another, but can be grouped based on the dNTP incorporation reaction mechanism. Sanger sequencing and Illumina sequencing involve reaction stop as each dNTP is inserted in order to obtain a read, whilst pyrosequencing and Ion torrent sequencing attain reads simultaneously with the incorporation of nucleotides (Kircher & Kelso, 2010).

1.2.2.1 Brief history of NGS

NGS revolutionised the field of genomics with a new library preparation strategy: random DNA fragmentation and subsequent adaptor ligation (J. Shendure & Ji, 2008). Launched in 2005 by 454 Life Sciences, the first NGS technology had a read length of 700 bp and an accuracy of 99.9% (L. Liu et al., 2012). The mechanism behind it was pyrosequencing on magnetic beads, which was fast, however, still expensive and prone to polymer errors as a result of accumulated light intensity variance (L. Liu et al., 2012; Luo et al., 2012). The technology was discontinued in 2013 (Slatko et al., 2018).

Almost two years after the release of 454 technologies, SOLiD (Sequencing by Oligo Ligation Detection) was released by Applied Biosystems. Based on the concept of polony sequencing, developed by George M. Church and his team at Harvard Medical School in Boston, Massachusetts, SOLiD also employed the use of beads, however they could be bound to a glass slide directly by their 3‘ end, instead of using a picotiter plate (Kircher & Kelso, 2010).

The next step in the NGS era was taken through the development of sequencing by detection of hydrogen ions. Ion Torrent sequencing technologies are still used in current times owing to their low cost, quality stability and acceptable read length of 200-600 bp (Gupta & Gupta, 2013; L. Liu et al., 2012). However, the most used short read technology nowadays is the Illumina sequencing by synthesis technology. Illumina sequencing uses repetitive bridge amplification of DNA sequences separated into clusters. Fluorescently labelled oligonucleotides are incorporated one at a time over the billions of clusters in the flow cell, pausing the reaction between each nucleotide detection (Figure 4) (Slatko et al., 2018). As a result, Illumina technologies achieve a higher throughput and lower indel error rate than Ion Torrent (Loman et al., 2012).

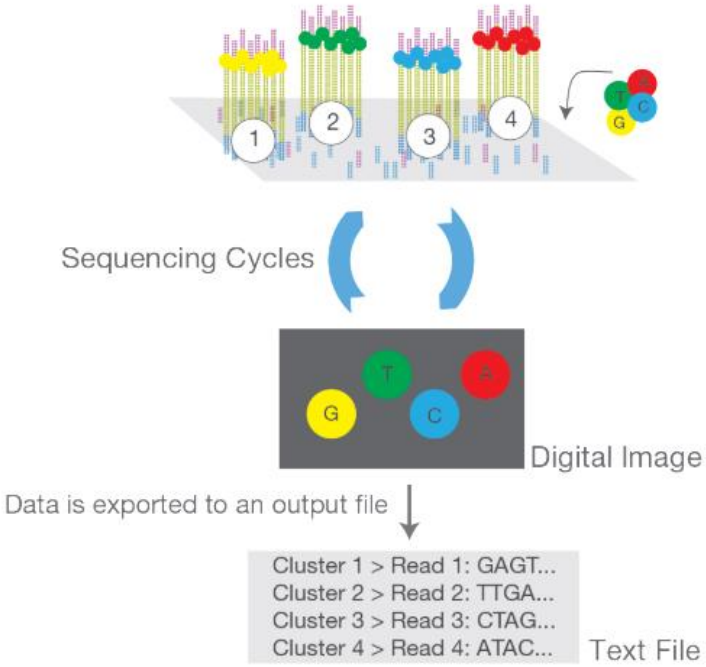


Figure 4 *Illumina sequencing chemistry overview* as described in (Illumina, 2017). Fluorescently labelled oligunucleotides are added and incorporated one at a time, producing light. After each step a digital image is produced and the wavelength and intensity of the light is documented. The nitrogenous base type is identified, (as A,T,C, or G) and stored in a text file. The process is repeated until all fragments have been processed.

Following the success of Illumina sequencing in genetics, two novel sequencing tools emerged. Pacific Biosciences and Oxford Nanopore currently offer the longest reads as output for their sequencing technology. Nevertheless, a cost decrease is required in order for these technologies to take over Illumina technology (further discussed in next chapter).

1.2.2.2 Strategies and applications of NGS

NGS is currently used in many scientific fields. The wide range of applications include studying the transcriptome and epigenome, capturing phylogenetic diversity in relation to disease (virology, epidemiology, oncolgy, etc.), metagenomic study of microbiomes, archaeological study of extinct species, and others (Barba et al., 2014; Morozova & Marra, 2008; Reis-Filho, 2009).

Depending on the starting material and on the desired result, NGS requires the preparation of customised sequencing strategies. Adequate sequencing strategies are necessary in order to avoid poor quality or un-mappable reads, PCR duplicates, and/or adapter reads, which are the most common reasons underlying a failed sequencing run (Genohub Inc., 2019).

The major concepts taken into account when organising a sequencing strategy are: sequencing libraries (single end, paired-end, or mate paired reads), read lengths, costs per base, and accuracy (J. A. Shendure et al., 2008). Each sequencing project begins with isolation of the target material, followed by library preparation. There are certain prerequisites for the selection of the sequencing library. The most important requirements are accurate establishment of genome size and targeted coverage, and selection of fitting informatic software for post-sequencing read assembly and annotation.

Although Illumina technologies produce short read scaffolds (35-300 base pairs per read), their sequencing services are the most used in current times (Levy & Boone, 2019). The reason for this general predilection is the effectiveness-low cost relationship. Illumina sequencing machines achieve an error rate lower than 1 % per run at a reagent cost of around $0.10 per million base pairs (Glenn, 2011; Reuter et al., 2015). Moreover, the paired-end sequencing feature grants clients the ability to perform a wide range of operations including whole genome resequencing, de novo whole exome sequencing (owing to the high coverage), and transcriptome related sequencing such as RNA, ChIP, and epigenetic modifications (Levy & Boone, 2019; Morozova & Marra, 2008). Illumina is also effective for repeat analysis (McCoy et al., 2014).

Long read sequencing technologies are currently employed by Pac Bio and Oxford Nanopore (Levy & Boone, 2019). They are particularly used for sequencing small bacterial and viral genomes, but may also be applied for larger telomere to telomere genome assemblies (Reuter et al., 2015). As Marks et al. concluded, contigs obtained via short read technologies (Illumina) exhibit an N50 32 times lower than those assembled via long read sequences, meaning the mean contig length is significantly higher (Marks et al., 2021). The genome coverage is increased (ranging from 1 kb to >1 Mb), which means the number of genome gaps is reduced (Logsdon et al., 2020). However, the increased read length comes at the expense of higher error rates and higher per-base sequencing costs. Error rates for Pacific Biosciences (Pac Bio) machines range from 2% to 8%, and the per-base sequencing cost may reach 86 $ per billion base pairs (T. Jiang et al., 2021). Nanopore sequencing technologies are cheaper than Pac Bio services (Branton et al., 2008).

For extremely large whole genomes, a widely employed strategy is combining Illumina sequencing technology with Pac Bio. Errors from reads obtained by Pac Bio are corrected by

applying Illumina only on the error regions, and assemblies obtained from both companies are then merged into one dataset (Y. Liu et al., 2017).

NGS techniques have their limitations (Erlich et al., 2008). Sequence assembly is less accurate than in Sanger sequencing, despite the considerable improvement in elapsed time and costs (Knyazev et al., 2021) Sequencing equipment can be very expensive, and has certain standards for the purity of DNA material (Gumińska et al., 2018).

Good sampling protocols are vital, as the sequenced material quality determines the integrity of the sequencing data (Van Dijk et al., 2014). Among the determining factors for quality control, sample ambient temperature, freshness, and level of exposure to contaminants are of utmost importance. In addition, fragment size has to be taken into account depending on the sequencing technology used. The fragment length usually targeted during the shearing step stands around a few hundred nucleotides (Slatko et al., 2018).

Sequencing library preparation is not bias free, as errors may result from unequal sonication of euchromatin and heterochromatin (in the case of chromatin sonication for chromatin immunoprecipitation) and from PCR preferential amplification of neutral guanine-cytosine (GC) percentage (Van Dijk et al., 2014). Also, NGS technologies result in an increase of base pair calling error as sequencing position progresses on the nucleotide strand (Mohsen et al., 2019). Correct classification of transposable elements (further described in chapter 1.3.2) is important, as the presence of coding regions in some of these repeats may result in erroneous gene annotation (Lerat, 2009). Another limitation is the low accuracy of the low diversity sequences at the first stages of Illumina sequencing (Mitra et al., 2015). Some of the errors have been reduced over the years by the development of error-correcting algorithms (Luo et al., 2012). Additionally, advancement in NGS techniques delivers longer read lengths, new bioinformatic instruments, and lower error rates each year, showing a promising future outlook.

1.3 Repetitive elements as integral components of the genome

The eukaryotic genome contains coding and non-coding regions accompanied by mitochondria and chloroplast DNA sequences acquired through endosymbiosis (Filip & Skuza, 2021). Parts of the eukaryotic genome which do not code for proteins are usually located between the coding regions or at the chromosome telomeres and centromeres. They encompass the intergenic regions, spacer sequences, and in some cases repetitive DNA sequences, and were shown to control gene expression regulation by transcription factor

binding (Wu & Brown, 2018). Mainly accountable for the genome complexity among the green plant clade are the repetitive DNA sequences, which occupy its majority in some species (Pisupati et al., 2018). Based on their arrangement and their type, repetitive sequences can be classified into tandem repeats and dispersed (or interspersed) repeats, which disperse throughout the entire genome (Lerat, 2009) (Figure 5).

Figure 5 *Classification of repetitive elements in the eukaryotic genome* as presented in (Paço, 2014)

Repetitive DNA, unlike unique DNA sequences, which directly encode amino acids for incorporation into proteins, may execute coding, as well as noncoding functions in the cell (Zdobnov et al., 2005). They play important roles in the chromosomal structure and are highly

relevant from an evolutionary, physiological, and epigenetic point of view, as they control gene expression (Mehrotra & Goyal, 2014).

Every known living organism retains repetitive DNA in its genome, in highly variable amounts (Favarato et al., 2017). Repetitive elements can be detected in even the smallest genomes and cause a genome size variation of 5 orders of magnitude in eukaryotes (Blommaert et al., 2019).

1.3.1 Tandem repeats

Tandem repeats represent successive copies of certain nucleotide sequences. Major tandem repeat types are mainly located in heterochromatic regions in the genome (Mehrotra & Goyal, 2014). They include all types of satellite sequences that range in length from a few up to 30000 base pairs. Microsatellites are the shortest type of tandem repeat and incorporate 1-6 nucleotides (Lerat, 2009). Larger elements are minisatellites, which consist of 6-100 nucleotides, followed by satellites.

Owing to their heterochromatinization, satellite elements are mainly not transcribable but play a role in the function and protection of the centromeric and telomeric regions in the chromosome. Centromeric repeats, for example pAL (150 bp) in plants and Alu in humans are considered key drivers of the evolution of chromosomes and genes (Hartley & O'neill, 2019). Despite their fast evolution, centromeric repeats tend to manifest themselves in a homogenous fashion in genomes. Unlike centromeric repeats, telomeric repeats diverged slowly throughout plant evolution (Lermontova et al., 2015). Their composition is usually defined in the form of $(T_xA_yG_z)_n$ creating motifs for the binding of specific proteins and enabling telomere extension by telomerase. Therefore, modifications of telomeric repeats have a considerable influence on chromosome end stability, namely on cell ageing (Bianchi & Shore, 2008; Lindsey et al., 1991; Peska & Garcia, 2020).

Ribosomal DNA (rDNA), especially the 5S and 45S rDNA, is a transcriptionally active representative of tandemly arranged repetitive elements. Its transcription results in the production of the majority of eukaryotic ribosomal RNA, the structural constituent of ribosomes The 5S rDNA encodes the 5S rRNA, while the 45S rDNA transcript is processed into the 18S, 5.8S and 25(28)S rRNAs (Rosato et al., 2016; Warmerdam & Wolthuis, 2018).

1.3.2 Dispersed repeats

Dispersed repeats comprise recurring DNA patterns resulting either from duplications or from DNA insertions (Rao et al., 2010). These repeats can account for up to more than 50% of DNA

and they are distributed throughout the genome, rather than being restricted to certain loci (Flavell, 1986; Mehrotra & Goyal, 2014). Dispersed repeats are mostly composed of transposable elements (TEs), repetitive sequences of DNA which can relocate and/or multiply themselves within the genome (Klein & O'Neill, 2018; Lerat, 2009). Transposable elements are the main source of variability in the plant genome, and their nucleotide composition is species specific (N. Jiang, 2013). Their ability to produce insertion or deletion mutations results in transcriptional reprogramming, a likely factor of plant adaptation to environment (Lisch, 2012).

A classification system for transposable elements was established in 2007 by Wicker et al. (Wicker et al., 2007). As a result, all dispersed repeats can be further classified into class I (retrotransposons), and class II (DNA transposons). Class I transposable elements use a "copy and paste" mechanism to transpose throughout the genome aided by RNA intermediates, resulting in an increased replication speed. Class II DNA transposons on the other hand, are mobile by means of a DNA intermediated "cut and paste" or "peel and paste/rolling circle amplification" (in the case of Helitrons) mechanisms (Bourque et al., 2018; Y.-J. Kim et al., 2012) (Figure 6).

