Utilization of complete chloroplast genomes for phylogenetic studies

Shairul Izan Binti Ramlee

Thesis committee

Promotor

Prof. Dr R.G.F. Visser

Professor of Plant Breeding

Wageningen University

Co-promotors

Dr M.J.M. Smulders

Senior researcher, Wageningen UR Plant Breeding

Wageningen University & Research

Dr T.J.A. Borm

Researcher, Wageningen UR Plant Breeding

Wageningen University & Research

Other members

Prof. Dr M.E. Schranz, Wageningen University

Dr G.F. Sanchez Perez, Wageningen University

Dr R. Vos, Naturalis Biodiversity Center, Leiden

Dr R. van Velzen, Wageningen University

This research was conducted under the auspices of the Graduate School of Production Ecology and Resource Conservation

Utilization of complete chloroplast genomes for phylogenetic studies

Shairul Izan Binti Ramlee

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 28 October 2016
at 11 a.m. in the Aula.

Shairul Izan Binti Ramlee
Utilization of complete chloroplast genomes for phylogenetic studies
186 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English

ISBN: 978-94-6257-935-4
DOI: 10.18174/390196

Table of Contents

Chapter 1

General Introduction

Phylogenomics

Modern evolutionary theory hypothesizes that all organisms have descended from a common ancestor, which means that all extant and extinct species are related. Phylogenetic relationships can be inferred using morphological, physiological and molecular characteristics. Molecular sequences such as DNA sequences play a key role in recent day molecular phylogenetic analysis. The structure and function of the DNA sequences and how they change over time are used to infer evolutionary relationships. As new DNA sequencing methods became available since 2000, the costs have been driven down (http://www.genome.gov/sequencingcosts/). As a result, large amounts of sequence data can now be generated cheaply for researchers to infer species relationships from. Rather than using one or a few genes to study species evolutionary relationship in the conventional approach, one can now study evolutionary relationships based on comparative analysis of genome-scale data called phylogenomics (Eisen and Fraser 2003; Lemmon and Lemmon 2013).

Phylogenetic relationships can be reconstructed based on comparisons of DNA sequences and genome organisation features. In addition, rare genomic changes such as insertions and deletions in introns, retrotransposon integration, changes in gene order in the organellar genome, gene duplications and genetic code changes of the entire genome can be used as molecular markers for a wide range of taxonomical levels (species, genus, family or higher) (Rokas and Holland 2000). Phylogenetic trees reconstructed based on the conventional approach of using just one or a few genes may show conflicts (Teichmann and Mitchison 1999) due to the fact that individual genes may have gone through different evolutionary lineages. In addition, lack of sufficient phylogenetic informative variation leads to the risk of stochastic errors and poorly resolved phylogenetic trees. In contrast, phylogenomics should be able to resolve difficult phylogenies and be able to verify or overturn proposed relationships (Delsuc et al. 2005). Several empirical studies have shown the robustness of phylogenomics to resolve difficult phylogenies. For example, phylogenomic analysis using plastid sequences was able to produce strongly supported phylogenies of *Araceae* as discussed in Henriquez et al. (2014). Likewise,

in the study of Ma et al. (2014) and Pyron et al. (2014) a series of phylogenomic analyses was conducted to infer difficult phylogenies at low taxonomic levels in the temperate woody bamboo and snakes respectively.

Chloroplast phylogenomics

Beside the nuclear genome, plant cells contain up to two more genomes: The organellar genomes of the mitochondrion and the chloroplast (the plastome). Unlike the mitochondrial genome, chloroplast genomes or plastid genomes as referred to by some authors rarely show evidence of intra- or inter molecular recombination (Dong et al. 2012) and are therefore highly conserved in terms of gene order and content. These characteristics make the chloroplast genome an attractive tool for phylogenetic studies. Phylogenetic studies in plants mostly employ chloroplast genome sequences along with a few sequences on the nuclear genome, such as internal transcribed spacer DNA (ITS). Chloroplast DNA has been shown to provide a wealth of information on molecular variation for molecular phylogenetic studies. Early molecular phylogenetic studies using chloroplast DNA sequences were based on the comparison of restriction site polymorphism and gene order changes at a wide range of taxonomic levels (Olmstead and Palmer 1994; Jansen et al. 1998).

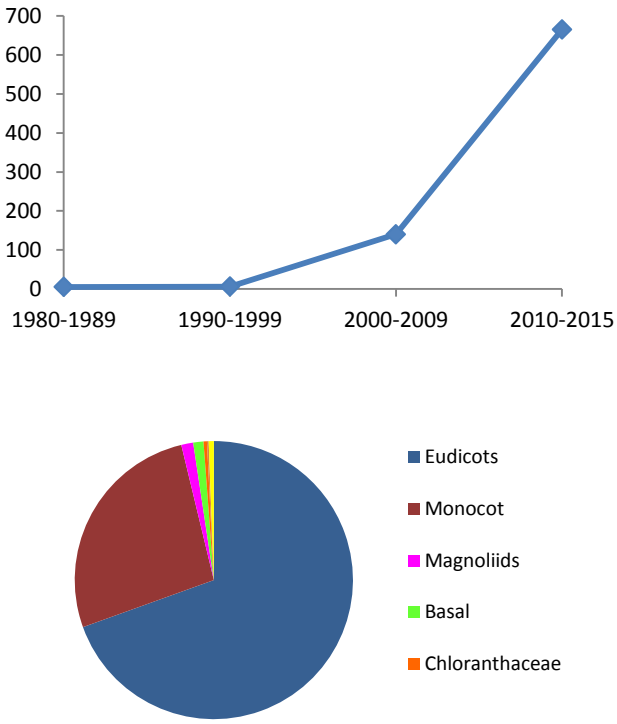


Figure 1: Numbers and taxonomic distribution of complete chloroplast genomes submitted to GenBank up to June 2016

Publication of the first complete chloroplast genome, that of *Nicotiana tabacum* (Shinozaki et al. 1986) was a defining moment in the study of chloroplast genome evolution as this enabled detailed nucleotide-level genome-wide comparisons to be made. Since then the number of complete chloroplast genomes sequenced for angiosperms deposited in the NCBI Organelle Genome Resources database has increased every year, reaching a total of 817 complete genomes in June 2016 (http://ncbi.nlm.nih.gov/genome/organelle/) (see also Figure 1). However, given the total number of angiosperm species of about 300,000 (Cowan et al. 2006), the fraction of published chloroplast genomes is concentrated on the eudicots and monocot class which is too low to fully understand chloroplast evolution (Figure 1).

Several groups of scientists have been focusing their efforts to develop and sequence complete chloroplast genomes to fill the most important gaps (Naito et al. 2013; Nikiforova et al. 2013; Ruhfel et al. 2014; Carbonell-Caballero et al. 2015) and the number of chloroplast genomes is expected to increase dramatically in the next few years.

Brief overview of chloroplast structure and evolution

The genes encoded in the chloroplast genome are generally conserved in content and in order among land plants. Genes can be categorised into functional groups (Kim and Lee 2004; Yi and Kim 2012; Li et al. 2013; Huang et al. 2014; Zhang et al. 2014). The first category includes genes that are involved in photosynthesis such as genes for photosystem I and II. Genes from the second category are involved in transcription, translation or self-replication such as transfer RNA, ribosomal RNA, ribosomal subunit genes and RNA polymerase genes (Mullet 1988). The third category comprises conserved open reading frames (ORFs) such as protein-coding genes like maturaseK (*matK)*, chloroplast envelope membrane protein (*cemA)* and hypothetical chloroplast open reading frame (*ycfs)*. The chloroplast genome of angiosperms is circular and the reported size varies from 120 to 220 kilobases (kb) (Wu et al. 2010; Wicke et al. 2011). The genome generally has a quadripartite structure with two copies of an inverted repeat (IR) region separated by small (SSC) and large single copy (LSC) regions (Saski et al. 2005; Yang et al. 2013; Yang et al. 2014). The IR size averages around 20-30 kb with some exceptions (such as *Pelargonium hortorum* (~76 kb) (Palmer et al. 1987; Chumley et al. 2006). The IRs are thought to act as stabilising regions and evolve ~2.3 times slower compared to the single copy region (Perry and Wolfe 2002). While the positions of boundaries between IR and single copy regions show some variability between species (this thesis), sometimes including some genes in the IR in one species that are present in a single copy region in another species, the IRs are exact reverse complemented duplicates, and hence both IR´s in a single chloroplast have the same gene content. The chloroplast genome is able to retain signatures of evolutionary history much longer than its nuclear counterparts due to the low level of mutation, which appears

at least in part due to the presence of these IRs. Mutations in the IR have been observed in species with chloroplasts lacking one copy of the IR, and here the synonymous substitution rate in this IR region is comparable to that in the SC region (e.g. *Medicago truncatula*, Ravi et al. 2008).

Chloroplast DNA sequences as molecular markers and their utility in phylogenomics

Phylogeny is the reconstruction of an evolutionary relationship history by comparing variation in homologous characters. Homologous characters are the characters that descend from a common ancestor, and are thus shared between organisms. These characters include morphological structures, ultra-structural characteristics of biological cells, biochemical pathways, genes, and the order of amino acids or nucleotides (Delsuc et al. 2005). The amount of difference between the homologous sequences in different organisms is used as a measure of their evolutionary relationship. However, homologous characters may evolve differently in terms of their rates of evolution, mutational saturation and compositional biases due to their own biological nature, thus not all character are suitable to be used as a phylogenetic markers and each character should be treated separately in a phylogenetic reconstruction (Gribaldo and Philippe 2002). The ideal marker should possess some features, for example the substitution rate should be optimum to provide enough informative sites, yet not be so high as to prevent comparison. A highly divergent gene may reach a state of saturation due to multiple substitutions yet it must be conserved enough to reflect the true ancestry (Galtier and Gouy 1995). Another important feature is that the markers must be acquired only by inheritance, not by transfer from another organism or horizontal gene transfer (HGT).

Chloroplast DNA sequences are a primary source of data in many plant phylogenetic studies. This is because the chloroplast genome is relatively conserved in its evolution making it an ideal molecule to retain phylogenetic signals. The chloroplast genome is also largely, but not completely free from evolutionary processes such as gene duplication, concerted evolution, pseudogene formation and genome

rearrangements whereas those are more common events in the nuclear genome (Palmer 1985). The conservation of the chloroplast genome also allows the design of primers targeting regions conserved well beyond species boundaries, and amplifications of molecular markers. Despite the low evolutionary rate in chloroplast genome compared to the nuclear and mitochondrial genomes, the small size together with their high copy number in leaf cells (as shown in Figure 2) greatly facilitates chloroplast genome sequencing.

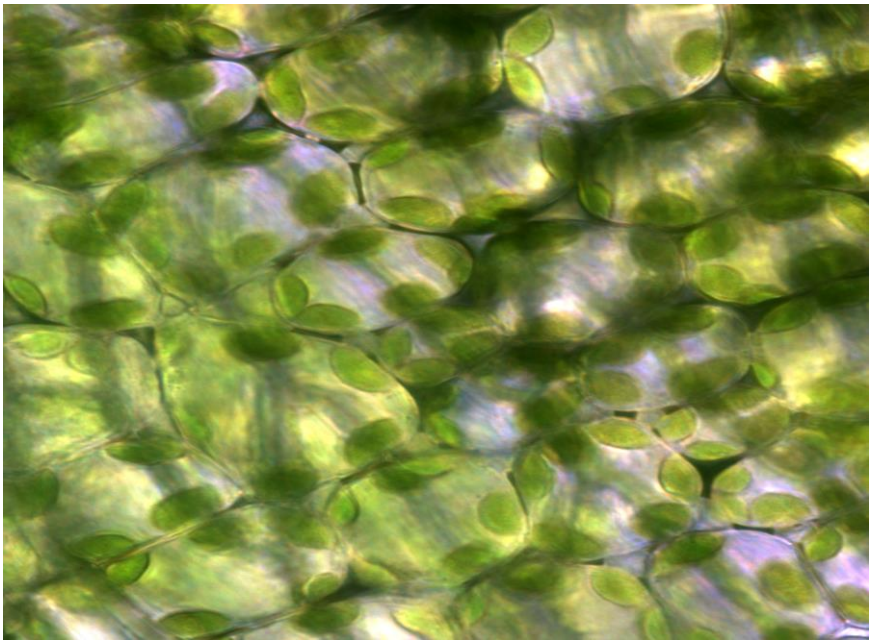


Figure 2: Example of chloroplasts visible in living mesophyll cells, observed with a light microscope (Norbert de Ruijter, Laboratory of Cell Biology)

Recently, sequencing technology breakthroughs have facilitated rapid sequencing of the entire chloroplast genome, making it possible to use complete chloroplast genomes for phylogenetics at genome scale (phylogenomics). This approach has become a universal method of providing evolutionary information for species identification (Wu et al. 2010; Nock et al. 2011), taxonomy and phylogenetic analysis in plants (Jansen et al. 2007; Moore et al. 2007). When using large datasets,

such as in a phylogenomic approach, the accumulation of phylogenetic signals normally overwhelms sampling errors, resulting in an improved statistical support (Blair et al. 2002; Wolf et al. 2004). However, a highly supported phylogeny tree does not necessary imply that the obtained tree is correct because systematic errors will also increase exponentially with the size of the data set (Philippe et al. 2005; Jeffroy et al. 2006; Brinkmann and Philippe 2008). Systematic errors are the result from violations of the model. In case of model violations, erroneous signal (noise) will be generated and compete with the genuine phylogenetic signal. If the genuine signal is weak or the noise level is high or non-random, the phylogenetic inference can be misled (Delsuc et al. 2005). As described in Rodríguez-Ezpeleta et al. (2007), there are several types of model violations such as across-site rate variation, heterotachy, site-interdependent evolution, compositional heterogeneity and site-heterogenous nucleotide/amino acid replacement. An example of the resultant systematic error is long-branch attraction (LBA). LBA is the phenomenon where two species that are more rapidly evolving than the rest of the taxa, were inferred to be closely related in the estimated tree (Felsenstein 1978). Strong support of artificial nodes occurs simply because of the accumulation of the systematic error with the addition of more data. The opportunity to examine all chloroplast genome features means that also any structural change in the genome can be detected and this may be informative in resolving certain intractable phylogenetic issues (Jansen et al. 2005; Philippe et al. 2005; Jansen et al. 2007). In a single gene phylogeny, the detection of systematic errors can be simply done by observing any incongruence between different genes. Unfortunately, this is not possible in a phylogenomic approach where genes are combined into a single supermatrix. Therefore, several approaches have been suggested to detect systematic errors including using different tree reconstruction methods and data partitioning strategy. Methods that are robust to violations of model assumptions are more preferable for tree reconstruction method whereas data partitioning strategies rely on the biological knowledge of a genes or sites [e.g., their relative substitution rates (Nishihara et al. 2007)] (Yang and Rannala 2012).

Methods to generate complete chloroplast genomes and their strategies

Researchers have been searching for new ways to obtain complete chloroplast genomes. As a result, many methods have been proposed to improve the accuracy and reduce the efforts in sequencing entire chloroplast genomes. These methods include the following:

i) Isolation of chloroplast DNA

Many methods were developed to isolate purified chloroplasts (Palmer, 1986). Most of these methods involve three basic steps: separation of plastids from other organelles and cell material, lysis of the chloroplast to yield intact chloroplast DNA, and subsequent purification of chloroplast DNA. Three methods to isolate intact chloroplasts are sucrose or Percoll gradients (Palmer, 1986), DNAse I treatment (Tewari and Kolodner 1979) and high salt buffers (Bookjans et al. 1984). Of those methods, sucrose gradients have been widely applied in land plants (Kim and Lee 2004; Samson et al. 2007). In general, all methods require a large quantity of fresh leaves, which will be difficult to achieve for herbarium samples or endangered species.

ii) Cloning the chloroplast genome for sequencing

The chloroplast genome can be cloned into a bacterial artificial chromosome (BAC) or a Fosmid vector. This method includes random shearing of the purified chloroplast DNA followed by cloning of the resulting long fragments in cloning vectors. These vectors allow easy production of large volumes of chloroplast DNA, amenable to sequencing. Clones containing fragments of the chloroplast genome can be either end-sequenced or shotgun sequenced using Sanger or next generation (NGS) sequencing such as Illumina. Details on this method were reviewed by Jansen et al. (2005). The approach is labour-intensive, technically demanding and time consuming. Nevertheless, this method is in some respects superior to direct high copy number plasmid cloning of the chloroplast, as the insert size is much larger (40-150 kb), allowing construction of a physical map spanning the IR region, allowing the orientation of all 4 compartments to be resolved. The first complete chloroplast, that of

tobacco (*Nicotiana tobacum*), was produced essentially using this strategy (Shinozaki et al. 1986).

iii) Designing long range PCR primers on conserved genes/regions in the chloroplast genome, followed by sub-cloning PCR fragments in a sequencing vector

This method involves PCR amplification of large fragments of the genome by using conserved primers to create a library for sequencing. Long-range PCR allows the amplification of much larger fragments of DNA than is possible with traditional PCR. Suitable primers can be designed on conserved regions or genes in the chloroplast genome. Amplified fragments ranging in size from 4 to 20 kb and covering the entire chloroplast genome can then be sequenced (Goremykin et al. 2003). Although the method is simple it requires a reference genome of a related species for designing the primers, which may not available for some non-model species. The primer combinations also may not work if there are changes in gene order such as for example in the *Campanulaceae* family (Cosner et al. 2004) or substantial divergence at the priming sites.

iv) Hybridization-based enrichment

Enrichment strategies include the use of molecular inversion probes and various DNA hybridization and sequence capture methods. Hybridization based organelle enrichments have been reported in several studies (Briggs et al. 2009; Cronn et al. 2012; Guschanski et al. 2013; Mariac et al. 2014). These methods are technically challenging and carry a high initial cost for laboratory protocol and reagents.

In conclusion, several methods have been developed to perform chloroplast isolation and chloroplast genome sequencing, but these have not led to simple protocols. Next generation sequencing, where whole genome shotgun sequences of chloroplasts are obtained as a by-product of whole genome shotgun sequencing, has the potential to make new steps in that direction.

Inferences of phylogenomic trees

Phylogenomics uses phylogenetic principles to infer evolutionary relationships, therefore it is necessary to assess only homologous sequences. In addition, one should use reliable characters for the phylogenetic inference, as the accuracy of tree reconstruction is strongly correlated with the reliability of the characters used. Delsuc et al. (2005), in their review on phylogenomics and the reconstruction of the tree of life, discussed two methods to assess which sequences are homologous: sequence-based methods and methods that are based on whole genome features. Of these two approaches the sequence-based method remains the method of choice because its properties have been intensively explored, tested and validated. Figure 3 shows a simplified figure describing methods of choice to infer phylogenetic relationships in a phylogenomic study.

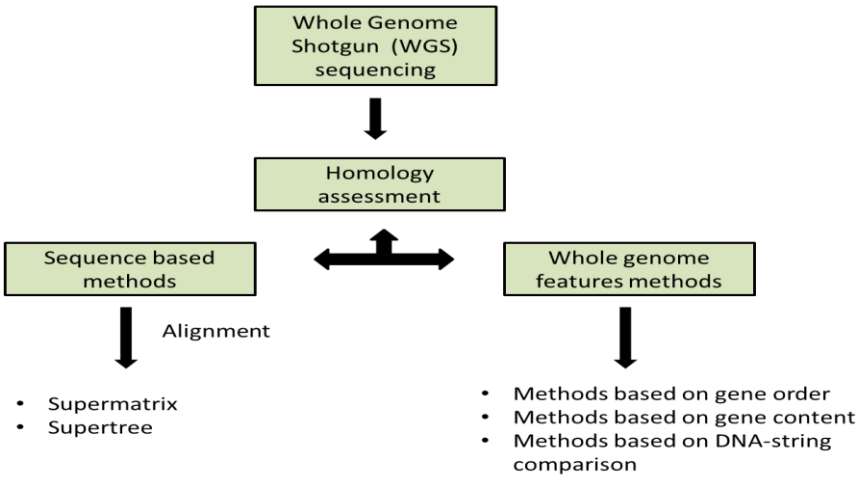


Figure 3: Simplified scheme of methods of choice for phylogenomic inferences as suggested by Philippe et al. (2005).

Sequence-based methods

In phylogenetic analysis using sequence-based methods, the construction of a phylogenetic tree starts with a multiple sequence alignment (MSA). Generally, a

MSA aims to arrange a set of orthologous genes from multiple organisms into an array in order to produce a table highlighting which variant of these genes each organism actually contains. Once constructed, the MSA is taken as input for the algorithm, thus an accurate MSA is essential to produce a reliable phylogenetic tree (Chan and Ragan 2013). For closely related species, each character entered in the same column is assumed to be homologous, super-posable, and to play a common functional role (Edgar and Batzoglou 2006). An MSA can be carried out manually as well as automated. An automated MSA is a more favourable approach as the increased throughput better matches the vastly increased throughput of next generation sequencing methods, bringing improvements in sensitivity. Various MSA methods such as ClustalW, MAFFT, MUSLE and T-COFFEE and methods with other types of input data such as PFAM were reviewed in by Edgar and Batzoglou (2006). Unfortunately assessment using alignment methods of which genes are orthologues actually requires rigorous and time consuming scrutiny, and the computational complexity and the – often implicit - choice and verification. of an appropriate evolutionary model are often overlooked in an automated procedure for the reconstruction of phylogenies. As a consequence, as more whole genome dataset are being generated and become available for constructing a phylogeny with, potentially, a much higher resolution, MSA will become insufficient in terms of quality control and affordable computation time (Delsuc et al. 2005).

The resulting MSA, usually containing sequences of unequal lengths from different sets of species, can be used to infer phylogenetic trees using either a supermatrix or a supertree approach (Delsuc et al. 2005). In the supermatrix approach as illustrated in Figure 3, MSA are first concatenated and then analysed as one set. In contrast, in the supertree approach the datasets are analysed individually and the resulting topologies are combined into a consensus (Delsuc et al. 2005). Up to now the supermatrix approach has been the usual practice in phylogenomics because the power of the approach is high and reliable. Furthermore, comparisons of the two approaches indicated that, with the size of the datasets up to now, the topology of the trees resulting from the two approaches was comparable as observed in Philippe et

al. (2005). Whether that will still be the case if datasets become very large based on next generation sequencing data, remains to be seen.

Alignment-free methods

Methods based on an atlas of specific features rather than an alignment of sequences present in the whole genome form a potential alternative to sequence alignment-based methods for analysing large amounts of sequence data in phylogenomic studies. These methods are also known as alignment-free methods because they completely avoid the MSA step. Generally there are two types of approaches for an alignment-free method. The first approach recognizes the need for an assessment of which characters are homologous, whereas the second approach completely avoids a homology assessment step. In the first approach, phylogenetic inference is constructed based on the comparison of gene order and gene content, but not gene sequence. Gene content and gene order do not require a MSA step yet they still depend on homology assessment. This type of method is capable of producing good phylogenetic markers that are less prone to homoplasy than sequence polymorphisms (Gribaldo and Philippe 2002). The methods are under continuous development. As an example, constructing phylogenies based on gene order was first introduced by Sankoff et al. (1992) using the complete genome of 16 mitochondria of fungi and other eukaryotes. Their method determined the evolutionary distance by the number of inversions, transpositions, deletions or insertions required to change gene order of one genome to another (Otu and Sayood 2003). Later, the method also has been used to test phylogenetic hyphotheses in Proteobacteria (Kunisawa 2001), in Gram-positive bacteria (Kunisawa 2003), and among various prokaryotic genomes (Wolf et al. 2001). Subsequently, more studies were carried out to develop an improved algorithm using gene order to infer phylogenetic relationships as described in (Lin and Moret 2010; Zhang et al. 2010; Hu et al. 2011). Recently, the method was applied on a genome-wide basis as reported by Lin et al. (2013) and Shifman et al. (2014). Latest work on the development of gene order based phylogenies was discussed in House et al. (2015) who developed a simple computational method to estimate a genome-wide gene

order of 143 and 172 prokaryotic genomes. They successfully demonstrated the robustness of gene order by uniting two phyla groups together. Nevertheless, such methods may be less suitable for chloroplast genomes as gene order changes are much rarer in chloroplast than in nuclear/mitocondrial genomes.

Word usage frequency

Approaches in which also homology assessments are completely avoided, would be the most practical way to construct a phylogenetic inference using an alignment-free method. One implementation visualizes DNA sequences or protein sequences as strings of letters, and every word of an exact subsequence of defined length extracted from those strings can be defined as a word of k length, commonly referred to as a k-mer. To be used in phylogeny reconstruction, k-mers are extracted and their counts and frequency distribution are then used to compute a pairwise distance matrix. The relatedness between sequences is then calculated based on the number of k-mers counted and the fraction that they share (Chan and Ragan 2013) .This approach does not suffer from the limitation of aligning sequences when there is too much sequence difference, or that alignments become arbitrary in case of gene duplications, recombinations, rearrangements and other biological events. Yang & Zhang (2008) claimed that the k-mer method would be capable of producing more accurate phylogenetic trees compared to trees computed from MSA. Phylogeny reconstruction using k-mers or derivative approaches is becoming increasingly popular with the increasing availability of genome sequences as evidenced by several studies that employ it (Edwards et al. 2002; Qi, Wang, et al. 2004; Höhl and Ragan 2007; Sims et al. 2009). However, although this approach sounds promising, the distances measured by word usage typically do not have a clear biological meaning and the distances rarely show a linear increase with evolutionary time. Up till now, several alignment-free methods have been proposed based on word frequency approaches, such as composition vector (Qi et al. 2004), feature frequency profile (Sims et al. 2009), chaos game representation (Joseph and Sasikumar 2006), return time distribution (Kolekar et al. 2012) and no doubt other refinement methods are on their way.

Research aims and thesis outline

Having the complete chloroplast genome could provide comprehensive data sets that are superior for inferring relationships at intraspecific, interspecific and genus level. Yet, the prospect of having complete chloroplast genomes for all angiosperms especially in the non-model species is still far away given the current state of chloroplast genome assembly methods as well as their data analyses. This thesis explores methods to obtain the chloroplast genome sequence and analyse it based on next generation sequencing data.

Chapter 2 describes the results of performing *de novo* assemblies of chloroplast genomes of *Solanum lycopersicum*, *Aegilops tauschii* and *Paphiopedilum heryanum* based on whole genome sequencing data. The chosen species were different in their nuclear size genome ranging from ~1 Gbp to 35 Gbp. Most methods of assembly rely on mapping against a reference genome, but this may leave out some of the differences from the assembly, including structural changes (rearrangements). The approach used here started with a statistical analysis of the k-mer frequency distribution of shotgun sequencing data to identify potential reads from the chloroplast genome in the mixture of paired-end reads from genomic DNA, followed by *de novo* assembly and several subsequent refinement steps. The importance of the interaction between the amount of data used and the k-mer size is also highlighted.

In Chapter 3 the results of creating a flexible assembly quality comparison tool is described. This tool combines and visualizes read mapping and alignment results in a two-dimensional plot without breaking any sequence connectivity. Correspondingly, the ability of this tool using the *de novo* assemblies of *Solanum lycopersicon* (tomato) and *Paphiopedium henryanum* (orchid) chloroplasts obtained from Whole Genome Shotgun (WGS) Illumina short read sequencing datasets in combination with specifically made alternative assemblies was evaluated.

In **Chapter 4**, the chloroplast genomes from whole genome sequencing data of 83 accessions of tomato and its related species were extracted and analysed. The 83 accessions covered the *Lycopersicon* section within the genus *Solanum* including wild accession, old cultivars and domesticated cultivars. The aim is to show the versatility of the approach for resolving the phylogenies of these closely related species of tomatoes.

Chapter 5 seeks to gain insight into the utility of complete chloroplast genomes to resolve conflicts concerning the division of the orchid subgenus *Paphiopedilum* into several sections. The study focused on two sections of the subgenus *Paphiopedilum*; *Coryopedilum* and *Perdalopetalum*. It has been suggested that these two should be combined as the section was shown to be paraphyletic to the monophylectic section *Perdalopetalum* based on ITS data (Cox et al., 1997). This is in conflict with the taxonomy of Cribb (1998) in his monograph based on their morphological characters. The *Coryopedilum* section includes species that can be found in Malaysia. Most of them are endemic to single islands. In contrast, species of section *Perdalopetalum* are more widespread and distributed through mainland Southeast Asia.

I conclude this thesis with a summary and discussion of the results in **Chapter 6**. The chapter also discusses and proposes a new direction to efficiently use genome-scale data to infer plant relationships at intraspecific, interspecific and genus level.

Chapter 2

De novo assembly of complete chloroplast genomes from non-model species based on a k-mer frequency-based selection of chloroplast reads from total DNA sequences

Shairul Izan, Danny Esselink, Richard G.F. Visser, Marinus J.M. Smulders, Theo Borm (submitted)

Abstract

Whole Genome Shotgun (WGS) sequences of plant species often contain an abundance of reads that are derived from the chloroplast genome. Up to now these reads have generally been identified and assembled into chloroplast genomes based on homology to chloroplasts from related species. This re-sequencing approach may select against structural differences between the genomes. The risk of missing such differences increases when reconstructing chloroplast genomes from non-model species for which no close relative genome is available. The alternative approach is to *de novo* assemble the chloroplast genome from total genomic DNA sequences. Although the chloroplast genome has a simple structure and conserved gene content, this is still a challenge. The Bruijn graph based assembly has been widely used to analyse short read sequences from next generation Illumina sequencers. Underlying the Bruijn graphs are tables consisting of counts of individual short sub-reads of length K as found in the WGS dataset. These so-called k-mer frequency tables have many other uses. In this study, we used k-mer frequency tables to identify and extract the chloroplast reads from the WGS reads and assemble these using a highly integrated and automated custom pipeline. This pipeline includes steps aimed at optimizing assemblies and filling gaps that are left due to coverage variation in the WGS dataset. We have successfully *de novo* assembled three complete chloroplast genomes from plant species with a range of nuclear genome size to demonstrate the universality of our approach; i.e. *Solanum lycopersicum*, *Aegilops tauschii* and *Paphiopedilum heryanum*. We also highlight the need to optimize the choice of k and the amount of data used. This new and cost-effective method for *de novo* short read assembly may facilitate the study of complete chloroplast genomes with more accurate analyses and inferences, especially in non-model plant genomes.

Introduction

Chloroplast genomes are frequently used in systematics and phylogeography because of the simplicity of the structure of its circular genome, its predominantly clonal inheritance along the maternal line, as well its high copy number in the cell (Palmer and Stein 1986; Moore et al. 2006; Ma et al. 2013). The chloroplast genome is often perceived to have a low amount of sequence variation, and the use of the genome has therefore been mostly confined to studies at the interspecific and interfamilial levels (Jansen et al. 2007; Moore et al. 2007; Xi et al. 2012; Barrett et al. 2013). Recently some studies involved in comparative analyses of complete chloroplast sequences showed that the perception of low variation of chloroplasts within species is wrong when looking at the genomic scale (Whittall et al. 2010; Besnard et al. 2011; Kane et al. 2012). Kane et al. (2012) suggested that the whole chloroplast genome could be used as a ultra-barcode for identifying plant varieties. Furthermore, using one or few regions of the chloroplast genome is not the appropriate approach to describe the level of variability of the chloroplast genome. Therefore, using the complete chloroplast genome will undoubtedly be the best way to exploit the information in this organelle genome.

Chloroplast DNA can traditionally be obtained by a chloroplast enrichment strategy using a sucrose gradient (Moore et al. 2006) or high salt method (Bookjans et al. 1984). These strategies require large amounts of starting materials (~5 g tissue), which may be challenging for endangered plant species or herbarium samples. Some plant groups may have a high content of polysaccharides, polyphenols, and/or terpenoids, which also poses a challenge to obtain high quality chloroplast DNA (Vieira et al. 2014). Using PCR the complete chloroplast genome can be amplified in the form of a series of long, overlapping PCR fragments. This approach requires appropriate primer design as well as high quality DNA to ensure successful long range amplifications. The primers for these reactions have been designed on conserved gene sequences (Goremykin et al. 2003; Jansen et al. 2005), which work

reasonably well across species. The implementation suffers from differences in gene organization among plant species (Atherton et al. 2010).

Next generation whole genome shotgun (WGS) sequences of plant species often contain 5% or more reads that are derived from the chloroplast (Bakker et al. 2016). This offers an alternative way to obtain chloroplast genomes. These reads are generally identified from the WGS reads and aligned into chloroplast genomes based on homology to chloroplast genome from reference genome. Such an alignment-based method has been a method of choice to do the sequence comparison during recent years. A comprehensive review about this method was written by Vinga et al. (2012). However, as structure and function in a genome may diverge over evolutionary time, such alignment-based methods may become unreliable for taxa for which no close relative exists with a high quality chloroplast genome. They may also become computationally unaffordable when dealing with very large datasets of sequences (Vinga et al. 2012 but see Bakker et al. 2016). Several alignment-free methods have been proposed to tackle those limitations and one of them is an approach based on k-mer frequency tables. The k-mer based approach may be the most developed alignment-free method (Chan and Ragan 2013). A k-mer is an exact substring of DNA sequence of defined length (k), whose frequency in a set of DNA sequences can simply be counted (Marçais and Kingsford 2011). Applying statistics on the sharing of k-mers between samples provides an estimate of genetic distance (Bonham-Carter et al. 2013). K-mer frequency tables are also used to distinguish sequencing errors from genuine sequences (Kelley et al. 2010) as sequencing errors are presumed to be random in nature thereby generating unique or low-frequency k-mers, while genuine sequences occur at a certain k-mer frequency, depending on the frequency of sequences in the target genome and the depth of sequencing in the WGS dataset. K-mer frequency tables have also been used to detect repeated sequences in the genomes (Kurtz et al. 2008), employing the fact that k-mers derived from a particular repeat of a certain copy number in the genome will have a similar frequency.

From the k-mer frequency tables, k-mer frequency distribution histograms can be derived (Chikhi and Medvedev 2014) which show the volume of k-mers occurring at each frequency in the dataset. If a particular, highly abundant (extrachromosomal) sequence occurs at a certain frequency in the dataset, this leads to a (broad) peak in this histogram. If another highly abundant sequence occurs at twice that frequency in the dataset, then there will be another peak in the histogram – at twice the frequency. Chloroplasts generally contain an Inverted Repeat (IR) region, and naturally k-mers obtained from reads in this IR region will occur at twice the frequency of k-mers obtained from Single Copy (SC) regions of the chloroplast, so we expect chloroplast-derived k-mers to be contained in two peaks in the histogram – the second at exactly twice the frequency of the first. In this study we have used k-mer frequency histograms to identify the two peaks corresponding to chloroplast-derived k-mers, and used their approximate frequencies to select the corresponding k-mers from the underlying k-mer frequency table. These k-mers were subsequently used to select reads containing them, which were then used in a first round of assembly. After the first round of assembly, subsequent rounds of assembly and refinement lead to an automated semi-finished assembly of a chloroplast genome.

This chapter demonstrates the feasibility of a procedure that employs a k-mer frequency table and derived k-mer frequency histogram to extract the chloroplast sequences from whole genome sequencing data without the use of a reference genome prior to *de novo* assembly of shotgun sequences obtained with the Illumina platform. We used WGS data obtained from three species notably a Solaneaceous species, a grass species and an orchid species with a range of nuclear genome sizes (950 Mb - 35 Gb) to demonstrate the universality of our approach. One of our cases is a novel chloroplast genome for an orchid species from the genus *Paphiopedilum*, which have a very large nuclear genome size (25-35 Gb).

Materials and methods

Source of sequencing data sets

Whole genome paired-end sequences of *Solanum lycopersicum* and *Aegilops tauschii* were downloaded from the sequence read archive of Genbank (http://www.ncbi.nlm.nih.gov/sra). The WGS dataset for *Paphiopedilum heryanum* was generated for this study (Table 1) using fresh leaves of *Paphiopedilum heryanum* obtained from Hortus Botanicus in Leiden, the Netherlands. The DNA isolation was carried out by combining a DNA extraction using the protocol as described in Fulton et al. (Fulton et al. 1995) with a DNEasy Plant Mini Kit (Qiagen), using the kit's DNA binding column to bind and clean-up DNA. A barcoded sequencing library was constructed by BGI, China, who also performed the 100 bp paired-end sequencing on an Illumina Hiseq2000 platform in a single lane along with 10 other samples from a separate experiment. For simplicity, from here onwards we will refer to the analysis of WGS datasets obtained from *Solanum lycopersicum*, *Aegilops tauschii* and *Paphiopedilum heryanum* as case study 1, 2 and 3 respectively.

Table 1: Species used in the study and their SRA number

Species (n)	Haploid genome size (bases)	Group	NCBI SRA number
1) Solanum lycopersicum (2n)	950 Mb	Dicot	SRR404081
2) Aegilops tauschii (2n)	4-5 Gb	Monocot	SRR124187
3) Paphiopedilum heryanum (2n)	25-35 Gb	Monocot	Own data

Bioinformatic analyses

Overview of the approach

Our assembly approach comprises five stages as illustrated in Figure 1. As the nuclear genome complement of different genomes results in differently shaped k-mer frequency distribution histograms, and as chloroplast DNA concentrations in WGS samples vary considerably, a visual inspection of k-mer frequency histograms is required between stages 1 and 2, where the user decides which k-mer frequency range to include in the analysis. While no human intervention is explicitly required between the other stages (2-5) of the pipeline, many optional parameters can be varied should the user require so, and the staging offers a convenient way for the user to monitor progress and output (assemblies) after each stage of the pipeline. Each stage is implemented as a separate PERL script, calling upon a large library of secondary PERL scripts, compiled C programs and external software (e.g. SOAPdenovo, BLAST) to perform its tasks. Access to the software pipeline can be granted on request.

Data preparation

Prior to stage 1 the user has to prepare the dataset by putting all sequence reads in fastq format files in a single directory. In order to allow the program to figure out which files contain matching paired-end reads and which files contain single end reads, the user has to adhere to a simple file naming convention**.**

Stage 1: Obtaining k-mer frequency tables and k-mer frequency histograms from WGS datasets

The script implementing stage 1 produces alphabetically sorted k-mer tables with k-mer size 31 by default. In these k-mer tables, k-mers and their exact reverse complement are counted as a single ordinal k-mer. This ordinal k-mer is chosen from the two options in such a way that the middle nucleotide is always either 'A' or 'C' – if it is not then the k-mer is reverse complemented before being counted. After counting, a k-mer frequency histogram is produced from the tables. The k-mer

frequency histograms are converted to histograms representative of the underlying data volume by multiplying the number of different k-mers occurring at each frequency by said frequency. We will refer to these histograms as k-mer volume histograms. To aid visualisation, a series of binned histograms is produced with frequency bin-sizes of 10, 25, 100 and 250.

