universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Computational analysis of plastid DNA insertions into nuclear and mitochondrial genomes in non-photosynthetic parasitic plants from next generation sequencing data

verfasst von / submitted by

Andreas Bachmayr, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 877
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Genetik und Entwicklungsbiologie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Thomas Rattei for his guidance and feedback and especially for his patience and continued support during a difficult time.

I would also like to thank Professor Gerald Schneeweiss, who initiated the project, for his input and direction. A big thanks also goes to Thomas Eder, PhD student at CUBE, on whose initial work the remainder of the project is based and who helped a novice in bioinformatics to overcome the occurring technical difficulties. The same goes for the various members of CUBE, who helped me with occasional problems.

Finally, I would like to thank my friends and my family. In particular, I want to express my deep gratitude to my parents, without whose unswerving support and encouragement it would not have been possible to finish my studies. I also would like to give thanks to my aunt and uncle, who were there for me, and to my brother, for his advice and for proofreading the final thesis.

Contents

Chapter 1

Introduction

Plastids, the organelles that, among other functions, enable plants to carry out photosynthesis, developed by primary endosymbiosis, the engulfment of a cyanobacteria-like prokaryote by an ancient eukaryote (see Figure 1.1). A century after its development, this endosymbiotic theory is now widely accepted. Relatively recent evidence narrowed the ancestor of the plastids down to a heterocyste-forming, nitrogen-fixing cyanobacterium (Deutsch et al., 2008).

Like mitochondria, plastids have their own genome and form their own compartment inside the eukaryotic cell. In contrast to free-living prokaryotes, however, their genome is strongly reduced in size, a consequence of the symbiotic relationship with their host cells (Martin et al., 2002). This apparent loss of genetic information in the plastids is caused by both complete loss of sequences no longer needed and by the transfer of DNA fragments from one cellular compartment to the other. This in turn leads to a dependence of the organelles on gene products imported from the nucleus. This process was discovered in the early 1980s and such fragments have been termed "promiscuous DNA" (Ellis, 1982; Stern and Lonsdale, 1982).

Most genetic material from the plastid to the nucleus seems to have been transferred in big amounts during early evolution (Deutsch et al., 2008; Martin et al., 2002). Why organelles still contain a core set of genes, or in other words, what sets the limit for transfer and/or loss of genetic material is not quite clear yet. There are different approaches to explain this phenomenon, for example the "hydrophobicity-improbability hypothesis", originally proposed by von Heijne (1986), which states that many of the transferred organellar gene products are too hydrophobic to allow efficient retargeting and import of the gene product from the nucleus to the organelle.