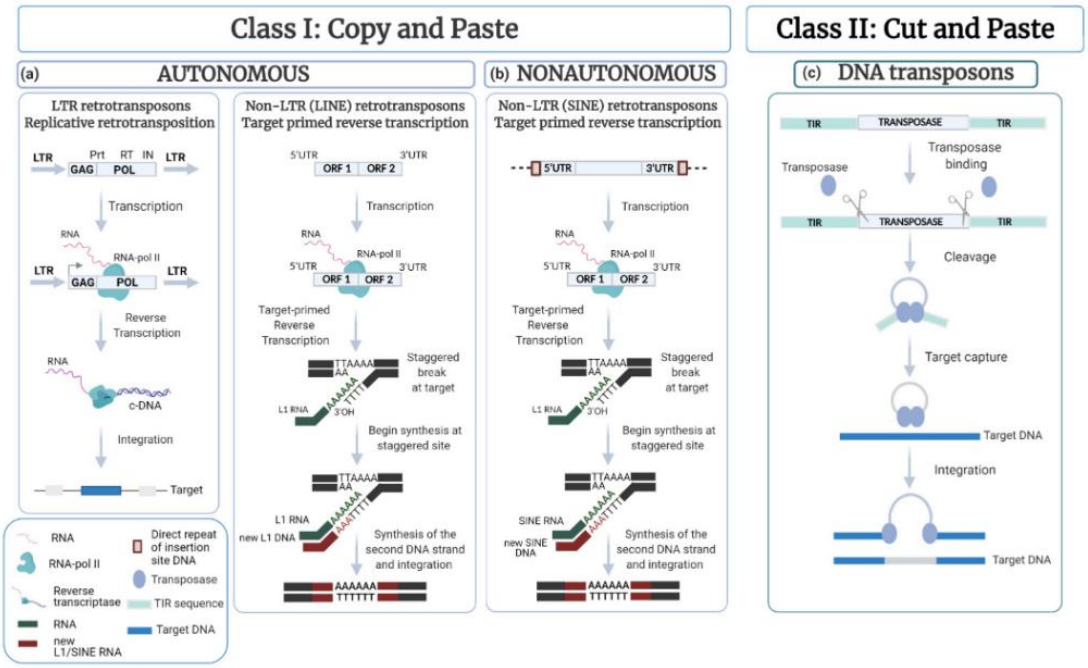


Figure 6 *Transposition mechanisms described in* (Colonna Romano & Fanti, 2022)

1.3.2.1 Retrotransposons

Based on their relocation mechanisms, retrotransposons (Class I repeats) are divided into five orders: Long Terminal Repeats (LTR), Dictyostelium Intermediate Repeat Sequences (DIRS), Penelope like elements (PLEs), and long and short interspersed nuclear elements (LINEs and SINEs) (Wicker et al., 2007). Retrotransposons may function autonomously, meaning they can transpose using their own encoded proteins, or non-autonomously, requiring reverse transcriptase from other transposable elements (Janicki et al., 2011). SINEs are retrotransposons which depend on other transposable elements to gain mobility (Figure 6), as they rely on LINE for transposition (Berthelier et al., 2018). LTRs are not frequent in animals, but are abundant in plants, constituting the majority of repetitive elements in most studied plant genomes (N. Jiang, 2013; Wicker et al., 2007).

1.3.2.2 DNA transposons

Class II transposable elements are present in almost all eukaryotes, usually in relatively small numbers, and can be grouped into 13 superfamilies (Jurka et al., 2007). They are composed of a transposase gene flanked by Terminal Inverted Repeats (TIRs) (Muñoz-López & García-Pérez, 2010). On the basis of the number of DNA strands cut to facilitate their transposition, DNA transposons include subclass 1 (double-stranded cleavage) and subclass 2 (mono-stranded cleavage) (Wicker et al., 2007). The cleavage reactions are catalyzed by the enzyme transposase (Jurka et al., 2007).

1.4 Identifying repeats and transposable elements in genomes

For a non-erroneous structure annotation, masking repetitive elements is conventionally performed prior to a plant genome annotation (N. Jiang, 2013). Genome masking implies the comparison of the analyzed genome to a repetitive elements database, identifying its repeat content (Tarailo-Graovac & Chen, 2009). The most popular tool used for this purpose is RepeatMasker. Also, due to the poor conservation and species specificity of TEs, a species specific repeat library should be created and used (N. Jiang, 2013; Yandell & Ence, 2012).

The overall DNA repeats contents in sequenced genomes can be identified using software such as Repeat Explorer 2 and RepeatModeler. The most important difference between these two software programs is the input data. Whilst RepeatModeler uses genome assemblies as input data, RE2 is assembly-free, requiring only short read Illumina data (Berthelier et al., 2018; Novák, Neumann, et al., 2020). Since repetitive elements create complications for assembly

algorithms, not many high-quality assemblies exist for eukaryotic species. This creates an advantage for assembly-free repeats identification software programs, as they compare reads directly with repetitive elements databases (Weilguny & Kofler, 2019).

On another hand, a variety of tools exist for specifically identifying Transposable Elements in genomes. Two major strategies exist for TE identification and annotation: homology-based and structure-based. The first one, uses the reads similarity either with TE databases, or within themselves, and requires no prior knowledge about the sequences of interest. The structure-based strategy detects TEs by similarity between their own structural components (Bergman & Quesneville, 2007). This implies specifically selected models for each TE type as a prerequisite and does not require TEs to be repetitive for their detection. Recently, a third strategy has been developed for TEs *de novo* identification that is repetitiveness-based (N. Jiang, 2013). The new versions of the software RepeatModeler 2 and Repeat Explorer 2 (RE2) are de novo identifying software programs for TEs we mention (further described in chapter 1.6.3). Moreover, the Tandem Repeat Analyzer (TAREAN) software can be used for identifying satellite repeats. TAREAN uses a clustering algorithm to detect satellite repeats from unassembled short reads (Novák et al., 2017).

2. Aims

The aim is to explore the genome evolution and diversity of repetitive elements in the green lineage (*Viridiplantae*) species. To achieve this aim, I will analyze the distribution of repetitive elements in the genomes of 12 organisms ranging from single-cells green algae to land plants. Moreover, I formulate the following specific aims for this thesis:

I. Determine the proportion of repetitive elements in the genomes of 12 green lineage species, including single-celled green algae, moss and seed plants using RepeatExplorer 2.

II. Characterize and annotate the identified repetitive elements.

III. Compare repetitive element amount and composition in the different species of the green lineage.

3. Materials and Methods

3.1 Selected species and data sources

The main criterion for species selection was the phylogenetic position of the organisms in the green lineage tree of life. Since the most early diverged Viridiplantae species are the chlorophytes, I identified the genomes for five representatives of this green algae phylum. I selected two prasinophytes (*Desmodesmus quadricauda* and *Scenedesmus sp.*), one Chlorophyceae species (*Chlamydomonas reinhardtii*) and two Trebuxiophyceae representatives (*Chlorella vulgaris* and *Volvox africanus*).

The next organisms documented in evolutionary order were the green algae representatives belonging to Streptophyta (Gitzendanner et al., 2018). This group is represented by *Klebsormidium nitens*, *Chara braunii* and *Zygnema circumcarinatum*.

Among the embryophytes, bryophytes are the earliest branching land plants. As a result, I analyzed the genome of *Physcomitrium patens*, a wetland model moss whose life cycle is dominated by the gametophytic stage. In addition, *Marchantia polymorpha*, a thallose liverwort belonging to the *Marchantiophyta* clade was analyzed.

The second criterion for species selection was data availability. Since short reads were required for the analysis, all the selected datasets were attained by Illumina sequencing. Illumina sequencing raw data were obtained from the Sequence Read Archive (SRA) database, NCBI (National Center for Biotechnology Information). A narrowing of the possible number of candidates was caused by the lack of genomic data available, as out of the approximately 450.000 plant species, less than 0,2% have been completely sequenced (Kress et al., 2022). For example, the whole genome assembly of *Selaginella moellendorffii is* not available in NCBI. *S. moellendorffii* is the common model organism from the *Selaginella* genus and represents *Lycophyta*, the early diverged vascular plant. Surprisingly, I could not locate Illumina genomic sequencing raw data in the SRA database for *S. moellendorffii* but Illumina sequencing raw data for *Selaginella tamariscina* was located. However, the origin of all the other land plants has phylogenetically been proven to be a distinct branch, *Euphyllophyta*. *Amborella trichopoda*, the most primitive angiosperm is a member of this clade included in the selection. Characterized by absence of vessels in its vasculature, *A. trichopoda* uses only simple tracheids for transport of water and assimilates. The more diverged land plant selected from the euphyllophytes is *Arabidopsis thaliana*, perhaps the species with the most well-

studied genome. This organism is a crucifer flowering plant widely used as a model in genetic laboratories.

Table 1 List of the Studied Organisms includes NCBI Taxonomy ID and genome sizes information coupled with accession for genomic NGS reads used in the analyses.

Taxon	Class/Order	Taxonomy ID	Estimated genome size (Mb)	Accession number
Chlamydomonas reinhardtii	Chlorophyceae/ Chlamydomonadales	3055	293	SRR1734617
Volvox africanus	Chlorophyceae/ Chlamydomonadales	51714	293	DRR240879
Chlorella vulgaris	Trebouxiophyceae/ Chlorellales	3077	26,3	SRR14100011
Desmodesmus quadricauda	Chlorophyceae/ Sphaeropleales	3089	12,6	Bíšová et al (unpublished)
Scenedesmus sp. NREL 46B-D3	Chlorophyceae/ Sphaeropleales	2650976	151	SRR12974924
Klebsormidium nitens	Klebsormidiophyceae/ Klebsormidiales	105231	269	DRR103728
Zygnema circumcarinatum	Zygnemophyceae/ Zygnematales	35869	873	SRR10413429
Chara braunii	Charophyceae/ Charales	69332	1200	DRR054005
Physcomitrium patens (accession Kaskaskia)	Bryopsida/ Funariales	3218	518	SRR4408325
Marchantia polymorpha	Marcantiopsida/ Marchantiales	3197	802	SRR1800536
Selaginella tamariscina	Lycopolyopsida/ Selaginellales	137178	301	SRR7135413
Amborella trichopoda	Magnolipsida/ Amborellales	13333	868	SRR954308
Arabidopsis thaliana	Magnoliopsida/ Brassicales	3702	176	SRR7890551

3.2 Repetitive elements identification

RE2 is a computational tool designed to detect all similarities between short reads of 100-300 base pair length and reads from small proportions of the genome (1x), resulting in highly accurately isolated repetitive sequences. The use of graph-based clustering of next-generation sequencing reads yields a better classification algorithm. Furthermore, the Galaxy platform creates a comprehensible user-friendly interface for the operation of the procedures and for the interpretation of the results (Novák, Neumann, et al., 2020).

By virtue of the multitude of features RE2 displays, it is widely employed in the bioinformatics field. The most prominent characteristic of the RE2 tool is the easy-to-use interface

implemented on the Galaxy platform. The analysis does not require command line skills and the Galaxy instance of RE2 provides other utility tools such as quality checking and filtering and adapter removal. Also, RE2 is designed specifically for short sequence reads, which are currently predominant in genome databases, and is suitable for detection of all repetitive elements types.

3.2.1 Pre-processing of the data

The Illumina genomic sequencing raw data of all organisms were retrieved from the SRA database, together with their organelle (plastid and mitochondria) genomes sequences, when they were available. If organelle genomes for a species were not available, organelle genomes of another species from the same genus were used. If organelle genomes of another species from the same genus was not available, the analysis was performed without removing organelle DNA.

The organelle genomes sequences were retrieved in fasta format and merged to obtain one single file containing both the chloroplast and mitochondrial DNA sequences for each organism. This organelle genome file served as a template for the removal of all the reads that do not belong to the nuclear DNA.

To obtain organelle reads-free input files for the RE2 tool, we developed an in-house script which implements four analyses steps (listed in appendix):

1. The first step constitutes an alignment using the “Minimap2” program. During this process, the program uses organellar genomes sequences as a template. These reads are then stored by the “samtools” software in a mapping file in bam format.
2. The second step separates the reads into three files consisting only of the unmapped reads: one containing the unmapped paired reads, one containing the unmapped forward reads, and another one with the unmapped reverse reads.
3. In the third step, the three unmapped reads files are merged into one mapping file.
4. The fourth step extracts all the unmapped reads sequences according to the merged mapping file.

3.2.2 Repeat Explorer 2 analysis

A series of the RE2 utilities tools were employed in the analysis. The first tool used was the FastQC, which offers a detailed visualization of the read quality report including nucleotide composition among the read datasets. This report aids the decision-making process for

establishing the most appropriate trimming range, or even for discarding the dataset altogether if the per base sequence content and GC content are inadequate. The following tool is a preprocessing of the reads to filter any adaptors and bad quality reads by trimming the inadequately distributed read areas (most often at the beginning and at the end of the dataset due to adaptors used by sequencing facilities). Another RE2 utility tool is Preprocessing of FASTQ reads. This tool makes memory efficient preprocessing of fastq files. Output of this file can be used as input of Repeat Explorer clustering. There are two versions of this tool: for paired-end and single-end reads.

The submission form requires to set a read sampling parameter. Read sampling is set to 1x genome values and maximum100 nt length per read (Table 2). Then the clustering analysis can be done on the produced sequences from the preprocessing utility tool. In half of the analyzed organisms (*M. polymorpha, S. tamariscina, K. nitens, C. braunii, Z. circumcarinatum*, and *A. trichopoda*) the sampling reads was less than 1x, most probably because of high-copy number repeats (Novák, Neumann, et al., 2020) – see results. However, the read coverage range followed recommendations by developers is (0.5x - 1x).