Visual inspection of k-mer frequency histograms

As each plant cell contains multiple chloroplasts, unless special precautions are taken during DNA sample preparation, molar concentration of chloroplast DNA in the WGS sample will be higher than that of nuclear DNA. Moreover, because chloroplasts most often contain an exactly duplicated Inverted Repeat (IR), the chloroplast DNA derived k-mers will give rise to a pair of peaks in the k-mer frequency histogram that can be easily distinguished from any other peaks because of their fundamental relation: The second (IR) peak occurs at twice the frequency of the first Single Copy (SC) region peak. The user then imports these k-mer frequency histograms into his/her favourite graphing package, and on he basis of the location of the peaks representing chloroplast sequence read derived k-mers decides where to set k-mer frequency boundaries.

Stage 2: Obtaining chloroplast specific reads and initial assembly

The frequency boundaries set by the user are used in stage 2 to select, from the original k-mer frequency table, those k-mers occurring in this frequency range. These k-mers will, besides chloroplast derived k-mers, also contain k-mers derived from nuclear repeat-regions that coincidentally occur at the same frequencies. This k-mer table is then used to select, from the full WGS dataset, those reads that contain them. These selected reads are then sub-sampled into a series of batches of increasing size (by default starting at 100,000 read-pairs, with 100,000 read-pair increments), and automatically assembled using SOAP-denovo (v1.05) (Luo et al. 2012). SOAPdenovo is a the Bruijn graph-based assembler that can use a range of values for the k-mer size (K), and results have previously been found to be highly dependent on the value of K (Chikhi and Medvedev 2014). Therefore we employed

a range of different values for K (all odd values between 63 and 99). This yields a multitude of separate assemblies which are then filtered (by default using BLAST against the tobacco chloroplast genome) to remove any contig or scaffold that does not seem to be chloroplast-related (putatively repeats from the nuclear genome), and size-selected to remove any contig or scaffold smaller than twice the size of K (as used in the assembly). The resulting filtered assemblies are subsequently subjected to a sanity check where excessively short or excessively long assemblies are discarded. This filter is by default based on previously observed length ranges for SC and IR regions, and is user-configurable. The remaining assemblies are then ranked according to: a) the number of scaffolds they consist of (fewer is better), b) the number of gaps they contain (fewer is better) and c) the total length of the assembly (longer is better). The best assembly is used in the next stage.

Stage 3: Iterative refinement of read selection and assembly

As discussed, the selection of k-mers in a set frequency range means that k-mers derived from nuclear genomic repeats coincidentally occurring at these frequencies are also selected. While enrichment of the dataset for chloroplast-derived reads is certainly achieved, the repeat region-derived reads co-selected because of this k-mer table contamination can be considered problematic. In the previous stage we tried to alleviate this by using BLAST and a size filter, but this carries the risk that some small fragments of genuine chloroplast sequence or highly deviant chloroplast sequences are lost. Stage 3 iteratively uses the putatively pure chloroplast derived assembly obtained in a previous iteration (or stage 2 for the first round) to select reads and re-assemble. To this end, a k-mer table is obtained from the chosen assembly, which is then used as described in the description of stage 2 to select reads, which are then assembled and filtered as described previously. Assemblies are ranked to produce a new best assembly until either no better assembly is produced or until a set limit on the number of iterations is reached. In addition to the assembly performed by SOAPdenovo, this stage employs its own assembly algorithm that looks for remaining overlap between scaffolds and contigs produced by SOAPdenovo, and where possible assembles these, taking into account the fact that

a circular genome with an inverted repeat is expected (two aspects that existing assembly programs are unaware of). The final output of stage 3 is a new best assembly that is used in the next stage, and which may consist of linear or circular fragments. As the read-pair insert sizes attainable with current short read technology do generally not span a complete IR region, the exact relative orientation of the Short Single Copy (SSC) and Long Single Copy (LSC) regions cannot be determined. The internal assembly algorithm can (in case a circular assembly can be made) output either a set of three linear fragments (putatively representing LSC, IR and SSC), two separate assemblies for both possible circular configurations OR just one (randomly chosen) circular assembly. Stage 4 and 5 require the last option, and it is left to the user to find the correct relative orientation of the LSC and SSC (to be validated for instance using long range PCR).

Stage 4: Scaffold extension and spanning-read based re-scaffolding

The newly assembled genome resulting from stage 3 may or may not be circular, and if not circular it may or may not consist of multiple unconnected scaffolds, each of which may or may not contain gaps. The purpose of step 4 is to iteratively connect linear scaffolds remaining from stage 3 by extending and connecting scaffolds with additional sequence reads until scaffold ends overlap or by finding read-pairs spanning gaps between scaffolds. Stage 4 is skipped if stage 3 delivered a circular assembly. Briefly, all the raw reads are aligned back to the assembly using BWA and those (paired-end or single) reads that extended outside the gaps are picked. Each scaffold-end will produce a separate set of (paired-end) reads which are then assembled to obtain new scaffolds. These new scaffolds are added to the previous round best assembly and used as input to the internal sequence assembly algorithm and subsequently filtered as described under stage 3, producing a new assembly for use in the next iteration. Iterations are terminated if either a) the resultant assembly is circular OR b) the quality of the assembly does not improve (per the same criteria used to find the best assembly) OR c) until a set limit on the number of iterations is reached. After the last iteration, if the resultant assembly is not circular already, raw reads are mapped back (BWA) against the resultant

scaffolds and any read connecting scaffold-ends is selected and counted in a scaffold-end connectivity matrix. This scaffold-end connectivity matrix is combined with the scaffold sequences and used by the internal sequence assembly algorithm to produce a new assembly, placing N's in gaps that are bridged by gap-spanning reads. Again, this may in some cases lead to construction of a circular assembly.

Stage 5: gap filling

After stage 4 gaps may remain in the sequence. These gaps are putatively caused by systematic (sequence dependent) low coverage in such regions, which should be considered an artefact of the Illumina sequencing technology used (Minoche et al. 2011). As we have used variable sized batches and various settings for K during the assembly, sufficient reads covering these low coverage areas may still remain unused in the dataset. Stage 5 attempts to fill the gaps by focussing only on reads covering such gaps, again assembling (using SOAPdenovo) variable sized batches of reads with a range of values for K. To this end, gap-context sequences (default 500 bp on either side of the gap) are extracted from the previous best assembly (either the previous iteration or stage 4), and used to produce a k-mer table for positive selection of reads. The regions of the previous stage best assembly scaffolds that are outside the defined gap-context are used to produce a second k-mer table that, after comparison with the positive selection k-mer table, is exported as a negative selection k-mer table. Raw reads are filtered using the positive selection k-mer table, retaining any read containing a k-mer from this set. Subsequently this subset is filtered using the negative selection k-mer table, discarding any read containing a k-mer from this set. The resulting set of reads is then assembled in variable sized (default 1000 read (-pair)s, with 1000 read (-pair)s increment) batches with SOAPdenovo using a range of values for K (odd values between 63 and 99). This delivers a number of scaffolds, which are then re-scaffolded using the internal assembly algorithm before being size filtered, discarding any scaffold shorter than K base-pair. The remaining scaffolds are then, one by one, combined with each separate gap context sequence using the internal sequence assembly algorithm, and ranked (for each of the gaps separately) to find the best gap-closing assembly.

Finally, the best gap-closing assemblies (if any) are used to replace the gap context sequences in the original assembly, and the whole process repeats iteratively until either a) all gaps are closed OR b) until assemblies do no longer improve OR c) a set limit on the number of iterations is reached.

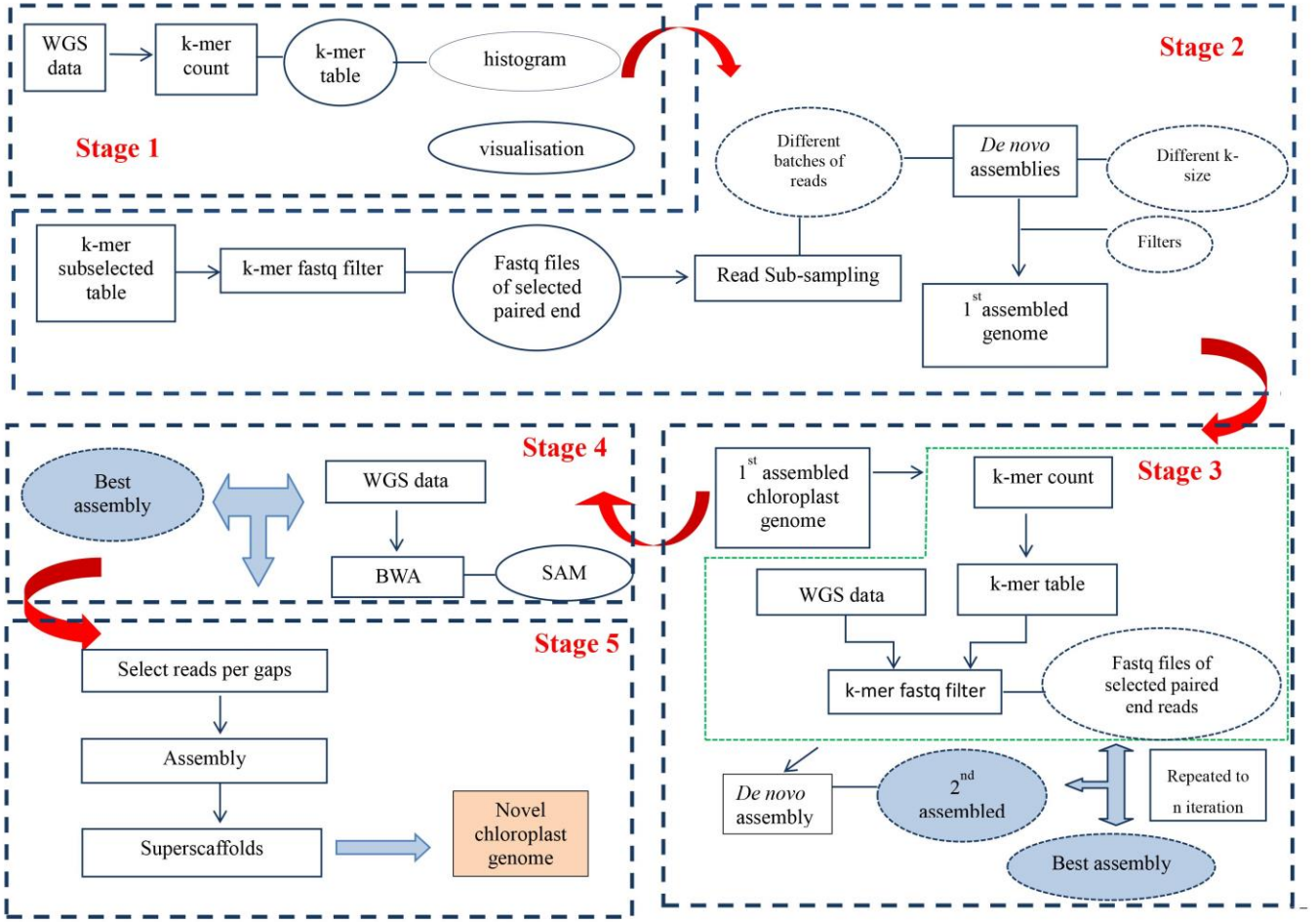


Figure 1. Work flow of the assembly pipeline.

Results

Determining chloroplast-derived k-mers based on the k-mer frequency distribution

Figure 2 (a, b, c) shows k-mer volume histograms (binned per 25 frequencies) of the raw reads of case study 1, 2 and 3. The two expected peaks for k-mers derived from the chloroplast genome sequences are clearly visible as sharp peaks in case study 1 (at 12000x and 24000x coverage), they were flatter in case 3 (at 170x and 350x coverage) (Figure 2 a, c), while in case 2 only one peak (at 1500x coverage) could be discerned (Figure 2 b). To see the effect of k-mer based read selection for chloroplast reads, we overlaid the k-mer volume histogram from the raw reads with the k-mer volume histogram of the reads picked out using the selected k-mers in the left part of Figure 2, In all datasets the volume of k-mers specific to erroneous sequences and to the nuclear genome were significantly reduced while the volume of kmers in both chloroplast peaks essentially remained the same. This indicates that our selection enriches for chloroplast sequences.

Extracting chloroplast reads and *de novo* assembly

Each case study contained between 15 million and 198 million raw read pairs. Following the k-mer based extraction of chloroplast reads from the raw reads of the case study, significant read reductions were seen across the stages. Table 2 presents the total number of read-pairs in a dataset as well as the number of read-pairs used in stages two and three. Across three case studies a reduction by almost 40% of the number of reads-pairs is seen in stage two.

To investigate the optimum assembly for each case study, *de novo* assembly with different batches of subsampled read pairs and k-sizes were performed. Basically, the pipeline gave a candidate best assembly at the end of stage 3 based on 1) the lowest number of scaffolds, 2) the fewest gaps and 3) the longest assembly length (within the allowed range). In case studies 1 and 2, inspection of the assembly statistics of all assemblies produced in stage 2 revealed that the automatically chosen

assembly with the fewest number of scaffolds was either too long or contained an excessive number of gaps. Therefore, in these cases we manually selected an alternative best assembly based on minimal number of scaffolds plus gaps, with the longest length in the allowed range. In contrast, the automatically selected best assembly was a reasonable choice in case study 3 and thus did not need manual selection. In addition, we also investigated the efficacy of stage 4 and 5 for scaffold expansion or re-scaffolding and the gap filling. Table 3 shows the statistics of the best assembly after stage 4 and 5. From our observation, all case studies showed that the stage 4 and 5 helped to merge scaffolds and fill the gaps. As example, in case study 3, eight scaffolds were merged and two gaps resolved in stage 4 and 5 compared to the underlying SOAPdenovo assembly (12 scaffolds with 3 gaps).

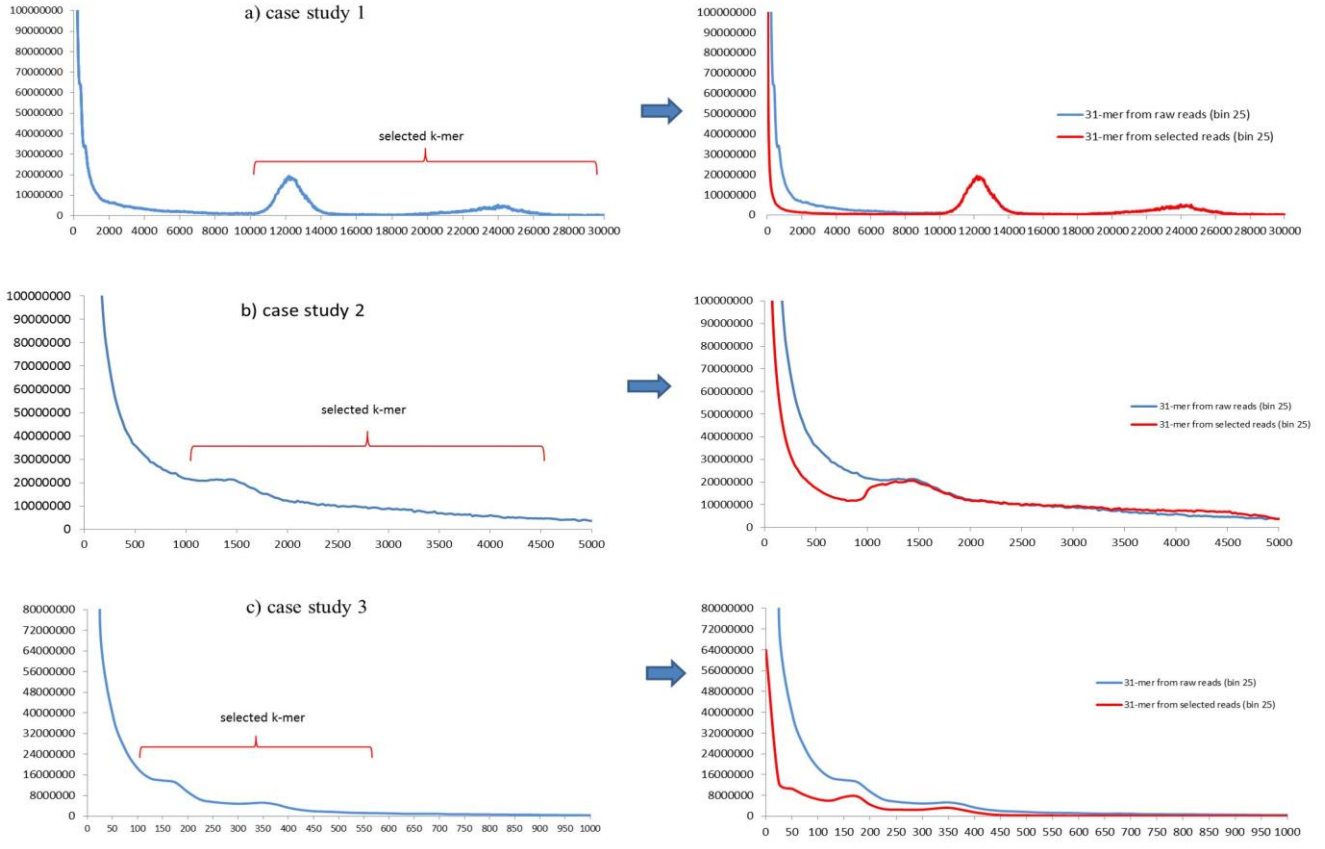


Figure 2. K-mer volume histograms of the raw reads (left panels) and an overlay of raw reads and selected reads after selection (right panels). X-axis: k-mer frequency; *y*-axis: k-mer counts. In case study 1 (tomato) the nuclear haploid genome size is 950 Mbp, in case study 2 (*Aegilops*) it is 4-5 Gbp, in case study 3 (*Paphiopedilum*) it is 25-35 Gbp.

Table 2: Summary statistics before and after the fetching of the chloroplast reads

	Case study 1	Case study 2	Case study 3
Genome size	950 MB	4-5 GB	25-35 GB
Total no of raw reads (pairs)	198 264 041	86 067 571	15 142 939
Total no of reads after stage 2 (pairs)	32 701 410	51 717 173	6 172 495
Total no of reads after stage 3 (pairs)	14 855 294	1 582 279	213 669

Table 3: Comparison of the SOAP*denovo* assembly and de novo assembly derived after stage 4 and 5 from the proposed pipeline

Case study	No of scaffold	No of gap	Total assembly length	Total reference length
Case study 1				
SOAPdenovo	3	0	130 181*	
Our approach	1	0	155 461	155 461 [a]
Case study 2				
SOAPdenovo	9	4	114 806*	
Our approach	2	2	135 760	135 685 [b]
Case study 3				
SOAPdenovo	12	3	122 051*	
Our approach	4	1	156 087	174 417 [c]

*Contained only one copy of IR

a *:Solanum lycopersicum* chloroplast, complete genome (NC 007898.3)

b: *Aegilops tauschii* cultivar AL8/78 chloroplast, complete genome (KJ 614412.1)

c: *Cypripedium japonicum* chloroplast, complete genome (KJ 625630.1)

Mummer analysis of reference and *de novo* genomes

To detect any large structural variants such as inversions, insertions or deletions in the *de novo* assembled genomes, dot plot analyses were using MUMmer (Delcher et al. 2003). Figure 3 displays the dotplots comparing all three *de novo* genomes as well as three reference genomes in all 15 possible combinations Appropriate reference chloroplast genomes were downloaded from Genebank, NCBI with accession number NC_007898.3, KJ_614412.1 and KJ625630.1 respectively. As no reference genome is available for case study 3, we used a complete chloroplast genome from a related species.

From the dotplot analyses of only the reference genomes against each other (Fig. 3a, b and c), we noted that the chloroplast of *Aegilops tauschii* (KJ_614412.1) has an inversion in the LSC region of about 13 860 bp length. The structure of the other two reference genomes was comparable without large structural variants. The inversion in the *Aegilops tauschii* reference genome was also detected in our *de novo* assembly of case study 2 (as shown in Fig. 3k). Moreover, we concluded the inversion in *Aegilops taushii* chloroplast genome was a genuine event as it was also supported by read mapping of the raw reads against the *de novo* assembled genome.

Interestingly, we also found two large structural changes in the *de novo* chloroplast assembly of case study 3 (Fig. 3m). These structural variants in the *Paphiopedilum* species chloroplast genome are reported here for the first time. The first structural variation is an inversion in the LSC region. This inversion is absent in the reference genome of a related orchid species (*Cypripedium japonicum*). Secondly, we observed an IR expansion into the whole SSC region. Both these structural variations are absent in the other genomes including the orchid species *Cypripedium japonicum*. In addition, we conclude that all inversions are genuine events as they are supported by the read mapping (not included in this thesis).

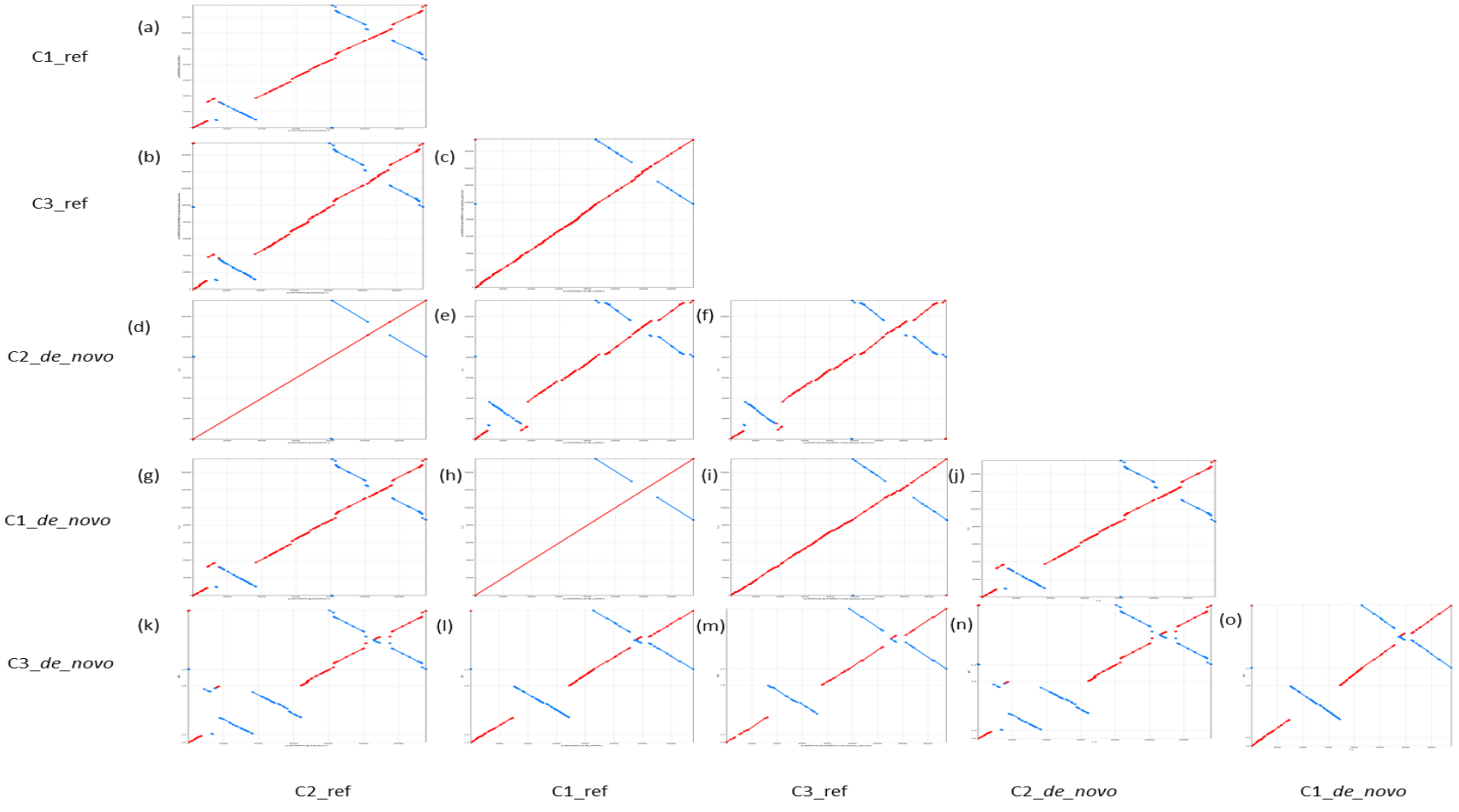


Figure 3: Dotplot analyses against reference genomes and *de novo* assembled genomes for case study 1 (C1, tomato), 2 (C2, *Aegilops tauschii*) and 3 (C3, *Paphiopedilum heryanum*)

Mapping and *de novo* assembly of sequence reads

The raw reads were aligned against the *de novo* assembled genomes to verify the detected structural variation as well as to detect any miss-assemblies in the *de novo* assembled genomes. The read alignments were performed using BWA with default parameters. The mean coverage of the reads varied considerably among these three case studies (17822x, 4396x and 497x coverage for case study 1, 2, and 3 respectively) illustrating that different DNA sequencing datasets contain different numbers of chloroplast reads. Figure 4 shows comparison coverage plots of genomes assembled using our pipeline and unaltered assembly from the SOAPdenovo assembler. The assembly that SOAPdenovo produced only contained one copy of IR. The read coverage (*y*-axis) was plotted against the genome position and has been averaged using a window of 100 bp (*x*-axis).

In general, read coverage was sufficient to detect any miss-assemblies. Coverage plot comparison between the genome assemblies in each case study also demonstrated that our pipeline successfully assembled the scaffold across the low coverage regions. In contrast, SOAPdenovo assembler left gaps in the scaffolds (black boxes). This illustrated the power of the scaffold expansion, re-scaffolding and gap filling implemented in our pipeline leading to better quality of chloroplast genome assembly. Worth to mention, the zero coverage at the start and end of the genome (circular) of scaffolds (linear) characterized by red arrow was due to the pseudo-circularization – addition of a copy of the first N basepairs to the end of the assembly. This was done to facilitate the read mapping of the overhanging reads that used to connect two scaffolds. Beside the artefact because of pseudo-circularization, we also found several positions (indicated by the yellow arrows in the assembly of case study 2 and 3 in Fig. 4) with zero read coverage, representing gaps in the genome assembly. This also suggests that the assembly will not improve anymore with this particular dataset.

.

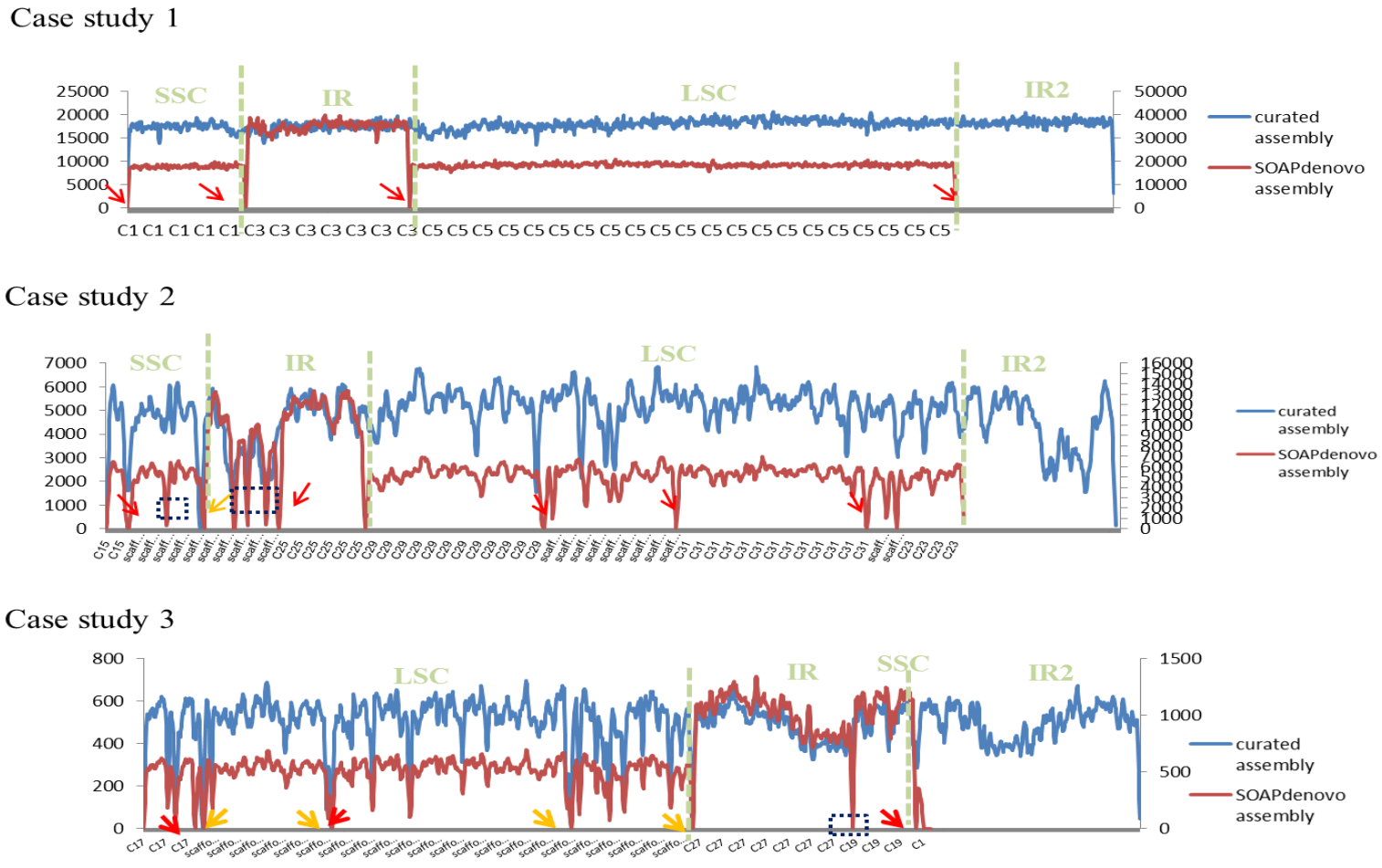


Figure 4: Comparison of read coverage (*y*-axis) against the genome position (*x*-axis) between the assembly from SOAPdenovo and the curated assembly. Yellow and red arrow: occurrence of zero read coverage; black dotted box: gap in the scaffold.

Variant calling

Pairwise alignments for *de novo* assembled genome with their reference genome were conducted to call for variants. The result of variant calling is represents in Table 4. We do not present the pairwise alignment from case study 3 because we encountered a large number of variants across the genome, including two large structural variations. This large difference is due to the fact that the reference was from a related species and clearly the two species were too far diverged. We investigated the pairwise alignment from both other case studies and variants that were called included insertions or deletions (INDELs) and mismatches (SNPs). Remarkably, we only found only one mismatch in the alignment of case study 1 at the position 127 404 bp, which was located in the IR region. On the other hand, we successfully called 13 variants in the case study 2 consisting of 10 INDELs and three mismatches. Looking at those locations, we found five length variants of a homopolymer region.

Table 4: Variant calling for case study 1 and 2

Case study	Type	Variants	Position in the assembled genome
Case study 1	Mismatch	G (ref) > T (ass)	127404
Case study 2	Insertion	AGGTACCTAA	7653-7662
	Insertion	homopolymer T region	18272-18274
	Insertion	homopolymer A region	18614
	Insertion	homopolymer A region	34160
	Insertion	CT	43329-43330
	Insertion	homopolymer A region	56672-56673
	Mismatch	CTCTC (ref) > TCTCT (ass)	76298-76302
	Deletion	homopolymer A region	78860
	Insertion	TTTACTTTTATGTTTTATTTG	107322-107342
	Insertion	GCAATAATCTACTAAAAAAA	109678-109697
	Mismatch	G (ref) > N (ass)	109894
	Mismatch	T (ref) > N (ass)	109893
	Mismatch	T (ref) > N (ass)	109899

Discussion

Chloroplast genomes from next generation sequencing datasets

A chloroplast genome sequence provides information for addressing various biological questions, including phylogenetic analysis (Oxelman et al. 1997; Goremykin et al. 2003; Capella-Gutierrez et al. 2014). Furthermore, since the chloroplast genome is inherited uniparentally and is not subject to recombination during gametogenesis like the nuclear genome, it is an ideal locus for barcode analyses (Austerlitz et al. 2009; Hollingsworth et al. 2009; Li et al. 2015). The present study shows that it is possible to assemble high quality complete chloroplast genomes from whole genome shotgun (WGS) sequencing datasets using a largely automated pipeline. As next generation sequencing technology advances, more WGS data will become available to the researcher. Those data could be exploited using the approach outlined here in order to provide an easy and cost-effective way to construct complete chloroplast genomes. In this way it will be possible to reliably mine these resources for information on the chloroplast genome. We also hope that our approach can help to increase the number of available chloroplast genomes. This will open up the possibility to do comparative analyses. In spite of the small size of the chloroplast genome, many fundamental characteristics such as functional sequences outside the coding sequences (promoter, terminator, replication origin), detection of selective signatures in gene sequences as well as mutational rates and their mechanism (Raubeson et al. 2007) are poorly described. Those hypotheses can be critically addressed by comparative studies.

K-mer frequency distribution, sequencing error, coverage bias and genome size

The distribution of k-mer frequencies in a whole genome DNA sequence dataset includes information on the underlying genomes as well as on characteristics of the sequencing run. Unlike other protocols to assemble chloroplast genomes, which either require a protocol to either physically (e.g. specific isolation of chloroplast DNA) or in-silico (alignment of WGS reads to a chloroplast reference) enrich the dataset for target sequences, our method fetches chloroplast sequences from WGS

sequencing reads without prior knowledge about the sequence and without additional effort during DNA isolation, and use those in a *de novo* assembly. This takes advantage of the known (LSC-IR-SSC-IR) chloroplast structure and the resulting, predicted, structure in the k-mer frequency distribution: as there is a large inverted repeat in the chloroplast, a bimodal k-mer frequency distribution is expected, with one peak (representing the inverted repeat) occurring at exactly twice the frequency of the other peak. This allows identification of these peaks in a k-mer frequency distribution. However, as there are other (e.g. genomic) sequences present in the dataset, there may be a significant background present of k-mers derived from these other sequences at similar frequencies as the choloroplast derived k-mers, and the amount of background is clearly influenced by the nuclear genome size, as can be observed in our three case-studies. Several studies investigating the link between k-mer frequency distribution and sequencing errors have been carried out (Liu et al.; Kurtz et al. 2008; Kelley et al. 2010). Random sequencing errors will generate a high peak with low coverage, and as the rate of sequencing errors increases, this "error-peak" on the left side of the frequency plot will increase in size, while other peaks will become smaller and also decrease in frequency, thus move to the left. Of course, if there are highly repetitive regions in the genome, with correspondingly higher k-mer frequencies, errors in the sequences generated from these repetitive regions will also occur at a larger rate, consequently giving rise to a widening of the error-peak. For large, complex genomes it is expensive to generate sufficient coverage of the nuclear genome to be able to easily separate the peak corresponding with genomic DNA ("nuclear genome peak") in the k-mer frequency histogram from the error-peak, and as a consequence, the "nuclear genome peak" may overlap the "error peak" and become an inseparable, very wide combined peak, even overlapping the "chloroplast peaks", as can be seen in case study 3, and to a bit lesser degree in case study 2. On the other hand, for case study 1 the "nuclear genome-peak" is well-separated from both the "error peak" and the "chloroplast peak". Case study 1 is an excellent example of the desired separation of the sequencing error, while the datasets of case studies 2 and 3 might benefit from more sequencing data – better separation between the desired "chloroplast peaks" and the

undesired “error peak” and “nuclear genome peak” would improve the selectivity of the k-mer frequency based filtering of reads. As was intended, we noticed in all cases that the coverage of k-mers specific to error and nuclear genome were reduced significantly after the k-mer selection while the coverage of peaks belong to chloroplast sequences remained the same or slightly reduced as seen in case study 3. Wherever frequencies of k-mers obtained from the nuclear genome overlap the “chloroplast peaks”, reads derived from the corresponding, evidently repetitive regions, from the nuclear genome will also be selected and included in the assembly process. The effect that this might have on the chloroplast assembly depends on several factors. First of all it depends on the lengths of the repeating units – if these are small (e.g. <500bp), the resulting assemblies will be also be small, and may be removed on the basis of their size alone. If the repeating units are large (e.g. > 10K) and high frequency, then this would be a novelty and mean that a large proportion of the nuclear genome would be contained in such repeats. Such long repeats are also very easy to remove as long as they don’t bear any resemblance to known chloroplast genomes. Insertions of parts of a chloroplast genome into the nuclear genome might be an interesting problem if these insertions would happen be large and would happen within repetitive regions – in such cases chimeric scaffolds may be expected. Outside the repetitive regions the non-repetitive nuclear genome will give rise to relative low frequency k-mers, which would therefore not be selected, and which would therefore not lead to inclusion of larger regions of nuclear genome derived reads into the assembly process. While this may, depending on overall sequence coverage, lead to some confusion in the assembler, this should not lead to many problems in the downstream analysis. Incidental insertion of parts of a chloroplast genome into the nuclear genome should also not lead to detection of SNP’s in the chloroplast – the SNPs will give rise to k-mers occurring at frequencies corresponding to the nuclear genome, and the underlying reads will either not be selected on the basis of their k-mer frequencies or, if they happen to be selected, add little coverage in the assembly process, and be consequently treated as sequencing errors and be removed.

The relative positions in the k-mer frequency histograms of the peaks corresponding to the nuclear genome and the chloroplast, in combination with their respective genome sizes can give us some insights into the number of chloroplast genomes per cell. From the perspective of chloroplast genome assembly, a fixed ratio between the number of nuclear genomes (1) and chloroplast genomes is a worst case scenario: In WGS datasets of larger genomes the percentage of chloroplast derived reads would then be lower, necessitating disproportionally more sequencing in larger genomes to obtain a usable coverage of the chloroplast genome. In some cases it may even be appropriate to combine our method with a chloroplast DNA enrichment strategy.Our data seem to indicate that the percentage chloroplast reads in a WGS dataset is not constant, but decreases when the nuclear genome size increases. This could be expected if the number of chloroplasts per cell is more or less constant, or regulated between tissues in the same way regardless of nuclear genome size, but it was not what Bakker et al. (2016) observed. This may be related to the fact that they only tested a limited range of genome sizes. On the other hand, the anecdotic case studies that we present here may be the ones deviating from the general trend.