Table 2 *Read sampling*

Taxon	Sampling
Chlamydomonas reinhardtii	0.33x
Volvox africanus	1x
Chlorella vulgaris	1x
Desmodesmus quadricauda	1x
Klebsormidium nitens	1x
Zygnema circumcarinatum	1x
Chara braunii	1x
Physcomitrium patens	0.8x
Marchantia polymorpha	1x
Selaginella tamariscina	1x
Amborella trichopoda	0.72x
Arabidopsis thaliana	1x

4. Results

Genomes were analyzed and processed using the RE2 computational tool, to isolate and provide a basic characterization of repetitive elements for each organism, followed by annotation and graphical representation. RE2 employs an effective graph-based clustering algorithm, which discards the organelle and contamination reads and groups each repetitive element into clusters by comparison to specific nucleotide databases. Following this algorithm, the tool produces a summary clustering graph portraying the unique elements of the genome (singletons) coupled with the repetitive genome elements grouped into clusters and superclusters. Clusters depict all repetitive sequences which belong to one particular repeat family, while superclusters show a more in-depth characterization of repeats by grouping clusters together on the basis of paired-end read similarity between repetitive elements, resulting in annotation (Novák, Neumann, et al., 2020). For the purpose of clarity, superclusters will from this point forward be referred to as annotated repeats, and clusters without superclusters as non-annotated repeats.

4.1 Pipeline setup, testing and data pre-processing

Chloroplast DNA was removed in the preprocessing in 11/12 organisms. Mitochondrial DNA sequence was only available for 9/12 organisms, for which reads corresponding to mtDNA were filtered away (Table 3). Since organelle reads were detected by RE2 even after removing them, they were not taken into account in the final results.

Table 3 *Organelle DNA removed from each dataset*

Taxon	Chloroplast DNA accession number	Mitochondrial DNA accession number
Chlamydomonas reinhardtii	NC_005353.1	NC_001638.1
Volvox africanus	NC_039755.1	EU760701 *Volvox carteri s. nagariensis*
Chlorella vulgaris	NC_001865.1	NC_045362.1
Desmodesmus quadricauda	-	AF204057 *Scenedesmus obliquus*
Scenedesmus sp. NREL 46B-D3	-	AF204057 *Scenedesmus obliquus*
Klebsormidium nitens	DF238762	DF238763
Zygnema circumcarinatum	NC_008117.1	-
Chara braunii	AP018555.1	AP018556.1
Physcomitrium patens	NC_005087.2	NC_007945.1

Marchantia polymorpha	NC_042505.1	FigShare link in appendix *Marchantia inflexa*
Selaginella tamariscina	NC_041646.1	-
Amborella trichopoda	NC_005086.1	-
Arabidopsis thaliana	NC_000932.1	NC_037304.1

To confirm the robustness of RE2, a test run was conducted using *Arabidopsis thaliana* with read sampling to reach genome coverage 1x and 0.5x. Repeat Explorer was able to successfully analyze the complete number of reads for both sampling methods, displaying an even higher number of analyzed reads than the input sample for the 0.5x sampling method. However, the percentage of reads representing repetitive elements was 12% lower in the 0.5x sampling than in the 1x sampling (Figure 7).

In contrast to the 50%-50% GC distribution, the genome of *Scenedesmus sp*. NREL 46B-D3 showed a 30%-70% ratio, suggesting contamination with a foreign organism. Hence, it was removed from further analysis.

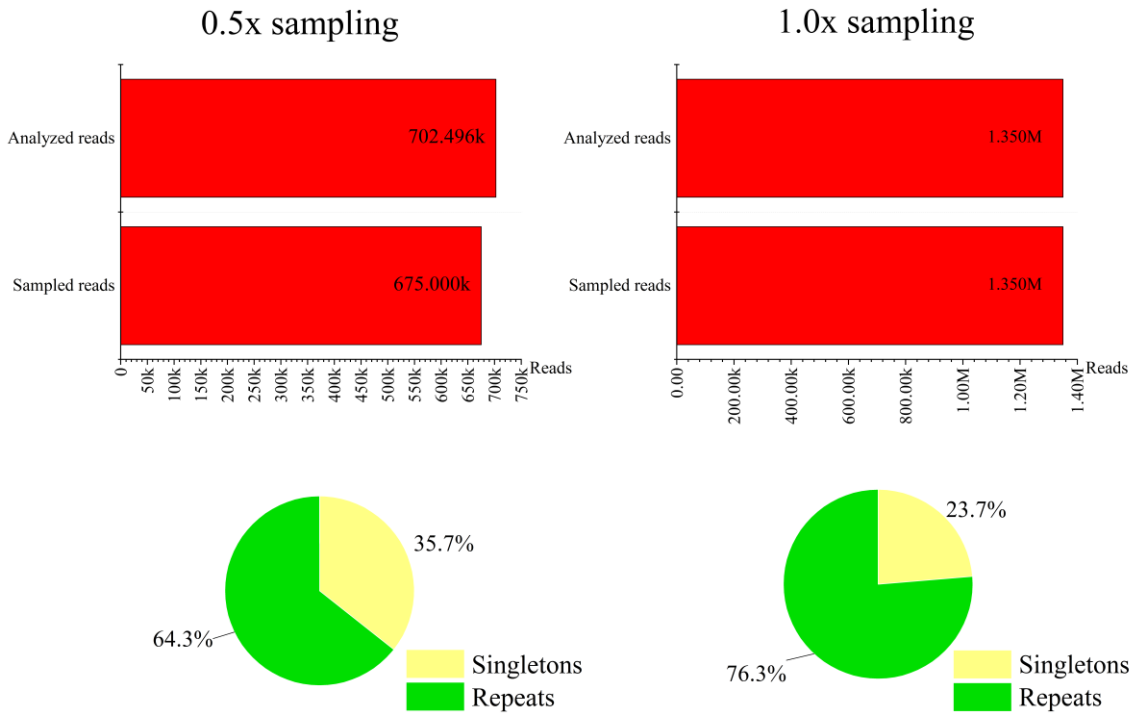


Figure 7 *Results of sampling of 1x and 0.5x of the Arabidopsis thaliana genome* showing the number of analyzed reads and the amount of single-copy and repetitive reads in percentage. The 0.5x genome sampling is shown on the left. The two charts on the right represent the 1x

genome sampling, where all sampled reads were analyzed ($N_{sampled\ reads} = N_{analyzed\ reads}$). A difference of 12% can be observed between the number of repeats in the 1x sampling and 0.5x sampling method.

Reads were trimmed to less than 100 nucleotides in *P. patens* (80), *A. trichopoda* (72), and *C. reinhardtii* (33) due to poor read sequencing quality. Due to the bad quality of the available WGS paired-end reads, single-end reads were used to analyze the repeats in *K. nitens* genome.

4.2 Diversity of repeat content in genomes of the green lineage

Following the pre-processing, I next analyzed the repeat content in the 12 plant genomes (Figure 8) and classified the repetitive elements based on super-cluster annotation (Figures 8 and 9). Following the annotation of repetitive elements, a summary of repeat proportions was retrieved from the RE2 Galaxy platform, indicating the percentage of repeat classes for each genome (represented in chapters 4.2, 4.3 and 4.4).

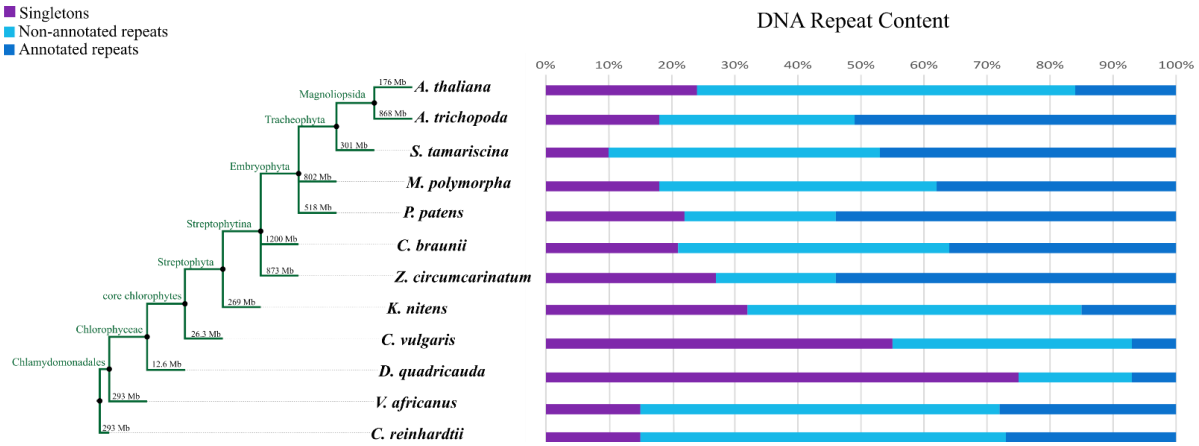


Figure 8 *DNA repeat content*: bar chart shows the repetitive sequence content in the 12 studied organisms. The proportions in percentage out of the entire genome for non-repetitive sequences (singletons), non-annotated repeats, and annotated repeats are depicted on the right side. A cladogram (on the left side) depicts the evolutionary relationship between each organism, and the genome size of each organism is indicated in Mbp.

To assist result interpretation, both repetitive and non-repetitive genome elements were summarized in a series of pie charts (Figures 9 and 10). The first step was obtaining a general overview of all genome components analyzed in RE2. Repetitive elements (including non-annotated repeats) and singletons are depicted as genome percentages in Figure 9. In order to quantify the contribution of each repeat class to the total amount of repeats in each genome,

the repetitive elements in the 12 species are depicted in Figure 10. Intriguingly, satellite repeats were not detected in *K. nitens*, and no transposable elements were detected in *S. tamariscina*.

The repeat content ranged between 25 – 90 % of the genomes (Figures 9 and 10), with the highest proportion in *S. tamariscina* and lowest in *D. quadricauda*. The annotated repeats constituted approximately 44% of the total repeat content. Among the annotated repeats, mobile elements dominated in all species, but *S. tamariscina*, *A. thaliana* and *C. vulgaris* (Figure 10).

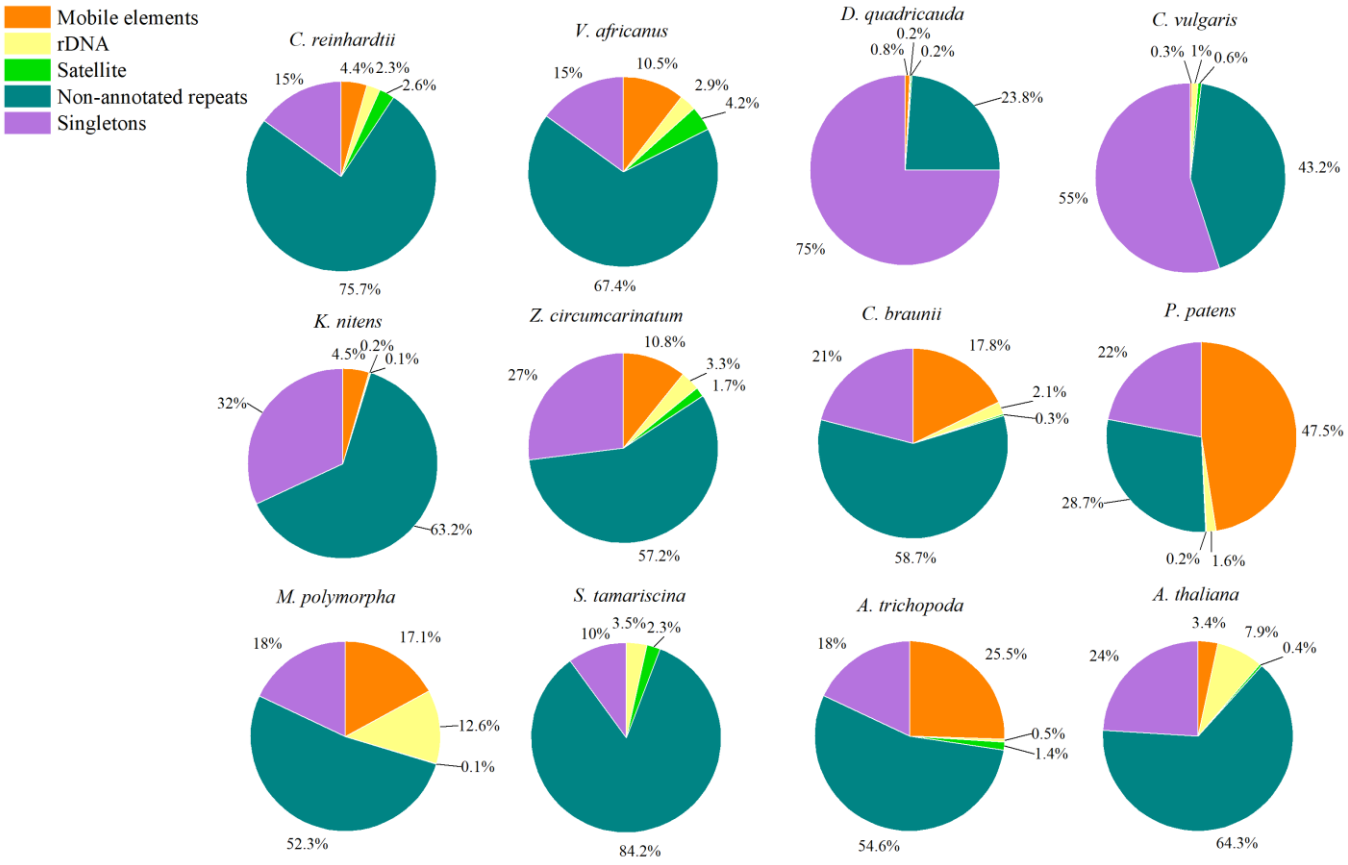


Figure 9 *Proportion of repetitive and singleton reads in each genome.* The percentages of singletons, annotated repetitive elements (mobile elements, rDNA, satellite) and non-annotated repeats are depicted.

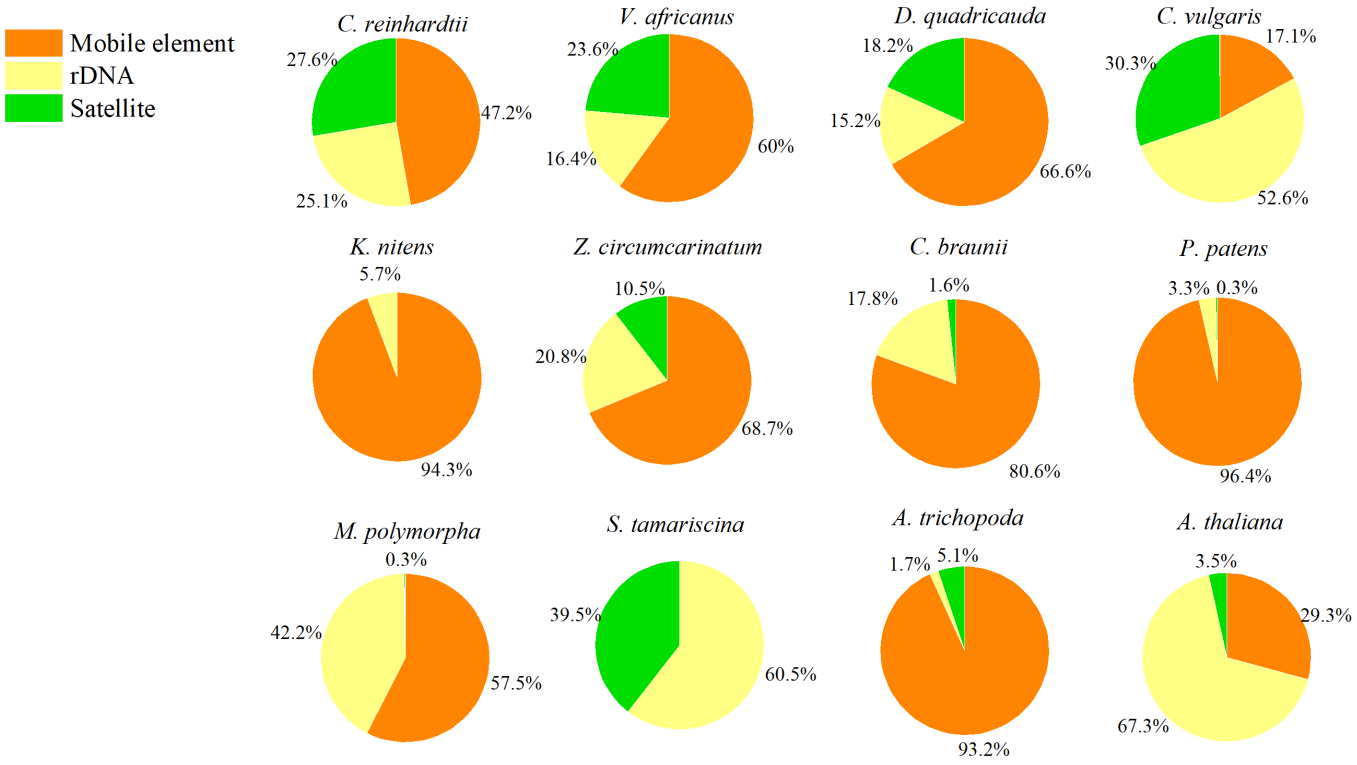


Figure 10 *Repetitive element families as percentage of total annotated repeats.*

Noteworthy is the fact that some elements were annotated to some extent as certain repetitive sequences classes, but not precisely classified further ("not completely classified repeats") (Figure 11).

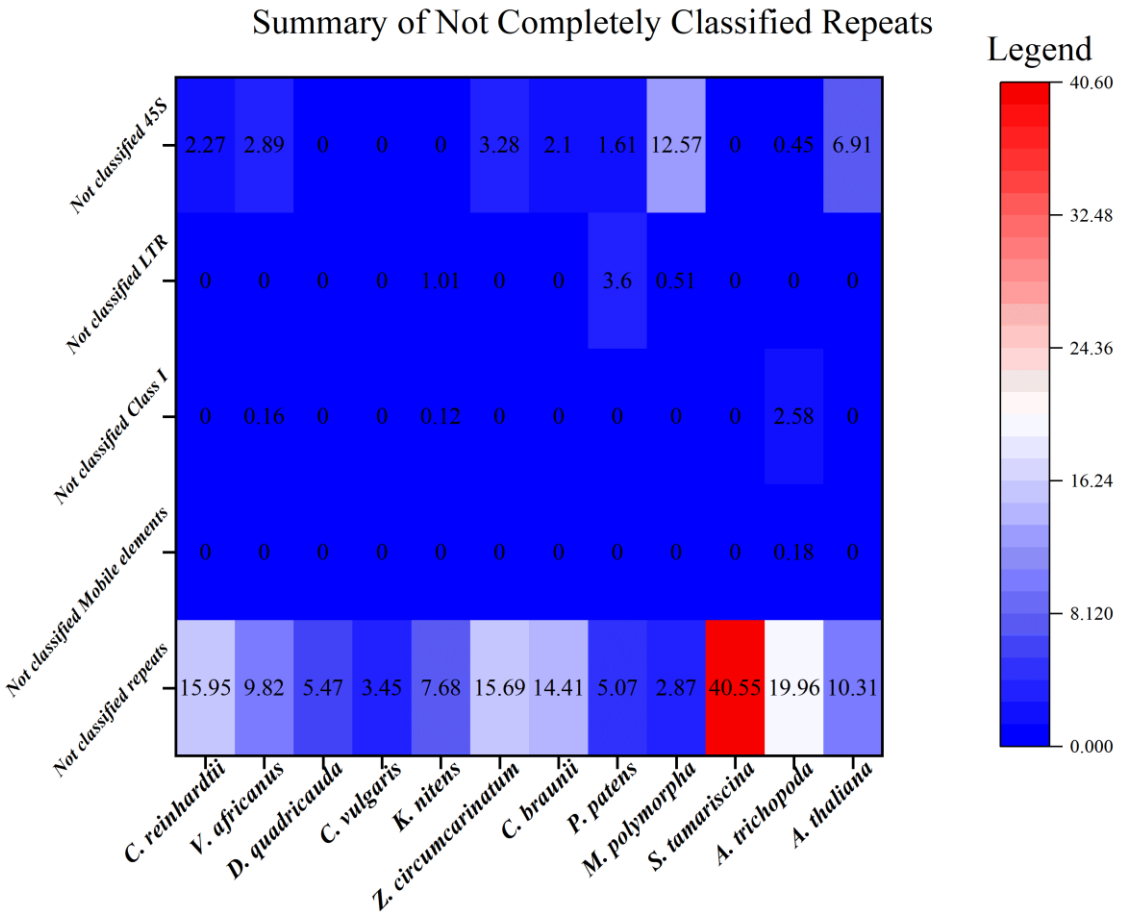


Figure 11 *Summary of not completely classified repeats*: Each number represents the percentage of repetitive elements not assigned to a more specific class of genomic repeats.

I next addressed the question of the contribution of repeat content to genome size (Figure 12). The lowest proportion of total (annotated and non-annotated) repetitive sequences corresponds to the smallest genome size. Repeat percentages show a steep increase positively correlated with genome size, before reaching a plateau of repeat proportion of around 75 – 90 % of the genome, which is observed in genomes of sizes 293 – 1200 Mbp.

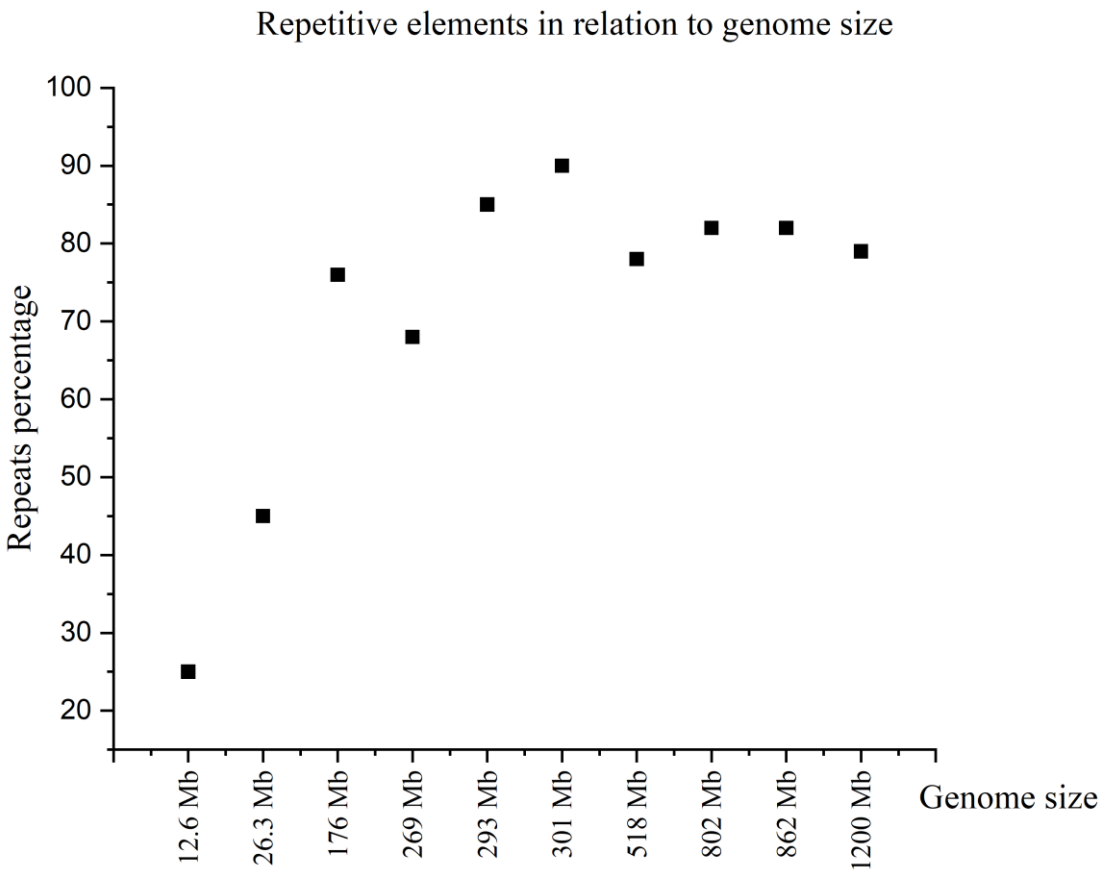


Figure 12 *Repetitive elements in relation to genome size*: Genome sizes are showed on the x-axis in numerical order, not in the order of evolution of the species that carry these genomes.

4.3 Mobile elements proportions

I next addressed the relative proportion of repeats within different annotated categories in each species. Figures 13 - 19 show the separate distribution of members in each class of annotated repetitive elements over the 12 organisms. Figures 13 - 16 provide a graphical comparison of the amount of different mobile elements detected in all genomes, Figures 17 and 18 portray all rDNA repeats in the organisms, and Figure 19 shows the distribution of satellite repeats.

4.3.1 Class I elements dominate the mobile elements

The RE2 results show that the proportion of mobile elements of Class I dominates those of Class II in all species, constituting 85 (*A. trichopoda*) - 99% of all annotated mobile elements (Figure 13). Annotated mobile elements were detected in all organisms except *S. tamariscina*.

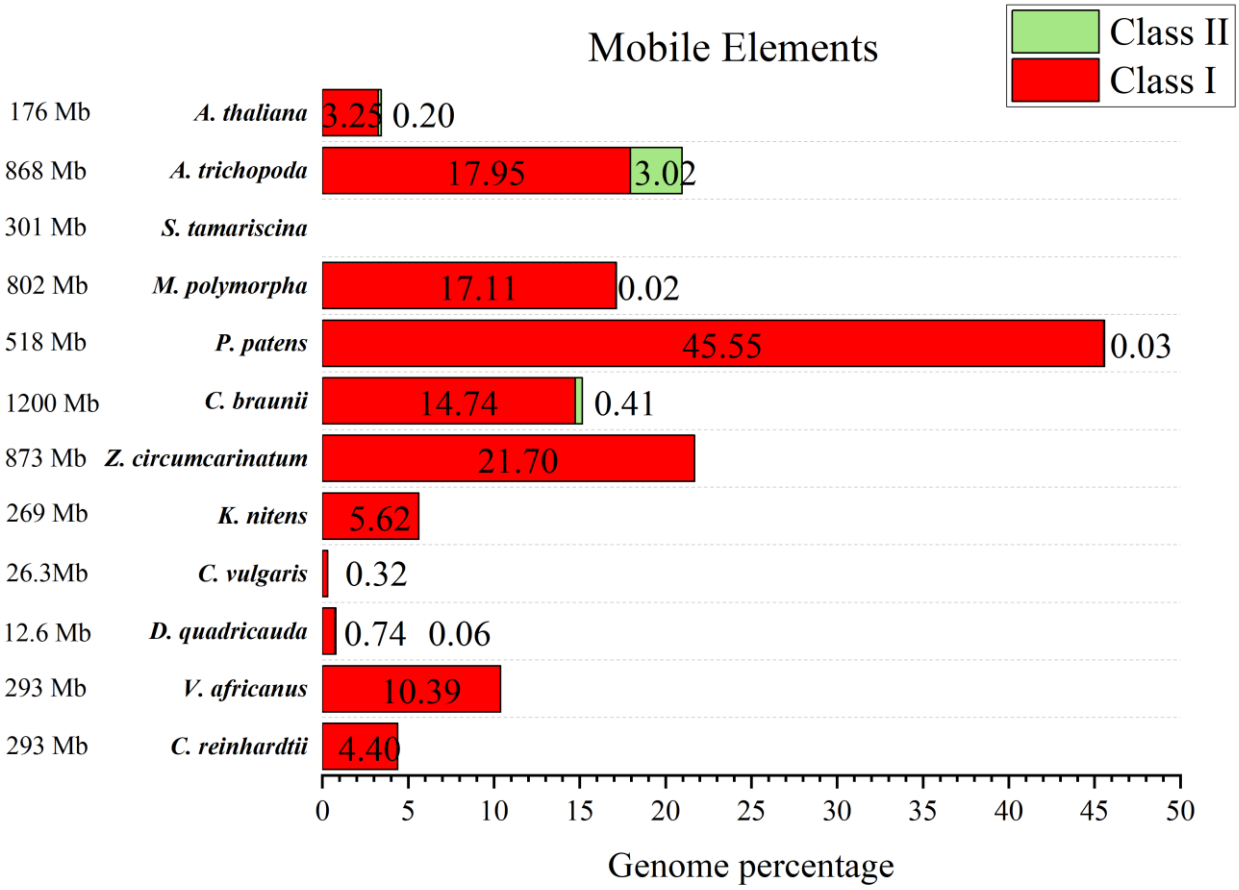


Figure 13 *Proportion of class I and class II mobile elements*: Numbers on the left side of the y-axis represent the genome size of each organism. The labels with the genome percentage of class II mobile elements (when present) are listed to the right of the class one mobile elements. Absence of bar and value indicates no elements were detected in the genome.