K-mer size and assemblies

The SOAPdenovo assembler is based on a de Bruijn-based graph, which breaks the reads into k-mers of defined size before assembling them into contigs (Pevzner et al. 2001). After initial k-mer based graph construction, several steps refer back to the original underlying data to resolve some of the issues caused by the short length of K – most notably resolution of knots caused by repeat units smaller than the length of the reads yet larger than K. The robustness of the SOAPdenovo assembler relies on several competing effects that are difficult to quantify. One important parameter is the k-mer size K. For instance, K smaller than some repeat sequences may cause tangling up in the de Bruijn graph, which, if very complex and unresolvable with the raw-read-data, may lead to contigs being broken up. Thus, we need large K. However, larger K will reduce the number of k-mers that can be extracted from a given sequence read – and as a consequence lead to fewer k-mers being extracted from a dataset overall and hence lowering of k-mer frequencies. Lower k-mer frequencies may make it difficult to distinguish good sequence from sequencing errors, and may eventually lead to problems in de Bruijn graph construction. Also, assuming random distribution of sequencing errors, the probability of a longer k-mer containing a sequencing error is larger, which will lead to more k-mers being included in the error-peak. Another effect is that if two contigs overlap by less than k-1 characters, this will create a coverage gap resulting in the break-up of a contig (Chikhi and Medvedev 2014). Another factor influencing the assembly process is the amount of data being used. More data does not necessarily improve assembly quality. Especially for extreme coverage data, and for non-random sequencing errors, assembly of larger datasets may give rise to alternative assemblies, one with the "proper" sequence, and one containing a "SNP". Having alternatives for regions is not easily representable in FASTA format assembly output, and in SOAPdenovo it generally leads to fragmentation. In the algorithm of the pipeline presented here we employed a range of different values for K in order to minimize the trade-off effects. We also employed a range of dataset sizes by including different numbers ("batches") of reads in the assembly process. This yields a multitude of separate assemblies, which are then filtered out using some filters. The remaining assemblies

are then ranked accordingly and putatively best assembly was selected automatically. As seen in case study 1 and 2 the automatic selection of a best assembly based on maximum assembly length and minimal (number of scaffolds plus gaps) may be more appropriate than maximum assembly length and minimal number of scaffolds alone. In contrast, in case study 3 the automatic selection of a best assembly based on maximum assembly length and minimal (number of scaffolds and gaps) was sufficient. This indicates that intelligent inspection of intermediary results for every stage in the pipeline is useful.

Assemblies and sequencing bias

Compared to other studies that use reference sequences to extract chloroplast reads, the approach proposed here extracts the reads derived from the chloroplast solely based on the fact that they occur at the certain frequency in the k-mer frequency distribution of WGS data. By utilizing such an approach, we obtained reasonably high coverage of chloroplast genome across the case studies. Nevertheless, there are several gaps in *de novo* assembled genome compared to the reference genome in case study 2 and 3. Those gaps in the assembled genome may be caused by sequencing bias in the sequencing library. For instance, bias in the pre-sequencing amplification step could result in poor or no sequencing coverage in certain regions of the genome. Generally, a GC content sequencing bias has been observed. In accordance with our results, several studies (e.g., Dohm et al. 2008; Li et al. 2010; Minoche et al. 2011) claim that even though there is sufficient average depth of sequence coverage within sequencing datasets, sequencing bias leads to region of no sequence coverage within sequencing datasets, resulting in multiple gaps in the assemblies, and hence a larger number of contigs and scaffolds even in small sized genomes such as bacteria and the chloroplast genome.

INDEL detection and homopolymers length polymorphism

The selected reads were assembled *de novo* instead of taking an alignment or reference guided *de novo* assembly approach. This approach offers additional possibilities for detecting structural differences that may be missed in other

approaches. Moreover, the approach uses the read coverage information which provides a reliable detection of sequence variation. We detected several structural differences in two out of three case studies. Even considering the general conservation of chloroplast genome, several structural differences were reported for nine grass species (Golenberg et al. 1993), Korean ginseng (Kim and Lee 2004) and Pinus (Parks et al. 2009). Hence, it may be inappropriate to assemble the chloroplast genome for non-model species by alignment to a reference sequence of a related species because it may miss important structural differences but also because reads from repeated or homologous regions can generally not be distinguished in a mapping based approach – which may lead to identification of false SNP's in such regions. Another issue to be aware of is that half of variants detected in case study 2 were homopolymer length polymorphisms. This may due to the fact that the reference genome of *Aegilops tauschii* (KJ_614412.1) was sequenced using the SOLiD platform while WGS dataset of case study 2 was sequenced using Illumina. It is known that Illumina sequencing is less affected by homopolymer length variation (Harismendy et al. 2009). It is also a known issue that SOLiD shows low coverage of AT-rich regions, while Illumina sequencing has been observed to have more problems with GC-rich regions (Morozova and Marra 2008; Harismendy et al. 2009).

Conclusion

The chloroplast genome certainly is a great resource of molecular markers in many studies including parentage analysis, hybridization, population and genetic structure and phyleogeography. The pipeline described here provides a tool to extract chloroplast sequences from Whole Genome Shotgun (WGS) sequences of plant species. Our newly developed pipeline was able to efficiently assemble the chloroplast genome across a range of nuclear genome sizes, and using it we discovered several structural rearrangements compared to published reference chloroplast genomes. This cost-effective approach will be particularly useful for exploring in the increasing number of WGS sequences from non-model species. In

principle, our pipeline in combination with high throughput short read sequencing can greatly expand the scope of comparative genomics of the chloroplast genome in plants.

Chapter 3

Visual comparison of the quality of chloroplast assemblies

Shairul Izan, Peirong Li, Theo Borm, Richard G.F. Visser, Marinus J.M. Smulders
(to be submitted)

Abstract

So far, no single sequence assembly algorithm and no parameter setting has emerged as a gold standard that reliably produces perfect assemblies, and hence there is a need for objectively measuring the quality of an assembly. Often this quality is measured in derivative terms such as the number and length of contigs and scaffolds that a program produces. Because parameter-tweaking may be used to optimize assemblies, which may favour erroneous albeit longer assemblies, this may not always give appropriate results. Also, sometimes much more detailed information on the exact differences between assemblies is desirable, and while programs that highlight specific problems in individual assemblies exist, these results tend to be difficult to compare. To address these issues, we have created a flexible assembly quality comparison tool. This tool combines and visualizes read mapping and alignment results in a two-dimensional plot without breaking any sequence connectivity. We have evaluated the ability of this tool using the *de novo* assemblies of *Solanum lycopersicon* (tomato) and *Paphiopedium henryanum* (orchid) chloroplasts obtained from Whole Genome Shotgun (WGS) Illumina short read sequencing datasets in combination with specifically made alternative assemblies. The results show that not only we can immediately select the best of two options for a purpose, but also determine the location of specific artifacts.

Introduction

Even though sequencing costs and costs of computational power have significantly dropped, making high quality sequence assemblies is still a challenge. While *de novo* sequence assembly programs like SOAPdenovo (Luo et al. 2012) or VELVET (Zerbino and Birney 2008) follow well-established and often very similar procedures to create assemblies from Whole Genome Shotgun (WGS) data, results can vary considerably, not only because of different parameter settings and general sensitivity to parameter settings, but also because underlying assumptions may differ somewhat. Often the quality of assemblies is primarily measured in terms of the number and length of contigs and scaffolds that a program produces, and parameter-tweaking may be used to optimize assemblies, which may favor erroneous albeit longer assemblies. Some tools have been published that attempt to capture and visualize assembly quality using different parameters (Kelley et al. 2010; Barthelson et al. 2011; Earl et al. 2011; Salzberg et al. 2012; Bradnam et al. 2013)

For large genomes, due to human constraints, the level of detail in the quality assessment must be extremely limited, whereas for smaller (e.g. bacterial or organellar) genomes, where manual finishing is still an option, a higher level of detail may be appropriate. The most detailed level is that showing the individual reads in the assembly. For genomes larger than a few kilobases, any visualisation showing individual reads quickly becomes unwieldy, and one will want to us a visualisation showing aggregate data such as sequence coverage along the assembly rather than individual reads. Sequence coverage along the assembly alone may not be enough as some artifacts in assemblies may have only limited impact on coverage, and some available software is able to show a variety of types of data [e.g. Tablet (Milne et al. 2013)].

Because it is quite common to optimize assemblies by tweaking parameters and making multiple assemblies from the same dataset, it would be useful to be able to directly compare these assemblies. This presents a problem, because in order to

yield comparable graphs of the desired aggregate data along the assemblies, the assemblies must be co-linear. Making assemblies co-linear may present a challenge if repeats and/or structural variation are present. In such cases it may be necessary to break scaffolds to restore co-linearity. The chloroplast genome is circular and generally has a quadripartite structure consisting of a Long Single Copy region (LSC), an Inverted Repeat region (IR), a Short Single Copy (SSC) region and another copy of the IR, so co-linearity issues are to be expected. To address these issues, we have created a flexible assembly quality comparison tool, employing mummer to visualize which segments in two separate assemblies are homologous, while reads mapped back to the sequence are used to extract several types of aggregate data that may be diagnostic for assembly problems. This tool combines and visualizes read mapping and alignment results in a two-dimensional plot without breaking any sequence connectivity.

Materials and methods

Our goal is to demonstrate a visualisation tool allowing direct comparison of pairs of assemblies on the basis of read mapping data. To make these comparisons, we used the *de novo* assemblies of *Solanum lycopersicon* (tomato) and *Paphiopedilum henryanum* (orchid) chloroplasts previously (Chapter 2) obtained from Whole Genome Shotgun (WGS) Illumina short read sequencing datasets in combination with specifically made alternative assemblies. These alternative assemblies were obtained using an iterated reference backed read-mapping procedure consisting of: a) mapping back all paired-end reads in a WGS dataset to an appropriate chloroplast reference genome using the sampe module of BWA (Li and Durbin 2009) (with default parameters). b) Calling variants using the mpileup and call modules of SAMTOOLS (Li et al. 2009) and VCFTOOLS (Danecek et al. 2011) respectively (both with default parameters). c) Incorporating the variants thus found into the reference sequence using the consensus module of BCFTOOLS (Danecek et al. 2011), generating a derived chloroplast reference. d) Iterating steps a), b) and c), replacing the reference with the derived reference until convergence is reached.

The reference chloroplast sequences used in this experiment as a starting point for iterations were *Cypripedium japonicum* (NC_027227.1) (Kim et al. 2014) for *Paphiopedilum henryanum*, and *Nicotiana tabacum* (NC_001879.2) (Kunnimalaiyaan and Nielsen 1997) for tomato. For validation purposes, the previously obtained *de novo* assemblies were themselves also used as reference sequences in this iterated procedure. For *Paphiopedilum henryanum*, the *de novo* assembly consisted of four disconnected linear segments, and prior to analysis, these segments were ordered and oriented as far as possible (without breaking up any existing scaffolds) according to the order and orientation of the homologous sequences in the *Cypripedium japonicum* reference, and concatenated into a single scaffold with spacers consisting of 100 N's. The tomato *de novo* assembly consisted of a single circular scaffold already, and was left as it was.

Pairwise assembly alignments were made between *de novo* and iterated read mapping assemblies using mummer (Delcher et al. 2003) (specifically nucmer) and the ".coords" output file this produces was combined with the read-mapping data of the underlying WGS reads against both assemblies and visualized using the custom perl script described here. This script makes a highly configurable combined graphical plot that contains the mummer alignment and a combination of other tracks, including, but not limited to the read mapping density of normal and discordant mapping reads, the insert size average and the number of clipped nucleotides found in partially mapped reads. Specific (local) regions of assemblies and read mapping data were inspected using IGV (Robinson et al. 2011).

Results

For validation and to check if the *de novo* assemblies could be improved upon, the *de novo* assemblies were also used as reference in our iterated read mapping assembly procedure. No differences whatsoever were observed between input and the derived assemblies for tomato. For *Paphiopedilum henryanum*, simple convergence was never reached. Rather, after 3 iterations, derived sequences cyclically repeated through a total of 6 different variants. Five separate sites were found to participate in cyclical convergence; four sites alternating between two and one site alternating between three options, resulting in a combined 6-state cycle. Close inspection (data not shown) revealed that two sites were located in low diversity, low coverage AT-rich regions surrounding gaps in the *de novo* assembly. One site was found to be located in a nearly exact (albeit unresolved) repeat. The remaining two unstable sites, both with adequate coverage suggest that the sample was heterogeneous. Close inspection of the regions surrounding the five gaps remaining in the *de novo* assembly reveals that these are mostly reduced complexity, AT rich regions with GC fraction in the 100 bp directly adjacent (on either side) found to be 7%, 4%, 0%, 22%, 3%, 2%, 35%, 34%, 34% and 42% respectively. Overall GC fraction was 36.1%. While some sequences could be retrieved from the read mapping extending into the 100 N's spacers placed there during concatenation of scaffolds, these were low complexity AT-rich sequences, and none of the remaining gaps could be resolved.

The iterated read mapping assembly against an alien chloroplast reference converged to an alternation of two sequences after 7 (tomato WGS data with a tobacco reference) and 21 (*Paphiopedilum henryanum* WGS data with a *Cypripedium japonicum* reference) iterations, and 2 and 105 nucleotide differences were observed between the resultant variants. Close inspection (not shown) in tomato suggests that the variant sites (SNPs) are an artifact caused by truncated mapping of reads at sites where larger structural variation between tomato and tobacco exists. In *Paphiopedilum henryanum* many of the variant sites were found to

be located either near low complexity, AT rich regions or near regions without sequence coverage.

Figure 1 and Figure 2 show the example visual assembly comparisons made using the perl script presented here. In Figure 1, some standard graph elements are annotated; lower case characters in both figures are used to show particular highlights described later in text. In these figures, tracks associated with one assembly are plotted along the *x*-axis while the tracks associated with the other assembly are plotted along the *y*-axis, with a mummer-plot linking the two assemblies shown between these tracks. Figure 1 and Figure 2 show comparisons between the *de novo* (X-axis) and the iterated read mapping (*y*-axis) assemblies of tomato and *Paphiopedilum henryanum* respectively. In the mummer-plot a yellow background denotes areas with paired-end coverage of less than 10% of the average, while green and red line segments represent homologous sequence fragments found in either the same or opposite orientation in both assemblies respectively. On the basis of segment length, duplication and relative position, Long Single Copy (LSC), Short Single copy (SSC) and (IR) Inverted Repeat sections can be annotated in both the mummer plot and along the axes (as was done in Figure 1). The tracks shown in each figure are:

I. Average fragment length (blue) of the paired-end data. These fragment length are averaged over all read-pairs centred on 200 bp bins, and where no average could be calculated, none is shown. The fragment lengths are based on the mapping positions of the reads, and (where appropriate) include clipped bases.
II. Relative number of non-mapped basepairs in reads with clipped mapping per 200 bp bin (cyan).
III. Coverage plot showing coverage by paired-end sequenced DNA fragments (green). Coverage in this track includes any un-sequenced nucleotides located between the paired sequences, so that gaps where no sequenced nucleotides are available (N's) will still be covered, indicating proof of connectivity between separate sequence contigs.

IV. Nucleotide coverage by discordant read-pairs (magenta). This graph only includes read-pairs mapping on the same scaffold but in unexpected orientations, so with both reads in a pair mapping in the same direction or with a negative insert size between them, but excludes read-pairs mapping in the correct orientation but with an unexpectedly large insert size.

V. Nucleotide coverage by "linking" read-pairs (red). Linking read-pairs either link different scaffolds together or link different areas within a scaffold together. The latter category only includes read-pairs mapped in the correct relative orientation on the scaffolds with an aberrant insert size as otherwise they would either be considered normal read pairs (with a normal insert size) or discordant reads (with aberrant orientation).

VI. The number shown next to each track is the scale of each division (thin grey lines) of that track.

Most of the elements in Figure 1 (tomato) are notable for their presence along a single axis: the iterated read mapping assembly. This indicates that these elements are mostly defects present in this assembly and absent from the *de novo* assembly. We observe a short low coverage region (a) coinciding with an excess (f) of "linking" reads indicating an insertion in this assembly not supported by the WGS data, with no DNA fragments bridging this gap. Peaks (b) indicating short regions with longer than average insert size, with two peaks coinciding with peaks in the "linking" read pairs graph (g), without evidence of complete lack of coverage suggesting insertions of relatively small extraneous sequences. Many scattered peaks (c) in the track showing the number of nucleotides clipped during mapping putatively indicating many small deviations. Peaks (d) in the discordant coverage track near the SSC/IR/LSC junctions, suggesting that these thee junctions are not precisely conserved between tomato and tobacco. Peaks (e) in the linking reads track, suggesting structural differences between tomato and tobacco. A single peak (h) in the discordant reads track for the *de novo* assembly and peaks (i) near the end of the assembly may represent an artifact caused by the circular nature of the chloroplast genome and the presence of the inverted repeats. Erratic coverage (j)

may be due to coverage lost in discordant mapping and clipping. Subtle differences (k1 and k2) between tracks at the sites of the inverted repeats are putatively caused by random fluctuations in read assignments to these regions and the exact boundaries of the 200bp wide sequence bins employed in visualisation. Overall it should be noted that the tomato WGS dataset provides enormous coverage (~40 000x) across the chloroplast genome, which means that some level of (background) artifacts is expected in all tracks except the insert size and coverage tracks. Auto-ranging means that the scales of tracks may differ significantly.

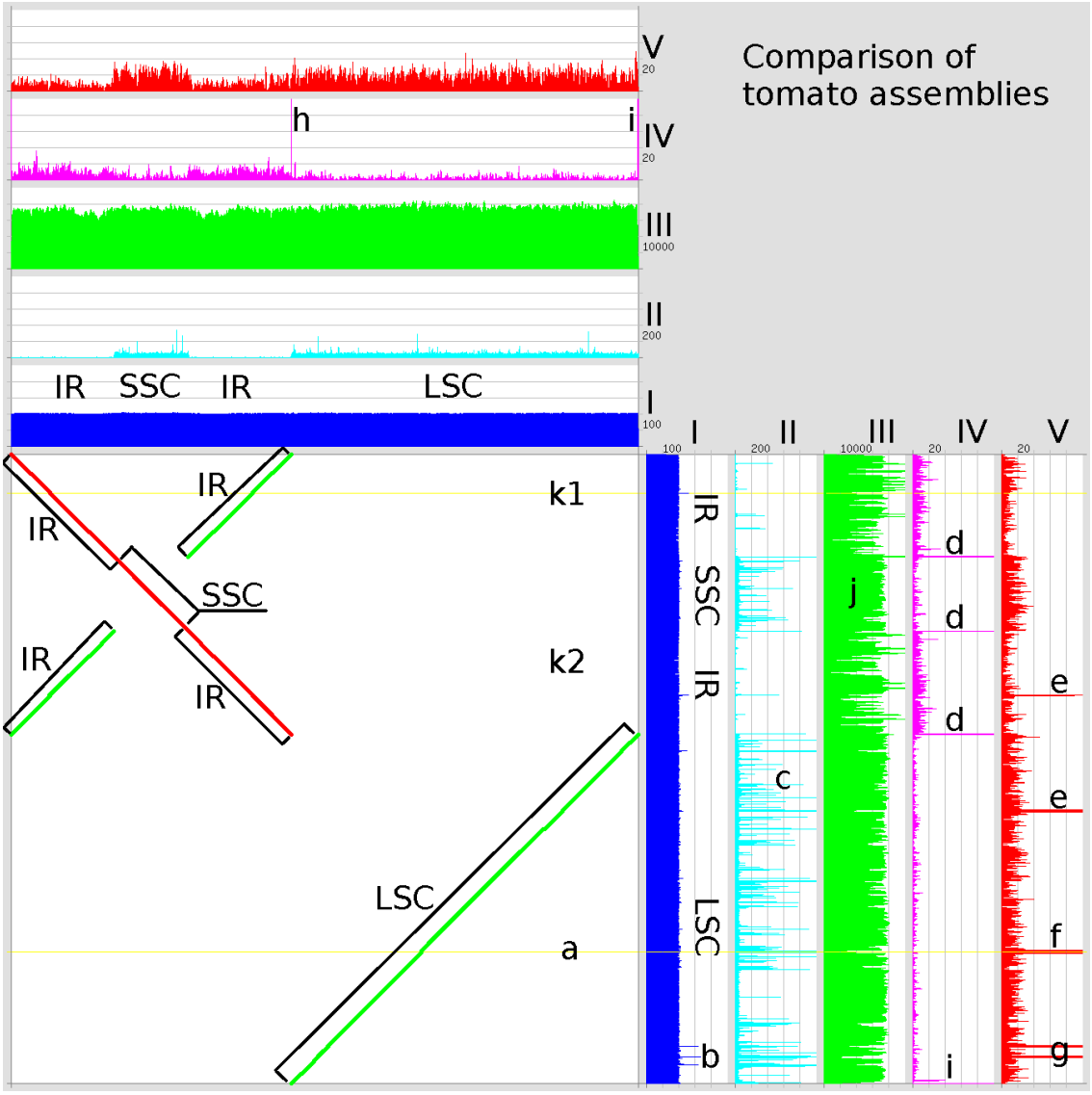


Figure 1. Comparison of the *de novo* assembly (along the *x*-axis) with the iterated read mapping assembly (along the *y*-axis) of tomato. The bottom left shows the mummer alignment of the genomes, while different tracks to the right and on top show various aggregate data types. Further explanation in the text.

In Figure 2, the iterated read mapping assembly of *Paphiopedilum henryanum* using *Cypripedium japonicum* also contains most of the defects. From our observations, many low or absent coverage regions (a) (only one shown) in the iterated mapping assembly, indicating that these regions, making up a considerable fraction of the chloroplast genome, are completely missing from the *Paphiopedilum henryanum* chloroplast genome. The structure of the short single copy region (b) is completely different between the two assemblies, with the iterated read mapping assembly in the Short Single Copy (SSC) region being considerably longer and the whole region appearing twice (in inverted repeat) in the *de novo* assembly. Coverage data (d) indicates that the SSC in the read mapping assembly attracts approximately twice the average coverage (note the scale on the axis), supporting the hypothesis that the SSC in *Paphiopedilum henryanum* has been incorporated into the IR. Excessive discordant read mapping (e) and "linking" reads (f) at the SSC boundaries add support to this hypothesis. Approximately in the middle of the SSC (a) a sequence coverage gap is not flanked by discordant coverage (g) but only by "linking" coverage (h) suggesting that this is an actual insertion unaffected by rearrangements. The region (c) in the LSC that appears inverted is flanked on both sides by dips in coverage in the *de novo* assembly. The left-hand coverage dip (j) coincides (i) with a change in average read length and close inspection (not shown) of this region reveals a low complexity AT rich sequence and an N-filled gap in the underlying scaffold and support of sequence linkage across this gap is scanty, with only 19 read pairs supporting it. If (data not shown) one manually rearranges the contigs in the *de novo* assembly so that the apparent inversion (c) is removed, only 3 read-pairs linking the scaffolds across this junction are found, all of which end, with numerous sequence differences, in the low complexity AT rich sequences found in this junction. At the position of the junction between SSC and IR, a high number of discordant mapping (k) and "linking" (l) reads is observed, indicating that the assembly in this region is incorrect. The peak (m) in the number of clipped nucleotides coincides with a peak (n) in the "linking" reads, and there are indications (from close inspection, not shown) that this region contains an unresolved nearly exact duplication. Overall, the tracks on the *y*-axis show many problems, in

particular large numbers of “linking” read-pairs around areas without coverage in this assembly and larger numbers of clipped reads. Also, there are many short mummer alignments (not annotated in Figure 2) that are dispersed throughout the iterated mapping assembly, that are generally near the areas without coverage.

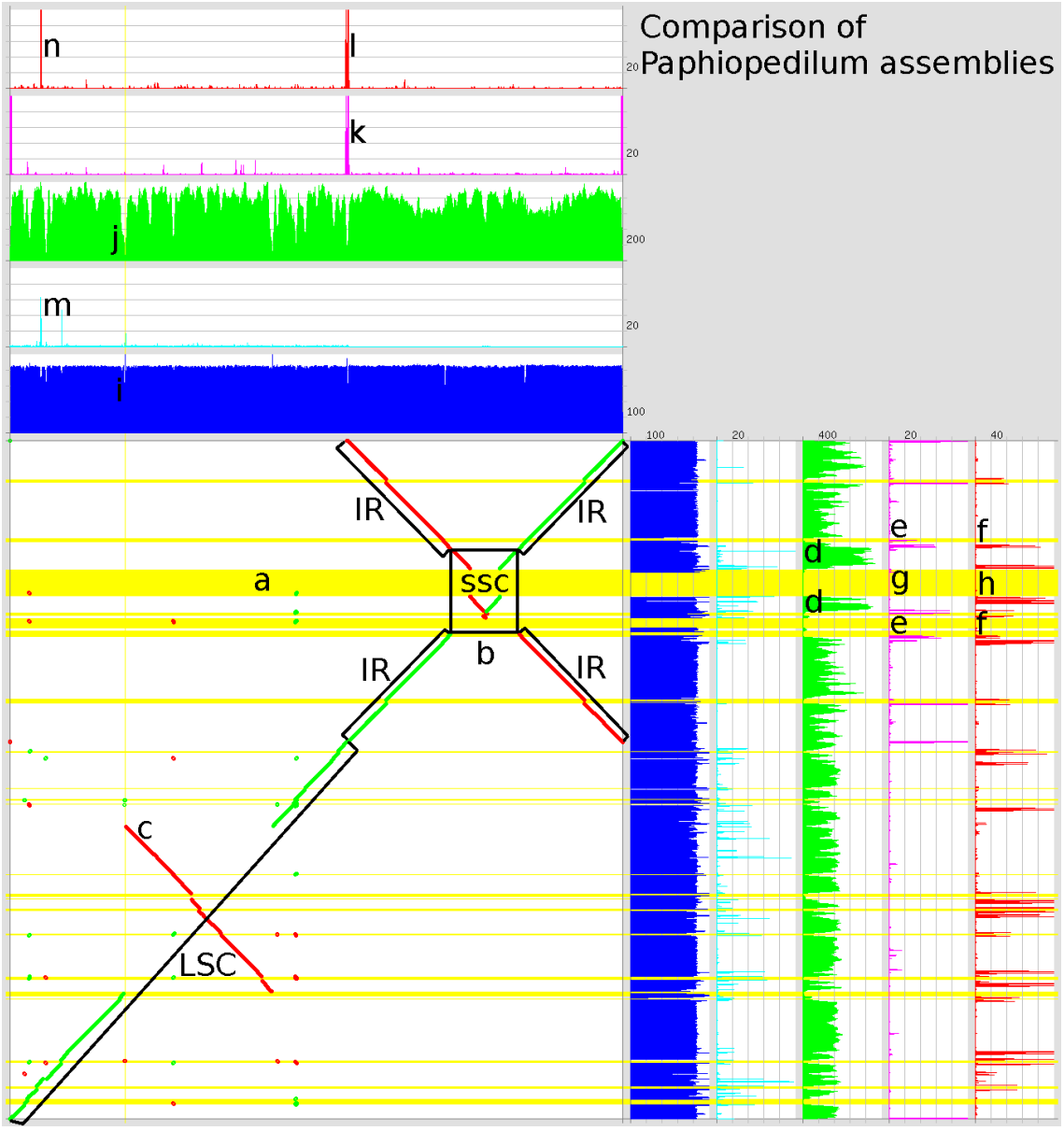


Figure 2. Comparison of the *de novo* assembly (along the *x*-axis) with the iterated read mapping assembly (along the *y*-axis) of *Paphiopedilum henryanum*. The bottom left shows the mummer alignment of the genomes, while the tracks to the right and on top show various aggregate data types. Further explanation in the text.

Discussion

Since none of the current sequence assembly algorithms is perfect, it is difficult to choose which assembly is the best. When considering the quality of an assembly, data such as read coverage of various types of reads can be useful for a researcher. In this study, we showed the effectiveness of this novel visualisation tool. This tool combines and visualizes read mapping and alignment results in a two-dimensional plot without breaking any sequence connectivity. Based on this study, we can draw a number of conclusions about the capabilities of the visualisation tool and of the iterative read mapping procedure that was used to construct alternative assemblies for comparison purposes.

In addition to the five tracks (I-V) shown, the tool can also show tracks for GC content, simple nucleotide coverage (which is different from the DNA fragment coverage for paired end data shown in the examples), mapping quality, single end mapping of paired reads and preferred read orientation. The tool can also be used to visualize magnified sections to aid more detailed analysis, can handle multiple tracks of the same type and is highly configurable: color, graph type, order, scaling and binning can all be configured. Deviations observed in each of the tracks offer various clues about what may be wrong with assemblies: Deviations in the insert size track may indicate misassembly. Clipping can indicate misassembly, larger structural differences between reference genome and the WGS sequences, and mapped WGS sequences extending into (N) gaps. Higher error rates in reads can also cause clipping. Peaks in the discordant mapping track may either be an indication of a misassembly or be an artefact caused by the circular nature of the chloroplast genome. Any peaks in the “linking” track may be an indication of linkage between separate scaffolds in an assembly or an indication that some areas may suffer from a false insertion. The coverage track may help resolve repeats and indicates where sequences absent from the target genome are found.

While there are no larger structural rearrangements between tobacco and tomato, smaller structural variations exist and can be detected through the read mapping results. The structural differences between *Paphiopedilum henryanum* and *Cypripedium japonicum* are much larger, and data offers good support for the hypothesis (Chapter 2) that the IR in orchids expanded to include the complete SSC region. Also the comparison supports the hypothesis that the *Cypripedium japonicum* chloroplast genome is significantly expanded, with novel sequences as well as dispersed repeats compared to the *Paphiopedilum henryanum* chloroplast genome. The evidence for the inverted section in the LSC region is scanty, with low and virtually absent support for connectivity on either side. Inverting this inversion (data not shown) did not resolve the issue. Many problems appear to be caused by the AT rich low complexity regions near gaps in the assembly. While difficulty sequencing through this particular template may be the cause for low sequence coverage, the problem is aggravated by difficulty assembling the lower complexity regions they represent.

Arguably the iterated read mapping-based assembly strategy is naive, ignoring structural variation, focusing on the regions where genomes are sufficiently similar and potentially introducing "SNP"-type errors in regions with structural variation. If phylogenetically sufficiently close material is used as a reference these problems may be acceptable, however, from the data presented here it is evident that neither *Cypripedium japonicum* nor tobacco are suitable references for such an approach. In addition, use of a reference that is closer to some accessions in a phylogeny and further away from some other accessions may introduce an unwanted phylogenetic bias. The exact order and orientation of all scaffolds in the *Paphiopedilum henryanum de novo* assembly is not yet known, and in particular resolving the issues surrounding the AT-rich regions may require additional work. The cyclical behavior of iterated read-mapping can in part be attributed to problems mapping reads (and indeed assembling reads) in low complexity regions. Another factor that appears to play a role is apparent heterogeneity at two loci in the *Paphiopedilum henryanum*

chloroplast genome. The exact cause of this is unclear and this warrants further investigation.

In our examples one of the assemblies was the output of our *de novo* chloroplast genome assembly pipeline (Chapter 2), and the other assembly was made using an iterated read mapping procedure. The tool is not limited to these specific comparisons; it can compare any pair of chloroplast assemblies, and a possible use might be visual assessment of several of (the multitude of) different assembly variants produced by our *de novo* assembly pipeline. Currently our chloroplast assembly pipeline only uses overall assembly size and the number of scaffolds and gaps as optimization criterion, but is may also be useful to quickly visually compare top contenders in order to pick the best option.

While other visual means for assessing assembly quality exist, for instance IGV (Robinson et al. 2011) and Tablet (Milne et al. 2013) these tools do not offer a direct in-context visualisation of a comparison between two assemblies. In addition, some tools (e.g: IGV) are more suitable for very detailed work because they either do not offer graphs showing aggregate data other than simple sequence coverage or do not show any useful data until sufficiently zoomed in. Consequently, with some tools it is very easy to lose track of the larger sequence context.

Conclusions

Here we demonstrate a visualisation tool that allows us to make a comparative assessment of the quality of two different assemblies. This comparison may be immediately useful to quickly select the best of two options for a purpose, but is also useful as it provides hints as to where specific artifacts are located. While it is targeting chloroplast genomes, it will also handle larger genomes and can produce magnified versions of specific regions, allowing efficient comparisons at any scale.

Chapter 4

Phylogenetic analysis of tomato (*Solanum* section Lycopersicon) based on various complete chloroplast genomes and subsets thereof.

Shairul Izan, Marinus J.M. Smulders, Richard Finkers, Richard G.F. Visser, Theo Borm (to be submitted)

Abstract

Whole Genome Shotgun (WGS) sequencing from total DNA has been shown to offer potential for chloroplast phylogenomics. Thanks to advances in high-throughput data handling methods, chloroplast sequences that were previously often discarded in the bioinformatics analyses now can be used to determine phylogenetic and taxonomic relationships. In the present study, we explore and evaluate a chloroplast phylogenomic approach in various species within *Solanum* section *Lycopersicon* utilizing the available WGS data from the Tomato Genome Sequencing Consortium. This enabled the alignment of 84 chloroplast genomes with several protein coding genes and noncoding regions that are potentially useful to the molecular systematic community. In particular, more than 50% of all phylogenetically relevant information was present in just four genes (*ycf1, ndhF*, *ndhA*, and *ndhH)*. Moreover, when one would only use *ycf1* one would already use 34% of all information available in the chloroplast genomes of the accessions used in this study. The topology of phylogenetic trees inferred from *ycf1* was the same as that of trees based on all other protein coding genes, although with lower bootstrap values. Moreover, we also saw that the non-coding regions contained approximately twice as many polymorphic sites per basepair compared to the coding regions. These revealed additional regions of non-coding DNA that may be explored and exploited for intraspecific phylogenetic studies. Phylogenetic analyses using different subsets (protein coding, noncoding, Single Copy and Inverted Repeats) of the chloroplast genomes successfully recovered major groups in the section with some taxon placement discrepancies. Incongruences between chloroplast genome and nuclear genome-derived phylogenies suggest ancient hybridization events or incomplete lineage sorting as the most likely explanation.

Introduction

Molecular phylogenetic analysis aims to reconstruct evolutionary relationships using DNA or amino acid sequences. Chloroplast DNA sequences have been used extensively to study plant species divergence. Accordingly, as next generation sequencing (NGS) methods advance, chloroplast phylogenomics has emerged as an effective approach to clarify phylogenetic relationships among plant species. Chloroplast phylogenomics based on the whole chloroplast genome may enhance our confidence in the phylogenies produced by increased resolution and support for relationships that remained unresolved in earlier studies based on data of a single gene or intergenic region, or a small number of genes. In plants, chloroplast phylogenomics has been reported to resolve difficult phylogenetic relationships across low and high taxonomic levels (Moore et al. 2007; Moore et al. 2010; Jansen et al. 2007; Parks et al. 2009; Xi et al. 2012; Yang et al. 2013; Barrett et al. 2013). In addition, the whole chloroplast genome has been used to characterize chloroplast evolutionary dynamics such as genome rearrangements (Cosner et al. 2004), gene loss or structural changes (Magee et al. 2010; Yi et al. 2013) and changes in gene content, order and function (Wicke et al. 2011).

Tomato (*Solanum lycopersicum*) belongs to the section *Lycopersicon* of the genus *Solanum L.* (Peralta et al. 2008). The section *Lycopersicon* is a monophyletic clade, which consists of wild and domesticated species (Moyle 2008). The taxonomic relationship of the species in the *Lycopersicon* section is controversial with many different classifications proposed (Zuriaga et al. 2009). Recently, the section was further divided into four groups: *Lycopersicon*, *Neolycopersicon*, *Eriopersicon*, and *Arcanum* (Peralta et al. 2008). Various molecular markers have been used to elucidate the evolutionary relationship among species in the *Lycopersicon* section, including chloroplast DNA (Palmer and Zamir 1982), mitochondrial DNA (McClean and Hanson 1986), nuclear RLFPs (Miller and Tanksley 1990), AFLPs (Spooner et al. 2005) and the combination of the Internal Transcribed Spacer (ITS) sequences in the ribosomal RNA genes and nuclear genes (Marshall et al. 2001;

Zuriaga et al. 2009). Nevertheless, the problem regarding their relationship and classification remained unresolved. This is largely due to the lack of informative characters available, as many species within the section *Lycopersicon* are relatively recently derived. Chloroplast phylogenomics has the potential to maximise the phylogenetic signal leading to accurate classification and increased resolution of the relationship of species in *Lycopersicon* section.

The chloroplast genome has a quadripartite structure with two single copy regions (the large single copy region, LSC, and the small single copy region, SSC) and two copies of an inverted repeat (IR) region. A simple approach in chloroplast phylogenomics would be to analyse the genome as a whole. However, using the whole chloroplast genome might be inaccurate because coding regions and intergenic regions may have a different rate of evolution (Curtis and Clegg 1984; Wolfe et al. 1987; Gaut 1998). Although noncoding sequences have been suggested to provide maximum phylogenetic signal when inferring phylogenies at lower taxonomical levels, the performance of noncoding versus coding sequences has not been well evaluated in chloroplast phylogenomic studies (Jansen et al. 2007; Moore et al. 2007; Xi et al. 2012; Barrett et al. 2013).

Tomato is an excellent system to address the various challenges of analysing the chloroplast genome in a phylogenetic framework as well for understanding the evolution of chloroplast genome. Thus, we set out to study 84 complete chloroplast genomes of species within the section *Lycopersicon*, created from WGS sequence data generated by the Tomato Genome Sequencing Consortium (Aflitos et al. 2014) using Illumina paired-end sequences. We applied our method of k-mer frequency distribution analysis (Chapter 2) to extract the chloroplast reads and perform *de novo* assemblies. These *de novo* assembled chloroplast genomes were subsequently used to conduct phylogenetic analyses. This study had two objectives. First, we aimed to explore the potential of complete chloroplast genomes in resolving the species relationship at lower taxonomic levels. Second, we investigated whether the

complete chloroplast genome could increase the phylogenetic resolution compared to subsets of sequence information from the genomes.

Materials and methods

Taxa sampled and sequence data

Paired-end genome sequencing data for 84 *Lycopersicon* accessions were obtained from the Tomato Genome Sequence Consortium and the EU-SOL project (Aflitos et al. 2014). The WGS sequencing was performed using the Illumina HiSeq 2000 platform with sequencing libraries prepared as per manufacturers' instructions, targeting an average 500 bp insert size (Aflitos et al. 2014). The 84 accessions covered the *Lycopersicon* section within the genus *Solanum L*., including wild accessions, old and modern cultivars (Table 1).