4.3.2 Non-LTR mobile elements

Annotated non-LTR elements were found in all but three organisms (*Z. circumcarinatum, P. patens* and *S. tamariscina*), which exhibited no annotated non-LTR elements within their genome (Figure 14a). LINE mobile elements dominated the non-LTR elements, while Penelope, Pararetrovirus and DIRS were each detected in one organism only and usually occurred with a lower abundance (Figure 14b).

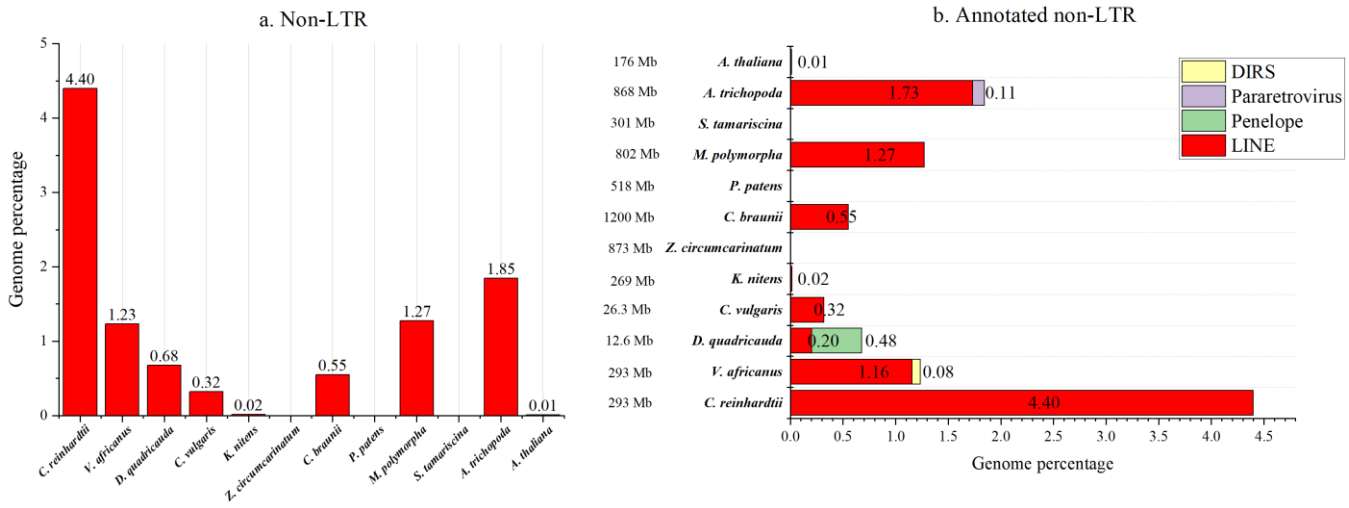


Figure 14 *Non-LTR **(a)** and annotated non-LTR elements **(b)***: The label on the y-axis represents the percentage of non-LTR elements out of the entire genomes. The 0.01 and 0.02 percentage values labeled for *A. thaliana* and *K. nitens* in Figure 14b both correspond to LINE mobile elements. Absence of bar and value indicates no elements were detected in the genome.

4.3.3 LTR mobile elements

The amount of LTR repeats ranges from 0.06% to 47.54% in the analyzed organisms. Ty1_copia was not detected in *D. quadricauda*, however, Ty3_gypsy retrotransposons were present. Also interesting is the absence of LTR elements in the organisms *C. reinhardtii*, *C. vulgaris*, and *S. tamariscina* (Figure 1a). Gypsy repetitive elements have a higher proportion in comparison to copia retrotransposons in each genome in which LTR retrotransposons were detected (Figure 15b).

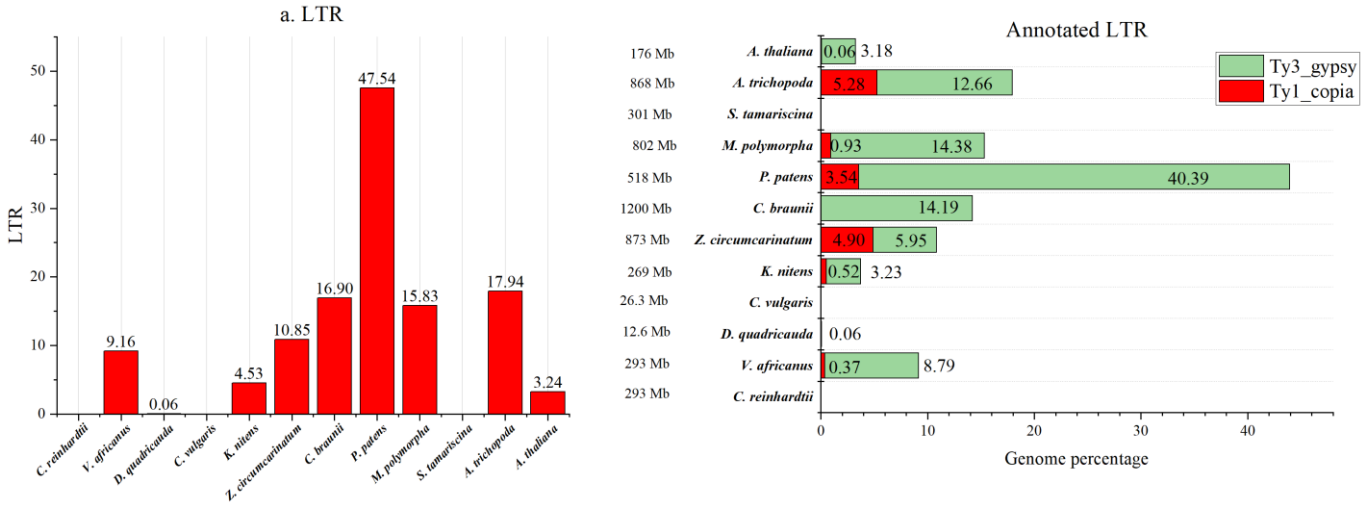


Figure 15 *Proportion of **(a)** all and **(b)*** annotated *LTR mobile elements in the studied genomes*.

The value of 0.06 in *D. quadricauda* corresponds to Ty3_gypsy type repeats. Absence of bar and value indicates no elements were detected in the genome.

4.3.4 Class II mobile elements

Class II transposons were detected in half of the organisms analyzed (Figure 16a). Class II repeats ranged from 0.02 % to 3.02 % in the 12 analyzed organisms. Helitron repeats were detected only in *C. braunii* (0.16 %) and *A. trichopoda* (0.01 %), constituting 39 % and 0.3 % of class II repeats, respectively (Figure 16b).

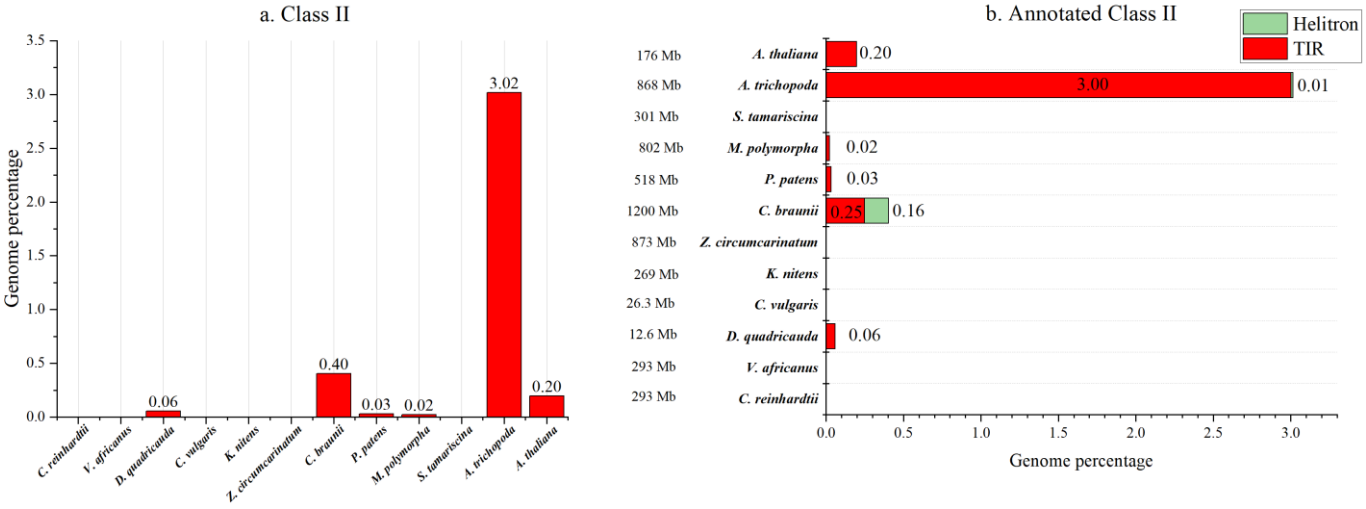


Figure 16 *Class II* and *Annotated Class II* show the total class II repetitive elements percentage out of the entire genomes **(a)** and the number of annotated TIR and Helitron class II repeats given as percentage for each genome **(b)**.

4.4 rDNA and satellite repeats

Even though not further annotated into distinct subclasses, rDNA was detected in all analyzed organisms (Figure 17a). The analysis results show that 45S rDNA is more abundant than 5S rDNA (Figure 17b). The highest percentage of 45S rDNA was detected in the genome of *Marchantia polymorpha* (12.57 %). 5S rDNA was detected in half of the organisms. The only organism genome which contained 5.8S rDNA (*D. quadricauda*) is also the only one where all types of rDNA repeats were annotated (Figure 18). 25S rDNA occupied the largest genome percentage out of all 45S rDNA repeats.

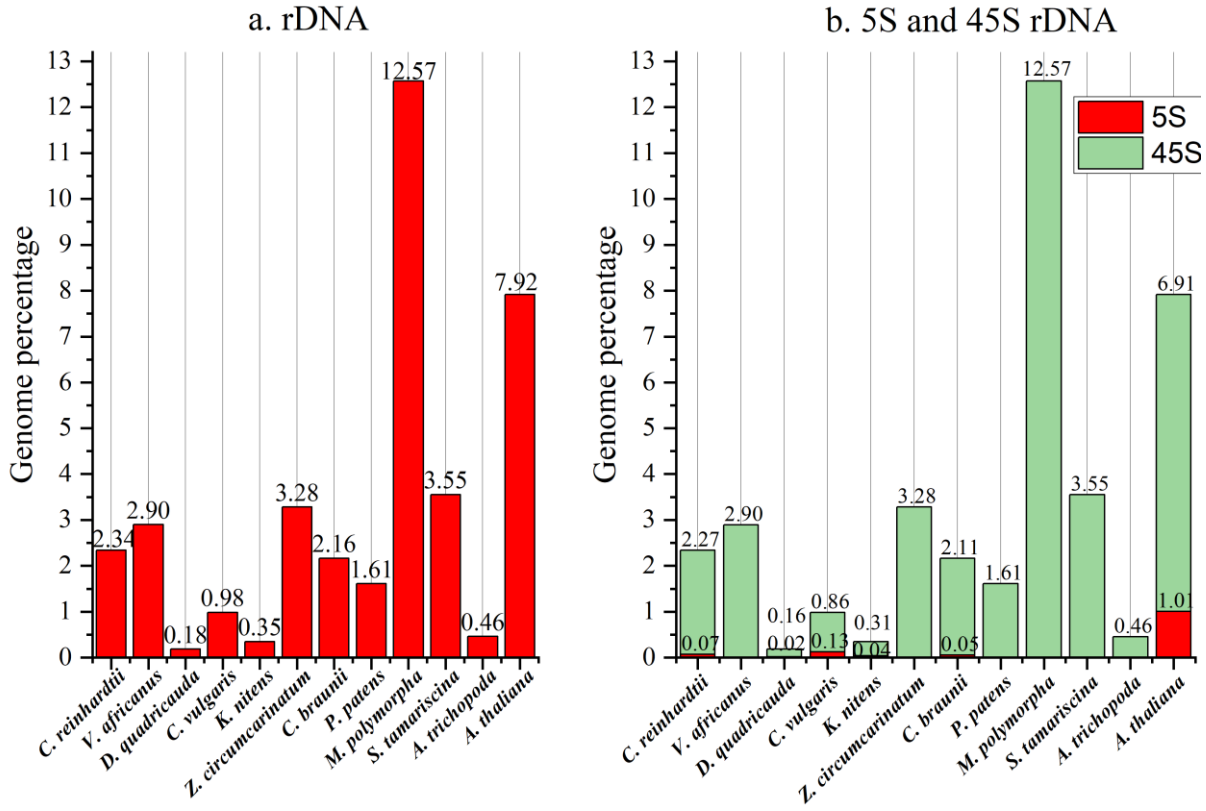


Figure 17 *rDNA* and *5S and 45S rDNA*: **(a) Bar chart** shows all repetitive elements assigned to the rDNA category. **(b)** shows the rDNA percentages specifically annotated as 5S or 45S, including the ones which could be further assigned to the 25S, 18S, or 5.8S subcategories of 45S rDNA.