Reconstructing Chloroplast Genomes from the DNA sequences

Paired-end chloroplast reads were extracted from the raw sequence reads using a k-mer frequency-based selection, using the procedure described in chapter 2 and assembled. The procedure consists of five stages. Briefly, the first stage was to generate the k-mer table for each accession using a k-mer size of 31. Subsequently, the chloroplast paired-end reads were collected from the whole set of reads on the basis of a selection of k-mers from the k-mer table. In the second stage, the selected reads were assembled using SOAPdenovo (Li et al. 2010) with combinations of different parameters (k-mer size and amount of data). The third stage involved repeating the steps in stage two in order to refine the assembly. Stage four is to iteratively connect linear scaffolds remaining from stage three by extending and connecting scaffolds with additional sequence reads until scaffold ends overlap or by finding read-pairs spanning gaps between scaffolds. Finally, in the fifth stage gaps in the newly assembled chloroplast genome were filled.

Chloroplast Genome Annotation and Genomic Feature Extraction

The newly generated chloroplast genomes were annotated with the web-based program CPGAVAS (Liu et al. 2012) using the default parameters. The annotation program produces an output of chloroplast genome annotation in the Generic/General Feature (GFF) file format. The annotations were checked and

curated for missing annotations using Geneious software version 8 (Kearse et al. 2012). Genomic features for each of the 84 chloroplast genomes were extracted using a custom perl script.

Data subsets

The chloroplast genome has a quadripartite structure with two single copy regions (LSC and SSC), and two copies of an inverted repeat (IRs). We extracted five datasets by making selections from the complete genomes: A) The single copy regions (LSC and SSC combined). B) The IR region. C) All coding sequences (exons of protein-coding genes) concatenated. D) All non-coding sequences (intergenic regions and introns) concatenated E) All coding sequences and non-coding sequences concatenated. Chloroplast sequences in all datasets were concatenated using FasConcat-G software (Kück and Longo 2014) and aligned separately under linux using with MAFFT version 5 (Katoh et al. 2005). All alignments were performed using default settings and were visualized using Mesquite software version 3.03 (Maddison and Maddison 2008).

Phylogenetic analyses and phylogenetic tree visualisation

The phylogenetic analyses were carried out using a Maximum Likelihood (ML) analysis (a 100 iteration bootstrap (BS) through an heuristic search) which was run with RAxML-HPC2 (Stamatakis 2014) on XSEDE version 8.2.4 via the CIPRES Science Gateway Web Portal at the San Diego Supercomputer Center (Miller et al. 2010). Analyses were done with GTR+GAMMA model and default parameter settings. The produced trees were visualized with FigTree v1.3.1 (http://tree.bio.ed.ac.uk/ software/figtree).

Results

De novo assembly of chloroplast genomes

The chloroplast reads present among the raw shotgun Illumina paired-end sequences were identified and extracted based on the presence of k-mers with frequencies corresponding to the chloroplast single-copy and inverted repeat regions (as described in Chapter 2). To aid the visualisation, a series of binned k-mer volume histograms (bin size 100) were produced. Peaks corresponding to single copy chloroplast regions were present at k-mer frequencies ranging from 400 to 2000. Peaks representing the inverted repeat regions were located at k-mer frequencies ranging from 4000 to 20000 in the 84 WGS data sets analysed. After assembly, individual chloroplast genomes were obtained that were either in a single scaffold representing a complete, circular genome or in one or more scaffolds representing several (potentially incomplete) linear fragments (Table 1). All 84 genomes followed the typical quadripartite structure of flowering plants. The correct orientation of the SSC and LSC relative to each other cannot be determined using the short read sequences from Illumina. That would require an extra validation with e.g. a long range PCR.

Table 1: *De novo* assembly for selected tomato and wild species used in this study

Species	Group	Accession No	No of contigs	Circular (C)/ Linear(L)	Genome Size (bp)	LSC Length	SSC Length	IR Length	GC Content
S. lycopersicum	Lycopersicon	LA2706	1	C	155431	86026	18358	25522	36.11%
S. lycopersicum	Lycopersicon	LA2838A	6	L	155328	85679	18616	25515	36.70%
S. lycopersicum	Lycopersicon	PI406760	3	C	155306	85964	18305	25517	36.84%
S. lycopersicum	Lycopersicon	LA1090	4	C	155247	85729	18291	25612	36.39%
S. lycopersicum	Lycopersicon	EA00325	2	C	155299	86091	18287	25459	36.52%
S. lycopersicum	Lycopersicon	EA00488	4	C	155217	85891	18293	25515	36.54%
S. lycopersicum	Lycopersicon	EA00375	25	L	154265	85198	18253	25406	36.86%
S. lycopersicum	Lycopersicon	EA00371	3	C	155323	85782	18314	25612	37.12%
S. lycopersicum	Lycopersicon	LA2463	3	L	155274	85691	18358	25611	36.38%
S. lycopersicum	Lycopersicon	LYC1969	2	C	155349	85766	18358	25611	36.92%
S. lycopersicum	Lycopersicon	LYC1738	3	C	155309	85790	18292	25612	36.48%
S. lycopersicum	Lycopersicon	LYC3476	5	C	155200	85672	18301	25612	36.75%
S. lycopersicum	Lycopersicon	TR00003	6	L	154873	85546	18488	25418	36.79%
S. lycopersicum	Lycopersicon	LYC1343	2	C	155331	85801	18303	25612	36.60%
S. lycopersicum	Lycopersicon	LYC3306	4	L	155117	85531	18359	25612	36.89%
S. lycopersicum	Lycopersicon	EA01155	5	L	155149	85480	18622	25522	37.33%
S. lycopersicum	Lycopersicon	EA01049	3	C	155315	85766	18322	25612	36.76%
S. lycopersicum	Lycopersicon	LYC3153	3	C	155323	85798	18298	25612	35.84%
S. lycopersicum	Lycopersicon	EA03222	3	C	155321	85792	18302	25612	35.28%
S. lycopersicum	Lycopersicon	PI129097	4	C	155247	85881	18317	25523	36.71%
S. lycopersicum	Lycopersicon	PI272654	3	C	155278	85742	18309	25612	36.67%
S. lycopersicum	Lycopersicon	EA00990	4	L	154937	85604	18296	25517	37.38%
S. corneliomulleri	Lycopersicon	LA0118	2	L	155359	85921	18357	25539	37.48%
S. lycopersicum	Lycopersicon	EA00157	5	L	155124	85607	18290	25612	37.41%
S. lycopersicum	Lycopersicon	EA02054	4	L	155388	85824	18523	25519	37.12%
S. lycopersicum	Lycopersicon	PI303721	11	L	153051	84291	17675	25541	39.02%
S. lycopersicum	Lycopersicon	LA4451	6	C	155251	85700	18324	25612	36.20%
S. lycopersicum	Lycopersicon	V710029	5	C	155227	85700	18300	25612	36.42%
S. lycopersicum	Lycopersicon	PC11029	4	C	155248	85661	18362	25611	36.71%
S. lycopersicum	Lycopersicon	PI93302	3	C	155280	85697	18358	25611	36.44%
S. lycopersicum	Lycopersicon	SG16	5	L	154990	85448	18315	25612	36.88%
S. lycopersicum	Lycopersicon	EA01088	3	L	155247	85721	18299	25612	37.16%

S. lycopersicum	Lycopersicon	PI203232	6	L	154931	85638	18398	25446	37.06%
S. lycopersicum	Lycopersicon	PI311117	4	C	155205	85696	18290	25608	36.65%
S. lycopersicum	Lycopersicon	LA1324	13	L	154632	85292	18298	25519	36.53%
S. lycopersicum	Lycopersicon	PI158760	4	C	155242	85722	18293	25612	36.77%
S. lycopersicum	Lycopersicon	LA0113	5	L	154879	85735	18305	25418	35.67%
S. lycopersicum	Lycopersicon	LYC1410	1	L	155409	86043	18313	25525	37.02%
S. lycopersicum	Lycopersicon	PI169588	7	L	155051	85685	18335	25514	38.09%
S. lycopersicum	Lycopersicon	LYC2962	6	C	155194	85671	18296	25612	36.85%
S. lycopersicum	Lycopersicon	LYC2910	4	C	155231	85683	18321	25612	35.96%
S. pimpinellifolium	Lycopersicon	LYC2798	4	L	155057	85649	18359	25523	37.71%
S. pimpinellifolium	Lycopersicon	LYC2740	1	C	155387	85833	18327	25612	36.61%
S. pimpinellifolium	Lycopersicon	LA1584	2	C	155325	85801	18297	25612	35.90%
S. pimpinellifolium	Lycopersicon	LA1578	4	C	155314	85796	18293	25611	37.52%
S. peruvianum	Eriopersicon	LA1278	1	C	155465	85884	18354	25612	37.46%
S. chmiemlewskii	Arcanum	LA2663	2	C	155428	85842	18355	25614	37.27%
S. chmiemlewskii	Arcanum	LA2695	27	C	155422	85783	18372	25632	37.47%
S. cheesmaniae	Lycopersicon	LA0483	4	L	155115	85620	18357	25521	37.11%
S. lycopersicum	Lycopersicon	CGN15820	3	C	155317	85917	18359	25519	37.35%
S. cheesmaniae	Lycopersicon	LA1401	5	L	155152	85757	18427	25423	37.10%
S. neorickii	Arcanum	LA2133	3	C	155438	86063	18374	25499	37.76%
S. neorickii	Arcanum	LA0735	22	L	150538	83609	16617	25426	38.48%
S. arcanum	Arcanum	LA2157	2	C	155381	85970	18354	25527	37.92%
S. arcanum	Arcanum	LA2172	2	C	155243	85864	18316	25530	36.87%
S. peruvianum	Eriopersicon	LA1954	1	C	155498	86103	18316	25538	37.58%
S. huaylasense	Eriopersicon	LA1983	1	C	155451	85864	18356	25614	37.60%
S. huaylasense	Eriopersicon	LA1365	1	C	155477	85855	18363	25628	37.61%
S. chilense	Eriopersicon	CGN15532	1	C	155541	85885	18375	25639	38.33%
S. chilense	Eriopersicon	CGN15530	1	C	155496	85922	18317	25627	36.61%
S. habrochaites	Eriopersicon	CGN157591	2	C	155336	85746	18363	25612	35.77%
S. habrochaites	Eriopersicon	PI134418	3	L	154653	85176	18262	25606	36.46%
S. habrochaites	Eriopersicon	CGN157592	3	C	155279	85691	18361	25612	36.93%
S. habrochaites	Eriopersicon	LA1718	1	C	155279	85691	18361	25612	36.67%
S. habrochaites	Eriopersicon	LA1777	2	C	155283	85774	18282	25612	36.39%
S. habrochaites	Eriopersicon	LA0407	4	L	148214	85802	18333	25643	37.55%
S. habrochaites	Eriopersicon	LYC4	1	C	155421	86021	18363	25517	36.02%
S. pennellii	Neolycopersicon	LA1272	3	L	155517	86074	18362	25539	38.33%

S. pennellii	Neolycopersicon	LA0716	1	C	155254	85873	18346	25516	37.25%
S. huaylasense	Lycopersicon	LA1364	8	L	155453	86044	18356	25525	38.64%
S. lycopersicum	Lycopersicon	TR00018	3	L	155309	85924	18332	25525	37.32%
S. lycopersicum	Lycopersicon	EA00940	17	L	154463	84915	18493	25525	36.87%
S. lycopersicum	Lycopersicon	TR00019	1	C	155408	85820	18361	25612	36.09%
S. lycopersicum	Lycopersicon	EA01019	4	C	155313	85785	18301	25612	37.37%
S. lycopersicum	Lycopersicon	TR00020	15	L	155029	85625	18330	25526	37.21%
S. lycopersicum	Lycopersicon	EA01037	1	C	155416	85873	18318	25611	36.80%
S. lycopersicum	Lycopersicon	TR00021	4	C	155199	85659	18313	25612	37.88%
S. lycopersicum	Lycopersicon	TR00022	4	C	155265	85709	18329	25612	35.66%
S. lycopersicum	Lycopersicon	TR00023	2	L	155034	85712	18289	25515	37.17%
S. lycopersicum	Lycopersicon	EA01640	1	C	155412	85872	18313	25612	36.25%
S. lycopersicum	Lycopersicon	LA4133	1	L	155091	85746	18300	25521	37.79%
S. lycopersicum	Lycopersicon	LA1421	5	L	154896	85575	18288	25515	37.47%
S. galapagense	Lycopersicon	LA1044	5	C	155077	85553	18297	25612	35.08%
S. lycopersicum	Lycopersicon	LA1479	7	L	155191	85756	18596	25418	37.18%

Properties of the tomato chloroplast genomes

The assembled 84 complete chloroplast genomes have assembly sizes ranging from 148 214 to 155 541 bp long in total. The length varied from 83 106 to 86 103 bp in the LSC region, from 16 617 to 18 622 bp in the SSC region and from 25 406 to 25 643 bp in IR regions. In addition, the GC contents of the assembled genomes ranged from 36.11% to 39.02%. Furthermore, the gene content and gene order were conserved among the tomato species and accessions. The tomato chloroplast genome encodes 114 unique genes including 80 protein-coding genes, 30 transfer RNA genes and four ribosomal RNA genes (Table 2). Seventeen of these unique genes were present and duplicated in the IR, giving a total of 133 genes. 11 protein-coding genes had an intron of which the *clpP* and *ycf3* genes contained two introns. The *rps12* gene had three exons with the first exon is located in the LSC region while the second and the third exon are located in the IR region.

Contraction and expansion of IRs

In several accessions the junction between LSC and IR was located within the *rps*19 gene, resulting in partial duplication of this gene in the IR. This partial duplication consisted of various lengths (87 bp, 90 bp, 93 bp, 96 bp), differing by three nucleotides each (one amino acid). Partial duplications were observed in accessions of *S. pimpinellifolium* (LYC2798), *S. corneliomulleri* (LA0118), *S. cheesmaniae* (LA1401), *S. arcanum* (LA2157 and LA2172), *S. habrochaites* (LYC4), *S. pennelli* (LA1272 and LA0716), *S. huaylanse* (LA1364) and 26 *Solanum lycopersicum* accessions. In the other tomato species *rps 19* was not duplicated.

Table 2: Genes present in the chloroplast genome of tomato species.

Category	Group of genes	Name of genes
Self-replication, transcription and translation	Large subunit of ribosomal proteins	*rpl*2*,14,16,20,22,23,32,33,36
	Small subunit of ribosomal proteins	*rps*2,3,4,7,8,11,12,14,15,16*
	DNA dependent RNA polymerase	
	rRNA genes	*rpo*A,B,C1*,C2
	tRNA genes	*rrn*4.5,5,16,23 *trn*A-UGC,*trn*C-GCA,*trn*D-GTC,*trn*E-TTC,*trn*F-GAA, *trn*fM-CAT, *trn*G-GCC, *trn*G-TCC, *trn*H-GTG, *trn*I-CAT, *trn*I-GAT, *trn*K-UUU, *trn*L-CAA, *trn*L-TAA,*trn*L-TAG, *trn*M-CAT,*trn*N-GTT,*trn*P-TGG, trnQ-TTG,trnR- ACG, *trn*R-TCT,*trn*S-GCT,*trn*S-GGA, *trn*S-TGA, *trn*T-GGT,*trn*T-TGT,*trn*V-GAC, *trn*V-TAC, *trn*W- CCA,*trn*Y-GTA
Photosynthesis	Photosystem I	*psa*A,B,C,I,J,*ycf*3*,*ycf*4
	Photosystem II	*psb*A,B,C,D,E,F,H,I,J,K,L,M,N,T,Z
	NADH oxidoreductase	*nadh*A*,B*,C,D,E,F,G,H,I,J,K
	Cytochrome b6/f complex	*pet*A,B*,D*,G,L,N
	ATP synthase	*atp*A,B,E,F*,H,I
	Rubisco	*rbcl*
Other gene	Maturase	*matK*
	Protease	*clpP**
	Envelop membrane protein	*cemA*
	Subunit Acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
Unknown gene	Conserved Open Reading Frames	*ycf*1,2,15

*Genes containing one or two introns

Sequence divergence of structural and functional units

Regions in the chloroplast genome may evolve at a different rate. To investigate this hypothesis for structural units, we extracted each region (LSC, SSC and IR) from the 84 tomato genomes and performed alignments. The alignments of LSC and SSC were also concatenated to produce a synthetic Single Copy (SC) region. Table 3 shows number of variation sites, parsimony informative sites and nucleotide diversity found among different genomic regions. The number of variable sites refers to the total number of polymorphic sites in the region examined. In contrast, parsimony informative sites include only those variants which were detected in at least two of the sequences under study. In general, half of the variation sites were parsimonious among genomic regions. Few mutations were detected in the IR region (25 parsimony informative sites) compared to the SC region (508 parsimony informative sites). This supports the notion that the IR region is more conserved than the single copy regions.

With regards to functional units, the total protein coding sequences contained data for only 49 protein coding genes as other genes had no variable sites reported. The proportion of sequence variation (parsimony informative) in the noncoding (intron plus intergenic) sequences (356 parsimony informative sites) was twice that of the protein coding regions (211 parsimony informative sites), translating into parsimony informative site densities of 5 and 2 per kilobase for non-coding and coding regions respectively. The ratio of nucleotide diversity between functional region (protein coding: intergenic: intron) was 1:6.5:3, indicating that intergenic sequences evolved faster than the protein coding and the intron sequences, and that intron sequences evolved slower that intergenic sequences in these tomato species.

Table 3: Comparison of sequence divergence in different genomic regions of chloroplast genomes within section Lycopersicon.

Region/dataset	Number of sites	Variable sites	Parsimony informative sites	Nucleotide diversity
Structural				
Single copy region	103832	1104	508	1.06/0.49
LSC	85711	818	353	0.95/0.41
SSC	18121	286	155	1.58/0.86
Inverted repeat (one copy IR)	25460	37	25	0.15/0.1
Functional				
Coding (49 genes only*)	72807	441	211	0.61/0.29
Noncoding	70130	676	356	0.96/0.51
Intergenic	59704	659	347	1.10/0.58
Introns	10422	17	9	0.16/0.09
Total chloroplast	154753**	1117	567	0.72/0.36

* The other protein-coding genes had no variable sites.
** Average length of chloroplast genome assembly size

Parsimony informative polymorphisms

Given the privilege of having the complete chloroplast genome for all taxons included in this study, we set out to identify the most variable genes. Figure 1 shows the nucleotide variation of all 49 variable protein-coding genes in the order in which they occur in the genome. Interestingly, the *ycf1* gene is by far the fastest evolving gene in the chloroplast genome of the tomato species, containing 34% of the total amount of variation in protein-coding genes. It is followed by *ndhF*, *ndhA* and *ndhH*, which together contain another 26%. Hence, studying only these four genes means that 60% of the parsimony informative sites are already accessed. All four genes are located in the SSC region.

Non-coding chloroplast DNA sequences are an important data source for phylogenetic studies at a lower taxonomy level. Non-coding chloroplast DNA sequences are mostly intergenic, as there are only eight introns in six genes in the tomato chloroplast genome. All 121 intergenic and intron regions were ranked by parsimony informative sites as above in Figure 2. Based on this analysis we identified the top 10 variable regions. Intergenic region *rps16-trnQ-TTG* was ranked 1st, followed by *trnH-GTG_psbA*, *petN_psbM*, *psbM_trnD-GTC*, *atpH_atpI*, *rbcL_accD*, *rps4_trnT-TGT*, *ndhF_rpl32*, *matK_rps16*, and *trnS-GCT_trnR-TCT*. These regions were all located in the LSC region.

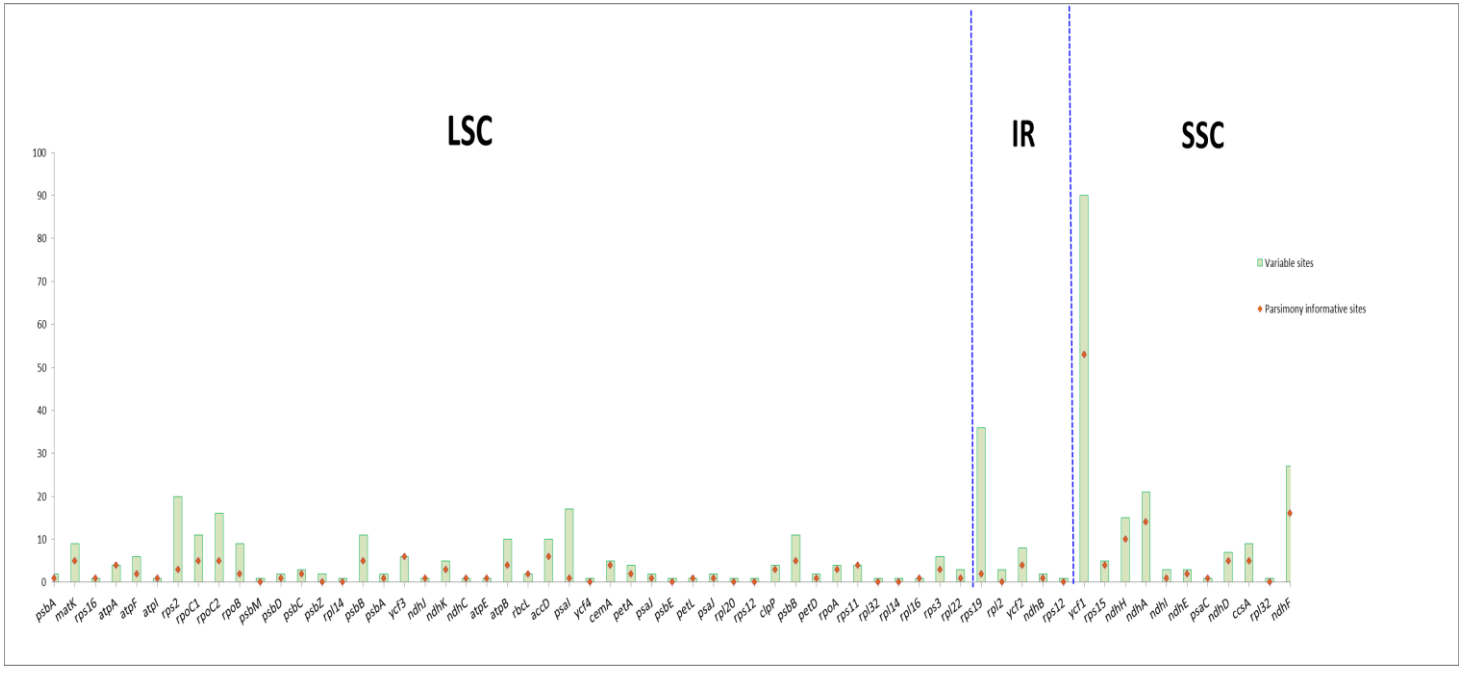


Figure 1. Number of variable and parsimony informative sites (*y*-axis) of 49 protein coding genes in the order of their position in the chloroplast genome (*x*-axis). The green bars represent the number of variable sites and the diamonds the number of parsimony informative sites. Boundaries between LSC and IR and between IR and SSC are indicated. Only one copy of the inverted repeats is included.

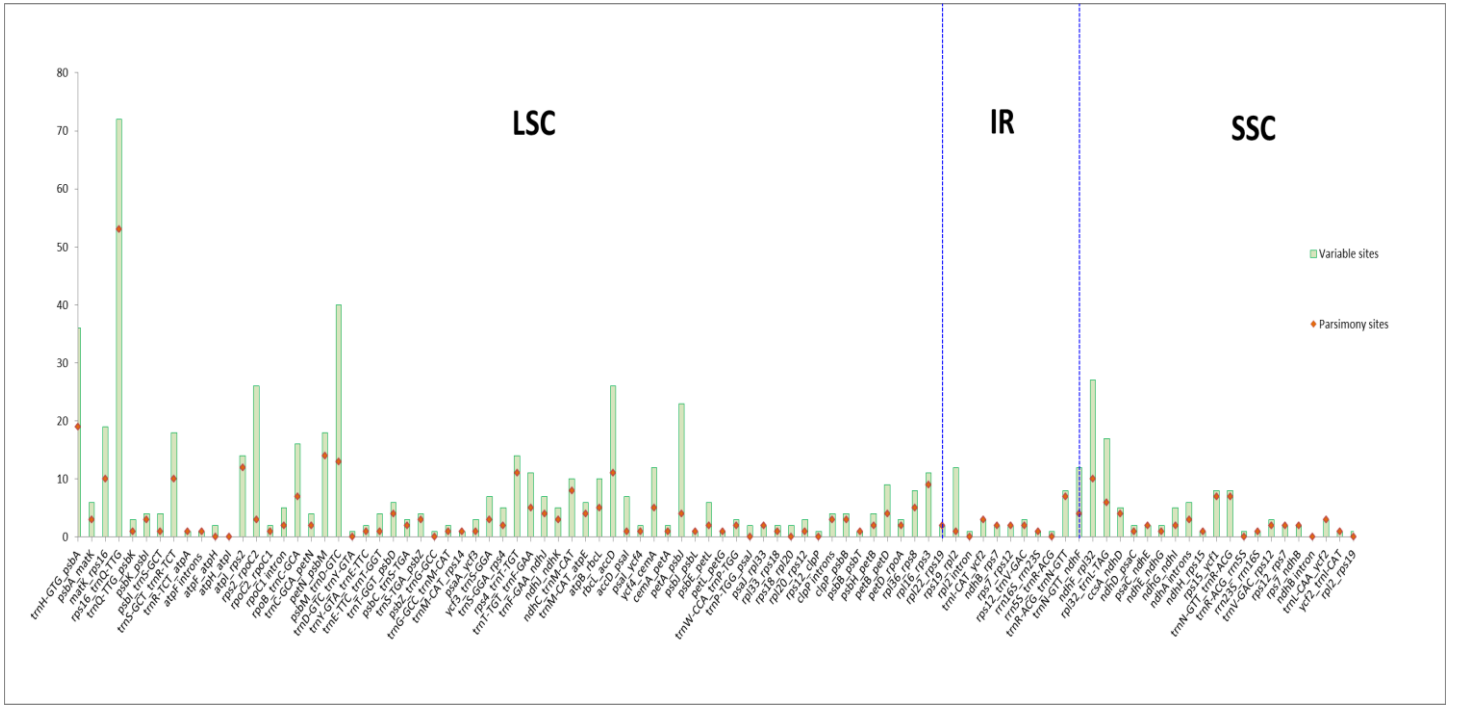


Figure 2. Number of variable and parsimony informative sites (*y*-axis) of 121 noncoding sequences (the intergenic regions and a few introns in genes) in the order of their position in the chloroplast genome (*x*-axis). The green bars represent the number of variable sites and the diamonds indicate the number of parsimony informative sites. Boundaries between LSC and IR and between IR and SSC are indicated, with only one copy of the inverted repeats included.

Phylogenetic analyses

Overall, the ML analyses of single copy, coding and noncoding data sets produced congruent phylogenetic trees in terms of topology and resolution. A comparable resolution was recovered across data sets with all major clades receiving high bootstrap support (BS 73-100%). In contrast, the IR data set was not able to resolve most of the taxa. Generally, phylogenetic resolution is correlated with the number of informative sites. The IR data set contained the least number of parsimony informative sites compared to other data sets with only 37 sites (Table 3), so we decided to ignore the IR from here on.

Phylogenetic trees of all data sets confirmed that tomatoes and their wild relatives are monophyletic (Fig. 3 – Fig. 6). Despite the long branches that connected the in-group with the out-group, a clear grouping within species was detected with some discrepancies in taxon placements across data sets. This concerned several accessions of *S. lycopersicum* that were grouped together with a red-orange fruited clade containing *S. cheesmaniae*, *S. galapagense*, *S. pimpinellifolium*, in the *Lycopersicon* group. Specifically, *S. cheesmaniae* and *S. galapagense* were clustered into a subgroup. Accession LA0483 (*S. cheesmaniae*) appears to have a close relationship with a group containing *S. lycopersicum* cultivars.

In all datasets, ML analysis resolved the *Lycopersicon* group as sister to a monophyletic clade (BS=100) containing five accessions of *S. habrochaites*. *S. habrochaites* belongs to the *Eriopersicon* group according to previous classification (Peralta et al. 2008). Nonetheless, another two accessions (LA1777 and PI134418) consistently appeared at the base of the tree and formed a highly supported clade with *S. pennelli* in data sets of protein coding (BS=100) and single copy gene (BS=73). In contrast the non-coding data set placed *S. habrochaites* (LA1777) as a sister to all groups within the *Lycopersicon* section. Peralta et al. (2008) placed *S. pennelli* in its own group (*Neolycopersicon*) rather than sister to *S. habrochaites.*

We also recovered a well-supported green-fruited clade including species from the *Arcanum* group (*S. chmielewskii*, *S. neorickii* and *S. arcanum*) which is divided into two sister groups with species within the *Eriopersicon* group (except *S. habrochaites*) across three data sets illustrating a very close relationship between both groups. While the *S. chmielewskii* and *S. neorickii* are resolved into two monophyletic groups and cluster together with *S. arcanum* (LA2172), accession *S. arcanum* (LA2157) was placed more distantly.

In all datasets, *S. arcanum* (LA2157) was a sister to the *Lycopersicon* group and *S. habrochaites* clade. Furthermore, our results also show that the *Eriopersicon* group formed two subgroups. The first subgroup consisted of *S. peruvianum* (northern) and *S. chilense* that formed a sister group with *S. peruvianum* (southern) across three data sets. The second subgroup included *S. corneliomulleri*, *S. huaylasense*, *S. chilense* and *S.spp*.

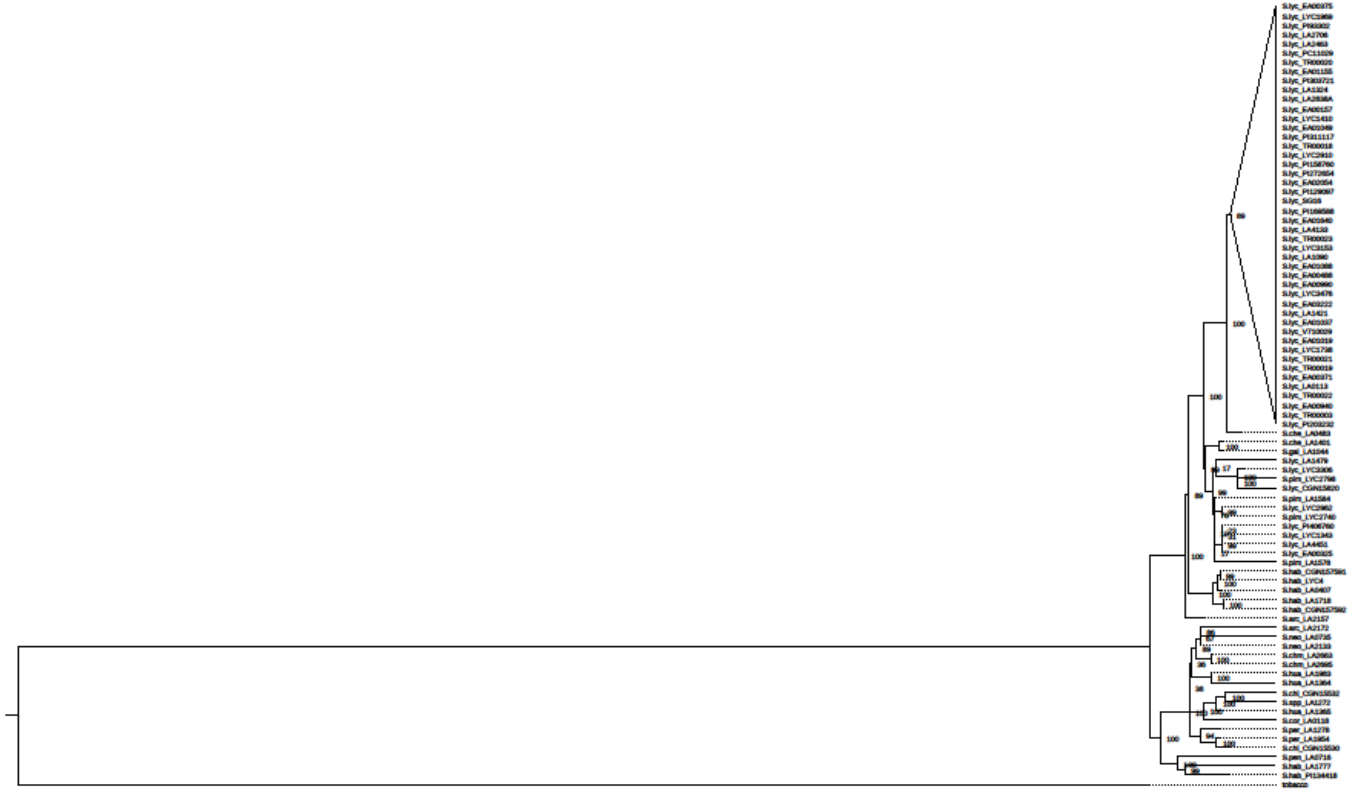


Figure 3. Maximum likelihood tree based on 49 protein coding genes. Numbers at each node indicate bootstrap support values and branch scale is shown at the base of the tree

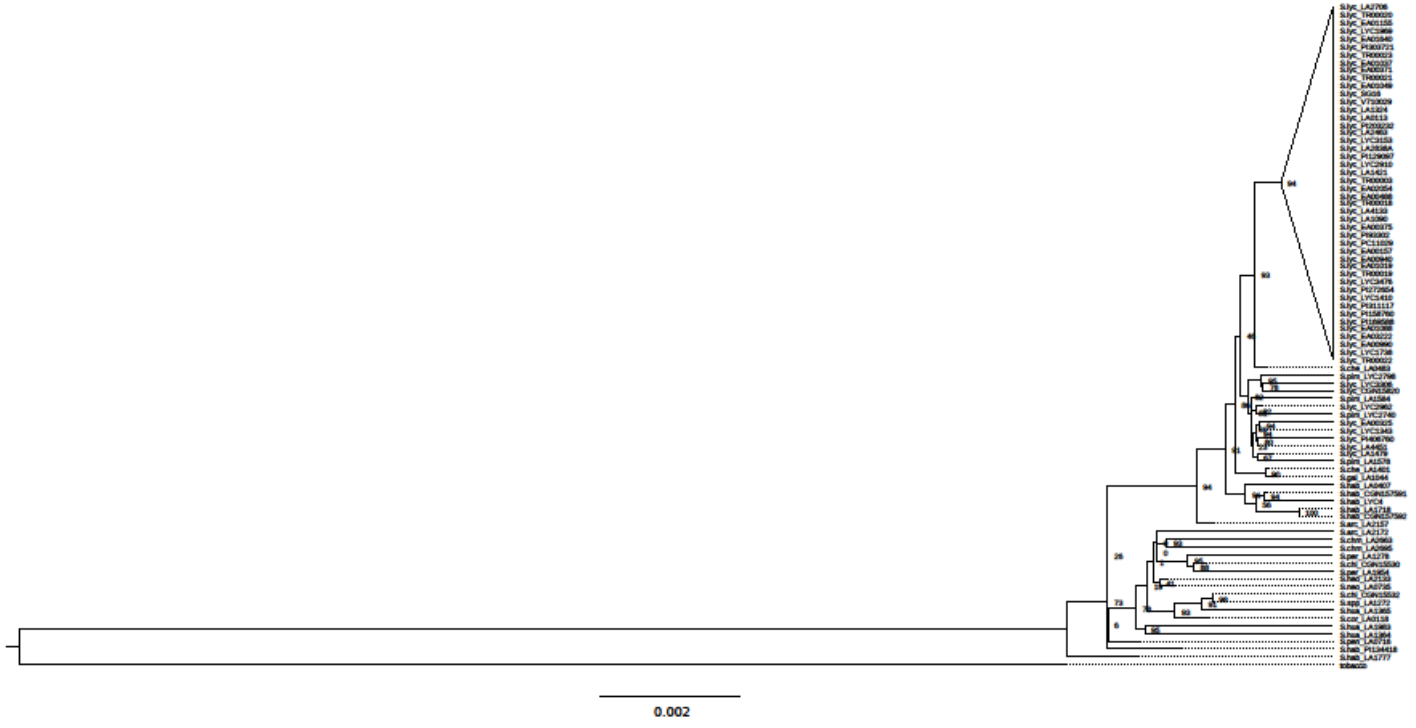


Figure 4. Maximum likelihood tree based on 121 non-coding sequences of complete chloroplast. Numbers at each node indicate bootstrap support values and branch scale is shown at the base of the tree

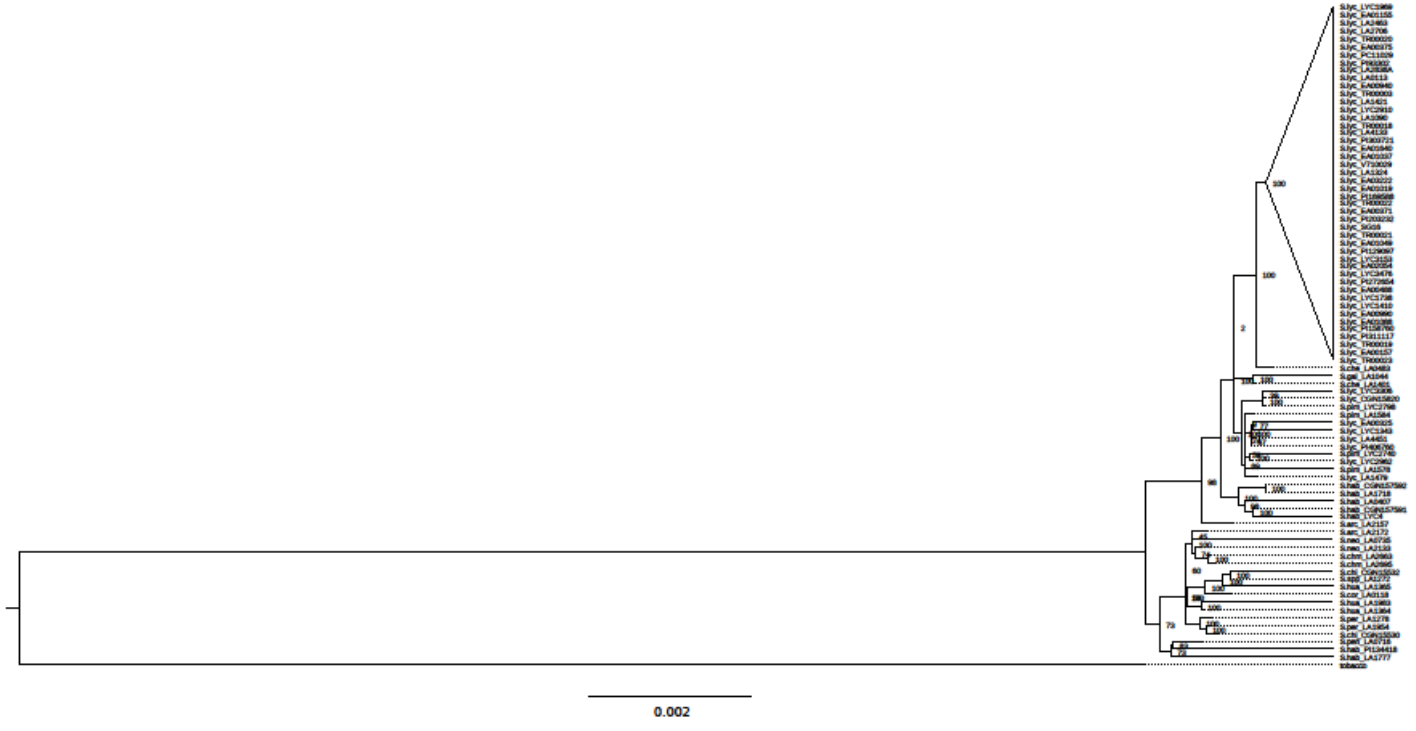


Figure 5. Maximum likelihood tree based on a synthetic single copy region (LSC and SSC concatenated). Numbers at each node indicate bootstrap support values and branch scale is shown at the base of the tree

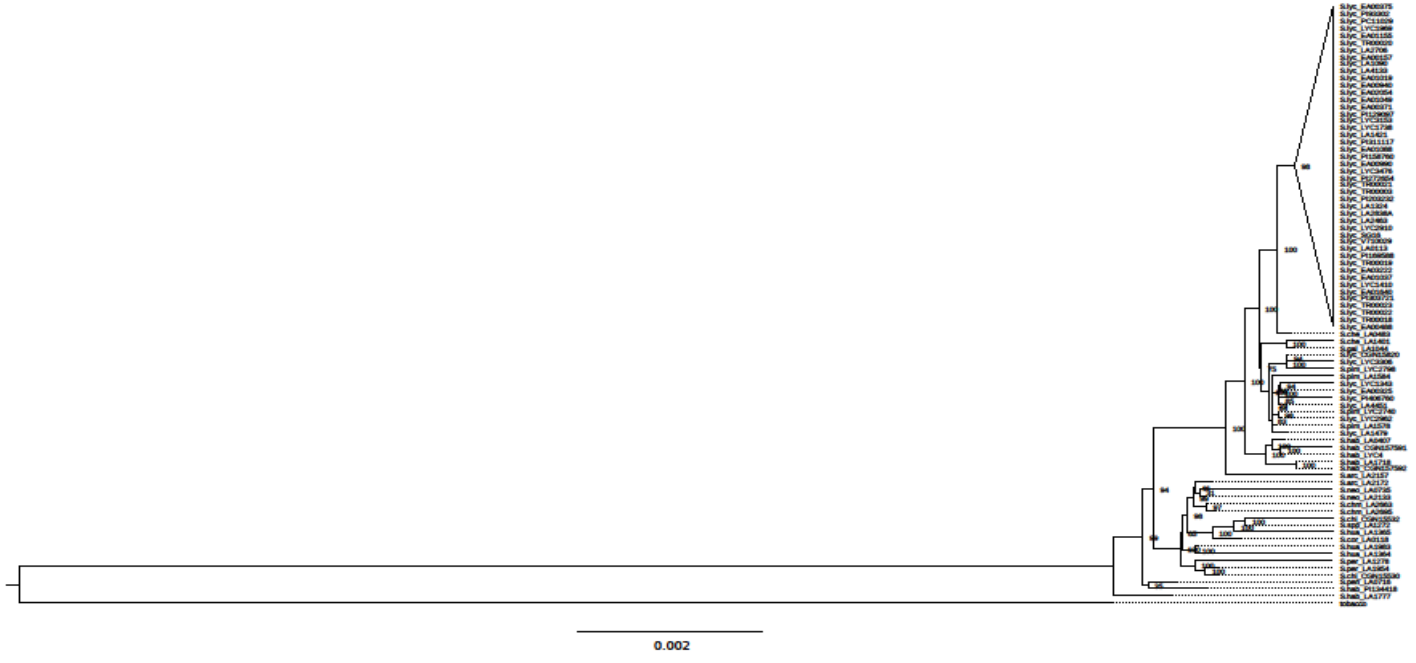


Figure 6. Maximum likelihood tree based on a matrix containing 49 protein coding genes and 121 non-coding sequences concatenated. Numbers at each node indicate bootstrap support values and branch scale is shown at the base of the tree.

Incongruent coding and noncoding data sets and the phylogeny estimation based on single loci

Visual comparison of phylogenetic topologies between protein coding sequences and non-coding sequences indicate four incongruences among phylogenies derived from protein coding and non-coding sequences data sets. Nevertheless, phylogenetic resolution in analyses of these data sets are comparable despite the fact that the non-coding data set contained nearly twice as many parsimony informative sites than the protein coding data set (356 vs 211) as shown in Table 3. These incongruences may be due to differences in the evolutionary histories of protein coding and non-coding sequences in the chloroplast genome. The non-coding phylogenetic tree appears to strongly reject the placement of *S. habrochaites* (LA 1777 and PI 134418) and *S. pennelli* (LA 0716) at the base of phylogenetic tree. Instead, only *S. habrochaites* (LA 1777) was placed at the base of the tree while accessions *S. habrochaites* (PI 134418) and *S. pennelli* (LA 0716) were placed in the polytomy with the Eriopersicum and Arcanum groups (Fig. 4). In addition, the *Arcanum* group was not recovered in the non-coding data set compared to protein coding data. One of the species from *Arcanum* group, *S. arcanum* (LA 2172) was unexpectedly conformed in the polytomy with *S. chmielewskii* and a clade consisting of *S. peruvianum* and *S. chilense.* It is worth to mention that both coding and noncoding datasets did not resolve the relationship between members of the *Eriopersicon* group (*S. huaylasense, S. chilense, S. corneliomulleri* and *S. peruvianum)* except *S. habrochaites* in one clade but instead we find them in three sub-groups. Furthermore, phylogenetic estimation based on the single *ycf1* locus was successful in recovering the major nodes when compared to the phylogenies derived from other datasets such as the protein coding sequences (Fig. 7). The major nodes in the single locus *ycf1* gene tree suffered from lack of resolution compared to the multi-locus based trees, as expected. The topology of the phylogenetic tree also did not indicate significant conflicts, but rather a polytomy resulting from unresolved relationships among members of the *Arcanum* group and some discrepancies in taxon placement.

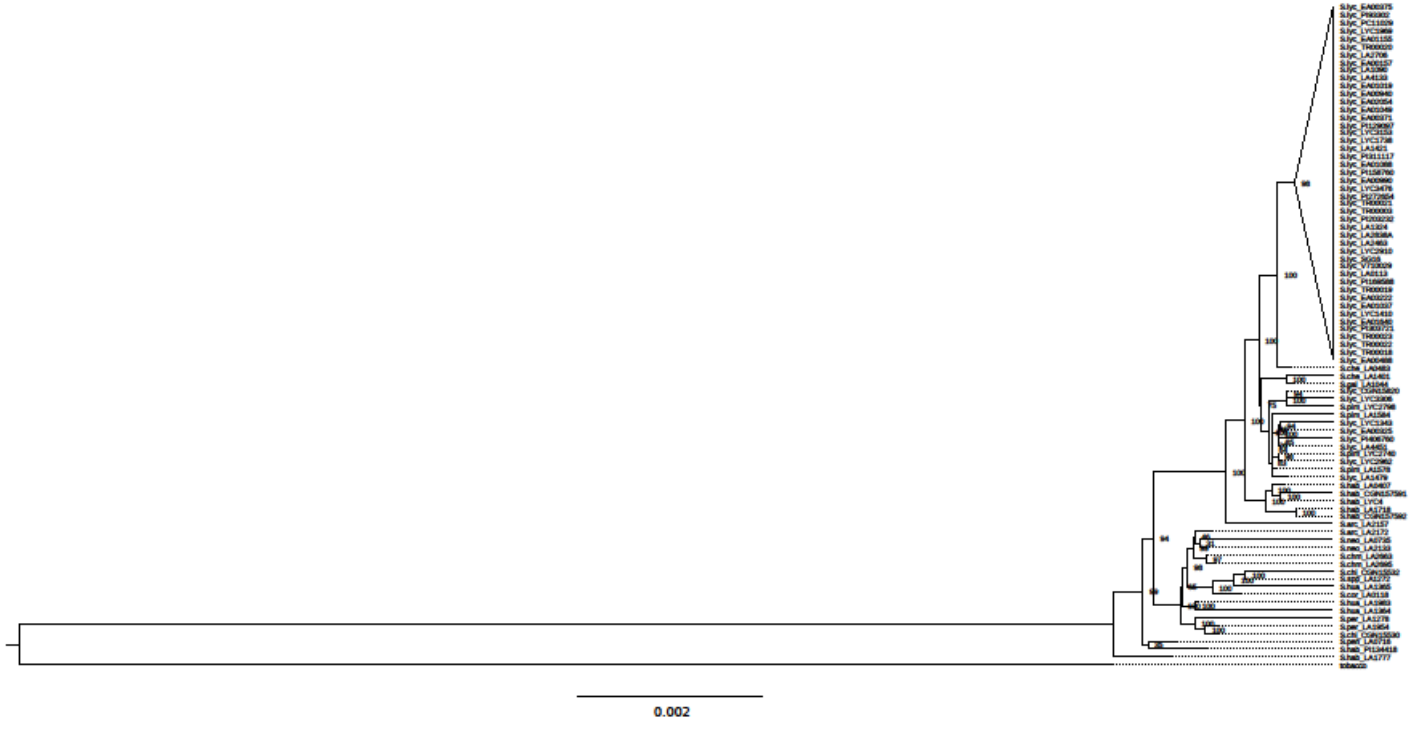


Figure 7: Maximum likelihood tree based on single gene loci, *ycf1* gene. Numbers at each node indicate bootstrap support values and branch scale is shown at the base of the tree.

Table 4. Comparison of hypotheses of relationships in the swction *Solanum* as proposed in four earlier studies with the bootstrap support values from our results (subdivided into four data subsets).

Hypotheses of relationships	I	II	III	IV	Bootstrap values in the present study			
					A	B	C	D
i) Monophyletic nature of the *Lycopersicon* group	√	√	√	√	100	86	100	100
ii) Clade recovery of *Arcanum* group includes *S. arcanum*, *S. chmielewskii* and *S. neorickii*		√		√	89	-	100	99
iii) Green fruited clade recovery (except *S. habrochaites*)		√			100	78	100	100
iv) Red-orange fruited clade recovery		√		√	99	82	100	75
v) Clade recovery consist *S. habrochaites* sister with *S. pennelli* at basal of the phylogenetic tree		√	√	√	100	-	73	-
vi) *S. habrochaites* sister to the *Lycopersicon* section	√				-	73	-	85
vii) *S. arcanum* (LA 2172) and (LA 2157) are phylogenetically distant		√			√	√	√	√
viii) *S. cheesmaniae* (LA 0483) is a sister to tomato cultivars					√	√	√	√
ix) Recovery of *S. habrochaites* clade (monophyletic and has sister relationship with the *Lycopersicon* group)					100	91	100	100
x) Northern and southern *S. peruvianum* separate into different groups		√			94	88	100	100
xi) *S. chilense* groups with southern *S. peruvianum* (LA 1954)					√	√	√	√

The earlier studies: I = Dodsworth et al. (2016); II = Aflitos et al. (2014); III = Grandillo et al. (2011); IV = Rodriguez et al. (2009)
Our datasets: A = protein coding data set (72 807 bp, 211 pasimony informative sites), B = non-coding data set (70 130 bp, 356 parsimony informative sites), C = single copy genes data set (103 832 bp, 508 parsimony informative sites), D = protein coding and noncoding data set (142 937 bp, 567 parsimony informative sites)

Discussion

In this study, we showed that shotgun DNA sequences include more than sufficient chloroplast-derived reads to be able to assemble complete chloroplast genomes. Based on the k-mer frequency distribution we took chloroplast-derived reads and *de novo* assembled 84 chloroplast genomes of tomato within the section *Lycopersicon*. No extra costs for a separate library preparation of chloroplast DNA and laboratory work were needed, nor PCR amplification prior to sequencing. In our opinion, this strategy will increase the number of available complete chloroplast genomes for all angiosperm species, especially for non-model species. Hence, chloroplast phylogenomics can be exploited to its full potential to resolve many unsettled issues in plant evolutionary studies.

Chloroplast genome organization comparison

The structure of the chloroplast genome of tomatoes was well conserved as no structural rearrangement or loss or gain of genes was detected. This is consistent with the recent divergence of these lineages. We did observe an expansion of the single copy rps19 gene from LSC into the IR, but this is quite common, especially in monocots (Wang et al. 2008). A study by Shinozaki et al. (1986) described two junctions between LSC and IR called as JLA and JLB and the DNA region surrounding JLA evolved rapidly in Nicotiana while JLB region was much more conserved (Goulding et al. 1996). In the current study, JLA lies within or near rps19 gene since we found various lengths of partial duplication of rps19. This expansion of IR may have shifted the JLA junctions. Intraspecific variation in the position of LSC and IR was first discovered in Eucalypts using RLFP by Vaillancourt and Jackson (2000). Our result suggests that the junction of LSC and IR, specifically JLA, will provide useful chloroplast genome polymorphism for the future study of intraspecific variation. This is because, according to the previous study, major changes in position of IR junctions lead to structural rearrangement elsewhere in the chloroplast genome (Perry and Wolfe 2002; Chumley et al. 2006; Haberle et al. 2008; Wicke et al. 2011).

Exploiting the whole chloroplast genome sequence for phylogenetic studies of closely related species

The need to generate large amounts of data in order to have sufficient variations among closely related plant species has hampered reconstruction of chloroplast-based phylogenies. This is partly due to high cost to produce such data. In this study, we showed this limitation can be resolved by exploiting the data from WGS DNA sequencing. However, despite the use of complete chloroplast, low intraspecific variation was observed among tomato species. The inference of phylogenetic relationships among closely related tomato species based on the complete chloroplast genome demonstrated that protein-coding genes were more conserved while non-coding sequences (introns and intergenic regions) evolved faster, as they showed 2 versus 5 parsimony informative sites per kilobases respectively. We confirmed the utility of *ycf1, rpoB, matK, rpoC1, rps16-trnQ, trnH-psbA* and *psbK-psbI* genes (Shaw et al. 2007; Neubig et al. 2009; Dong et al. 2013). Surprisingly, it turned out that several other protein-coding genes and noncoding regions contain phylogenetic information but these have never been used. Protein genes (*ndhA,B,D,H, ycf3,accD,rpoC2,psbB and ccsA*) and non-coding sequences (*petN-psbM,rbcl-accd,matK-rps16,trnS-trnR,rpl16-rps3*) have previously not been reported, but they may provide an additional set of informative protein-coding and noncoding regions in the chloroplast genome for the molecular systematics community.

Interestingly, it was possible to include more than 50% of all phylogenetically relevant information by just using four genes (*ycf1, ndhF, ndhA*, and *ndhH)*. Moreover, when one would only use *ycf1* one would already use 34% of all information available in the chloroplast genomes of the accessions used in this study. We saw in our analysis that the topology of phylogenetic trees inferred from *ycf1* was the same as that of trees based on all other protein coding genes, although the *ycf1*-only tree had lower bootstrap values (50-100%). This shows that *ycf1* alone reflects the "true" phylogenetic relationship even in closely related species, in this

case tomato. This is consistent with the study of Neubig et al. (2009). They showed that the *ycf1* gene is highly variable and phylogenetically informative at the species level. In contrast, at high taxonomy levels the *ycf1* gene has been proposed as a promising plastid DNA barcode for distantly related plant groups (Dong et al. 2015). Our results indicate that the *ycf1* gene has also great phylogenetic utility at low taxonomic levels.

Towards a refined phylogenetic classification of tomato species within *Lycopersicon* section

Various molecular markers from both chloroplast and nuclear genomes have been used to construct phylogenetic relationships within the section *Lycopersicon* (Peralta and Spooner 2001; Spooner et al. 2005; Peralta et al. 2008; Grandillo et al. 2011; Aflitos et al. 2014; Dodsworth et al. 2016). The results of our study based on 84 complete chloroplast genomes largely support the informal classification suggested by Peralta et al. (2008) (but with several discrepancies of taxon placement as discussed below). They proposed four groups within the section: (i) *Lycopersicon* group with *S. lycopersicum, S. cheesmaniae, S. galapagense* and *S. pimpinellifolium,* (ii) *Arcanum* group with *S. arcanum, S. chmielewskii,* and *S. neorickii*, (iii) *Eriopersicon* group with *S. habrochaites, S. huaylasense, S. chilense, S. corneliomulleri* and *S. peruvianum* and (iv) *Neolycopersicon* group with *S. pennelli.*

Several of these proposed relationships are supported by our study, including the monophyly of section *Lycopersicon* (Table 4). We recovered the *Arcanum* clade, a red-orange fruited clade, and a basal clade consisting of *S. habrochaites* and *S. pennelli*. Whole genome sequence data robustly suggested that the phylogenetic relationships for a large number of tomato accessions and their wild species are correlated with their geography (Aflitos et al. 2014). Indeed, our chloroplast data place the northern and southern *S. peruvianum* species in separate groups. These two geographical groups were known to possess moderate breeding barriers (Rick 1986). Furthermore, it is important to note that data in our study demonstrated a

close relationship between southern *S. peruvianum* species and *S. chilense.* This close relationship has been observed in previous studies of tomatoes species using microsatellite (Alvarez et al. 2001) and AFLP (Spooner et al. 2005).

Additionally, we observed the presence of polytomies in the reconstructed phylogenies. Polytomies or phylogenetic bushes result from poor resolution of true bifurcating relationships (soft polytomies) or from rapid speciation (hard polytomies) (Maddison 1989). Soft polytomies can be easily resolved, often by increasing the number of characters analysed, while hard polytomies cannot be resolved into bifurcating relationships (Humphries and Winker 2010). Recently Zou et al. (2008) fully resolved the relationships among diploid genome of *Oryza* using 142 single copy genes, while the relationships had remained unresolved in previous studies because of rapid speciation. They suggested that rapid speciation in an angiosperm genus can be resolved as long as a sufficient number of unlinked genes are sampled. If rapid speciation is the cause of polytomies among our reconstructed phylogenies the using unlinked genes would be able to resolve those lineages into bifurcating relationships. For this a follow-up study could combine our chloroplast information with that of several nuclear genes extracted from the genome sequences of these plants, which are being assembled.

Topological incongruence

Phylogenetic incongruence between phylogenies derived from different data sets is common in plant systematics. Various explanations have been proposed and these can be divided into two categories. The first category is incongruence that occurs because of non-biological artefacts and the second category that causes incongruence is different underlying phylogenetic histories (hybridization and introgression, lineage sorting and horizontal gene transfer).

Within this study, we detected well-supported incongruences (BS $\geq$ 70) between two data sets (protein coding and non-coding). The topologies of these data sets were contradicting with several discrepancies in taxon placements. This may be

largely explained by the different rate of evolution in combination with coding sequences being subject to selection (Muir and Filatov 2007). For example, positive selection in *rbcl* was reported to be widespread in the Hawaiian endemic genus *Schiedea* and it has been shown that the adaptive selection in *rbcl* may have driven the spread and fixation of adaptive cytotypes in several *Schiedea* species inhabiting the same island of the archipelago. This in turn created a strong incongruence between the chloroplast gene tree and the phylogeny of the genus (Kapralov and Filatov 2006). In our analysis we recovered *S. huaylasense* (accession LA1983 and LA1364) as sister to the *Arcanum* group rather than being a member of the *Eriopersicon* group like another accession of *S. huaylasense* (LA 1365). Although our finding is incongruent with genome-wide SNP data (Aflitos et al. 2014), it is congruent with the study of Rodriguez et al. (2009) that reconstructed tomato phylogeny using 19 conserved orthologous set (COSII) nuclear loci. Based on this, the incongruent *S. huaylasense* placement in our chloroplast phylogeny may be due to lack of phylogenetic signal in the chloroplast genome.

Generally, the results of our chloroplast sequence analyses are comparable with the results based on whole genome sequencing SNP data (Aflitos et al. 2014), but with some well-supported incongruences, including a distant relationship between two *S. arcanum* species where accession LA 2157 was placed in the highly supported clade of its own and has sister relationship with a clade consisting of the *Lycopersicon* group and *S. habrochaites* cluster, while accession LA 2172 was grouped together with members of the *Arcanum* group (*S. neorikii* and *S. chmielewskii*). The same goes with *S. cheesmaniae* (LA 0483) which was placed as a sister with the group containing all *Solanum lycopersicum* cultivars instead of in the red fruited clade.

The incongruences between chloroplast and nuclear phylogeny trees are often explained as indicative of hybridization and introgression events. For easier hybridization and introgression identification and detection, chloroplast markers provide additional information complementary to nuclear markers due to their maternal inheritance and non-recombinant nature. However to support these

hypotheses, and to distinguish them from the effects of insufficient phylogenetic information, as discussed above, comprehensive analyses of incongruence such as incongruence length difference test (ILD) need to be carried out to determine the cause of observed incongruences (Farris et al. 1995).

Conclusion

This study presents the first study of phylogenetic relationships among species within the section *Lycopersicon* using a chloroplast phylogenomics approach. The results of this study show that indeed shotgun sequencing data from total genomic DNA have vast potential for chloroplast phylogenomics. Our chloroplast phylogenomics approach based on 84 chloroplast genomes produced strongly supported phylogenies of the main groups within the section, with few inconsistences in taxon placement compared to phylogenies based on nuclear sequences. These differences may indicate species hybridisation event or incomplete lineage sorting, which events would not be apparent based on nuclear data only. Moreover, the non-neutrality in chloroplast genes may significantly affect tree structure and bias the inferences of phylogenetic relationship based on chloroplast DNA polymorphisms. Interestingly, most information was present in just a few genes and regions. Even better, the variation in the *ycf1* gene only was already sufficient to generate the same tree as based on the whole chloroplast.

Chapter 5

Gene loss and inversions in the chloroplast of subgenus *Paphiopedilum* (Orchidaceae) based on 32 *de novo* assembled complete organellar genomes

Shairul Izan, Theo Borm, Jing Wei Yap, Yung-I Lee, Freek T. Bakker, Barbara Gravendeel, Rogier van Vugt, Michael F. Fay, Richard G.F. Visser, Marinus J.M. Smulders (to be submitted)

Abstract

The genus *Paphiopedilum* comprises about 100 species. They occur throughout South-East Asia. While previous studies of intersectional phylogenetic relationships of subgenus *Paphiopedilum* based on selected markers failed to provide sufficient information, the analysis of complete chloroplast genomes could provide better resolution and support for the species in sections *Coryopedilum* and *Pardalopetalum*. Here, we *de novo* assembled 32 complete chloroplast genomes of slipper orchid species of the *Paphiopedilum* genus based on Whole Genome Shotgun (WGS) sequencing data. Phylogenetic analyses based on subsets of the chloroplast genomes confirm that the genus *Paphiopedilum* is monophyletic, and that the division of the genus into three subgenera *Parvisepalum*, *Brachypetalum* and *Paphiopedilum* is well supported. The division of subgenus *Paphiopedilum* into five sections was also supported. Our assemblies show that the *Paphiopedilum* chloroplast genomes contain rearrangements including gene loss and inversions. In addition, the chloroplast genome of *Paphiopedilum* has experienced IR expansion that has included part of or, in some taxa, the entire SSC region, resulting in a larger IR region compared to other monocots. These rearrangements became visible as we used *de novo* assemblies rather than mapping (or aligning) reads to a reference genome, and we advise to make this the preferred method of analysing chloroplast genomes.

Introduction

With around 25,000 species worldwide, Orchidaceae are one of the largest families of flowering plants (Chase 2005). Orchidaceae comprise five recognized subfamilies namely Apostasioideae, Cypripedioideae, Epipendroideae, Orchidoideae and Vanilloideae (Chase et al. 2003, 2014). Among the subfamilies, species of Cypripedioideae are known as slipper orchids because the lower half lip of the flower is converted into a pouch which gives the flower a lady slipper-shaped appearance. This subfamily is further divided into five genera (*Selenipedium*, *Phragmipedium*, *Cypripedium*, *Mexipedium* and *Paphiopedilum*) occupying individual geographical ranges as described in Cox et al. (1997). Of these genera, *Paphiopedilum* is the largest, comprising 96 species (Guo et al. 2015). The genus is distributed from Southeast Asia, Northern India, southern China, Myanmar, Thailand up to New Guinea (Cribb, 1998). Most species in this genus are terrestrial, but some are epiphytic or lithophytic. Over-collection of *Paphiopedilum* spp. from the wild and destruction of its habitat have brought several of the species near extinction (IUCN 2016). Recent Red List assessments have shown that 98% of *Paphiopedilum* spp. are threatened with extinction. In 2012, the Convention on International Trade of Endangered Species (CITES) listed all *Paphiopedilum* spp. on Appendix 1.

Early studies concluded that *Paphiopedilum* can be divided into three subgenera; *Brachypetalum*, *Parvisepalum* and *Paphiopedilum*. The first comprehensive study of the molecular phylogenetics of the genus was based on the nuclear ribosomal DNA internal transcribed spacer (nrITS) by Cox et al. (1997) followed by a study by Morrison et al. (2005) based on more samples. The molecular data supported the division of subgenus *Paphiopedilum* into five sections and were congruent with infrageneric treatments (Cribb 1998). However, the studies did not provide adequate resolution for phylogenetic relationships at the intersectional level. In their study, Cox et al. (1997) considered a few suggestions for the genus *Paphiopedilum*, including combining sections *Coryopedilum* and *Pardalopetalum*. The reason was that in their study *Coryopedilum* was paraphyletic to section *Pardalopetalum*. Later, Chochai et al.

(2012) constructed the phylogenic relationships of the genus based on nuclear nrITS plus four chloroplast regions and successfully recovered five sections of subgenus *Paphiopedilum* (*Coryopedilum, Pardalopetalum, Cochlopetalum, Paphiopedilum and Barbata*), consistent with the previous studies, but with better resolution. However, they observed a discordance of tree topologies based on nrITS and chloroplast DNA sequences among sections in subgenus *Paphiopedilum*. A recent study on the genus *Paphiopedilum* suggests that reticulate evolution and sea level fluctuation are important factors that contributed to the diversification of the genus (Guo et al. 2015).

Chloroplast DNA sequences are useful for plant phylogenetic and evolutionary studies. In fact, many molecular phylogenetic studies on orchids were based on chloroplast DNA sequences (Yang et al. 2013; Luo et al. 2014). The chloroplast genome is maternally inherited in orchid, thus allowing us to distinguish gene flow through seeds from that through pollen as well as the identification of species hybridization events (Bonatelli et al. 2013). The chloroplast genome is relatively conserved which permits it to be used to infer phylogenetic relationships of plant species. However, choosing appropriate molecular markers from the chloroplast genome for relevant taxonomic levels is still a challenge in phylogenetic studies. Research has consistently shown that the usefulness of the phylogenetic signal of various chloroplast DNA regions for species identification and phylogenetic studies can vary extensively among taxonomic groups (Spangler and Olmstead 1999; Wu et al. 2007; Neubig et al. 2009; Shaw et al. 2014). In the meantime, rapid developments in DNA sequencing technology have allowed researchers to generate an affordable genome-scale data collection especially for small genomes such as those of chloroplasts. As a result, chloroplast phylogenomics is emerging as an effective approach for clarifying phylogenetic relationship in plants at any taxonomic level (Philippe et al. 2005).

Whereas previous studies of intersectional phylogenetic relationships of subgenus *Paphiopedilum* based on selected markers failed to provide sufficient information, the analysis of complete chloroplast genomes could provide better resolution and support for the species in sections *Coryopedilum* and *Pardalopetalum*. Here, we used

complete chloroplast genomes, *de novo* assembled from whole genome shotgun sequences using a novel strategy, in order to improve phylogenetic resolution and increase our understanding of the molecular evolutionary of species in the sections of subgenus *Paphiopedilum*. The utilities of various chloroplast genes as phylogenetic marker for the reconstruction of the phylogenetic relationships in *Paphiopedilum* will be discussed in this chapter as well as the challenges associated with this approach.

Materials and methods

Ingroup sampling and outgroup selection

Species analysed in this study were selected to represent sections *Coryopedilum* and *Perdalopetalum* of subgenus *Paphiopedilum*. At least two species from each of the other sections (*Cochlopetalum, Paphiopedilum* and *Barbata*) were also included. We obtained 32 samples of currently recognized *Paphiopedilum* spp. Leaf material of 27 spp. was obtained from the Hortus botanicus of Leiden University, The Netherlands (HBL), Royal Botanic Gardens, Kew, United Kingdom (RBGK) and the National Museum of Natural Science, Taichung, Taiwan (NMMS). In addition, raw DNA sequencing data for seven other species were provided by RBGK and NMMS. A list of the taxa analysed, including voucher information, is given in Table 1.

DNA extraction and Illumina sequencing

Fresh leaves from 27 accessions of *Paphiopedilum* spp. were collected for DNA extraction. The leaves were either wrapped in wet tissues, snap frozen with liquid nitrogen, or stored in RNALater prior to DNA extraction. For *Paphiopedilum gigantifolium* fresh tissue from the ovary was used. DNA extractions were conducted following the Fulton method (Fulton et al. 1995) and all samples were further purified using the Qiagen spin miniprep kit (Qiagen, Venlo, The Netherlands). Although DNA extraction was generally unproblematic, it was noted that leaf tissues which were stored in RNALater turned brown during storage and yielded comparatively less DNA than the other types of samples. The total DNA of 27 samples was pair-end sequenced in two batches of Next Generation Sequencing on an Illumina HiSeq 2500. The first sequencing with 11 species was conducted in BGI, HongKong and the remaining set of species was sequenced at ServiceXS in The Netherlands, as indicated in Table 1. Library preparations were made by the sequencing company. All DNA samples were sequenced in one Illumina lane, which produced variable amounts of 2 × 125 bp paired end reads.

Chloroplast genome assemblies

The procedure for assembling chloroplast genomes was as described in Chapter 2 with some modification with regard to extracting chloroplast reads from the whole genome dataset for *Paphiopedilum* spp as many samples did not exhibit easily identifiable peaks in the k-mer frequency histogram. To overcome this issue, our first step was to start the pipeline as usual, with a k-mer size of 31 (see Chapter 2 for details), make a wide selection of k-mers for each individual accession and running the pipeline until completion of the second stage (initial assemblies), followed by extraction of a "metagenome k-mer table" from the combination of all stage 2 assemblies of all accessions As it was observed that some samples had short insert sizes, reads where then pre-processed to (i) completely remove any read-pairs with insert sizes less than 125bp – which would putatively contain adapter sequences and (ii) merge any read-pairs with internal overlap, eventually producing files with (i) merged pseudo-single-end reads, (ii) remaining unmerged forward reads and (ii) remaining unmerged reverse reads Using the "meta-genome k-mer table", putative chloroplast reads were positively selected from merged and remaining single end reads for each individual accession. At the same time, to filter out any contaminant reads, a k-mer table with contaminants was created by combining k-mers from the Phix genome sequence combined with k-mers that occurred more than 7 million times in genotype 28 and used to remove putatively contaminant reads from the dataset. The Phix genome is used as an internal QC standard in Illumina sequencing and genotype 28 is the sample from the herbarium that already failed in the initial assembly. The reads remaining after this filtering were used to assemble chloroplast genomes using the newly developed pipeline that consists of five stages. Briefly, in the first stage a k-mer table was generated for each genotype and chloroplast reads were collected on the basis of a selection of k-mers from this table. In the second stage, the chloroplast genome was *de novo* assembled using the SOAPdenovo assembler (Li et al. 2010) using combinations of different parameters (k-mer size and amount of data). The third stage involved repeating the steps in stage two in order to refine the assembly. Stage four is to iteratively connect linear scaffolds remaining from stage three by extending and connecting scaffolds with

additional sequence reads until scaffold ends overlap or by finding read-pairs spanning gaps between scaffolds. Finally, in the fifth stage gaps in the newly assembled chloroplast genome were filled.

Chloroplast genome annotation and genomic feature extraction

The newly generated chloroplast genomes were annotated with the web-based program CPGAVAS (Liu et al. 2012) using default parameters. The annotation program produces a chloroplast genome annotation in the Generic/General Feature (GFF) file format. The annotations were checked and curated by eye for missing annotations using Geneious software version 8 (Kearse et al. 2012). Genomic features for each of the 32 chloroplast genomes were extracted using a custom Perl script.

Data subsets

The general chloroplast genome has a quadripartite structure with two single copy regions (LSC and SSC), and two copies of an inverted repeat (IRs) (Saski et al. 2005). We extracted five datasets by making selections from the complete genomes: A) The LSC region; B) The SSC region; C) The IR region; D) All coding sequences (exons of protein-coding genes) concatenated; and E) All non-coding sequences (intergenic regions and introns) concatenated. Chloroplast sequences in all datasets were concatenated using FasConcat-G software (Kück and Longo 2014) and aligned separately under Linux using MAFFT version 5 (Katoh et al. 2005). All alignments were performed using default settings and visualized using Mesquite software version 3.03 (Maddison and Maddison 2008).

Phylogenetic analyses and phylogenetic tree visualisation

The phylogenetic analyses were carried out using a Maximum Likelihood (ML) criterion (including a 100 iteration bootstrap through a heuristic search) which was run with RAxML-HPC2 (Stamatakis 2014) on XSEDE version 8.2.4 via the CIPRES Science Gateway Web Portal at the San Diego Supercomputer Center (Miller et al. 2010). Analyses were done with the GTR+GAMMA model and

default parameter settings. The trees produced were visualized with FigTree v1.3.1 (http://tree.bio.ed.ac.uk/ software/figtree).

Results

Assemblies of *Paphiopedilum* chloroplast genomes and their quality

The Illumina sequencing runs delivered variable numbers of reads as shown in Table 1. After the pre-processing step, which was included to eliminate overlapping reads, the number of reads was reduced significantly. Read pairs that we used as an input in the assembly ranged from 22903 to 5196111. Assemblies of *Paphiopedilum* chloroplast genomes based on reads that were filtered resulted in better assemblies in terms of the number of scaffolds generated. We observed that the improvements were more pronounced for the samples in the second batch of sequencing (which generally had poorer quality in terms of length, quality and amount of reads). Moreover, we noticed that genotypes sequenced from the tissues stored in RNA*Later* had low quality assemblies compared to the assemblies from tissues that were snap-frozen in liquid nitrogen or transported in wet tissue after harvesting. Genotypes identified as low quality assemblies included *Paphiopedilum randsii, P. adductum, P. gigantifolium, P. ooii, P. sanderianum, P. barbatum,* and *P. primulinum*. Aside from *P. barbatum* and *P. primulinum*, all of these genotypes were sequenced from tissue stored in RNA*Later*. Moreover, DNA concentrations of four genotypes (*P. adductum, P. gigantifolium, P. ooii, P. sanderianum)* had a low concentration of 2 to 17 nM, whereas the average concentration was 50 nM.

Table 1. List of *Paphiopedilum spp* included in this study and associated sequencing data and assemblies

Species	Verification P/V	Source/ Reference	Sample preparation	Raw reads (pairs)	Read pairs after pre-process	Filtered raw reads (pairs)	# scaff	# bases	# gaps	# non N-bases
P. adductum	V	NMNS/Yung-I Lee 201351	fresh in RNAlater	13700672	5541031	69902	19	134247	3	133959
P. appletonium	V	NMNS/Yung-I Lee 201209	WGS data	169159653	164143259	2319386	16	115727	18	115282
P. armenicum	V	NMNS/Yung-I Lee 201101	WGS data	144745237	136654109	5196111	5	129872	22	129621
P. barbatum	V	HBL/HBL 200070142	fresh in liquid nitrogen	15097287	14144700	348315	6	121844	2	121612
P. barbatum	n.a	HBL /L 0717340	herbarium samples	12104266	3637199	53266	0	0	0	0
P. concolor	V	NMNS/Yung-I Lee 201201	WGS data	146030085	144891128	2372607	4	177291	18	176468
P. druryi	P	HBL/HBL 201130012	fresh in liquid nitrogen	15595975	15370584	185195	7	118895	7	118433
P. fairrieanum	V & P	HBL/HBL 20110265	fresh in liquid nitrogen	17785274	16384615	229565	6	159693	3	159464
P. gigantifolium	V	NMNS/Yung-I Lee 201352	Ovary in RNAlater	15305745	4859365	37449	35	167922	38	162530
P. glanduliferum A	V & P	HBL/HBL 1071	fresh in liquid nitrogen	14263865	14133124	193879	11	135055	2	134837
P. glanduliferum B	V	HBL/KAS 2132371	fresh in liquid nitrogen	14410950	13861450	249755	11	133145	4	132869
P. haynaldianum	n.a	RBGK/1956-1830	fresh in wet tissue	17559966	14469864	330770	4	136647	5	136144
P. henryanum	V & P	HBL/HBL 20110270	fresh in liquid nitrogen	15142939	14926320	258560	2	155644	1	155634
P. hirsutissimum	P	RBGK/2000-1376	fresh in wet tissue	12277791	8003851	109101	14	150951	3	150921
P. kalopakingii	n.a	RBGK/1983-5478	fresh in wet tissue	34660971	31858576	686700	14	156809	4	156604
P. lowii	V	HBL/HBL 30629	fresh in liquid nitrogen	16821602	16417626	192175	6	121408	10	121300
P. micranthum	n.a	RBGK/1990-192	fresh in wet tissue	13185768	8895672	477078	15	153731	13	152646
P. ooii	V	NMNS/Yung-I Lee 201353	fresh in RNAlater	18352797	4975425	52119	29	163418	5	162978
P. parishii	P	RBGK/1986-1038	fresh in wet tissue	12312236	8461016	156797	17	134530	5	133879
P. phillipinense	n.a	RBGK/1956-1830	fresh in wet tissue	36557062	33091749	464656	12	131105	4	130747

P. praestan	n.a	RBGK/2010-1562	fresh in wet tissue	24491155	21970594	491154	8	161920	6	161815
P. primulinum	P	RBGK/K19811628	WGS data	1805388	1741678	22903	9	163364	19	161724
P. purpuratum	V & P	HBL/HBL 20110252	fresh in liquid nitrogen	14404250	14022616	295859	11	157178	0	157178
P. randsii	V	NMNS/Yung-I Lee 201355	Fresh in RNAlater	32563796	9318689	64578	40	149061	1	148954
P. rothschildianum	V	NMNS/Yung-I Lee 201107	WGS data	88095574	86863413	1238413	5	123386	4	123039
P. sanderianum	V	NMNS/Yung-I Lee 201108	fresh in RNAlater	40031573	6453545	176973	69	105953	45	99117
P. spp_1	n.a.	HBL/KAS 7960762	fresh in liquid nitrogen	15310352	14978748	399158	8	124007	3	123819
P. spp_2	n.a		fresh in liquid nitrogen	13828124	13623736	389871	8	125135	2	125124
P. stonei	n.a	RBGK/2001-198	fresh in wet tissue	19644224	13774001	442149	14	134142	15	132591
P. supardii	V	NMNS/Yung-I Lee 201354	Fresh in RNAlater	13210763	8566266	108745	16	157213	6	156587
P. victoria-regina	n.a	RBGK/1984-3498	fresh in wet tissue	21527366	19185126	272763	13	150113	15	149185
P. villosum	V	NMNS/Yung-I Lee 201012	WGS data	140006936	137576718	1973494	2	157769	17	157600
P. wilheminiae	P	RBGK/1953-38501	fresh in wet tissue	14430411	9942840	352852	19	108485	8	107827
Phrag. longifolium	V & P	HBL/HBL 20110235	fresh in liquid nitrogen	13286105	13067512	477902	5	162363	11	162253

Legends:

first batch sequencing		HBL	Hortus Botanicus of Leiden University, The Netherlands
second batch sequencing		RBGK	Royal Botanic Gardens, Kew, United Kingdom
WGS data from whole sequencing		NMMS	National Museum of Natural Science, Taichung, Taiwan
Herbarium sample			

Genome features and loss of the *ndh* gene complex in the chloroplast genomes of *Paphiopedilum* species

The *Paphiopedilum* chloroplast assembly size ranged from 143 529 to 167 381 bp. The chloroplast genomes of *Paphiopedilum* encoded a set of 109 of genes comprising 68, 27 and four protein coding, transfer RNA and ribosomal RNA genes, respectively. Six protein coding genes, namely *atpF, rpl2, rpoC1, clpP, rps12* and *ycf3* contained introns. In general, the genome features of the *Paphiopedilum* genomes analysed in this study were similar in terms of gene content, gene order, introns and intergenic spacers.