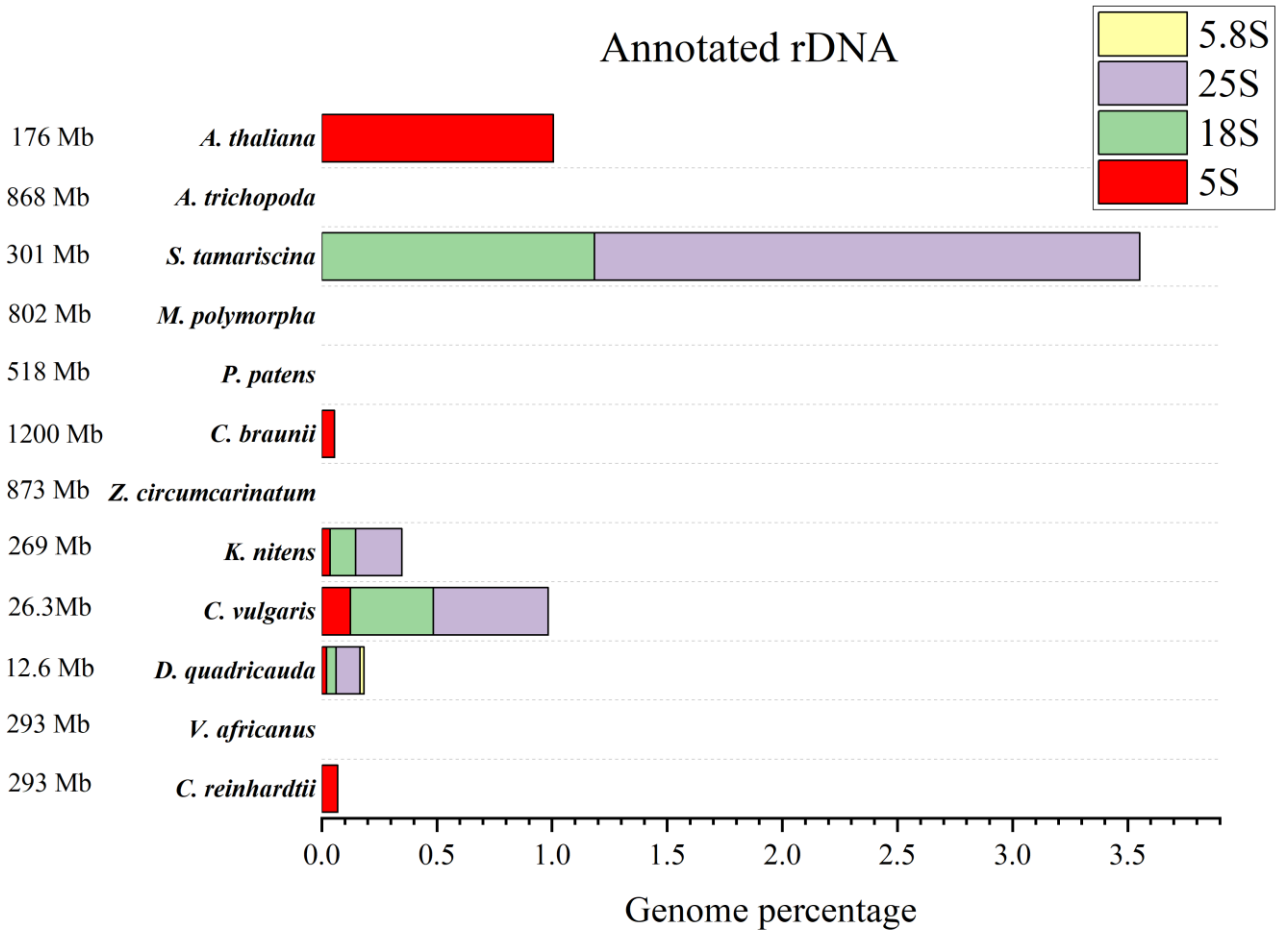


Figure 18 *Annotated rDNA* displays the genome percentages of repetitive elements distinguished as 5.8S, 25S, 18S and 5S rDNA.

When compared with Figure 17-b, Figure 18 shows that only a small proportion of 45S rDNA repeats are classified into the 25S, 5.8S and 18S subtypes. Despite the fact that 45S rDNA is present in each organism, only 4.2 - 66.47% of it was annotated as either 5.8S, 25S, or 18S.

Satellite repeats were identified in all but one of the 12 organisms, the exception being the genome of *K. nitens* (Figure 19). The organism with highest abundance of satellite DNA identified in *V. africanus*, with satellite sequences representing 4.15% of its genome.

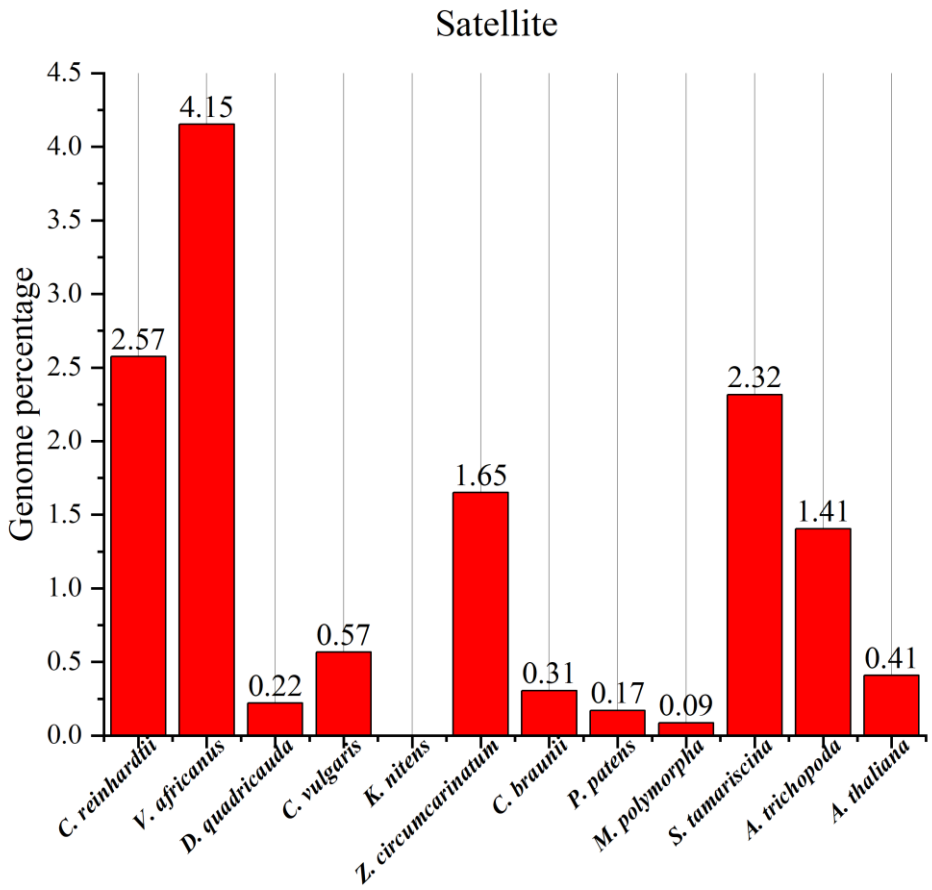


Figure 19 *Proportion of satellite repeats in the studied genomes.* Organism names are listed on the x-axis.

5. Discussion

Repetitive elements are a major component in plant genomes. However, not much is known about their distribution and functions across the 450000 plant species. The aim of this study was identifying and characterizing repetitive elements present in 12 plant genomes using Repeat Explorer 2. The Repeat Explorer 2 tool is a computational pipeline integrated in the Galaxy platform.

Mobile elements were detected in 11 out of 12 organisms, class I transposons dominating class II transposons in every case. These results confirm the standing hypothesis in the field, that class I mobile elements are more abundant in plant genomes than class II mobile elements (Dooner & Weil, 2007). A hypothesis why this is so could be the euchromatic organization of class II mobile elements (Dooner & Weil, 2007). The association of class II mobile elements with genic regions may impede their detection.

The software detected rDNA repetitive elements in all genomes, and non-LTR mobile elements in 9 out of 12 organisms. These findings confirm previous hypothesis that non-LTR transposable elements are almost ubiquitous in plants, especially in closely related species (Noma et al., 1999). Some class I mobile elements appear to be species specific. DIRS, Pararetrovirus and Penelope were each only detected in one organism out of the 12 we analyzed. Also, Helitron repeat families appear to be species specific, confirming the previously postulated theory that class II mobile elements are accountable for gene modifications by transposing into euchromatin regions and therefore, plant allele divergence (Han et al., 2013).

5.1 Amount of repeats in relation to genome size and phylogenetic position

We observe that *A. trichopoda* and *M. polymorpha* with similar genome size (802 and 868 Mb respectively) have a comparable proportion of repeats, but they belong to different evolutionary groups (Figure 8). *C. reinhardtii* and *V. africanus* belong to the same class (Chlorophyceae) and show highly similar genome sizes (293 Mb) and amounts of repetitive elements (87%) in their genome (Figure 8). However, the proportion of clustered repeats slightly differs between them (Figure 8). A similarity of repetitive elements percentage can also be observed between *C. braunii*, *P. patens* and *M. polymorpha*, which succeed each other in the cladogram (Figure 8). *D. quadricauda* has the smallest genome size and at the same time the least number of repeats out of all organisms, followed by *C. vulgaris* which has the second smallest genome and the second least number of repeats (Figure 8). The highest number of repeats (90%) was detected in *S. tamariscina*, an organism with a medium size genome (301 Mb) of the analyzed species (Figure 8). Despite this, however, the repeats percentage plateaued in the range of 67%-83% as genome size increased from 293 Mb to 1200 Mb (Figure 12).

Overall, the number of repetitive elements increased as the genome size became larger in the organisms we analyzed. This result is in line with standing hypotheses that major drivers of genome size expansion in plants are whole-genome duplication and repeat expansion (Macas et al., 2015; Wendel et al., 2016). A previously conducted study concluded that repeat content has a positive correlation to genome size in plants (Michael, 2014). Furthermore, Novák et al. determined that up to the genome size of 5 Gbp, repetitive elements have a sharp increasing trend in the genomes of seed plants (Novák, Guignard, et al., 2020). Still, the reason why the repetitive element percentage reaches a plateau of 80%-85% remains questionable. One

possibility may be the requirement of certain coding DNA percentage in the plant genomes. The highest amount of repeats in all plant genomes is 90%, suggesting that plants may need at least 10% coding DNA in order to function (Mehrotra & Goyal, 2014).

S. tamariscina has the highest amount of repeats (90%) despite the fact that no annotated mobile elements were found in its genome. This value is three times higher than the repeat content in the *S. moellendorffii* species, which has a considerably smaller genome size (105 Mbp) (Xu et al., 2018). As described by Xu et al., this may be caused by the expansion of repeats which produced desiccation tolerance in *S. tamariscina* (Xu et al., 2018).

5.2 Annotation of repeats

Despite the fact that the 45S rDNA cistron (in the standard tandem organization 18S-5.8S-25S) is well conserved in evolution (He et al., 2022), RE2 did not identify the individual subgenes precisely. Less than 15% of the 45S rDNA detected by RE2 was attributed to either 5.8S, 25S or 18S categories (Figure 18). However, the RE2 algorithm functions by selecting a "node" (either mobile element, rDNA or satellite), followed by classification attempts based on thresholds (Novák, Neumann, et al., 2020). When the threshold values are not satisfied, the further classification does not take place. As a result, 45S rDNA was identified first, and then could not be completely attributed to either 5.8S, 25S, or 18S rDNA. However, the developers recommended a manual classification for such cases.

Satellite repeats are not annotated in *K. nitens* despite being found in all other analyzed organisms (Figure 19). This may arise from difficulties of the software to detect highly variable tandem repeats in genomes not previously analyzed (Novák, Neumann, et al., 2020). It is possible, that satellite repeats in *K. nitens* carry a significant amount of sequence variation that prevents their annotation. Moreover, mobile elements were not detected in *S. tamariscina* (Figure 13). However, since only 33% of the repetitive elements were annotated further in this species, it may be that mobile elements are actually present, but not annotated as such.

5.3 Class II mobile elements in algae and land plants versus other organisms

DNA transposons were scarcely detected in the 12 organisms, ranging between 0.02% and a maximum value of 3% out of the entire genome (Figure 13). These values are significantly lower than the percentages occupied by class I transposons (0.32 % - 45.55 %) (Figure 13). This result may be explained by the fact that DNA transposons occupy highly variable proportions within plant genomes. Out of the total number of transposable elements, class II

mobile elements represent 87% in *Oryza sativa*, but 1% in *Zea mays*, both being monocots (Feschotte & Pritham, 2007; W. Li et al., 2004). Furthermore, class II mobile elements tend to occupy smaller percentages of mammalian, green algae and plants genomes than those of parasites. Around 10% of transposable elements in human are class II, 5% in mouse, but, interestingly, as much as 70% in *Aedes aegypti* (yellow fever mosquito) or 98% in *Trichomonas vaginalis* (Feschotte & Pritham, 2007).