The newly assembled *Paphiopedilum* chloroplast genomes exhibited varying degrees of *ndh* gene family losses. Generally, chloroplast genomes contain 11 *ndh* genes that encode for NADH dehydrogenase subunits (Kim et al. 2015). Using blast of the reference gene sequences annotated from the *Cypripedium japonicum* (NC_027227) chloroplast genome, we found five *ndh* genes (*ndhB, ndhC, ndhD, ndhK, ndhJ*). However, these genes possessed incomplete protein sequences or premature stop codons yielding non-functional genes. Table 3 shows the length variation of gene sequences obtained from the blast. The remaining six *ndh* genes (*ndhA, ndhE, ndhF, ndhG, ndhH, ndhJ*) were completely absent, suggesting that they were fully deleted from the chloroplast genome of *Paphiopedilum* spp. This result suggests that there are no functional *ndh* genes left in the chloroplast genome of *Paphiopedilum* spp.

Table 2. Length of various regions in the chloroplast genomes of *Paphiopedilum* spp.

Species	LSC	SSC	IR	Total length*
P. adductum	80968	10266	27548	118782
P. appletonianum	77883	10434	26945	115262
P. armeniacum	91392	11224	26003	128619
P. barbatum	83596	10773	27475	121844
P. concolor	81851	10807	25760	118418
P. druryi	80489	9732	28151	118372
P. fairrieanum	84835	10757	25563	121155
P. gigantifolium	82312	10122	26037	118471
P. glanduliferum A	81111	10389	25481	116981
P. glanduliferum B	81115	10530	25189	116834
P. haynaldianum	86413	10659	25027	122099
P. henryanum	84902	5513	25427	115842
P. hirsutissimum	79527	10683	24386	114596
P. kolopakingii	83151	10231	25431	118813
P. lowii	84127	10398	26883	121408
P. micranthum	81335	9706	28008	119049
P. ooii	82127	9979	25699	117805
P. parishii	82300	9877	24867	117044
P. phillipinense	80495	9059	25438	114992
P. praestan	78730	10018	25227	113975
P. primulinum	84613	11016	25873	121502
P. purpuratum	83788	10514	25230	119532
P. randsii	70024	9723	27581	107328
P. rothschildianum	86279	8111	25787	120177
P. sanderianum	42235	6137	23098	71470
P. spp_1	83786	10773	28719	123278
P. spp_2	85940	10660	26619	123219
P. stonei	81192	10114	25468	116774
P. supardii	82117	10364	27332	119813
P. victoria-regina	80749	10904	27377	119030
P. villosum	85688	11178	24417	121283
P. wilheminiae	69615	9736	25771	105122
Phrag longifolium	91397	12646	25059	129102

*total chloroplast genome including only one copy of IR

Table 3. Presence and length variation in *ndh* genes among sequenced *Paphiopedilum* spp.

Species	*ndhA*	*ndhB*	*ndhC*	*ndhD*	*ndhE*	*ndhF*	*ndhG*	*ndhH*	*ndhI*	*ndhJ*	*ndhK*
Cypripedium japonicum	2,386	2,242	363	1,506	306	2,214	531	1,182	495	477	699
P. abductum		1,508		1,150						355	200
P. appletonium		2,237		6,103						479	204
P. armenicum		2,225	363	3,875						489	704
P. barbatum		2,219		1,220						486	204
P. concolor		2,132	363	1,209						489	703
P. druryi		2,237		1,220						488	200
P. fairrieanum		2,234		1,220						377	200
P. gigantifolium		1,508		1,150						355	200
P. glanduliferum A		1,347		1,150						359	
P. glanduliferum B		1,508		1,150						355	
P. haynaldianum		1,490		1,150						331	200
P. henryanum		2,237		1,220						352	200
P. hirsutissimum		2,244		1,218							
P. kalopakingii		1,507		1,150						355	200
P. lowii		1,490		1,150						350	200
P. micranthum		2,238		1,224						489	703
P. ooii		1,508		2,712						355	200
P. parishii		1,495		616						349	207
P. phillipines		1,508		1,202						355	200
P. praestan		2,237		1,220						488	200
P. primulinum		2,237		1,220						489	200
P. purpuratum		2,237		1,220						487	204
P. randsii		1,508	363	1,150						307	200
P. rothschildianum		1,508		1,150						355	200
P. sanderianum		1,508		1,198						88	200
P. spp_1		2,238		1,218						489	132
P. spp_2		2,238	363	1,219						489	706

P. stonei		1,508		1,150						355	200
P. supardii		1,508		1,150						355	200
P. victoria-reginae		2,237		1,218						488	200
P. villosum		2,237		1,220						488	200
P. wilheminiae		1,508		1,150						355	

*Numbers in cells refer to gene or pseudogene length. Colors: White = full length, in frame; yellow = pseudogenized; red = completely deleted.

Structural rearrangements in the chloroplast genomes of *Paphiopedilum* spp.

To identify the possible occurrence of structural rearrangements in the chloroplast genomes, a combination of dotplots and read coverage graphs was generated for each sample. For each sample, the raw reads were mapped against the original assembly and edited assembly. The original assembly is the assembly that resulted from the *de novo* assembly pipeline and the edited assembly is the assembly that is co-linear with the reference genome *Phragmipedium longifolium* (KM032625). Assembly editing was done to facilitate the phylogeny analysis in the latter stage. Examples of the dotplots for one *Paphiopedilum* chloroplast genome as shown in the Appendix 1 and the summary of each dotplot is shown in Table 4. Table 4 shows a summary of read coverage of the assemblies, structural rearrangements and evidence of misassemblies generated from the dotplot and the read coverage graphs that were produced to detect structural rearrangements in the *Paphiopedilum* chloroplast genomes analysed in this study. Low read coverage was detected from the assemblies of *Paphiopedilum randsii, P. adductum, P. gigantifolium, P. ooii* and *P. sanderianum,* which made it difficult to conclude if any structural rearrangements had occurred in the chloroplast genome of those samples. We also identified a low quality sequencing library for those samples with low DNA concentration as shown in the excessive variation of coverage along the original and the edited assemblies. The read coverage of other *Paphiopedilum* spp. included in this study was within medium and high coverage.

Remarkably, within this study, inversions in the LSC region and IR expansions were detected in the new assemblies of *Paphiopedilum spp*. The first rearrangement was an inversion in the LSC region. The inversions were detected by aligning the original assembly against the edited assembly. The inversions in the LSC region could be detected by the presence of a reverse complement match - a diagonal line from higher left to lower right in the dotplots within the LSC region boundaries. Alignments against the edited assemblies revealed that *P. micranthum*, *P. parishii*, *P. haynaldianum, P. stonei*, *P. phillipinense*, *P. praestans, P. druryi, P. fairrieanum, P. glanduliferum B, P. spp_1, P. concolor and P. rothschildianum* have inversions in the LSC between 3177 to 28737 bp in size. Additionally, although the inversions were

supported by the read mapping, the boundaries of the inversions were not consistent throughout the *Paphiopedilum* samples analysed.

Another structural rearrangement detected by the read mapping was an expansion of the IR. IR expansions into SSC region of about 8 kb to 11 kb were observed in all *Paphiopedilum* spp. analysed. The read coverage of the SSC region doubled when there was an IR expansion into the SSC region in both the original and edited assemblies of most of the species. However, the IR expansion could not be confidently determined for the assemblies of *Paphiopedilum adductum, Paphiopedilum gigantifolium, Paphiopedilum ooii* and *Paphiopedil*um *sanderianum* due to low coverage in the overall read mapping.

Table 4. Summary of the dotplots and read coverage of *Paphiopedilum* species analysed in this study.

Species	Coverage			Structural rearrangements						Evident misassemblies	
				Inversion in LSC		IR expansion into whole SSC		IR expansion into part SSC			
	low	medium	high	Original assembly	Supported	Original assembly	Edited assembly	Original assembly	Edited assembly	Original Assembly	Edited Assembly
P. adductum	√										
P. appletoniaum			√	x		√	√	x	x	x	x
P. armeniacum		√		x		√	√	x	x	√	x
P. barbatum		√		x		√	√	x	x	x	x
P. barbatum(herbarium)											
P. concolor		√		√	√	√	√	x	x	x	x
P. druryi			√	√	√	√	√	x	x	x	x
P. fairrieanum			√	√	√	√	√	x	x	x	x
P. gigantifolium	√										
P. glanduliferum A			√	x		√	√	x	x	x	x
P. glanduliferum B			√	√	√	√	√	x	x	x	x
P. haynaldianum			√	√	√	√	√	x	x	x	
P. henryanum			√	x		√	√	x	x	x	x
P. hirsutissimum			√	x		√	√	x	x	x	

P. kolopakingii		√		x		√	√	x	x	x	x
P. lowii			√	x		√	√	x	x	x	x
P. micranthum		√		√	√	√	√	x	x	x	
P. ooii	√										
P. parishii			√	√	√	x	x	√	√	x	√
P. phillipinense		√		√	√	√	√	x	x	x	x
P. praestans			√	x		√	√	x	x	x	x
P. primulinum		√		x		√	√	x	x	x	x
P. purpuratum			√	x		√	√	x	x	x	x
P. randsii	√			x		x	x	x	x	x	x
P. rothschildianum		√		√		√	√	x	x	x	x
P. sanderianum	√										
P. spp_1			√	√	√	√	√	x	x	x	x
P. spp_2			√	x		√	√	x	x	x	x
P. stonei			√	√	√	√	√	x	x	x	x
P. supardii			√	x		√	√	x	x	x	x
P. victoria-reginae		√		x		√	√	x	x	√	√
P. villosum		√		x		√	√	x	x	x	x
P. wilheminiae			√	x		√	√	x	x	x	√
Phrag. longifolium			√	x		x	X	x	x	x	x

*red highlighted: read coverage was too low or the assembly failed

Phylogenetic relationships among *Paphiopedilum* spp. analysed

Phylogenetic relationships were analysed using five data subsets, representing structural and functional regions of the chloroplast genomes of the *Paphiopedilum spp*. analysed. The relationships found were largely congruent between the data subsets, as is visible in the maximum likelihood (ML) trees of the five datasets as shown in Figure 1 and Figure 2. All trees successfully recovered the monophyly of the genus *Paphiopedilum* with a bootstrap (BP) value of 100. In all datasets, subgenus *Parvisepalum* diverged first, followed by subgenus *Brachypetalum* and finally subgenus *Paphiopedilum*, always supported with 100 BP. In subgenus *Paphiopedilum*, two major lineages could be observed. The first lineage was composed of three sections: *Coryopedilum*, *Pardalopetalum* and *Cochlopetalum*. In three datasets (protein coding, non-coding and LSC region) we successfully recovered a clade for section *Cochlopetalum* with moderate bootstrap support (64 to 100). A close relationship between sections *Coryopedilum* and *Pardalopetalum* was observed as together they formed a subclade with high bootstrap support. The second lineage in subgenus *Paphiopedilum* was formed by species of sections *Paphiopedilum* and *Barbata*. The division of the second lineage into sections *Paphiopedilu*m and *Barbata* was strongly supported (82-100 BP) in all ML analyses in all datasets except for the SSC dataset (16 BP). However, across datasets we observed topological incongruences of the accessions of *P. fairrieanum* and *P. hirsutissimum*, which were grouped in section *Paphiopedilum* by Cribb (1998). For example, in the protein coding dataset, *P. fairrieanum* formed a subclade with other species from section *Paphiopedilum* whereas in the non-coding dataset, it was placed as sister clade to section *Barbata.*

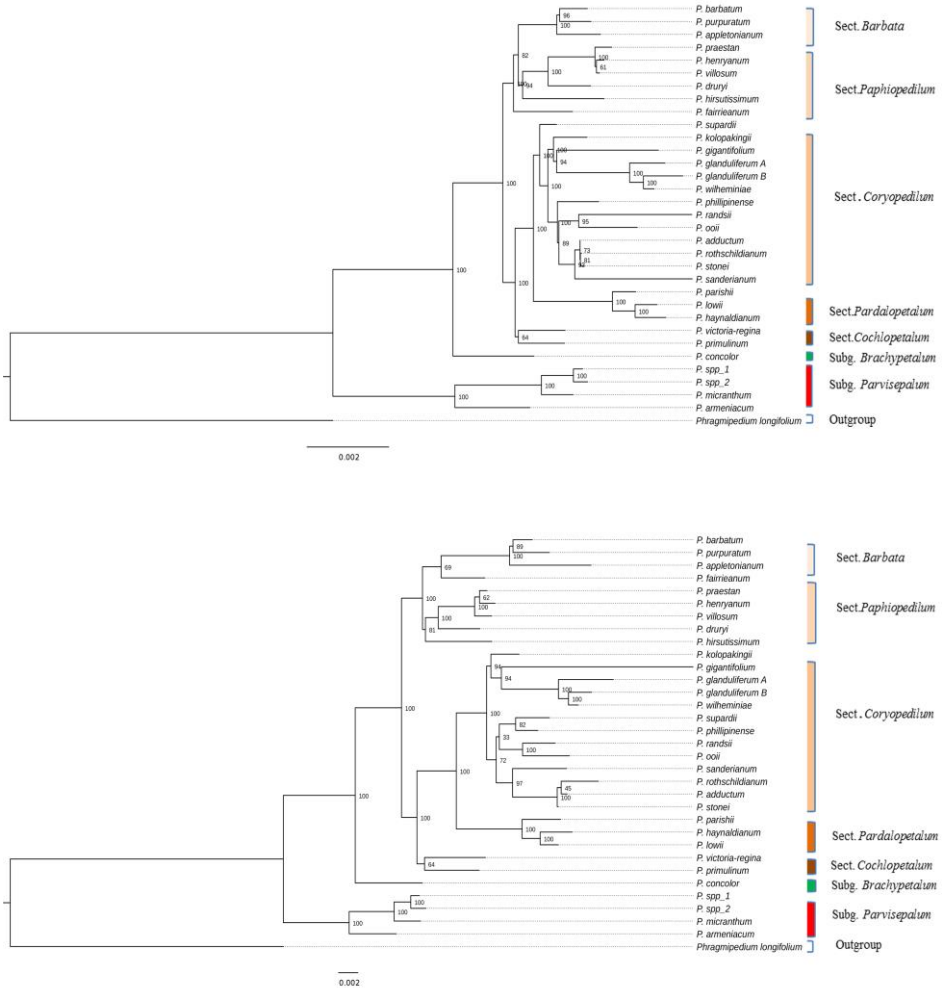
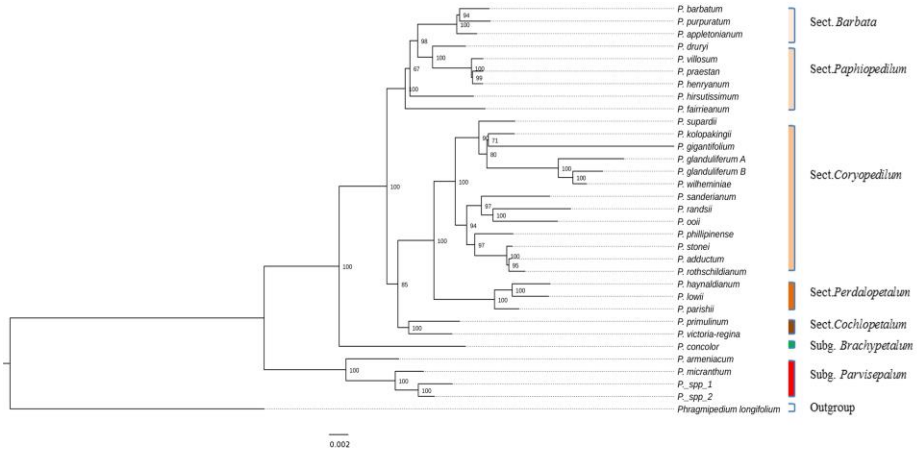
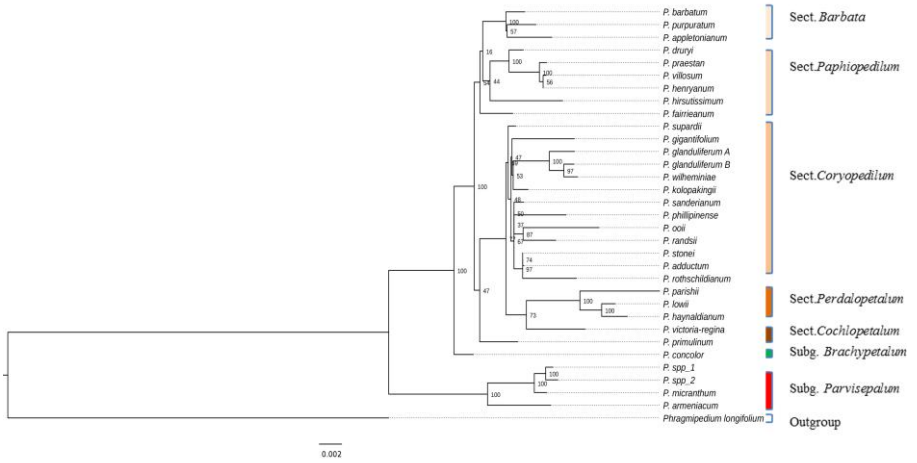


Figure 1. Maximum likelihood trees based on protein coding (above) and non coding (bottom) chloroplast DNA sequences. Numbers along branches indicate bootstrap values. The infrageneric treatment follows Cribb (1998).

a) LSC


P. barbatum
P. purpuratum
P. appletonianum
P. druryi
P. villosum
P. praestan
P. henryanum
P. hirsutissimum
P. fairrieanum
P. supardii
P. kolopakingii
P. gigantifolium
P. glanduliferum A
P. glanduliferum B
P. wilhelminiae
P. sanderianum
P. randsii
P. ooii
P. phillipinense
P. stonei
P. adductum
P. rothschildianum
P. haynaldianum
P. lowii
P. parishii
P. primulinum
P. victoria-regina
P. concolor
P. armeniacum
P. micranthum
P._spp._1
P._spp._2
Phragmipedium longifolium
Sect. Barbata
Sect. Paphiopedilum
Sect. Coryopedilum
Sect. Perdalopetalum
Sect. Cochlopetalum
Subg. Brachypetalum
Subg. Parvisepalum
Outgroup
0.002

b) SSC


P. barbatum
P. purpuratum
P. appletonianum
P. druryi
P. praestan
P. villosum
P. henryanum
P. hirsutissimum
P. fairrieanum
P. supardii
P. gigantifolium
P. glanduliferum A
P. glanduliferum B
P. wilhelminiae
P. kolopakingii
P. sanderianum
P. phillipinense
P. ooii
P. randsii
P. stonei
P. adductum
P. rothschildianum
P. parishii
P. lowii
P. haynaldianum
P. victoria-regina
P. primulinum
P. concolor
P. spp_1
P. spp_2
P. micranthum
P. armeniacum
Phragmipedium longifolium
Sect. Barbata
Sect. Paphiopedilum
Sect. Coryopedilum
Sect. Perdalopetalum
Sect. Cochlopetalum
Subg. Brachypetalum
Subg. Parvisepalum
Outgroup
0.002

c) IR

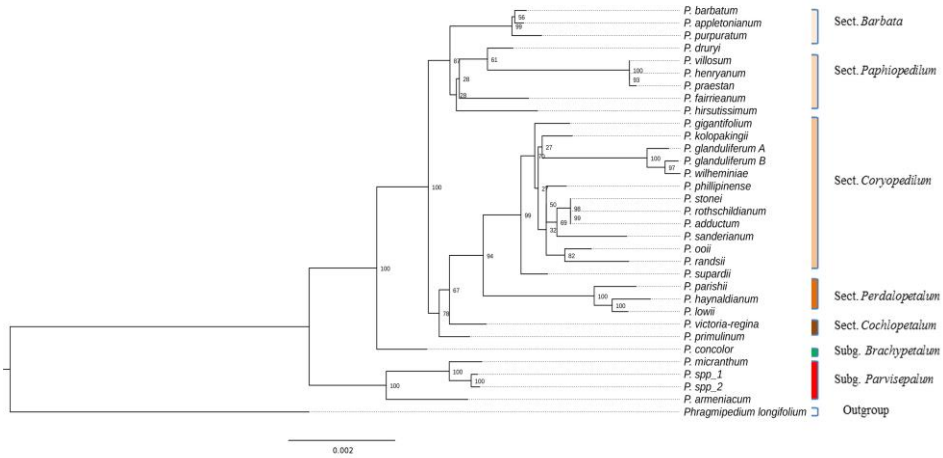


Figure 2: Maximum likelihood trees based on a) long single copy (LSC), b) short single copy (SSC) and c) inverted repeat (IR) regions of cpDNA. Numbers along branches indicate bootstrap values. The infrageneric treatment follows Cribb (1998).

Discussion

De novo assembly of chloroplast genomes of *Paphiopedilum spp*

The chloroplast genomes in this study have been assembled using an approach employing a k-mer frequency distribution to select chloroplast reads from WGS Illumina sequencing data, followed by *de novo* assembly. This allowed assembly of chloroplast genomes, including structural rearrangements with regards to the available reference sequence. The approach presented in this study not only helps to determine structural rearrangements in the chloroplast genome, but at the same time allows to assess the quality of the assemblies and sequencing.

Challenges to construct complete chloroplast genomes of *Paphiopedilum spp* included in this study

Since the first arrival in 2005 (Margulies et al. 2005), next generation sequencing (NGS) technologies have had a tremendous impact on chloroplast genomic research. More than 500 chloroplast genomes have been completely sequenced during the past 10 years. Apart from the computational advancements of NGS technologies, simplification of the sequencing process and lower costs have made chloroplast genome sequencing from total DNA preferable over conventional approaches to full chloroplast sequencing, which commonly involved purification or long range PCR amplifications of the chloroplast genome prior to sequencing (Jansen et al. 2005). Nevertheless, we encountered a couple of challenges in this study that may hinder or complicate the correct assembly and annotation of chloroplast genomes.

One of the challenges was associated with the quality of the raw reads using *Paphiopedilum spp.* DNA. Some sets of WGS sequences, especially in the second batch of the DNA sequencing, contained a high proportion of reads with an insert size shorter than the read size. Reads with an insert size less than the length of the separate paired ends are not useful because they do not extend reads beyond their original length, unlike reads with longer insert size. In addition, short insert size reads may contain the adapter on the opposite end of the reads. This may prevent a proper assembly. Therefore we inserted a

pre-processing and a filtering step after the k-mer frequency-based selection but prior to the *de novo* assembly (see Materials and Methods). Basically, the pre-processing step is where we eliminated the reads that have a smaller insert size. In the filtering step, we positively selected reads from a k-mer table representing the orchid meta-genome and negatively selected those from the k-mer table produced from contaminant sequences. These extra steps turned out to make the assembly easier and less prone to misassembly.

Another challenge is DNA preparation for Next Generation Sequencing. Illumina sequencing typically produces paired-end reads that have an insert size longer than the combined length of both reads. However, the variation of insert sizes is sometimes large and their average size can be difficult to control. The resulting reads with short insert size (as mentioned above) may be partly associated with the concentration of DNA and its quality (Turner 2014). We noticed that most of the low quality assemblies were from genotypes that were preserved in RNALater. RNALater was suggested as one of the alternative methods to preserving RNA and DNA contents in remote fieldwork locations (Gorokhova 2005), but we noticed on gel that some degradation of DNA had taken place. Possible causes of DNA degradation in the tissue preserved in RNALater include a (i) suboptimal volume of RNALater for the size of the tissue preserved, (ii) not ensuring that the tissue was fully submerged in RNALater, (iii) not immediately placing isolated tissue in RNALater, (iv) not storing tissue in RNALater at 4 degrees Celsius overnight prior to freezing, (v) too much RNALater residues in downstream applications. Specifically for *Paphiopedilum* spp. a possible cause may be the thick waxy layer on the leaves. This layer helps to minimize water loss for the plant, but it may have acted as a diffusion barrier for the RNALater, so that it did not fully penetrate the leaf tissue. Our results suggest that RNALater may not be a good method for sampling *Paphiopedilum* spp. in the wild, or for other plant species with waxy layers or thick leaves.

Last but not least, a challenge encountered during chloroplast genome assembly of *Paphiopedilum* spp. concerned areas and samples with a lower number of read (lower coverage). *De novo* assembly is the process to form long contiguous sequences (contigs) by merging individual sequence reads (Paszkiewicz and Studholme 2010). Therefore,

coverage is important because the assembler tends to break the contigs and introduces gaps in the assembly if there is not enough overlap in coverage of the reads. Such cases may give significant effects on the subsequent analyses and biological interpretations leading to wrong conclusions. The small size of the chloroplast genome (less than 180 kbp) as compared to the nuclear genome of *Paphiopedilum* spp. (25-30 Gigabases) means that it is easy to obtain 50× to 100× coverage of chloroplast genome, which is more than sufficient for a successful *de novo* assembly. However, we observed a large variation in number of reads and coverage across the samples. This indicated that the DNA used as an input to prepare the sequencing library varied in quality. For example, samples of *P. adductum*, *P. gigantifolium*, *P. ooii* and *Paphiopedilum sanderianum* had very low coverage (ranging between 0 and 10 of read pairs), evenly along the genome. This may be due to poor DNA quality (too short fragments, Healey et al. 2014) or contaminating substances such as polysaccharides or phenolics (Kasem et al. 2008). Although low coverage may still allow calling many SNP positions with sufficient probability, it does not allow for conclusions about structural rearrangement events for those samples. A poor quality sequencing library produces an uneven coverage, with very low coverage at specific sites, as was for instance observed for the sample of *P. randsii*.

Structural rearrangements in the chloroplast genomes of *Paphiopedilum* spp.

The characterization of 32 *de novo* assembled chloroplast genomes of *Paphiopedilum* spp. led to the identification of a number of structural rearrangements, which have not been reported before. Inversions in the LSC region were detected in several *Paphiopedilum* spp., ranging in size between 3177 bp to 28737 bp. These inversions were fully supported by read mapping against the newly assembled chloroplast genomes. Such inversions in *Paphiopedilum* spp. were first discovered in this study despite the fact that a few chloroplast genomes of the same genus have been published (Kim et al. 2015), possibly because these were based on mapping reads against a reference genome and not on a *de novo* assembly. Despite the fact that the chloroplast genome has been reported to be well conserved in terms of structure and contents, structural rearrangements in the LSC region have been found in various plant species For example, *Trachelium caeruleum* (Campanulaceae) and *Helianthus annuus* (Asteraceae) (Wu et al. 2011), *Vaccinium*

macrocarpon (Ericaceae) (Fajardo et al. 2013), and *Lactuca sativa* (Asteraceae) (Timme et al. 2007) were shown to have large inversions in the LSC region. It has been proposed that intramolecular recombination plays an important role in sequence rearrangement in the chloroplast genome (Ogihara et al. 1988; Ravi et al. 2008). Such sequence rearrangements that alter chloroplast genome structures in related species could provide useful phylogenetic markers for molecular classsification and evolutionary studies because they are readily polarized and lack homoplasy (Olmstead and Palmer 1994; Rokas and Holland 2000; Cosner et al. 2004).

In addition to the inversions we also discovered that the IR boundaries in the chloroplast genome of *Paphiopedilum* spp. have expanded compared with chloroplast genome of tobacco. IR boundaries among angiosperms are known to be dynamic and the IR may expand or contract (Chumley et al. 2006). In the case of most *Paphiopedilum* chloroplast genomes characterized here, the IRs have expanded into the whole SSC region resulting in a total loss of the SSC region. The SSC region of *Paphiopedilum* spp. ranged from 8,111 to 11,178 bp. Thus, the IRs expanded outside the normal size for angiosperms, where IRs range from 20–25 kb (Palmer et al. 1987). While most shifts in the IR boundaries that have been reported are small, others may encompass several kilobases (Zhu et al. 2015). Large IR expansions have been reported, such as an expansion of 12 kb in *Nicotiana acuminata* (Solanaceae) (Goulding et al. 1996), 11.5 kb in Berberidaceae (Kim and Jansen 1994) and 11 kb in *Lobelia thuliniana* (Campanulaceae) (Knox and Palmer 1999), and 50kb in *Pelargonium* × *hortorum* (Geraniaceae) (Guisinger & al. 2011). Martin et al. (2013) reported the expansion at the IR/SSC junction of the *Musa acuminata* (Musaceae) chloroplast genome, which was the largest observed in monocot IRs. Two additional genes (*rps15* and *ndhH*) plus the full sequence of *ycf*1 and 1030 bp of the *ndhA* gene were moved into the IR when compared to the IR structure of *Amborella trichopoda* (Amborellaceae). Expansions/contractions of the IR are probably mediated by intra-molecular recombination between two short direct repeat sequences that frequently occur within the genes located at the borders (Ravi et al. 2008). Goulding et al. (1996) proposed two distinct mechanisms for IR junction evolution: (a) gene conversion for the small stretches and (b) recombinational repair of double strand breaks for incorporation of large chunks of single

copy regions within the IR. The latter mechanism would operate rarely; whereas the former would be a continuous and random process maintaining the IR structure as a whole, but see Zhu et al. (2015). One possible scenario for *Paphiopedilum* spp. is that the loss of *ndh* genes has led to additional structural rearrangements. The loss of the *ndhF* gene was recently found to be correlated with instability of the IR/SSC junction in Orchidaceae (Kim et al. 2015). They observed in the *ndhF*-lacking orchid lineages that the IR/SSC boundaries were severely complicated, usually resulting in an IR expansion.

We confirmed that all 11 *ndh* genes were lost or pseudogenized in all sequenced *Paphiopedilum* spp. This result was comparable with a recent study of various orchid lineages including Epidendroideae, Orchidoideae, Cypripedioideae and Apostasioideae (Kim et al. 2015). In their study, they resolved deeper level phylogenetic relationships among major orchid groups and refined the history of gene loss in *ndh* loci across the orchid family. We also saw a variable pattern of gene loss among *Paphiopedilum* spp. by observing *ndh* gene/pseudogene length variation that supported the hypothesis of Kim et al. (2015) that *ndh* genes were present in common ancestors of orchids,but have undergone independent and significant losses at least eight times across four subfamilies.

Phylogenetic relationships within *Paphiopedilum*

In the present study, we used chloroplast genomic data to elucidate the evolutionary history of the genus *Paphiopedilum* with a main focus on the division of subgenus *Paphiopedilum* into sections. The sequence data were analysed in four sets, which were expected to evolve at a different evolutionary speed: the long single copy, the short single copy, the inverted repeat, the protein-coding sequences and the non-coding sequences. Our phylogenetic analyses are congruent with each other, and they also showed general congruence with previous studies (Albert 1994; Cox et al. 1997; Cribb 1998; Chochai et al. 2012; Guo et al. 2015) using morphological and molecular data. The trees from all four datasets indicate that the genus is monophyletic and differs extensively from *Phragmipedium,* which we used as outgroup. Our results also confirm that subgenus *Parvisepalum* is the first branch in the *Paphiopedilum* genus, followed by subgenera

Brachypetalum and *Paphiopedilum*. Subgenus *Brachypetalum* formed a sister relationship to subgenus *Paphiopedilum* with 100% bootstrap support.

At lower taxonomic levels, the phylogenetic trees derived from our chloroplast genomic sequences confirmed that subgenus *Paphiopedilum* can be divided into two lineages. In the morphological classification of Cribb (1998) these two lineages were characterized by one inflorescence character into a multi-flowered lineage and a single-flowered lineage. This was confirmed by recent studies using molecular sequences such as chloroplast and low copy nuclear genes (Chochai et al. 2012; Guo et al. 2015). In our study, all phylogenetic trees derived from subsets of chloroplast sequences were congruent with this classification. The multi-flowered lineage consists of species from sections *Coryopedilum*, *Pardalopetalum* and *Cochlopetalum*, whereas the single-flowered lineage includes species from sections *Barbatum* and *Paphiopedilum*. Only the phylogenetic tree derived from nuclear ITS sequences (Cox et al. 1997) appeared to place multi-flowered and single-flowered species in the same clade.

Within the first lineage, the results from our study strongly support all sections in subgenus *Paphiopedilum* (100% bootstrap values in all subsets of sequence data). Section *Coryopedilum* was discovered to be sister to section *Pardalopetalum*, whereas the species from this section formed a polytomy to the monophyletic section *Pardalopetalum*. Previously, Cox et al. (1997) proposed to combine these sections based on a phylogenetic tree based on nrITS data, but this suggestion was rejected by Cribb (1998). From his observations of floral morphology, species of section *Coryopedilum* can be clearly distinguished from species of section *Pardalopetalum*. Species of the *Coryopedilum* section have long tapering and unreflexed petals and a convex staminode without a basal protuberance and simple apex, whereas *Pardalopetalum* species have dorsal petals that are reflexed at the base and an obcordate staminode with a basal protuberance and tridentate apex (Chochai et al. 2012). In the phylogenetic trees derived from our chloroplast genome sequences, species of section *Pardalopetalum* grouped together in one clade with high bootstrap support (98-100 BP). Species of section *Pardalopetalum* occur throughout mainland South East Asia and in the Malay Archipelago to Sulawesi and the Philippines,

whereas species of section *Coryopedilum* are limited to the Malaysian islands and endemic to a single island only (Cribb 1998).

Although section *Pardalopetalum* is a well-supported unit, section *Coryopedilum* may not be monophyletic based on our results. This is consistent with previous studies conducted using four chloroplast regions and the nuclear genome ITS region (Chochai et al. 2012). Chochai et al. (2012) suggested that section *Coryopedilum* lacks sufficient molecular divergence to support monophyly of this section, possibly due to its selfing mode of reproduction. Chochai et al. (2012) further explained that the selfing mode of reproduction was the result of geitonogamy and that the species of section *Coryopedilum*, which are endemic to single Malaysian islands, are more prone to be geitonogamous. Including more variable regions such as low copy nuclear regions would possibly help in obtaining a clearer pattern. Guo et al. (2015) constructed phylogenetic trees based on eight chloroplast sequences and four unlinked low copy nuclear genes with more taxa sampled. Despite using more samples and more sequence information, the question of monophyly of section *Coryopedilum* remained unresolved. This is also what we observed in our study, with fewer taxa but much more sequence information. It may be that a much denser sampling of taxa is necessary, and this sampling would have to be based on the geographical occurrence of the species as well.

Our current study successfully recovered section *Cochlopetalum* in phylogenetic trees based on different sets of data (the protein coding, the non-coding and the LSC region), with high to moderate (100 to 64) bootstrap values. The monophyly of this section was not recovered in the tree from the two datasets (the SSC and the IR region) probably because of insufficient molecular polymorphisms. Likewise, the monophyly of this section was not attained in the trees using several loci of chloroplast sequences in previous studies (Chochai et al. 2012; Guo et al. 2015).

The second lineage, morphologically characterized by single-flowered inflorescences (Cribb 1998), includes species of sections *Barbata* and *Paphiopedilum*. The monophyly of section *Barbata* was fully supported with 99-100 BP value in all datasets. We only

included three accessions for this section, so we cannot resolve the existing issue of many internal branches collapsing into a polytomy, which may suggest a recent rapid radiation in the section (Cox et al. 1997; Chochai et al. 2012) or reticulation. Section *Paphiopedilum* was also resolved with 100 BP, consistent with the study of Chochai et al. (2012). Morphologically, species of section *Paphiopedilum* exhibit different leaf morphologies and chromosome numbers as compared with species of section *Barbata* (Cribb 1998; Cox et al. 1997).

Conclusions

Despite the general conservation of chloroplast genomes in most angiosperms, characterization of chloroplast genomes from *Paphiopedilum* spp. showed that it is highly rearranged in slipper orchids. The chloroplast genomes of *Paphiopedilum* spp. included in this study exhibit several structural rearrangements such as inversion in the LSC region, gene loss and duplication as well as IR expansion regardless of the limited number of sampling. The chloroplast genome of *Paphiopedilum* has experienced extreme IR expansion that included part of or the entire SSC region resulting in some of the larger IR regions among the monocots. The unusual features of the complete chloroplast genome of *Paphiopedilum* spp. as discovered in the current study make it an ideal genus in which to study chloroplast evolution in more detail.

Appendix 1

The figure shows the example visual assembly comparisons made using the perl script. Tracks associated with original assembly are plotted along the *x*-axis while the tracks associated with the edited assembly are plotted along the *y*-axis, with a mummer-plot linking the two assemblies shown between these tracks. The order of the region in the original assembly is LSC-IR-SSC-IR while in the edited assembly it is LSC-IR-SSC.

In the mummer-plot a yellow background denotes areas with paired-end coverage of less than 10% of the average, while green and red line segments represent homologous sequence fragments found in either the same or opposite orientation in both assemblies respectively. Above the mummer area and the right hand side are 5 tracks showing coverage with respect to the assembly. The tracks shown in figure are:

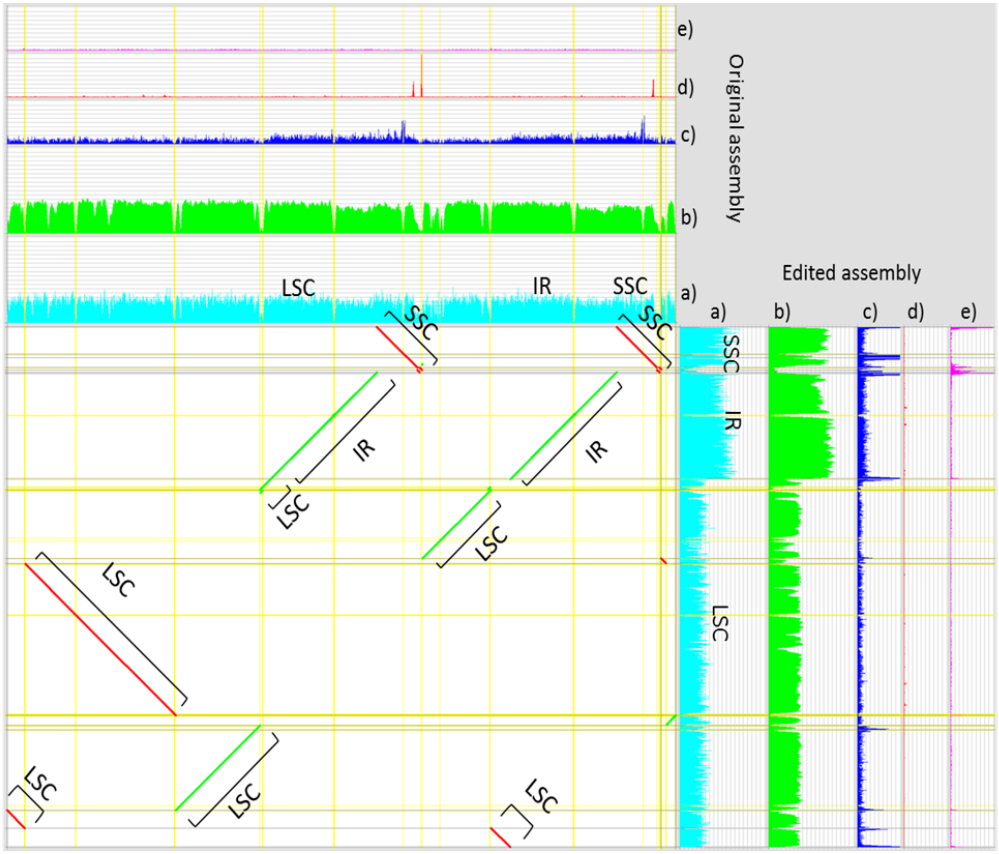
a) cyan: coverage by pseudo-single end fragments resulting from pre-overlapping read-
pairs (truncated at max coverage 200)

b) green: coverage by properly mapped paired end fragments (truncated at max coverage 200)

c) blue: coverage by single reads of pairs where the other read could NOT be placed (truncated at max coverage 100)

d) red: coverage by discordantly mapped paired end fragments (truncated at max coverage 100) - i.e. read-pairs mapping "--> -->", "<-- <--" or "<-- -->"

e) magenta: coverage by mapped paired end fragments (truncated at max coverage 100) where fragments link 2 scaffolds OR have an unusual insert size - i.e. "--><--" or "--> <--"


e)
d)
c)
b)
a)
LSC
IR
SSC
Original assembly
Edited assembly
a)
b)
c)
d)
e)
SSC
IR
LSC

Chapter 6

General Discussion

Next generation sequencing and the chloroplast genome

The study of plant molecular systematics has moved forward into the era of sophisticated, multigene analyses and, hopefully, significantly greater confidence in the inferences. This development was made possible by the development of fast and cheap next generation sequencing (NGS) technology. In the last decade, next generation sequencing technology platforms such as 454 Life Sciences' Genome Sequencer system (Margulies et al. 2005) and Illumina Genome Analyzer (Metzker 2010) have revolutionized plant phylogenetic research through increasing the size of data sets by orders of magnitude. Currently, more data that are phylogenetic informative can be obtained than ever before. It will be very interesting to see to what extent phylogenetically complicated situations can now be resolved, and if so, what other aspects need to be upgraded as well. For instance, more data (59 genes per taxon, produced using an Illumina Hiseq) enabled Zeng et al. (2014) to resolve the topology of the main clades in angiosperm evolution. Prum et al. (2015) combined a large set of sequences (259 nuclear genes with a total length of almost 400 kb per taxon) with very wide taxon coverage (198 species) to resolve the deep phylogeny of birds, but here the improved resolution compared to earlier studies was not due to the depth of sequencing but to the breadth of sampling (Thomas 2015).

In plants the chloroplast (plastid) genome is an invaluable resource for the study of evolution at a range of taxonomic levels. Both 454 and Illumina sequencers have already been successfully used to sequence chloroplast genomes (Wu et al. 2010; Zhang et al. 2011; J. Liu et al. 2012; Ma et al. 2014). The chloroplast genome is ideally suited for high-throughput next-generation sequencing because of its high copy number per cell, apparently highly conserved gene content and arrangement, and small size in comparison to plant mitochondrial and nuclear genomes (Jansen et al. 2005; Moore et al. 2006). From the single gene-based analysis to infer the phylogeny of a broad sampling of seed plant (Chase et al. 1993) to the now genome-scale phylogenetic analysis, this circular genome has been a mainstay to study plant relationships. In the angiosperms, various previously problematic deep-level relationships have recently been resolved, at least largely (Xi et al. 2012; Henriquez et al. 2014; Huang et al. 2014; Ma et al. 2014;

Yang et al. 2014; Carbonell-Caballero et al. 2015). The present study was conducted with the goal to evaluate the potential and limitation of generating chloroplast genomes for phylogenomic purposes from the huge amount of available sequences. The Illumina sequencing platform is one of the most powerful tools in sequence data analysis. There are several challenges associated with using the Illumina platform, which will be discussed in the following section. Next, we addressed the opportunities to understand the chloroplast evolutionary history as well as how this may affect lineage differentiation and phylogenomic discordance in phylogenies based on chloroplast genomes. In addition, as we employed *de novo* assembly rather than mapping against a reference genome, the assemblies and the underlying read data also enabled studying structural rearrangements in the chloroplast genome.

Next generation sequencing and the chloroplast genome; the challenges and pitfall

i) Errors and biases

Compared to Sanger sequencing, next generation technologies have a higher error rate. For instance, an Illumina Miseq paired-end sequencer produces errors at a rate of 0.1 substitutions per 100 bases sequenced (Loman et al. 2012). It is more susceptible to single nucleotide substitution errors than to erroneous insertions and deletions. Besides the errors that are inherent to the DNA sequencing platform, errors and biases can also arise from steps in the sample preparation such as in DNA fragmentation, adapter ligation, or selective amplification. Complications resulting from DNA sequencing errors include false positive variant calls and the detection of sequence polymorphism in regions of low sequence coverage, leading to incorrect interpretation of results. It is a challenge to distinguish true sequence variation from sequencing errors. In order to detect method-inherent errors and biases, a thorough characterization of NGS data is required.

Evaluation of high-throughput data from Illumina revealed several properties associated with the method of sequencing. One of the examples that we also observed within our study was coverage variation. Coverage variation in the sequencing data may partly be

due to the inherent bias of polymerase chain reaction (PCR) amplification during sample preparation (Kozarewa et al. 2009). A study by Stein et al. (2010) suggested that it is mainly caused by the formation of secondary structures in the single-stranded DNA. Lower coverage of sequencing reads have been reported for AT-rich repetitive sequences (Harismendy et al. 2009). Coverage and the variation therein are therefore important quality criteria. Coverage variation was low for the tomato dataset used in Chapter 4. In contrast, the *Paphiopedilum* data (Chapter 5), which were generated in two Hiseq runs, one of which produced fewer paired-end reads per sample, had a much larger variation in coverage (1.8 to 169 million reads), both within and among samples. This may be partly due to a low quality of the DNA (especially for samples stored in RNAlater prior to DNA extraction) and the sequencing libraries made from them.

ii) Assembly and reconstruction of the chloroplast genome

The advancements in next generation sequencing have accelerated the rapid sequencing of complete chloroplast genomes. Du et al. (2015) reported that the use of next generation sequencing technologies to obtain chloroplast sequences became predominant from 2011 onwards, replacing laborious methods that included chloroplast DNA extraction and long-range PCRs. Moreover, constructing complete chloroplasts from non-enriched libraries or whole genome sequencing (WGS) without further isolation or enrichment of cpDNA, became a popular strategy to obtain complete chloroplast genomes. This is possible as 5-15% of the DNA extracted from plant cells may be chloroplast DNA (depending on the type of tissue and the level of photosynthesis, and the extraction protocol used). It was calculated that pea cells contain almost 10,000 chloroplasts per cell (Lamppa and Bendich 1979). Genome coverage is possibly a significant issue when using such a strategy. For example, 7.5 to 15 GB data were used to construct the complete chloroplast of *Populus* (Huang et al. 2014). This is because the percentage of chloroplast reads in whole genome sequencing (WGS) data sets is not constant but appeared to decrease when the nuclear genome size increases (Chapter 2). Since the sequencing depth will be variable in DNA sequencing, the key value for successful assemblies of chloroplast genomes is the sequencing depth of chloroplast genomes rather than the overall sequencing depth. Complete chloroplast genomes were

successfully assembled from data sets that have ~ 25x to 40x sequencing depth for the chloroplast genome (this thesis and Sims et al. 2014).

The combination of using a k-mer frequency distribution to select chloroplast reads followed by *de novo* assembly, as we used in this thesis, represents a reliable option to assemble chloroplast genomes with structural rearrangements. Structural rearrangements such as inversions, insertions or deletions, IR expansion or contraction or loss, transpositions, and loss of genes have been reported in several species especially in monocot chloroplast genomes including *Acorus calamus* (Goremykin et al. 2005), *Trachelium caeruleum* (Haberle et al. 2008) and species of the *Campanulaceae* family (Cosner et al. 2004). In Chapter 2, we observed several structural rearrangements in the chloroplast genome in our *de novo* assembly of *Paphiopedilum* species, which had not been reported before. Therefore, with regard to the genome structure it is unreliable to assemble a chloroplast genome for a non-model species by aligning to a reference or related chloroplast genome because the information on changes in the structure will be ignored. Importantly, the resulting assembly may be incorrect but there is no information to flag this. Hence, it is well possible that the occurrence of structural rearrangement in chloroplasts genomes has systematically been underestimated (see also below), which may lead to problems during the subsequent phylogenetic analysis (Graham et al. 2000; Kelchner 2000) as well as to false SNP and INDEL calls in a purely mapping approach.

Our assembly pipeline requires paired end reads to further improve scaffolding. Chapters 2, 3, 4 and 5 showed that having paired end reads increases the effectiveness of these assemblies. Paired end reads and pseudo-single end reads (i.e., constructed by merging overlapping paired end reads into a pair) were also used to check any misassemblies by mapping the reads back to the assembled genome. In principle such checks may also be performed on assemblies based on mapping against a reference genome, but in that case the reads from regions that were misassembled may not be mapped back but remain in the 'basket' of unmapped reads, as would reads from regions that are entirely absent or too much diverged from the reference genome.

iii) Quality control

While all advancements in next generation sequencing are beneficial for the field of phylogenomics, there are several risks associated with huge amounts of data, some of which are encountered in this thesis. One of the examples is quality control. Quality control of WGS data is extremely important if WGS methods are to become part of a routine approach to generate large datasets for phylogenomic studies. Although the complete chloroplast genomes of many more species are available, most of them were published as "draft" assemblies whose quality is uncertain. In our opinion, to produce an accurate genome assembly and to correctly annotate them remains challenging. First, this is due to the properties of short read sequences itself. It is a challenge to completely assemble any genome whenever genomic sequences contain repeat sequences longer than the read length, as the assembler program may introduce gaps or produce misassemblies (Schatz et al. 2010; Ye et al. 2011; Treangen and Salzberg 2012). Third generation sequencing technology (such as PacBio) is foreseen to alleviate some limitations, as the reads are longer. However, researchers have been slow to adopt third generation sequencing because of relatively high error rates along with much higher costs.

Second, to determine which assembly is correct by comparing the quality of different assemblies of the same data set is also not straightforward. Although several methods have been proposed to assess the quality of a *de novo* assembly, none of them is broadly accepted because each study used a different collection of metrics and validation utilities, making it impossible to compare their respective results directly (Nagarajan and Pop 2013). There is therefore a crucial need for the scientific community to enforce standards of quality beyond nucleotide quality scores, that can be measured, and maintain and propagate these quality measures through downstream analyses and consistently store them in databases. Besides the aspect of disclosure of quality measures in final results, tools to evaluate and especially compare assemblies in detail are useful during assembly, and mapping back the sequence reads to different assembly variants they produce can provide useful insights (Chapter 3). In Chapter 3, we have created a flexible assembly quality comparison tool to address this issue. This tool combines and

visualizes read mapping and alignment results in a 2-dimensional plot without breaking any sequence connectivity. We have evaluated the ability of this tool using the *de novo* assemblies of *Solanum lycopersicon* (tomato) and *Paphiopedium henryanum* (orchid) chloroplasts obtained from Whole Genome Shotgun (WGS) Illumina short read sequencing datasets in combination with specifically made alternative assemblies.

iv) Chloroplast annotation

After the chloroplast genome has been assembled, accurate annotation of genome features such as genes coding for proteins, tRNAs as well as rRNAs need to be carried out before additional analyses can be made. Annotation of chloroplast genome is commonly performed using the Dual Organellar Genome Annotator (DOGMA) (Wyman et al. 2004), a web-based annotation tool that utilizes BLASTX and BLASTN of a chloroplast database. However, DOGMA is written so that the user chooses the start and stop codon for all genes and this requires manual inspection/curation for determining gene and intron/exon boundaries. Manual inspection/curation steps can be tedious and time consuming (Wyman et al. 2004). Therefore, this may easily become a bottleneck for bench scientists who want to correctly employ an abundance of chloroplast genome sequences. Therefore, within this study, we used another platform of chloroplast annotation that offers a semi-automatic and complete annotation of a chloroplast genome sequence. This web server, called CPGAVAS, includes the genome visualisation, editing and analysis of the annotation results (Liu et al. 2012). The CPGAVAS server uses a complete chloroplast genome sequence as input and output of the annotation results is in GFF3 format. Similarly to DOGMA but with additional functionalities, CPGAVAS integrates results from BLASTX, BLASTN, protein2genome and est2genome databases. The server also includes tRNAscan for tRNA genes and inverted repeats (IR) identification, calculates the summary statistics for the annotated genome, generates circular maps and extracts protein and mRNA sequences for a given list of genes and species. In case one has too many chloroplast genomes to be annotated in this way, we suggest using the Geneious software annotation program (Chapter 4 and 5). This software can transfer genome annotations on the basis of high sequence similarity.

The recently developed tools that we described above have been a great help in extracting the information from the chloroplast genomes we have assembled. Although in some cases it may not be so ideal we still could, with modest bioinformatics, extract different subsets of chloroplast sequences, and compare the phylogenetic information in them. This is similar to the strategy used by previous chloroplast phylogenomic studies including those in the Bamboo tribe (Ma et al. 2014), ginkgo (Wu et al. 2013) and the *Araceae* family (Henriquez et al. 2014).

Evolution of chloroplast genome: structure and genetics

i. Structural rearrangements in chloroplast genome

The chloroplast genome can be characterized by its quadripartite structure: two inverted repeats (IRs) separated by a long single copy region (LSC) and a small single copy region (SSC). The organization of the chloroplast genome is highly conserved over long evolutionary time scales. The arrival of NGS has significantly increased the number of complete chloroplast genomes available, creating the opportunity for comparative studies that led to new insights into the evolutionary history of chloroplasts in angiosperms (Jansen et al. 2007; Doorduin et al. 2011). Consistent with the presumed conserved nature of the chloroplast genome among angiosperms only a relatively small number of structural rearrangements have been reported. However, for some plant lineages large-scale structural rearrangements, gene loss and duplication events have been reported (Cosner et al. 1997; Cosner et al. 2004; Chumley et al. 2006a; Blazier et al. 2011; Dugas et al. 2015). Other comparative studies of chloroplast genomes did reveal changes of these regions including partial and complete loss of one IR copy (Chumley et al. 2006b), localized gene losses (Magee et al. 2010), a high number of dispersed repeats (Cai et al. 2008), and elevated rates of molecular evolution (Guisinger et al. 2008). This is an apparent contradiction, unless we assume that many structural arrangements have been overseen due to the habit of assembling against a reference genome. Structural rearrangements in chloroplast genomes result from intramolecular recombination events that may generate genetic diversity that is useful for molecular classification and evolution studies. The identification of the structural

rearrangements within this study was possible as we used *de novo* assembly for the chloroplast genomes. Examples of structural rearrangements that we discovered within this study include inversions, gene loss, gene duplication and the expansion of the IR region.

The first structural rearrangement that was identified was an inversion. Inversions have been reported occasionally, and they are associated with chloroplast gene order changes (Chumley et al. 2006b). Large inversions of 22.8 kb in *Asteraceae* (Jansen and Palmer 1987; Kim et al. 2005), 54 kb in *Oenothera* (Hachtel et al. 1991; Hupfer et al. 2008) and 50 kb in *Fabaceae* (Palmer et al. 1988; Doyle et al. 1996) have been previously reported. In Chapter 2, the inversion in the LSC region of the *Aegilops tauschii* chloroplast genome that was reported before was confirmed, and similar inversions in the LSC region were discovered in several *Paphiopedilum spp*. in Chapter 5. Furthermore, it was proven that the inversions found were genuine events by mapping the raw reads back to the newly assembled genome. In some instances, however, read coverage across the junctions between inversions is scant, and additional confirmation, for instance through PCR, is required.

Another structural rearrangement is the loss of *ndh* genes. The chloroplast genome usually encodes eleven chloroplast *ndh* genes (*ndhA*-*ndhK*) (Kim et al. 2015). The loss of the *ndh* gene complex from the chloroplast genome is not common in photosynthetic plants, as it has only been reported for Gnetales (McCoy et al. 2008; Wu et al. 2009), Pinaceae (Wakasugi et al. 1994; Cronn et al. 2008) and a large clade within the Orchidaceae (Neyland and Urbatsch 1996; Chang et al. 2006; Wu et al. 2010; Kim et al. 2015). In Chapter 5, chloroplast genomes of 32 *Paphiopedilum spp*. were generated and they all lack 11 intact *ndh* genes. The *ndh* genes from 32 *Paphiopedilum spp*. were either lost completely or pseudogenized by multiple stop codons and frameshifts, or short INDELs throughout their sequences. These results confirmed the *ndh* gene loss in *Paphiopedilum* and six other orchid lineages that had been recently described by Kim et al. (2015). Several other studies involving orchid chloroplast genomes belonging to the subfamily Epidendroideae including *Phalaenopsis* (Chang et al. 2006), *Oncidium*

(Wu et al. 2010), *Erycina* (Pan et al. 2012) and *Cymbidium* (Yang et al. 2013) demonstrated the loss of intact genes for all *ndh* genes. Among these orchids, only *ndhB* of *Oncidium "Grower Ramsey"* and *ndhE*, *J* and *C* of *Cymbidium* encoded functional *ndh* proteins.

Other deviations from the 'conserved' structure of the chloroplast genome detected in this study were typically the result of IR boundaries shifts. The IR boundaries are simply the points at which the single copy region in the chloroplast genome ends and the inverted repeat region starts, or vice versa, and shifts in the IR boundaries are usually in the form of expansions and contractions. In Chapter 4, a small IR expansion into the LSC region was observed, resulting in various lengths of partial duplication of the *rps*19 gene in several tomato accessions. Additionally, a large IR expansion was present in our sample of *Paphiopedilum* species (Chapter 5). In that chapter large IR expansions (8 kb to 11 kb) were detected concomitant to the shift of IR boundaries into the SSC region of *Paphiopedilum* species. Previously, large expansions into the SSC have been reported in some groups of plants such as in *Gramineae* (Hiratsuka et al. 1989; Maier et al. 1995), buckwheat species (Kishima et al. 1995), *Trachelium* (Cosner et al. 1997), and *Lobelia thuliniana* (Knox and Palmer 1999). It has been proposed that the large expansions of the IR observed in some groups may have been caused by double-strand DNA breaks and subsequent repair, which is different from the ordinary gene conversion mechanism (Goulding et al. 1996).

ii. Positive selection in the chloroplast genome

Genes in the chloroplast genome are shaped by the selective pressure to maintain the fundamental cellular functions during evolution. In Chapter 4 incongruences between phylogenies of protein coding data compared to those of non-coding data were observed. Positive selection can be one of the causes of this incongruence. Positive selection or variants that increase in frequency until they become fixed in the population (or, in this case, a species) are difficult to detect and analyse because neutral and deleterious mutations predominate in frequency (Ravi et al. 2008). In addition to positive selection, several coding regions have been shown to accumulate a higher number of variants. In

tomato three specific genes (*ycf1, ndhF and ndhH*) each accumulated more than 10 mutations in their coding region. These genes may function as general hotspots of natural genetic variation in tomato and it may be possible that several alleles are maintained under selective pressure because they provide some advantage (Carbonell-Caballero et al. 2015).

The chloroplast genome in plant phylogenetics

It is a challenge to obtain accurate phylogenies and effective species discrimination when using single or several chloroplast genes, because they contain few informative characters. This is even worse in evolutionary young lineages (Ruhsam et al. 2015). The application of WGS facilitates the reconstruction of complete genomes, and this in turn has made it possible to obtain dozens of polymorphic characters for molecular phylogenetic studies in plants, even among closely related ones. This can be observed by the number of studies applying phylogenomic approaches to WGS-generated chloroplast data (Zou et al. 2008; Sanderson et al. 2010; Capella-Gutierrez et al. 2014; Davis et al. 2014; Ma et al. 2014).

The use of nearly complete or complete chloroplast genomes results in complex data sets, and this may potentially increase sources of phylogenetic error (Philippe et al. 2005). There are two types of phylogenetic error: the stochastic error and the systematic error. The stochastic error or sampling error is caused by mechanisms such as gene duplication, horizontal gene transfer or lineage sorting (Rokas et al. 2003; Martin et al. 2005; Jeffroy et al. 2006; Zou et al. 2008). In contrast to the stochastic error, which decreases as the quantity of data increases, the systematic error may increase with data quantity because adequate modelling becomes increasingly difficult (Philippe et al. 2005; Kumar et al. 2012).

The use of the complete chloroplast genomes was evaluated to see if this increased species discrimination and phylogenetic resolution in a set of closely related tomato species (Chapter 4). Overall, the phylogenetic tree based on complete chloroplast genomes recovered the same clades as those that were previously defined by Peralta et

al. (2008). Tomato species from section *Lycopersicon*, section *Junglandifolia*, and section *Lycopersicoides* were grouped using a combination of recent morphological and molecular data from previous studies (Peralta et al. 2008). However, although the relationships among these clades were well resolved, several discrepancies in the placement of individual taxa were observed when comparing with nuclear phylogenies. Those samples might be the result of hybridization or have introgression events in their ancestry. Another explanation for the observed non-monophyly of the tomato chloroplast genome is the young evolutionary age of the tomato clade. Results of Särkinen et al. (2013) suggest that the split between tomato (*Solanum* section *Lycopersicon*) and potato (*Solanum* section *Petota*) was only around 8 Million years ago (Mya). This may have been insufficient time for species-specific mutations to accumulate or /and for complete sorting of ancestral polymorphism. Indeed, few variable sites (211) were detected that were informative among protein coding genes in the tomato chloroplast genome. In contrast, in a protein coding sequence dataset of 32 *Paphiopedilum* spp, which was dated back to 22.2±5.9 Mya (oldest age) (Guo et al. 2012) 1491 variable sites were detected that were informative. This supports the notion that low substitution rates contributed to a lack of complete monophyly of the important nodes in tomato species. Indeed, only few studies used whole chloroplast genomes to infer phylogenetic relationship at the intraspecific level among closely related species. For example, Bayly et al. (2013) demonstrated that this approach was useful to resolve phylogenetic relationships among eucalypt genera but not among closely related *Eucalyptus* species.

Genome-scale data and taxon sampling

Chloroplast-based phylogenies of recently diverged taxa were expected to yield limited sequence variation especially at low taxonomy levels species (example: Chapter 4 and Chapter 5). In general, both genome-scale data and dense taxon sampling may improve phylogenetic estimation by providing more data. In the past, molecular phylogenetic analyses were often hindered by DNA sequencing costs, which forced researchers to choose between dense taxon sampling with a small number of informative loci and wider sampling of the genome in a lower number of taxa. In studies that focus on

recently diverged taxa, taxon sampling needs to be sufficiently broad to detect interspecific variation and the phylogenetic depth of shared alleles (Whitfield and Lockhart 2007).

In Chapter 4, the difference in phylogeny resolution of a multilocus matrix (72,807 bp) and highly informative single loci (5738 bp) was highlighted using the same number of tomato taxa. The topologies of both phylogenies did not indicate a significant conflict but the single loci phylogeny suffered from lack of resolution. Similarly, the phylogenetic tree of *Paphiopedilum spp.* based on genome-scale data of chloroplast sequences (Chapter 5) was similar to the phylogenetic tree that was based on only eight chloroplast regions (Guo et al. 2015). Although in the study of Guo et al. (2015), that included a wider taxon sampling, the resolutions appeared better compared to the limited sampling taxon coverage in Chapter 5, the general relationships of species were in agreement. This suggests that the resolution of chloroplast-based infrageneric phylogenies does benefit from an increase of the data matrix length. However, it does not prove that a complete assembly is necessary, as we could also extract and use multiple genes from the NGS data. The complete assembly is useful if structural rearrangements can be uncovered that may be used as additional phylogenetic characters.

Research outlooks

i. K-mer selection for *de novo* assembly of chloroplast genome

Phylogenomics is a field of comparative biology that uses genomic data to infer relationships among organisms (Chan and Ragan 2013). Within this thesis, chloroplast phylogenomics was conducted using complete chloroplast DNA genomes obtained by a newly developed method of *de novo* assembly. The method was not only cost-effective but also has the potential to extract a wealth of useful information of thousands of chloroplast genomes from WGS data. This information is hidden in next generation datasets of whole genomic DNA, which often contains 5-15% chloroplast-derived reads. They can be identified based on their k-mer distribution, which shows two distinct peaks, one at the copy number of the chloroplasts in the cell and one at the double copy

number (for the reads from the inverted repeat). After extraction from the complete dataset, the pipeline developed in Chapter 2 can easily *de novo* assemble the chloroplast genome. In Chapter 2 and 5, it was demonstrated that this newly developed pipeline is able to discover structural rearrangements in the chloroplast genome. These structural rearrangements may be ignored or missed if the chloroplast genome would be assembled by alignment to a reference or related chloroplast genome (Chapter 3). Structural rearrangements or changes in chloroplast genome composition may have significant phylogenetic implications. Furthermore, the availability of genome-scale data of chloroplast sequences is a way for improving the resolution in phylogenetic studies. The chloroplast-based phylogenies reported in this study form a solid basis for future studies aimed to understand evolutionary relationships at low taxonomic levels. In doing so, the assembly pipeline may also mitigate the current reliance of relatively short sequences in phylogenetic research (Parks et al. 2009), such as in species identification, comparative studies as well as development in phylogenetic methods.

ii. Re-evaluation and discovery of molecular markers for phylogenetic analyses

The use of chloroplast molecular markers for phylogenetic analyses has significantly helped researchers in early years. However, most chloroplast molecular markers were identified before entire genomes were available, and they were selected for the possibility to be amplified using conserved primers flanking the genes or gene spacers, and the possibility to align the resulting sequences unequivocally. With the increasing number of complete chloroplast genomes available, it is time to re-evaluate the variability of chloroplast regions at low taxonomic levels. It was reported that many plant species evolved via adaptive radiations and possess only a few million years of evolutionary histories (Dong et al. 2012). The short evolutionary histories resulted in low sequence divergence. In order to resolve phylogenetic problems at the species level, we need to identify regions that have high evolutionary rates. The availability of complete chloroplast genomes as constructed within this study may increase our ability to resolve such identification problems. Furthermore, it also allows the discovery of new molecular markers that cannot be easily amplified by PCR but that can easily be

extracted from WGS data, optionally in the form of complete chloroplast genomes, and that are superior in information content.

iii. The chloroplast genome as a new way for species identification

DNA barcoding is one of the techniques used for species identification that are useful in plant biodiversity research. This technique uses particular DNA sequences to characterize the identity plant organisms by comparing it to a database of barcode sequences from various taxa (Hebert et al. 2003). A DNA barcode is a segment of DNA sequence that is sufficiently variable to be able to distinguish even closely related species. On top of that, the sequences flanking the barcode should be sufficiently conserved to facilitate amplification by PCR. Although the cytochrome c oxidase 1 (CO1) sequence has been developed as a universal barcode in animals, neither a single locus nor a single set of multilocus barcodes have been found that could efficiently discriminate plant species, due to lack of variation (Fazekas et al. 2008; Kress and Erickson 2008; CBOL Plant Working Group 2009; Chase and Fay 2009) in various organellar regions that could consistently be amplified across taxa. This has led several studies to propose the use of the whole chloroplast genome for species identification between closely related species (Parks et al. 2009; Nock et al. 2011), populations (Doorduin et al. 2011) and individuals (Kane et al. 2012; McPherson et al. 2013). Species identification using the whole chloroplast genome as a marker would make sequence variation in the genome accessible in regions that could not easily be amplified across species with conserved PCR primers (Huang et al. 2005) and would be more efficient in detecting gene loss and defining gene order than traditional DNA barcoding (Luo et al. 2008; Luo et al. 2009). However, to reconstruct the chloroplast genome for a number of taxa used to be costly. We anticipate that this limitation can be resolved using the pipeline in this thesis, and that it will lead to providing many more complete chloroplast genomes from total DNA shotgun sequences. Reconstruction of the whole chloroplast genome from WGS data is not only cost-effective but also less resource-intensive compared to other traditional methods such as obtaining it from purified chloroplast DNA (McPherson et al. 2013).

iv. Phylogenetic utility of structural rearrangement

Through comparative studies several structural rearrangements of the chloroplast genome such as inversion, gene or intron loss, loss of IRs and IR expansions/contractions have been found in certain plant lineages. The assembly method presented in this study (Chapter 2) and the examples of the structural rearrangements detected (Chapter 5) offer the possibility to use structural rearrangement data as informative characters in phylogenetic studies. Structural rearrangements data in chloroplast genomes are encountered more rarely than nucleotide mutations and they are considered to have less homoplasy (Rokas and Holland 2000). Although not all structural rearrangements are well understood, these characteristics can make a profound phylogenetic statement. For example, large inversions have been suggested to be extremely useful markers in phylogenetic inference (Doyle et al. 1996; Cosner et al. 1997; Perry et al. 2002; Timme et al. 2007). On the other hand, (Rokas and Holland 2000) expressed concern about the use of structural rearrangement data in phylogenetic studies as they lack statistical evaluation. They also said that such development is hampered by our limited understanding of the mechanism(s) causing the variation, which is important knowledge to be able to estimate the rate of production, character independence, mutational biases and reversibility of structural rearrangements. The only way to deal with such criticism is to generate sufficient information on structural arrangements, their types and frequency of occurrence across various taxonomic groups, in order to evaluate their characteristics. The *de novo* assembly and quality check procedures developed in this thesis will enable doing just that for the large amount of NGS data currently produced.

v. Alignments or Assembly-free phylogenetic analyses

Traditional sequence comparison using multi-sequence alignment (MSA) is often frustrated by the limitations of this method, including the necessity to manually adjust alignments, the fundamental problems in accuracy when arbitrary choices must be made, and their computational efficiency. The increasing availability of genome information has created a demand for alternative algorithms for fast and accurate phylogenetic inferences. Motivated to overcome the limitations of MSA, several alignment-free

methods have been proposed. Briefly, distances between pairwise organisms can be calculated using word frequency (reviewed by Bonham-Carter et al. 2013) , information theory (Li et al. 2001; Li and Vitâanyi 2009), average common length (Otu and Sayood 2003) and other methods. However, these alignment-free methods have their own problems. For instance, distances computed from information theory or word frequency do not usually have a biological definition and they are rarely linear with evolutionary time. As a matter of choice, one should consider what is the best alignment-free method that is suitable for one's own datasets and the desired end result from the phylogenetic analysis, but good comparions and evaluations of these methods are still missing.

In this thesis, basically an alignment-free method to extract chloroplast reads based on word frequency (Chapter 2) was used, the words being of arbitrary length k and therefore termed k-mer. Subsequently, the genomes were assembled, coding and non-coding regions were extracted, and comparions were made based on sequence alignments. Direct comparisons of the frequency of bits of sequence would certainly speed up this process, but it remains to be seen whether it would generate a similar level of information. Most certainly any information on larger structural variation would be lost.

In general, alignment-free methods are considered potentially attractive for phylogenomics because of the simplicity of their algorithms and the easier and faster computations, which require less resources and less time. As an interesting example, Yi and Jin (2013) proposed the Co-phylog approach specifically to take advantage of unassembled WGS data. This assembly-free approach creates micro-alignments, calculates pairwise distances, and then reconstructs the phylogenetic tree based on these distances. From a previous study, the approach was demostrated to be an efficient algorithm resulting in a high resolution and accurate phylogenic trees of several genera, especially for closely related organisms (Yi and Jin 2013). In their study they demonstrated that the phylogenic tree constructed using simulated and real NGS datasets with the Co-phylog approach was comparable to the benchmark tree produce by a traditional alignment-based method. However, the Co-phylog method did not perform as

well on distant organisms. It remains to be seen whether these methods can be extended to other genomes, but in terms of size (*Escherichia coli* and related taxa are 4-5 Mb) the plant chloroplast genomes would fall in the range in which such methods may perform adequately.

Conclusions

Overall, the application of WGS data offers opportunities to use partial or entire chloroplast genomes for phylogenetic studies. Species discrimination will be achieved already with partial data (subsets of genes), but the power will still be insufficient for evolutionarily young lineages, which may require more informative characters. Therefore, it is expected that the number of complete chloroplast genomes that become available, will increase in the years to come. While generating these genomes, the urge for *de novo* assembly of chloroplast genomes rather than mapping against reference genomes is adamant in order to also uncover structural rearrangements in chloroplast genome. Here, tools have been developed to perform such *de novo* assemblies, and important considerations discussed when using chloroplast genomes for phylogenetic analyses. Thus, I believe this thesis may fill an important gap towards producing robust and accurate chloroplast-based phylogenetic trees.

Summary

DNA sequences play a key role in modern molecular phylogenetic analyses. The structure and function of the DNA sequences and how they change over time are used to infer evolutionary relationships. Phylogenetic studies in plants mostly employ a number of chloroplast DNA sequences along with a few sequences of the nuclear genome, such as the internal transcribed spacer (ITS). The chloroplast genome has been shown to provide a wealth of information on molecular variation for phylogenetic studies, but rarely the whole genome has been used, as up to recently it was very laborious to generate full genomes of chloroplasts. Whole Genome Shotgun (WGS) sequences of plant species often contain 5-15% of sequence reads that are derived from the chloroplast genome, which is many times more than needed for the assembly of the chloroplast genome. In this thesis I have developed a method to extract the chloroplast reads from WGS datasets and to generate the complete chloroplast genome sequence, and explored how complete chloroplast genomes could provide comprehensive data sets that are superior for inferring relationships in several plant lineages.

Chapter 2 describes how *de novo* assemblies of chloroplast genomes of *Solanum lycopersicum, Aegilops tauschii* and *Paphiopedilum heryanum* were performed based on whole genome sequencing data. In this study, we used k-mer frequency tables to identify and extract the chloroplast reads from the WGS reads and assemble these using a highly integrated and automated custom pipeline for *de novo* assembly. This pipeline includes steps aimed at optimizing assemblies and filling gaps due to coverage variation in the WGS dataset. I used it to *de novo* assemble three complete chloroplast genomes from plant species with a 40-fold range of nuclear genome size to demonstrate the universality of our approach. This new and cost-effective method for *de novo* short read assembly may facilitate the study of complete chloroplast genomes with more accurate analyses and inferences, especially in non-model plant genomes.

The method developed is also suitable for studying structural variation in the chloroplast genome, as opposed to the common procedure of read mapping against a reference

genome. However, to support the putative rearrangements that were in the output of the assembly, a method had to be developed to visualise the support for the rearrangement in comparison to other regions in the chloroplast genome, and in contrast to a reference genome without rearrangements. This method was described in **Chapter 3**.

In order to explore and evaluate chloroplast phylogenomics, or phylogenetic analyses based on complete chloroplast genomes, the available WGS data of various species within the section *Lycopersicon* (from the Tomato Genome Sequencing Consortium) were used in **Chapter 4** to assemble 84 tomato chloroplast genomes and generate phylogenetic trees. These analyses revealed that next to the chloroplast regions and spacers traditionally used for phylogenetics, various additional regions of protein coding and non-coding DNA can be explored and exploited for intraspecific phylogenetic studies. In particular, more than 50% of all phylogenetically relevant information could be included by just using four genes (*ycf1*, *ndhF*, *ndhA*, and *ndhH)*. Moreover, when one would only use *ycf1* one would already use 34% of all information available in the chloroplast genomes of the accessions used in this study. The topology of the phylogenetic tree inferred from *ycf1* was the same as that of trees based on all other protein coding genes, although with lower bootstrap values. Although we successfully recovered major groups in the section, some topological incongruences for some taxa were observed from the phylogenetic analyses of different sub-sections [protein coding, noncoding, Single Copy (SC) and Inverted Repeats (IR)] of the chloroplast genomes. Incongruences between chloroplast genome and nuclear genome derived phylogenies suggest ancient hybridization events or incomplete lineage sorting (ILS) as the most likely explanation.