6. Conclusion

Here, I have characterized the proportion of repetitive sequences in the genomes of 11 representative genera of the green lineage, ranging from green algae to land plants using the RepeatExplorer 2 software. Repeats were identified and classified as mobile elements, rDNA, or satellites, and further into subcategories. Our findings were in line with standing theories of genome expansion, concluding that plant genomes differ considerably in size and repeat content, larger genomes having the tendency to contain more repeats. A large proportion of repeats was not annotated, suggesting a high sequence variance and/or a low standing repetitive element annotation ability in the green lineage. The results we obtained provide data with considerable potential implications in the plant genetics field. It can be inferred that both genome size and divergence may play a role in the repetitiveness of plant genomes, with larger genomes accumulating repeat sequences. Furthermore, certain subcategories such as DIRS, Pararetrovirus, Penelope and Helitron may be species specific. However, further work is required to gain a more in-depth understanding of the dominant mechanisms influencing plant genomes. Apart from the increased sequencing efforts necessary, considerable optimization is required for repetitive elements annotation software. As a future outlook, annotation results may be improved and verified by laboratory experimental analysis (such as fluorescence in situ hybridization).

7. References

Abu-Ghannam, S., & Gupta, N. (2010). An Assessment of the Antioxidant and Antimicrobial Activity of Six Species of Edible Irish Seaweeds. *International Food Research Journal*, *17*, 205–220. https://doi.org/10.21427/D7HC92

Ågren, J. A., & Clark, A. G. (2018). Selfish genetic elements. *PLoS Genetics*, *14*(11). https://doi.org/10.1371/JOURNAL.PGEN.1007700

Armisén, D., Lecharny, A., & Aubourg, S. (2008). Unique genes in plants: specificities and conserved features throughout evolution. *BMC Evolutionary Biology*, *8*(1). https://doi.org/10.1186/1471-2148-8-280

Avery, O. T., Macleod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE

OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of Experimental Medicine*, *79*(2), 137–158. https://doi.org/10.1084/JEM.79.2.137

Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses 2014, Vol. 6, Pages 106-136*, *6*(1), 106–136. https://doi.org/10.3390/V6010106

Bergman, C. M., & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, *8*(6), 382–392. https://doi.org/10.1093/BIB/BBM048

Berthelier, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., & Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga Tisochrysis lutea. *BMC Genomics*, *19*(1), 1–14. https://doi.org/10.1186/S12864-018-4763-1/FIGURES/4

Bianchi, A., & Shore, D. (2008). How Telomerase Reaches Its End: Mechanism of Telomerase Regulation by the Telomeric Complex. *Molecular Cell*, *31*(2), 153–165. https://doi.org/10.1016/J.MOLCEL.2008.06.013

Blommaert, J., Riss, S., Hecox-Lea, B., Mark Welch, D. B., & Stelzer, C. P. (2019). Small, but surprisingly repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome size in a metazoan species complex. *BMC Genomics*, *20*(1). https://doi.org/10.1186/S12864-019-5859-Y

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology 2018 19:1*, *19*(1), 1–12. https://doi.org/10.1186/S13059-018-1577-Z

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V, Ramsey, J. M., … Schloss, J. A. (2008). *The potential and challenges of nanopore sequencing*. https://doi.org/10.1038/nbt.1495

Broutin, J. (2019). *EMBRYOPHYTA*. https://www.eolss.net/Sample-Chapters/C03/E6-71-90-06.pdf

Cardona-Correa, C., Piotrowski, M. J., Knack, J. J., Kodner, R. E., Geary, D. H., & Graham, L. E. (2016). Peat moss-like vegetative remains from ordovician carbonates. *International Journal of Plant Sciences*, *177*(6), 523–538. https://doi.org/10.1086/686242/ASSET/IMAGES/LARGE/FG9.JPEG

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., & Fernández-Pozo, N. (2012). Why Assembling Plant Genome Sequences Is So Challenging. *Biology 2012, Vol. 1, Pages 439-459*, *1*(2), 439–459. https://doi.org/10.3390/BIOLOGY1020439

Colonna Romano, N., & Fanti, L. (2022). Transposable Elements: Major Players in Shaping Genomic and Evolutionary Patterns. *Cells 2022, Vol. 11, Page 1048*, *11*(6), 1048. https://doi.org/10.3390/CELLS11061048

Cowan, M. M. (1999). Plant Products as Antimicrobial Agents. *Clinical Microbiology Reviews*, *12*(4), 564. https://doi.org/10.1128/cmr.12.4.564

Crick, F. (1958). *Symp. Soc. Exp. Biol., The Biological Replication of Macromolecules*.

de Vries, J., & Archibald, J. M. (2018). Plant evolution: landmarks on the path to terrestrial life. *New Phytologist*, *217*(4), 1428–1434. https://doi.org/10.1111/NPH.14975

DeWitt, J. (2020). *Beginner's guide to next generation sequencing | IDT*. https://eu.idtdna.com/pages/education/decoded/article/what-is-next-generation-sequencing

Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Letter to the editor. *Cytometry Part A*, *51A*(2), 127–128. https://doi.org/10.1002/CYTO.A.10013

Dooner, H. K., & Weil, C. F. (2007). Give-and-take: interactions between DNA transposons and their host plant genomes. *Current Opinion in Genetics & Development*, *17*(6), 486–492. https://doi.org/10.1016/J.GDE.2007.08.010

Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R., & Hannon, G. J. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods 2008 5:8*, *5*(8), 679–682. https://doi.org/10.1038/nmeth.1230

FAO. (2009). *How to Feed the World in 2050*. https://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf

Favarato, R. M., Ribeiro, L. B., Feldberg, E., & Matoso, D. A. (2017). Chromosomal Mapping of Transposable Elements of the Rex Family in the Bristlenose Catfish, Ancistrus (Siluriformes, Loricariidae), from the Amazonian Region. *The Journal of Heredity*, *108*(3), 254–261. https://doi.org/10.1093/JHERED/ESW084

Feschotte, C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, *41*, 331. https://doi.org/10.1146/ANNUREV.GENET.40.110405.090448

Filip, E., & Skuza, L. (2021). Horizontal Gene Transfer Involving Chloroplasts. *International Journal of Molecular Sciences 2021, Vol. 22, Page 4484*, *22*(9), 4484. https://doi.org/10.3390/IJMS22094484

Flavell, R. B. (1986). Repetitive DNA and chromosome evolution in plants. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *312*(1154), 227–242. https://doi.org/10.1098/RSTB.1986.0004

Fleury, D., Baumann, U., & Langridge, P. (2012). Plant genome sequencing: Models for developing synteny maps and association mapping. *Plant Biotechnology and Agriculture*, 83–97. https://doi.org/10.1016/B978-0-12-381466-1.00006-7

Genohub. (2019). *Designing Next-Generation Sequencing Runs*. https://genohub.com/next-generation-sequencing-guide/

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., & Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany*, *105*(3), 291–301. https://doi.org/10.1002/AJB2.1048

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. https://doi.org/10.1111/J.1755-0998.2011.03024.X

Gumińska, N., Płecha, M., Walkiewicz, H., Hałakuc, P., Zakryś, B., & Milanowski, R. (2018). Culture purification and DNA extraction procedures suitable for next-generation sequencing of euglenids. *Journal of Applied Phycology*, *30*(6), 3541–3549. https://doi.org/10.1007/S10811-018-1496-0/FIGURES/4

Gupta, A. K., & Gupta, U. D. (2013). Next Generation Sequencing and Its Applications. *Animal Biotechnology: Models in Discovery and Translation*, 345–367. https://doi.org/10.1016/B978-0-12-416002-6.00019-5

Halagarda, M., & Popek, S. (2018). Consumer Response to Genetically Modified Foods. *Reference Module in Food Science*. https://doi.org/10.1016/B978-0-08-100596-5.21884-2

Han, Y., Qin, S., & Wessler, S. R. (2013). Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics 2013 14:1*, *14*(1), 1–10. https://doi.org/10.1186/1471-2164-14-71

Haq, S. H., Al-Ruwaished, G., Al-Mutlaq, M. A., Naji, S. A., Al-Mogren, M., Al-Rashed, S., Ain, Q. T., Al-Amro, A. A., & Al-Mussallam, A. (2019). Antioxidant, Anticancer Activity and Phytochemical Analysis of Green Algae, Chaetomorpha Collected from the Arabian Gulf. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-55309-1

Hartley, G., & O'neill, R. J. (2019). Centromere Repeats: Hidden Gems of the Genome. *Genes 2019, Vol. 10, Page 223*, *10*(3), 223. https://doi.org/10.3390/GENES10030223

He, J., Zhao, Y., Zhang, S., He, Y., Jiang, J., Chen, S., Fang, W., Guan, Z., Liao, Y., Wang, Z., Chen, F., & Wang, H. (2022). Uneven Levels of 5S and 45S rDNA Site Number and Loci Variations across Wild Chrysanthemum Accessions. *Genes*, *13*(5), 894. https://doi.org/10.3390/GENES13050894

Illumina. (2017). *An introduction to Next-Generation Sequencing Technology*. www.illumina.com/technology/next-generation-sequencing.html

Janicki, M., Rooke, R., Yang, G., Gregory, R., Bainard M Janicki, J. D., Rooke, R., Yang, G., & Janicki, M. (2011). Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Research 2011 19:6*, *19*(6), 787–808. https://doi.org/10.1007/S10577-011-9230-7

Jiang, N. (2013). Overview of repeat annotation and De Novo repeat identification. *Methods in Molecular Biology*, *1057*, 275–287. https://doi.org/10.1007/978-1-62703-568-2_20

Jiang, T., Liu, S., Cao, S., Liu, Y., Cui, Z., Wang, Y., & Guo, H. (2021). *Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation*. https://doi.org/10.1186/s12859-021-04422-y

Jiao, Y., & Guo, H. (2014). Prehistory of the Angiosperms: Characterization of the Ancient Genomes. *Advances in Botanical Research*, *69*, 223–245. https://doi.org/10.1016/B978-0-12-417163-3.00009-3

Jurka, J., Kapitonov, V. V., Kohany, O., & Jurka, M. V. (2007). Repetitive sequences in complex genomes: Structure and evolution. In *Annual Review of Genomics and Human Genetics* (Vol. 8, pp. 241–259). https://doi.org/10.1146/annurev.genom.8.080706.092416

Karami, A. (2013). *Largest and Smallest Genome in the World*. https://www.researchgate.net/publication/235907922_Largest_and_Smallest_Genome_in_the_World

Kersey, P. J. (2019). Plant genome sequences: past, present, future. *Current Opinion in Plant Biology*, *48*, 1–8. https://doi.org/10.1016/J.PBI.2018.11.001

Kim, J. K., Park, M., Kim, Y., Min, B., Hwang, E. K., Hwang, I. K., Park, E.-J., Il Yoo, H., Han, T., Kim, C.-H., & Yarish, C. (n.d.). *Macroalgae cultivation in Korea/Asia with emphasis on emerging technology trends*.

Kim, Y.-J., Lee, J., & Han, K. (2012). Transposable Elements: No More "Junk DNA." *Genomics & Informatics*, *10*(4), 226. https://doi.org/10.5808/GI.2012.10.4.226

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536. https://doi.org/10.1002/BIES.200900181

Klein, S. J., & O'Neill, R. J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research : An International Journal on the*

Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology, *26*(1–2), 5–23. https://doi.org/10.1007/S10577-017-9569-5

Knyazev, S., Hughes, L., Skums, P., & Zelikovsky, A. (2021). Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Briefings in Bioinformatics*, *22*(1), 96–108. https://doi.org/10.1093/BIB/BBAA101

Kress, W. J., Soltis, D. E., Kersey, P. J., Wegrzyn, J. L., Leebens-Mack, J. H., Gostel, M. R., Liu, X., & Soltis, P. S. (2022). Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), 2115640118. https://doi.org/10.1073/PNAS.2115640118/-/DCSUPPLEMENTAL

Kumar, S. (2021). Aspects of Genetically Modified Plants in Removing Heavy Metals From the Soil. *Policy Issues in Genetically Modified Crops*, 273–289. https://doi.org/10.1016/B978-0-12-820780-2.00012-1

Lambert, S. (2009). *Stoneworts: their habitats, ecological requirements and conservation Integrated catchment science programme Science report: SC030202*. www.environment-agency.gov.uk

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S. A., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Ullrich, K. K., Wickett, N. J., DeGironimo, L., … Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature 2019 574:7780*, *574*(7780), 679–685. https://doi.org/10.1038/s41586-019-1693-2

Leitch, I. J., & Leitch, A. R. (2013). Genome Size Diversity and Evolution in Land Plants. In *Plant Genome Diversity Volume 2* (pp. 307–322). Springer Vienna. https://doi.org/10.1007/978-3-7091-1160-4_19

Leliaert, F. (2019). Green algae: Chlorophyta and streptophyta. *Encyclopedia of Microbiology*, 457–468. https://doi.org/10.1016/B978-0-12-809633-8.20890-X

Lerat, E. (2009). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity 2010 104:6*, *104*(6), 520–533. https://doi.org/10.1038/hdy.2009.165

Lermontova, I., Sandmann, M., Mascher, M., Schmit, A. C., & Chabouté, M. E. (2015). Centromeric chromatin and its dynamics in plants. *The Plant Journal*, *83*(1), 4–17. https://doi.org/10.1111/TPJ.12875

Levy, S. E., & Boone, B. E. (2019). Next-Generation Sequencing Strategies. *Cold Spring Harbor Perspectives in Medicine*, *9*(7), a025791. https://doi.org/10.1101/CSHPERSPECT.A025791

Li, L., Wang, S., Wang, H., Sahu, S. K., Marin, B., Li, H., Xu, Y., Liang, H., Li, Z., Cheng, S., Reder, T., Çebi, Z., Wittek, S., Petersen, M., Melkonian, B., Du, H., Yang, H., Wang, J., Wong, G. K. S., … Liu, H. (2020). The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants. *Nature Ecology & Evolution 2020 4:9*, *4*(9), 1220–1231. https://doi.org/10.1038/s41559-020-1221-7

Li, W., Zhang, P., Fellers, J. P., Friebe, B., & Gill, B. S. (2004). Sequence composition, organization, and evolution of the core Triticeae genome. *The Plant Journal : For Cell and Molecular Biology*, *40*(4), 500–511. https://doi.org/10.1111/J.1365-313X.2004.02228.X

Libertín, M., Kvaček, J., Bek, J., Žárský, V., & Štorch, P. (2018). Sporophytes of polysporangiate land plants from the early Silurian period may have been photosynthetically autonomous. *Nature Plants 2018 4:5*, *4*(5), 269–271. https://doi.org/10.1038/s41477-018-0140-y

Lindsey, J., McGill, N. I., Lindsey, L. A., Green, D. K., & Cooke, H. J. (1991). In vivo loss of

telomeric repeats with age in humans. *Mutation Research DNAging*, *256*(1), 45–48. https://doi.org/10.1016/0921-8734(91)90032-7

Lisch, D. (2012). How important are transposons for plant evolution? *Nature Reviews Genetics 2013 14:1*, *14*(1), 49–61. https://doi.org/10.1038/nrg3374

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*. https://doi.org/10.1155/2012/251364

Liu, Y., Lan, C., Blumenstein, M., & Li, J. (2017). Bi-level error correction for PacBio long reads. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(3), 899–905. https://doi.org/10.1109/TCBB.2017.2780832

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597. https://doi.org/10.1038/S41576-020-0236-X

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology 2012 30:5*, *30*(5), 434–439. https://doi.org/10.1038/nbt.2198

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Correction: Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE*, *7*(3), e30087. https://doi.org/10.1371/ANNOTATION/64BA358F-A483-46C2-B224-EAA5B9A33939

Macas, J., Novak, P., Pellicer, J., Cizkova, J., Koblizkova, A., Neumann, P., Fukova, I., Dolezel, J., Kelly, L. J., & Leitch, I. J. (2015). In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. *PLOS ONE*, *10*(11), e0143424. https://doi.org/10.1371/JOURNAL.PONE.0143424

Mandal, A. (2020). *What is Genomics?* https://www.azolifesciences.com/article/What-is-Genomics.aspx

Manktelow, M. (2010). *History of Taxonomy*.