The phylogenetic analyses in **Chapter 5** based on 32 complete *Paphiopedilum* chloroplast genomes confirmed that the genus *Paphiopedilum* is monophyletic, and that the division of the genus into three subgenera *Parvisepalum*, *Brachypetalum* and *Paphiopedilum* is well supported. The division of five sections of subgenus *Paphiopedilum* was also recovered. The *de novo* assemblies revealed several structural rearrangements including gene loss and inversion. In addition, the chloroplast genome of *Paphiopedilum* has experienced extreme

IR expansion that has included part of or the entire SSC region, resulting in larger IR regions than commonly observed among monocots.

In **Chapter 6** the results produced in this thesis are summarized and placed into a broader context. Several challenges associated with using the Illumina platform for producing WGS sequences and the evolution of chloroplast genome structure and genetics that were discovered within this thesis were discussed. Furthermore, I also addressed the opportunities of the vast amounts of short reads produced nowadays to understand the chloroplast evolutionary history as well as how this may affect lineage differentiation and phylogenomic discordance in phylogenies based on chloroplast genome sequences. Finally, I make a pledge for *de novo* assembly based on chloroplast-derived reads rather than mapping against reference genomes, as this will most likely uncover a much larger extent of structural variation than commonly assumed.

References

Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, et al. 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J.:136–148.

Albert VA. 1994. Cladistic relationships of the slipper orchids (*Cypripedioideae:Orchidaceae*) from congruent morphological and molecular data. Lindleyana 9:115–132.

Alvarez AE, Van de Wiel CCM, Smulders MJM, Vosman B. 2001. Use of microsatellites to evaluate genetic diversity and species relationships in the genus *Lycopersico*n. Theor. Appl. Genet. 103:1283–1292.

Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods 6:22.

Atwood J. 1986. The size of Orchidaceae and the systematic distribution of epiphytic orchids. Selbyana 9:171–186.

Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, Veuille M, Laredo C. 2009. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10 Suppl 1:S10.

Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE, et al. 2016. Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. Biol. J. Linn. Soc. 117:33–43.

Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW. 2013. Plastid genomes and deep relationships among the commelinid monocot angiosperms. Cladistics 29:65–87.

Barthelson R, McFarlin AJ, Rounsley SD, Young S. 2011. Plantagora: Modeling whole genome sequencing and assembly of plant genomes. PLoS One 6.

Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, Bossinger G, Merchant A, Udovicic F, Woodrow IE, et al. 2013. Chloroplast genome analysis of Australian eucalypts - *Eucalyptus,Corymbia, Angophora, Allosyncarpia* and *Stockwellia* (Myrtaceae). Mol. Phylogenet. Evol. 69:704–716.

Besnard G, Hernández P, Khadari B, Dorado G, Savolainen V. 2011. Genomic profiling of plastid DNA variation in the Mediterranean olive tree. BMC Plant Biol. 11:80.

Blair JE, Ikeo K, Gojobori T, Hedges SB. 2002. The evolutionary position of nematodes. BMC Evol. Biol. 2:7.

Blazier CC, Guisinger MM, Jansen RK. 2011. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). Plant Mol. Biol. 76:263–272.

Bonatelli I a S, Zappi DC, Taylor NP, Moraes EM. 2013. Usefulness of cpDNA markers for phylogenetic and phylogeographic analyses of closely related cactus species. Genet. Mol. Res. 12:4579–4585.

Bonham-Carter O, Steele J, Bastola D. 2013. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Brief. Bioinform. 15:890–905.

Bookjans G, Stummann BM, Henningsen KW. 1984. Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic strength. Anal. Biochem. 141:244–247.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. Gigascience 2:10.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, et al. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science. 325:318–321.

Brinkmann H, Philippe H. 2008. Animal phylogeny and large-scale sequencing: progress and pitfalls. Journal of Systematics and Evolution. 46:274-286.

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl J V., Boore J, Jansen RK. 2008. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J. Mol. Evol. 67:696–704.

Capella-Gutierrez S, Kauff F, Gabaldón T. 2014. A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucleic Acids Res. 42:1–11.

Carbonell-Caballero J, Alonso R, Ibañez V, Terol J, Talon M, Dopazo J. 2015. A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic Species within the Genus *Citrus*. Mol. Biol. Evol. 32:2015–2035.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proc. Natl. Acad. Sci. U. S. A. 106:12794–12797.

Chan CX, Ragan MA. 2013. Next-generation phylogenomics. Biol. Direct 8:3.

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, et al. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol. Biol. Evol. 23:279–291.

Chase M, Cameron K, Barrett R, Freudenstein J V. 2003. DNA data and Orchidaceae systematics: a new phylogenetic classification. IN: Orchid Conservation (K.W. Dixon, S.P. Kell, R.L. Barrett and P.J. Cribb, eds), Natural History Publications, Kota Kinabalu, Sabah. Pp 69–89.

Chase MW, Cameron KM, Freudenstein J V., Pridgeon AM, Salazar G, van den Berg C, Schuiteman A. 2015. An updated classification of Orchidaceae. Bot. J. Linn. Soc. 177:151–174.

Chase MW, Fay MF. 2009. Barcoding of Plants and Fungi. Science (80-.). 325:682–683.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR,

Price RA, Hills HG, Qiu Y-L, et al. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcL*. Ann. Missouri Bot. Gard. 80:528–580.

Chase MW. 2005. Classification of Orchidaceae in the age of DNA data. Curtis's Bot. Mag. 22:2–7.

Chikhi R, Medvedev P. 2014. Informed and Automated k -Mer Size Selection for Genome Assembly. Bioinformatics 30:31–37.

Chochai A, Leitch IJ, Ingrouille MJ, Fay MF. 2012. Molecular phylogenetics of *Paphiopedilum* (Cypripedioideae; Orchidaceae) based on nuclear ribosomal ITS and plastid sequences. Bot. J. Linn. Soc. 170:176–196.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. 2006a. The complete chloroplast genome sequence of *Pelargonium* x hortorum: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol. Biol. Evol. 23:2175–2190.

Cosner ME, Jansen RK, Palmer JD, Downie SR. 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr. Genet. 31:419–429.

Cosner ME, Raubeson LA, Jansen RK. 2004. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol. Biol. 4:27.

Cowan RS, Chase MW, Kress WJ, Savolainen V. 2006. 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. Taxon 55:611–616.

Cox A V, Pridgeon AM, Albert VA, Chase MW. 1997. Phylogenetics of the slipper orchids (*Cypripedioideae, Orchidaceae*): nuclear rDNA ITS sequences. Plant Syst. Evol. 208:197–223.

Cribb P. 1998. The Genus *Paphiopedilum*. (2nd ed.). Kota Kinabalu: Natural History Publications (Borneo) in association with Royal Botanic Gardens, Kew

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring J V., Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99:291–311.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36.

Curtis SE, Clegg MT. 1984. Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. 1:291–301.

Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCF tools. Bioinformatics 27:2156–2158.

Davis CC, Xi Z, Mathews S. 2014. Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. BMC Biol. 12:11.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinformatics Chapter 10:Unit 10.3.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Dessimoz C, Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11:R37.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. 2016. Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae). Biol. J. Linn. Soc. 117: 96–105.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36:969–970.

Dong W, Liu J, Yu J, Wang L, Zhou S. 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS One 7:e35071.

Dong W, Xu C, Cheng T, Lin K, Zhou S. 2013. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of saxifragales. Genome Biol. Evol. 5:989–997.

Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015. ycf1, the most promising plastid DNA barcode of land plants. Sci. Rep. 5:8348.

Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Res. 18:93–105.

Doyle J, Ballenger JA, Palmer J. 1996. The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. Mol. Phylogenet. Evol. 5:429–438 EP – .

Du FK, Lang T, Lu S, Wang Y, Li J, Yin K. 2015. An improved method for chloroplast genome sequencing in non-model forest tree species. Tree Genet. Genomes 11:114.

Dugas D V, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT, et al. 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. Sci. Rep. 5:16958.

Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, et al. 2011. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. Genome Res. 21:2224–2241.

Edgar RC, Batzoglou S. 2006. Multiple sequence alignment.Bishop MJ, editor. Curr. Opin. Struct. Biol. 16:368–373.

Edwards S V, Fertil B, Giron A, Deschavanne PJ. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Syst. Biol. 51:599–613.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. Science (80-.). 300:1706–1707.

Fajardo D, Senalik D, Ames M, Zhu H, Steffan SA, Harbut R, Polashock J, Vorsa N, Gillespie E, Kron K, et al. 2013. Complete plastid genome sequence of *Vaccinium macrocarpon*: Structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genet. Genomes. 9:489–498.

Farris JS, Kallersjo M, Kluge a G, Bult C. 1995. Testing significance of incongruence. Cladistics 10:315–319.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH. 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS One 3:e2802.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Syst. Biol. 27:401–410.

Fulton TM, Chunwongse J, Tanksley SD. 1995. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol. Biol. Report. 13:207–209.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. U. S. A. 92:11317–11321.

Gaut BS. 1998. Molecular clocks and nucleotide substitution rates in higher plants. In: Evolutionary Biology Vol 30. Vol. 30. p. 93–120.

Golenberg EM, Clegg MT, Durbin ML, Doebley J, Ma DP. 1993. Evolution of a noncoding region of the chloroplast genome. Mol. Phylogenet. Evol. 2:52–64.

Goremykin V V., Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol.

22:1813–1822.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2003. The chloroplast genome of the “basal” angiosperm *Calycanthus fertilis*-structural and phylogenetic analyses. Plant Syst. Evol. 242:119–135.

Gorokhova E. 2005. Effects of preservation and storage of microcrustaceans in RNA*later* on RNA and DNA degradation. Limnol. Oceanogr. Methods 3:143–148.

Goulding SE, Olmstead RG, Morden CW, Wolfe KH. 1996. Ebb and flow of the chloroplast inverted repeat. Mol. Gen. Genet. 252:195–206.

Graham SW, Reeves PA, Burns ACE, Olmstead RG. 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. Int. J. Plant Sci. 161:S83–S96.

Grandillo S, Chetelat R, Knapp S, Spooner D, Peralta I, Cammareri M, Perez O, Termolino P, Tripodi P, Chiusano ML, et al. 2011. Wild Crop Relatives: Genomic and Breeding Resources. (Kole C, editor.). Berlin, Heidelberg: Springer Berlin Heidelberg

Gribaldo S, Philippe H. 2002. Ancient Phylogenetic Relationships. Theor. Popul. Biol. 61:391–408.

Guisinger MM, Kuehl J V, Boore JL, Jansen RK. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc. Natl. Acad. Sci. U. S. A. 105:18424–18429.

Guo YY, Luo YB, Liu ZJ, Wang XQ. 2012. Evolution and biogeography of the slipper orchids: Eocene vicariance of the conduplicate genera in the old and new world tropics. PLoS One 7: e38788.

Guo YY, Luo YB, Liu ZJ, Wang XQ. 2015. Reticulate evolution and sea-level fluctuations together drove species diversification of slipper orchids

(*Paphiopedilum*) in Southeast Asia. Mol. Ecol. 24:2838-2855.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, et al. 2013. Next-generation museomics disentangles one of the largest primate radiations. Syst. Biol. 62:539–554.

Haberle RC, Fourcade HM, Boore JL, Jansen RK. 2008. Extensive Rearrangements in the Chloroplast Genome of *Trachelium caeruleum* Are Associated with Repeats and tRNA Genes. J. Mol. Evol. 66:350–361.

Hachtel W, Neuss A, Stein J Vom. 1991. A chloroplast DNA inversion marks an evolutionary split in the genus *Oenothera*. Evolution (N. Y). 45:1050–1052.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10:R32.

Healey A, Furtado A, Cooper T, Henry RJ. 2014. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods 10:21.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270:313–321.

Henriquez CL, Arias T, Pires JC, Croat TB, Schaal BA. 2014. Phylogenomics of the plant family Araceae. Mol. Phylogenet. Evol. 75:91–102.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, et al. 1989. The complete sequence of the rice (Oryza sativa) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. 217:185–194.

Höhl M, Ragan M a. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst. Biol. 56:206–221.

Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM. 2009. Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol. Ecol. Resour. 9:439–457.

Hu F, Gao N, Zhang M, Tang J. 2011. Maximum likelihood phylogenetic reconstruction using gene order encodings. 2011 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol.:1–6.

Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138:1723–1733.

Huang DI, Hefer CA, Kolosova N, Douglas CJ, Cronk QCB. 2014. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). New Phytol. 204:693–703.

Huang H, Shi C, Liu Y, Mao S-Y, Gao L-Z. 2014. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol. Biol. 14:151.

Humphries EM, Winker K. 2010. Working through polytomies: Auklets revisited. Mol. Phylogenet. Evol. 54:88–96.

Hupfer H, Swiatek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears B. 2008. Complete nucleotide sequence of the *Oenothera* elata plastid chromosome, representing plastome I of the five distinguishable *Euoenothera* plastomes. Mol. Genet. Genomics 280:363.

Jansen RK, Cai Z, Raubeson L a, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. U. S. A. 104:19369–19374.

Jansen RK, Palmer JD. 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc. Natl. Acad. Sci. U. S. A. 84:5818–5822.

Jansen RK, Raubeson L a, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, et al. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 395:348–384.

Jansen RK, Wee JL, Millie D. 1998. Molecular systematics of plants II. In: Soltis DE, Soltis PS, Doyle JJ, editors. Springer US. p. 87–100.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Joseph J, Sasikumar R. 2006. Chaos game representation for comparison of whole genomes. BMC Bioinformatics 7:243.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. 2012. Ultra-barcoding in cacao (*Theobroma* spp.;Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am. J. Bot. 99:320–329.

Kapralov M V, Filatov D a. 2006. Molecular adaptation during adaptive radiation in the Hawaiian endemic genus *Schiedea*. PLoS One 1:e8.

Kasem, S., Rice, N., & Henry RJ. 2008. 14 DNA Extraction from Plant Tissue. Plant Genotyping II SNP Technology.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kelchner S a. 2000. The evolution of non-coding chloroplast DNA and its aplications in

plant systematics. Ann. Missouri Bot. Gard. 87:482–498.

Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11:R116.

Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim J-H. 2015. Seven New Complete Plastome Sequences Reveal Rampant Independent Loss of the *ndh* Gene Family across Orchids and Associated Instability of the Inverted Repeat/Small Single-Copy Region Boundaries. PLoS One 10:e0142215.

Kim JS, Kim HT, Kim J-H. 2014. The Largest Plastid Genome of Monocots: a Novel Genome Type Containing AT Residue Repeats in the Slipper Orchid *Cypripedium japonicum*. Plant Mol. Biol. Report.

Kim KJ, Choi KS, Jansen RK. 2005. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol. Biol. Evol. 22:1783–1792.

Kim KJ, Lee HL. 2004. Complete chloroplast genome sequences from Korean ginseng (Panax schinseng Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 11:247–261.

Kim Y-D, Jansen RK. 1994. Characterization chloroplast and phylogenetic rearrangement distribution of a in the Berberidaceae. Plant Syst. Evol. 193:107–114.

Kishima Y, Ogura K, Mizukami K, Mikami T, Adachi T. 1995. Chloroplast DNA analysis in buckwheat species: phylogenetic relationships, origin of the reproductive systems and extended inverted repeats. Plant Sci. 108:173–179.

Knox EB, Palmer JD. 1999. The chloroplast genome arrangement of *Lobelia thuliniana* (*Lobeliaceae*): Expansion of the inverted repeat in an ancestor of the *Campanulales*. Plant Syst. Evol. 214:49–64.

Kolekar P, Kale M, Kulkarni-Kale U. 2012. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular

phylogeny and subtyping. Mol. Phylogenet. Evol. 65:510–522.

Kozarewa I, Ning Z, Quail M a, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods 6:291–295.

Kress WJ, Erickson DL. 2008. DNA barcodes: genes, genomics, and bioinformatics. Proc. Natl. Acad. Sci. U. S. A. 105:2761–2762.

Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 11:81.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Kunisawa T. 2001. Gene arrangements and phylogeny in the class Proteobacteria. J. Theor. Biol. 213:9–19.

Kunisawa T. 2003. Gene arrangements and branching orders of gram-positive bacteria. J. Theor. Biol. 222:495–503.

Kunnimalaiyaan M, Nielsen BL. 1997. Fine mapping of replication origins (ori A and ori B) in *Nicotiana tabacum* chloroplast DNA. Nucleic Acids Res. 25:3681–3686.

Kurtz S, Narechania A, Stein JC, Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517.

Lamppa GK, Bendich a J. 1979. Changes in Chloroplast DNA Levels during Development of Pea (*Pisum sativum*). Plant Physiol. 64:126–130.

Lemmon EM, Lemmon AR. 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.

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

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

Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17:149–154.

Li M, Vitâanyi P. 2009. An Introduction to Kolmogorov Complexity and Its Applications. Texts in Computer Science. Springer New York.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010. The sequence and *de novo* assembly of the giant panda genome. Nature 463:311–317.

Li R, Ma PF, Wen J, Yi TS. 2013. Complete Sequencing of Five Araliaceae Chloroplast Genomes and the Phylogenetic Implications. PLoS One 8:1–15.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 20:265–272.

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015. Plant DNA barcoding: from gene to genome. Biol. Rev 90:157–166.

Lin Y, Hu F, Tang J, Moret BME. 2013. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. Pac. Symp. Biocomput.:285–296.

Lin Y, Moret BME. 2010. A new genomic evolutionary model for rearrangements, duplications, and losses that applies across eukaryotes and prokaryotes. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 6398 LNBI. p. 228–239.

Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W, et al. Estimation of genomic characteristics by analyzing k- mer frequency in *de novo* genome projects. arXiv preprint. arXiv:1308.2012.

Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X. 2012. CpGAVAS, an integrated web server for the annotation, visualisation, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13:715.

Liu J, Qi ZC, Zhao YP, Fu CX, Jenny Xiang QY. 2012. Complete cpDNA genome sequence of *Smilax china* and phylogenetic placement of Liliales - Influences of gene partitions and taxon sampling. Mol. Phylogenet. Evol. 64:545–562.

Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30:434–439.

Luo H, Shi J, Arndt W, Tang J, Friedman R. 2008. Gene order phylogeny of the genus *Prochlorococcus*. PLoS One 3.

Luo H, Sun Z, Arndt W, Shi J, Friedman R, Tang J. 2009. Gene order phylogeny and the evolution of methanogens. PLoS One 4: e6069.

Luo J, Hou B-W, Niu Z-T, Liu W, Xue Q-Y, Ding X-Y. 2014. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. PLoS One 9:e99016.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1:18.

Ma J, Yang B, Zhu W, Sun L, Tian J, Wang X. 2013. The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. Gene 528:120–131.

Ma PF, Zhang YX, Zeng CX, Guo ZH, Li DZ. 2014. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (poaceae). Syst Biol 63:933–950.

Maddison W. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

Maddison WP, Maddison DR. 2008. Mesquite: A modular system for evolutionary analysis. Evolution (N. Y). 62:1103–1118.

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20:1700–1710.

Maier RM, Neckermann K, Igloi GL, Kössel H. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. 251:614–628.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, Kougbeadjo A, Maillol V, Martin G, Sabot F, et al. 2014. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol. Ecol. Resour. 14:1109–1113.

Marshall JA, Knapp S, Davey MR, Power JB, Cocking EC, Bennett MD, Cox A V. 2001. Molecular systematics of *Solanum* section *Lycopersicum* (*Lycopersicon*) using the nuclear ITS rDNA region. Theor. Appl. Genet. 103:1216–1222.

Martin G, Baurens F-C, Cardi C, Aury J-M, D'Hont A. 2013. The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. PLoS One 8:e67350.

Martin W, Deusch O, Stawski N, Grünheit N, Goremykin V. 2005. Chloroplast genome

phylogenetics: Why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:203-209.

McClean PE, Hanson MR. 1986. Mitochondrial DNA Sequence Divergence among *Lycopersicon* and Related *Solanum* Species. Genetics 112:649–667.

McCoy SR, Kuehl J V, Boore JL, Raubeson LA. 2008. The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. BMC Evol. Biol. 8:130.

McPherson H, van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M. 2013. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol. 13:8.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11:31–46.

Miller JC, Tanksley SD. 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor. Appl. Genet. 80:437–448.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop, GCE 2010. p. 1–8.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shawand PD, Marshall D. 2013. Using tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. 14:193–202.

Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 12:R112.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. U. S. A. 104:19363–19368.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol. 6:17.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl. Acad. Sci. U. S. A. 107:4623–4628.

Morozova O, Marra M a. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–264.

Moyle LC. 2008. Ecological and evolutionary genomics in the wild tomatoes (*Solanum* Sect. *Lycopersicon*). Evolution (N. Y). 62:2995–3013.

Muir G, Filatov D. 2007. A selective sweep in the chloroplast DNA of dioecious silene (Section *Elisanthe*). Genetics 177:1239–1247.

Mullet JE. 1988. Chloroplast Development and Gene Expression. 1967:475–502.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. Nat. Rev. Genet. 14:157–167.

Naito K, Kaga A, Tomooka N, Kawase M. 2013. De novo assembly of the complete organelle genome sequences of azuki bean (*Vigna angularis*) using next-generation sequencers. Breed. Sci. 63:176–182.

Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M. 2009. Phylogenetic utility of *ycf1* in orchids: A plastid gene more variable than *matK*. Plant Syst. Evol. 277:75–84.

Neyland R, Urbatsch LE. 1996. Phylogeny of subfamily Epidendroideae (Orchidaceae) inferred from *ndhF* chloroplast gene sequences. Am. J. Bot. 83:1195–1206.

Nikiforova S V, Cavalieri D, Velasco R, Goremykin V. 2013. Phylogenetic Analysis of 47 Chloroplast Genomes Clarifies the Contribution of Wild Species to the Domesticated Apple Maternal Line. Mol. Biol. Evol. 30:1751-1760.

Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and

pitfalls of phylogenomics. Genome Biol. 8:R199.

Nock CJ, Waters DLE, Edwards M a, Bowen SG, Rice N, Cordeiro GM, Henry RJ. 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol. J. 9:328–333.

Ogihara Y, Terachi T, Sasakuma T. 1988. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proc. Natl. Acad. Sci. U. S. A. 85:8573–8577.

Olmstead RG, Palmer JD. 1994. Chloroplast DNA systematics: A review of methods and data analysis. Am. J. Bot. 81:1205–1224.

Otu HH, Sayood K. 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19:2122–2130.

Oxelman B, Lidén M, Berglund D. 1997. Chloroplast *rps16* intron phylogeny of the tribe Sileneae (Caryophyllaceae). Plant Syst. Evol. 206:393–410.

Palmer JD, Nugent JM, Herbon L a. 1987. Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proc. Natl. Acad. Sci. U. S. A. 84:769–773.

Palmer JD, Osorio B, Thompson WF. 1988. Evolutionary significance of inversions in legume chloroplast DNAs. Curr. Genet. 14:65–74.

Palmer JD, Stein DB. 1986. Conservation of chloroplast genome structure among vascular plants. Curr. Genet. 10:823–833.

Palmer JD, Zamir D. 1982. Chloroplast DNA evolution and phylogenetic relationships in *Lycopersicon*. Proc. Natl. Acad. Sci. U. S. A. 79:5006–5010.

Palmer JD. 1985. Chloroplast DNA and molecular phylogeny. BioEssays 2:263–267.

Pan I-C, Liao D-C, Wu F-H, Daniell H, Singh ND, Chang C, Shih M-C, Chan M-T, Lin C-S. 2012. Complete chloroplast genome sequence of an orchid model plant candidate: *Erycina pusilla* apply in tropical Oncidium breeding. PLoS One 7:e34738.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7:84.

Paszkiewicz K, Studholme DJ. 2010. *De novo* assembly of short sequence reads. Brief. Bioinform. 11:457–472.

Peralta IE, Spooner DM, Knapp S. 2008. Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersoides*, sect. *Juglandifolia*, sect. *Lycpersicon*; Solanaceae). Syst. Bot. Monogr. 84:1–186.

Peralta IE, Spooner DM. 2001. Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicon* [Mill.] Wettst. subsection *Lycopersicon*). Am. J. Bot. 88:1888–1902.

Perry AS, Brennan S, Murphy DJ, Kavanagh T a, Wolfe KH. 2002. Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res. 9:157–162.

Perry AS, Wolfe KH. 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. J. Mol. Evol. 55:501–508.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. U. S. A. 98:9748–9753.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Phylogenomics. Annual Rev. Ecol. Evol. Syst. Annu. Rev. 36:541–562.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

Pyron RA, Hendry CR, Chou VM, Lemmon EM, Lemmon AR, Burbrink FT. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol. 81:221–231.

Qi J, Luo H, Hao B. 2004. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32:W45–W47.

Qi J, Wang B, Hao B-I. 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58:1–11.

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics 8:174.

Ravi V, Khurana JP, Tyagi a. K, Khurana P. 2008. An update on chloroplast genomes. Plant Syst. Evol. 271:101–122.

Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008. An update on chloroplast genomes. Plant Syst. Evol. 271:101–122.

Rick C.M. 1986. Reproductive isolation in the *Lycopersicum peruvianum* complex. In Solanaceae: Biology and Systematics (D'Arcy, W.G.D., ed.). New York, NY: Columbia University Press, pp. 477–495.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat. Biotechnol. 29:24–26.

Rodriguez F, Wu F, Ané C, Tanksley S, Spooner DM. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? BMC Evol. Biol. 9:191.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Rokas A, Holland PWH. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15:454–459.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms - inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol. Biol. 14:23.

Ruhsam M, Rai HS, Mathews S, Ross TG, Sean W, Raubeson L a, Mei W, Thomas PI, Gardner MF, Ennos R a, et al. 2015. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22:557–567.

Samson N, Bausher MG, Lee S-B, Jansen RK, Daniell H. 2007. The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. Plant Biotechnol. J. 5:339–353.

Sanderson MJ, McMahon MM, Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10:155.

Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. U. S. A. 89:6575–6579.

Särkinen T, Bohs L, Olmstead RG, Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol. Biol. 13:214.

Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, Jansen RK. 2005. Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. Plant Mol. Biol. 59:309–322.

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20:1165–1173.

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast

genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The Tortoise and the hare III. Am. J. Bot. 94:275–288.

Shaw J, Shafer HL, Rayne Leonard O, Kovach MJ, Schorr M, Morris AB. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. Am. J. Bot. 101:1987–2004.

Shifman A, Ninyo N, Gophna U, Snir S. 2014. Phylo SI: A new genome-wide approach for prokaryotic phylogeny. Nucleic Acids Res. 42:2391–2404.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5:2043–2049.

Simmons MP, Richardson D, Reddy ASN. 2008. Incorporation of gap characters and lineage-specific regions into phylogenetic analyses of gene families from divergent clades: An example from the kinesin superfamily across eukaryotes. Cladistics 24:372–384.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15:121–132.

Sims GE, Jun S-R, Wu GA, Kim S-H. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. U. S. A. 106:2677–2682.

Spangler RE, Olmstead RG. 1999. Phylogenetic analysis of Bignoniaceae based on the cpDNA gene sequences *rbcL* and *ndhF*. Ann. Missouri Bot. Gard. 86:33–46.

Spooner DM, Peralta IE, Knapp S. 2005. Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon* (Mill.) Wettst.]. Taxon 54:43–61.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-

analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stein A, Takasuka TE, Collings CK. 2010. Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? Nucleic Acids Res. 38:709–719.

Teichmann S a., Mitchison G. 1999. Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. 49:98–107.

Tewari KK, Kolodner R. 1979. Inverted repeats in chloroplast DNA from higher plants. Proc. Natl. Acad. Sci. U. S. A. 76:41–45.

Thomas GH. 2015. Evolution: An avian explosion. Nature:10–11.

Timme RE, Kuehl J V., Boore JL, Jansen RK. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. Am. J. Bot. 94:302-312

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13:36–46.

Turner FS. 2014. Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. Front. Genet. 5:1–7.

Vaillancourt RE, Jackson HD. 2000. A chloroplast DNA hypervariable region in eucalypts. Theor. Appl. Genet. 101:473–477.

Vieira L do N, Faoro H, Fraga HP de F, Rogalski M, de Souza EM, de Oliveira Pedrosa F, Nodari RO, Guerra MP. 2014. An improved protocol for intact chloroplasts and cpDNA isolation in conifers.Theg SM, editor. PLoS One 9:e84792.

Vinga S, Carvalho AM, Francisco AP, Russo LM, Almeida JS. 2012. Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis. Algorithms Mol. Biol. 7:10.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl. Acad. Sci. U. S. A. 91:9794–9798.

Wang R-J, Cheng C-L, Chang C-C, Wu C-L, Su T-M, Chaw S-M. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol. Biol. 8:36.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol. Ecol. 19 Suppl 1:100–114.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol. Biol. 76:273–297.

Wolf YI, Rogozin IB, Grishin N V, Tatusov RL, Koonin E V. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1:8.

Wolf YI, Rogozin IB, Koonin E V. 2004. Coelomata and not ecdysozoa: Evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84:9054–9058.

Wu CS, Chaw SM, Huang YY. 2013. Chloroplast phylogenomics indicates that *Ginkgo biloba* is sister to cycads. Genome Biol. Evol. 5:243–254.

Wu CS, Lai YT, Lin CP, Wang YN, Chaw SM. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: Selection toward a lower-cost strategy. Mol. Phylogenet. Evol. 52:115–124.

Wu C-S, Wang Y-N, Liu S-M, Chaw S-M. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol. Biol.

Evol. 24:1366–1379.

Wu F-H, Chan M-T, Liao D-C, Hsu C-T, Lee Y-W, Daniell H, Duvall MR, Lin C-S. 2010. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol. 10:68.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc. Natl. Acad. Sci. 109:17519–17524.

Yang JB, Li DZ, Li HT. 2014. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. Mol. Ecol. Resour. 5: 1024–1031.

Yang JB, Tang M, Li HT, Zhang ZR, Li DZ. 2013. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol. Biol. 13:84.

Yang JB, Yang SX, Li HT, Yang J, Li DZ. 2013. Comparative Chloroplast Genomes of Camellia Species. PLoS One 8:1–12.

Yang K, Zhang L. 2008. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Res. 36:e33.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13:303–314.

Ye L, Hillier LW, Minx P, Thane N, Locke DP, Martin JC, Chen L, Mitreva M, Miller JR, Haub K V, et al. 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 12:R31.

Yi DK, Kim KJ. 2012. Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. PLoS One 7.

Yi H, Jin L. 2013. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Res. 41:e75.

Yi X, Gao L, Wang B, Su Y-J, Wang T. 2013. The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): evolutionary comparison of cephalotaxus chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. Genome Biol. Evol. 5:688–698.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat. Commun. 5:4956.

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zhang Y, Hu F, Tang J. 2010. Phylogenetic reconstruction with gene rearrangements and gene losses. In: Proceedings - 2010 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2010. p. 35–38.

Zhang Y, Ma J, Yang B, Li R, Zhu W, Sun L, Tian J, Zhang L. 2014. The complete chloroplast genome sequence of *Taxus chinensis var. mairei* (Taxaceae): loss of an inverted repeat region and comparative analysis with related species. Gene 540:201–209.

Zhang YJ, Ma PF, Li DZ. 2011. High-throughput sequencing of six bamboo chloroplast genomes: Phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS One 6: e20596.

Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2015. Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. New Phytol. 209:1747–1756.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008.

Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biol. 9:R49.

Zuriaga E, Blanca J, Nuez F. 2009. Classification and phylogenetic relationships in *Solanum* section *Lycopersicon* based on AFLP and two nuclear gene sequences. Genet. Resour. Crop Evol. 56:663–678.

Acknowledgements

This phD thesis may never seen the light without the help of many generous people.

My sincerest appreciation must go to my supervisor, René Smulders, who took the risk of supervising me even knowing that I was different background. Many thanks for his brilliance, guidance, advice, patience, and constant care which I am grateful to have you as my supervisor. It was not an easy project because all the work that had to be stopped after disappointing results. I appreciated his very positive personality to keep me up with the good spirit and motivation. I liked the freedom you gave me within the research, but steered me in the right direction whenever I needed it. I really value his consideration especially when I had to take care of my son alone in the last two years of my study.

I could have never reached this stage of my phD without Theo Borm. I encountered many difficulties to analyse my sequencing data related to the fact that I was lacked of the skill sets required in the field of bioinformatics. Theo, you were the first one to show me bioinformatics is the most useful tool to a biologist could have to understand genetics. During the whole process, he tried to provide me the tools that allow me to follow the right path. Theo was always very busy, but he managed to find time to discuss with me whenever I encounter technical problems. I admired his realistic views and reasonable approach that keep my feet on the ground when things get overwhelming. Moreover, I appreciate his help with the writing when writing happened to be difficult with all the technical terms and involved complex bioinformatics process.

I would like to extend my gratitude towards my promoter Prof. Richard Visser. You accepted me as part of Wageningen UR Plant Breeding, where you offered a very good scientific environment for my research and study. Thanks for all useful discussion on my phD progress and for the critical comments on my thesis.

Thank you to Barbara Gravendeel from Naturalis, Leiden, Freek Bakker from Wageningen University, Yung I from National Museum of Natural Science, Taichung, Taiwan, Mike Fay and Brian Yap from Kew garden, London for their contributions to

this thesis especially for providing sample of *Paphiopedilum* spp and for their attention reading my manuscripts.

I would like to express my appreciation to the technicians in the laboratory Plant Breeding from whom I received the most efficient technical support. I thank all of you Marian, Elly, Dorette, Dianka, Danny, Johan, Koen, and Gert. Special thanks to Danny Esselink, for your guidance in the lab at the beginning. Thank you for helping me with the DNA extractions and teaching me so many tricks in the lab. It was a pleasure to work with you.

It was really a great pleasure to make many friends from various backgrounds in Wageningen. Special thanks to my best friend forever, Xuan Xu for always being there and bearing with me the good and bad times during my wonderful days of phD. She was a true friend ever since we began to share an office in the start of our phD journey. Xuan, you are an amazing person in too many ways. I am grateful to have you as my happy pills, punching bag, and crime partner whenever I needed. I also thank Tim van der Weijde, Jordi Petit Pedro, and Mathilde Daniau, my other great office mates who have been supportive in every way. I appreciated the nice talks during the breaks with you guys. To my friends who I met in Wageningen, Peter (Quy Dinh), Johan Bucher, PingPing Huang, Michel Arts, Cheng Liu, Tom van Stein, Marcela, Mas Muniroh, Marian Oortwijn, Bart van Tuijl, Michela Appiano, Mylusk Carolina, Anne Giesbers, Valentina Bracuto, Arwa Shahin, thank you for listening, offering me advice and supporting me through the entire process.

Thanks to the Malaysian community here in Wageningen. To Uncle Alan and Auntie Tony, thank you for helping us and assisting us in many things. Special thanks to HudaK, Ani, Su for the weekend sleepover and picnics. To Razak and Nozie, Naim and Fatimah thank you for helping me to babysit Aryan especially in the summer holidays. Thanks to Yani, Due, Nuyu, Hafeez, Mas, Moritz, Aidil and Tihah for the help and kindness. And for the others, Nazri, Zul, Lini, Shikin, Shahrul, Arina, Razak, Shakila and Azie for sharing pleasant moments on various occasions.

Thanks to my family and my family in-laws in Malaysia for their constant support and prayers for me. They have sacrificed a lot to make sure I can finish my study, which I can never ever repay.

I close my personal acknowledgements expressing my gratitude to my dearest husband, Mohd Ameen for his enormous contributions to the completion of my phD. There is no way I can succeed without your flexibility, patience, love, support and attentiveness. I am thankful to my lovely sons, Aryan and Arman for being with us and giving us a lot of joy!

Finally, my acknowledgements go to the Ministry of Higher Education (MOHE) Malaysia and Universiti Putra Malaysia (UPM) for providing the funding for this thesis. The research was supported in part by the Netherlands' Ministry of Economic Affairs (KB-14-004-030).

Shairul Izan Ramlee
Wageningen, 19 September 2016

Curriculum Vitae

Shairul Izan Bt Ramlee was born in 12th September 1984 in Kuala Lumpur Malaysia. After obtaining her high school degree, she went to Universiti Putra Malaysia for her Bachelor study in Science of Horticulture. Then, in 2008 she was appointed as a tutor at Universiti Putra Malaysia. She continued with her Master study in Plant Science at the Wageningen University, The Netherlands. In September 2011, she was enrolled in a phD programme in the laboratory of Plant Breeding at Wageningen University under the supervison of Dr. René Smulders. Starting 1st of November 2016, she is a lecturer at the Faculty of Agriculture, Universiti Putra Malaysia.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (6 ECTS)

- Plastome assembly based on k-mer size selection methods

Writing of project proposal (4.5 ECTS)

- Novel methods for phylogenetic and phyleogeographic studies of slippers orchids using next generation sequencing (2012)

Post-graduate courses (5.6 ECTS)

- Current trends in phylogenetics; WUR (2011)
- Workshop: bioinformatics and statistical genetics and genomic s, using PLAZA; WUR (2012)
- Bioinformatics – an user's approach; WUR (2013)
- Next generation sequencing; MGC (2013)
- RNA-seq Data analysis; NBIC (2014)

Laboratory training and working visits (1.8 ECTS)

- Data analysis, sampling samples for DNA sequencing and meeting discussion for collaborations; Kew garden, UK (2013)
- Sampling and data analysis; Leiden University (2013)

Competence strengthening / skills courses (3 ECTS)

- Scientific writing; Wageningen in'to languages (2014)
- Efficient writing strategies; Wageningen in'to languages (2015)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.2 ECTS)

- PE&RC Day: optimization of science: pressure and pleasure (2014)
- WGS Symposium: WGS PhD workshop carousel (2014)
- WGS Symposium: 2nd Wageningen PhD symposium (2015)
- WGS Symposium: WGS PhD workshop carousel (2015)

Discussion groups / local seminars / other scientific meetings (12.3 ECTS)

- Biodiversity group meeting (2011-2015)

- Plant breeding Monday colloquium (2011-2015)
- Plant breeding 11 years conference (2012)
- Symposium: improving yield prediction: EU SPICY (2012)
- Experimental plant sciences meeting (2013)
- Symposium: omic advances for academia and industry-towards true molecular plant breeding (2014)
- B-WISE: Bioinformatics@wageningen seminar series (2015)
- IOGA Workshop (2015)
- Experimental plant sciences meeting (2015)

International symposia, workshops and conferences (2.9 ECTS)

- Symposium: all inclusive breeding; poster presentation; Wageningen (2014)
- Bioinformatics & system biology conference; poster presentation; the Netherlands (2015)

This research was conducted at the Laboratory Plant Breeding of Wageningen University and was financially supported by the Ministry of Higher Education Malaysia within the framework of the graduate school of Production Ecology & Resource Conservation

Cover and layout by: Shahira Aishah Bt Ramlee

Printed by Digiforce: || Proefschriftmaken.nl ||