Marks, R. A., Hotaling, S., Frandsen, P. B., & VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nature Plants 2021 7:12*, *7*(12), 1571–1578. https://doi.org/10.1038/s41477-021-01031-8

Marshall, E. (2000). Claim and Counterclaim on the Human Genome. *Science*, *288*(5464), 242–243. https://doi.org/10.1126/SCIENCE.288.5464.242B

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., Petrov, D. A., & Fiston-Lavier, A. S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PloS One*, *9*(9). https://doi.org/10.1371/JOURNAL.PONE.0106689

Mehrotra, S., & Goyal, V. (2014). Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics & Bioinformatics*, *12*(4), 164–171. https://doi.org/10.1016/J.GPB.2014.07.003

Michael, T. P. (2014). Plant genome size variation: Bloating and purging DNA. *Briefings in Functional Genomics and Proteomics*, *13*(4), 308–317. https://doi.org/10.1093/BFGP/ELU005

Mitra, A., Skrzypczak, M., Ginalski, K., & Rowicka, M. (2015). Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform. *PLOS ONE*, *10*(4), e0120520. https://doi.org/10.1371/JOURNAL.PONE.0120520

Mohsen, A., Park, J., Chen, Y. A., Kawashima, H., & Mizuguchi, K. (2019). Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks. *BMC Bioinformatics*, *20*(1), 1–10. https://doi.org/10.1186/S12859-019-3187-5/FIGURES/6

Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*. https://doi.org/10.1016/j.ygeno.2008.07.001

Muñoz-López, M., & García-Pérez, J. L. (2010). DNA Transposons: Nature and Applications in Genomics. *Current Genomics*, *11*(2), 115. https://doi.org/10.2174/138920210790886871

Noma, K., Ohtsubo, E., & Ohtsubo, H. (1999). Non-LTR retrotransposons (LINEs) as ubiquitous components of plant genomes. *Molecular & General Genetics : MGG*, *261*(1), 71–79. https://doi.org/10.1007/S004380050943

Novák, P., Guignard, M. S., Neumann, P., Kelly, L. J., Mlinarec, J., Koblížková, A., Dodsworth, S., Kovařík, A., Pellicer, J., Wang, W., Macas, J., Leitch, I. J., & Leitch, A. R. (2020). Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature Plants*, *6*(11), 1325–1329. https://doi.org/10.1038/S41477-020-00785-X

Novák, P., Neumann, P., & Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols 2020 15:11*, *15*(11), 3745–3776. https://doi.org/10.1038/s41596-020-0400-y

Novák, P., Robledillo, L. Á., Koblížková, A., Vrbová, I., Neumann, P., & Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, *45*(12), e111–e111. https://doi.org/10.1093/NAR/GKX257

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/SCIENCE.ABJ6987/SUPPL_FILE/SCIENCE.ABJ6987_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Paço, A. (2014). *Repetitive DNA Sequences in Rodentia Genomes. Its involvement in chromosome architecture reshuffling and in genome functionality*. https://www.researchgate.net/publication/279974133_Repetitive_DNA_Sequences_in_Rodentia_Genomes_Its_involvement_in_chromosome_architecture_reshuffling_and_in_genome_functionality

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

Peska, V., & Garcia, S. (2020). Origin, Diversity, and Evolution of Telomere Sequences in Plants. *Frontiers in Plant Science*, *11*, 117. https://doi.org/10.3389/FPLS.2020.00117/BIBTEX

Piganeau, G. (2020). A planktonic picoeukaryote makes big changes to the green lineage. *Nature Ecology and Evolution*, *4*(9), 1160–1161. https://doi.org/10.1038/S41559-020-1244-0

Pisupati, R., Vergara, D., & Kane, N. C. (2018). Diversity and evolution of the repetitive genomic content in Cannabis sativa. *BMC Genomics*, *19*(1), 156. https://doi.org/10.1186/s12864-018-4494-3

Podgorny, & Irina. (2019). Bureaucracy, Instructions, and Paperwork – The Gathering of Data about the Three Kingdoms of Nature in the Americas, 1770-1815. *Http://Journals.Openedition.Org/Nuevomundo*.

https://doi.org/10.4000/NUEVOMUNDO.75454

Rao, S. R., Trevedi, S., Emmanuel, D., Merita, K., & Hynniewta, M. (2010). DNA repetitive sequences-types, distribution and function: A review. *Journal of Cell and Molecular Biology*, *7*(2), 1–13. http://jcmb.halic.edu.tr

Reis-Filho, J. S. (2009). Next-generation sequencing. *Breast Cancer Research*, *11*(SUPPL. 3), 1–7. https://doi.org/10.1186/BCR2431/TABLES/1

Rensing, S. A. (2020). How Plants Conquered Land. *Cell*, *181*(5), 964–966. https://doi.org/10.1016/J.CELL.2020.05.011

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/J.MOLCEL.2015.05.004

Rosato, M., Kovařík, A., Garilleti, R., & Rosselló, J. A. (2016). Conserved Organisation of 45S rDNA Sites and rDNA Gene Copy Number among Major Clades of Early Land Plants. *PLOS ONE*, *11*(9), e0162544. https://doi.org/10.1371/JOURNAL.PONE.0162544

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., Cavalier-Smith, T., Guiry, M. D., & Kirk, P. M. (2015). A Higher Level Classification of All Living Organisms. *PLOS ONE*, *10*(4), e0119248. https://doi.org/10.1371/JOURNAL.PONE.0119248

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., … Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, *326*(5956), 1112–1115. https://doi.org/10.1126/SCIENCE.1178534

Schreiber, M., Rensing, S. A., & Gould, S. B. (2022). The greening ashore. *Trends in Plant Science*. https://doi.org/10.1016/J.TPLANTS.2022.05.005

Shah, S. A. A., Ul Hassan, S. S., Bungau, S., Si, Y., Xu, H., Rahman, M. H., Behl, T., Gitea, D., Pavel, F. M., Aron, R. A. C., Pasca, B., & Nemeth, S. (2020). Chemically diverse and biologically active secondary metabolites from marine phylum chlorophyta. *Marine Drugs*, *18*(10). https://doi.org/10.3390/MD18100493

Shampo, M. A., & Kyle, R. A. (2011). J. Craig Venter—The Human Genome Project. *Mayo Clinic Proceedings*, *86*(4), e26. https://doi.org/10.4065/MCP.2011.0160

Shendure, J. A., Porreca, G. J., & Church, G. M. (2008). Overview of DNA Sequencing Strategies. *Current Protocols in Molecular Biology*, *81*(1), 7.1.1-7.1.11. https://doi.org/10.1002/0471142727.MB0701S81

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/NBT1486

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1). https://doi.org/10.1002/CPMB.59

Soltis, D. E., & Soltis, P. S. (1999). Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution*, *14*(9), 348–352. https://doi.org/10.1016/S0169-5347(99)01638-9

State, P. (2017). *The history of Next Generation Sequencing (NGS). | Stem Cell and Regenerative Biology Program*. https://sites.psu.edu/stemcellhershey/2017/03/07/the-history-of-next-generation-sequencing-ngs/

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *Chapter 4*(SUPPL. 25). https://doi.org/10.1002/0471250953.BI0410S25

Turmel, M., & Lemieux, C. (2018). Evolution of the Plastid Genome in Green Algae. *Advances in Botanical Research*, *85*, 157–193. https://doi.org/10.1016/BS.ABR.2017.11.010

United Nations, DESA, P. D. (2022). *World Population Prospects - Population Division - United Nations*. https://population.un.org/wpp/Graphs/DemographicProfiles/Line/900

Valencia, C. A., Pervaiz, M. A., Husami, A., Qian, Y., & Zhang, K. (2013). *Sanger Sequencing Principles, History, and Landmarks*. 3–11. https://doi.org/10.1007/978-1-4614-9032-6_1

Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, *322*(1), 12–20. https://doi.org/10.1016/J.YEXCR.2014.01.008

Vu, G. T. H., Schmutzer, T., Bull, F., Cao, H. X., Fuchs, J., Tran, T. D., Jovtchev, G., Pistrick, K., Stein, N., Pecinka, A., Neumann, P., Novak, P., Macas, J., Dear, P. H., Blattner, F. R., Scholz, U., & Schubert, I. (2015). Comparative Genome Analysis Reveals Divergent Genome Size Evolution in a Carnivorous Plant Genus. *The Plant Genome*, *8*(3). https://doi.org/10.3835/PLANTGENOME2015.04.0021

Warmerdam, D. O., & Wolthuis, R. M. F. (2018). Keeping ribosomal DNA intact: a repeating challenge. *Chromosome Research 2018 27:1*, *27*(1), 57–72. https://doi.org/10.1007/S10577-018-9594-Z

Waterston, R. H., Lander, E. S., & Sulston, J. E. (2002). On the sequencing of the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(6), 3712–3716. https://doi.org/10.1073/PNAS.042692499/ASSET/A63D7976-5D32-47EB-851F-6C66AA8C2FDA/ASSETS/GRAPHIC/PQ0426924002.JPEG

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 1953 171:4356*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Weilguny, L., & Kofler, R. (2019). DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Molecular Ecology Resources*, *19*(5), 1346–1354. https://doi.org/10.1111/1755-0998.13030

Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biology 2016 17:1*, *17*(1), 1–14. https://doi.org/10.1186/S13059-016-0908-1

Wiart, C. (2006). Anti-Inflammatory Plants. In *Ethnopharmacology of Medicinal Plants* (pp. 1–55). Humana Press. https://doi.org/10.1007/978-1-59745-160-4_1

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, *8*(12), 973–982. https://doi.org/10.1038/NRG2165

Wu, J., & Brown, M. (2018). Epigenetics and Epigenomics. *Hematology: Basic Principles and Practice*, 17–24. https://doi.org/10.1016/B978-0-323-35762-3.00002-0

Xu, Z., Xin, T., Bartels, D., Li, Y., Gu, W., Yao, H., Liu, S., Yu, H., Pu, X., Zhou, J., Xu, J., Xi, C., Lei, H., Song, J., & Chen, S. (2018). Genome Analysis of the Ancient Tracheophyte Selaginella tamariscina Reveals Evolutionary Features Relevant to the Acquisition of Desiccation Tolerance. *Molecular Plant*, *11*(7), 983–994. https://doi.org/10.1016/J.MOLP.2018.05.003

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/NRG3174

Zdobnov, E. M., Campillos, M., Harrington, E. D., Torrents, D., & Bork, P. (2005). Protein coding

potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Research*, *33*(3), 946–954. https://doi.org/10.1093/NAR/GKI236

Zhang, Z., Ma, X., Liu, Y., Yang, L., Shi, X., Wang, H., Diao, R., Zhong, B., & Zhang Bojian Zhong, Z. (2022). Origin and evolution of green plants in the light of key evolutionary events. *Journal of Integrative Plant Biology*, *64*(2), 516–535. https://doi.org/10.1111/JIPB.13224

8. Appendix

Genomes on a Tree (GoaT) - accessible at https://goat.genomehubs.org/

FigShare link to mitochondrial DNA of *Marchantia inflexa*:

https://figshare.com/articles/dataset/Marchantia_inflexa_mitochondrion_and_chloroplast_genome/6639209/1

Script used in the data preprocessing:

```
#!/bin/bash

file_n=$(echo $2|sed "s/_1P.fastq.gz//g")

minimap2 -ax sr -t 30 $1 $2 $3 |samtools sort -@ 30 -o $file_n.bam -

samtools view -@ 30 -u -f 4 -F 264 $file_n.bam > tmps1.bam

samtools view -@ 30 -u -f 8 -F 260 $file_n.bam > tmps2.bam

samtools view -@ 30 -u -f 12 -F 256 $file_n.bam > tmps3.bam

samtools merge -@ 30 -u - tmps[123].bam | samtools sort -n - -o unmapped.bam

bamToFastq -i unmapped.bam -fq $file_n.1.fq -fq2 $file_n.2.fq

rm *.bam

gzip *.fq

echo 'DONE'
```