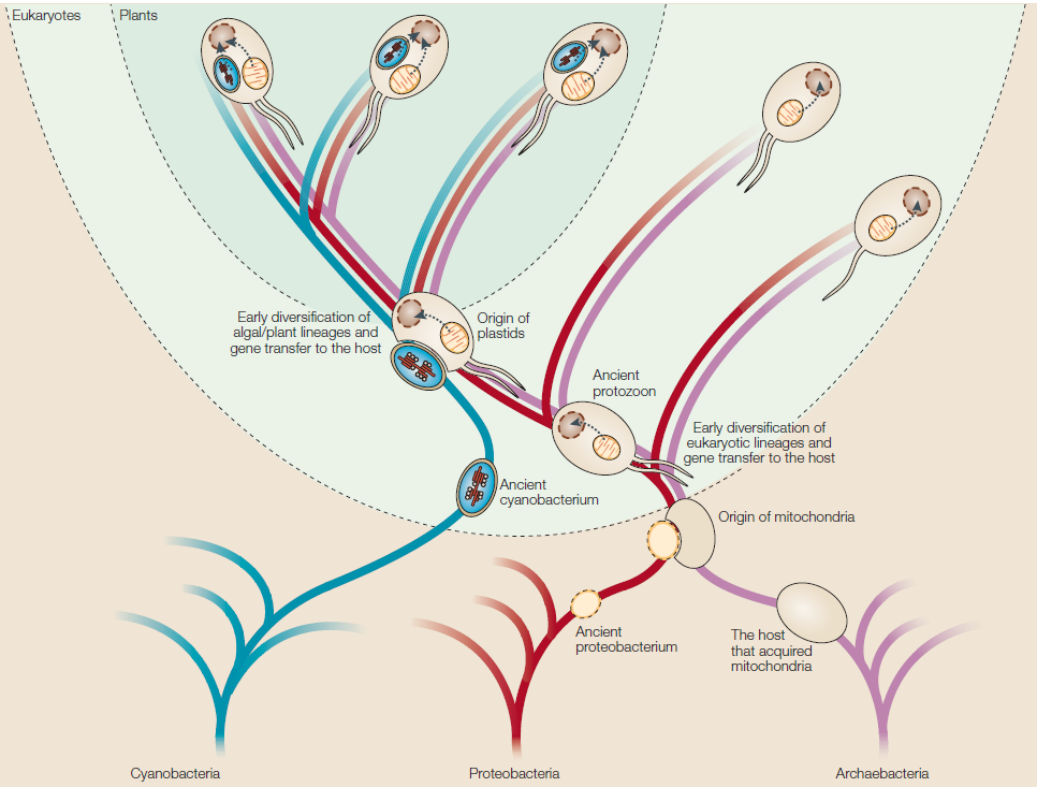


Figure 1.1: *A representation of the evolutionary origin of chloroplasts and mitochondria due to endosymbiosis. Adapted by permission from MacMillan Publishers Ltd: Nature Reviews Genetics 5, 123-135, Timmis et al., 2004, © 2004*

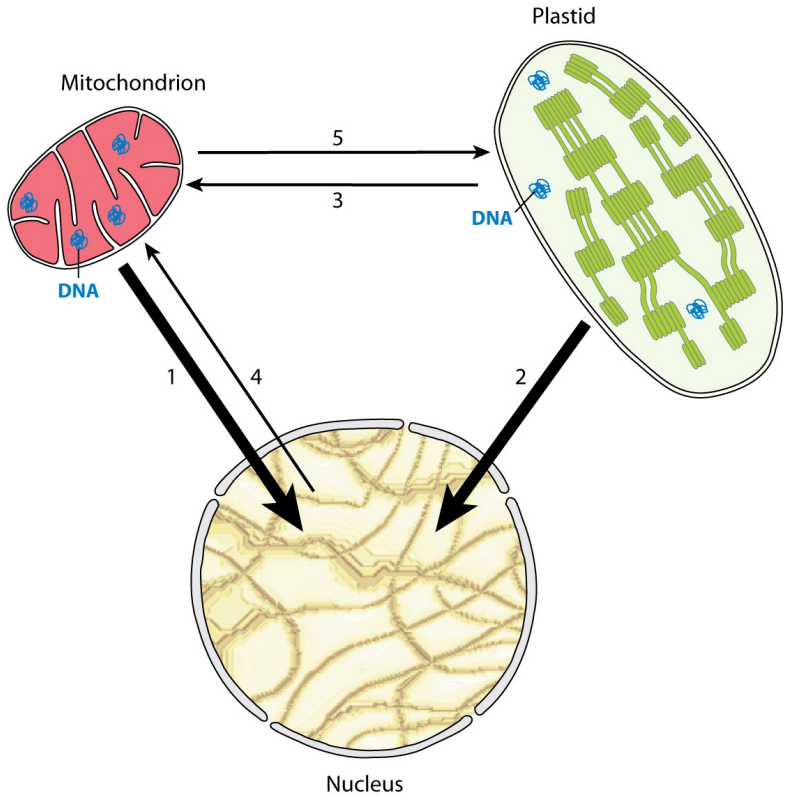
There are some exceptions to this hypothesis, however, for example the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase, which is usually encoded in the plastid genome, but which is water-soluble. It has been shown that functional transfer to the nucleus is possible for the large subunit (Kanevski and Maliga, 1994), albeit at a lower activity. An alternative explanation is the need of organelles to tightly control their electron transport chains. Import of required gene products from the nucleus might take too long, leading to dangerous accumulation of reactive oxygen species in the compartment (Allen, 1993a; Allen, 1993b).

Fragments from the plastid genome that have inserted into the nuclear genome have been termed NUPTs , which stands for "nuclear plastid DNA", while sequences homologous to mitochondrial DNA have been termed NUMTs (Michalovova, Vyskot, and Kejnovsky, 2013), for "nuclear mitochondrial DNA". A number of other terms have been introduced over the years,

such as MIPTs or MTPTs, for "mitochondrial plastid DNA" (Park et al., 2015; Wang et al., 2007), norgs as an umbrella term for NUPTs and NUMTs (Wang and Timmis, 2013), as well as more recently odins, which stands for "organelle-derived inserted sequences" (Samaniego Castruita et al., 2015), a concept very similar to norgs.

NUPTs and NUMTs are abundant in plant nuclear genomes (Matsuo et al., 2005; Michalovova, Vyskot, and Kejnovsky, 2013; Richly and Leister, 2004; Shahmuradov et al., 2003) and frequency of transfer appears to be very high and ongoing, according to studies using transgenic plastids performed on tobacco (Huang, Ayliffe, and Timmis, 2003; Stegemann et al., 2003), which allowed to study the process of organelle DNA transfer under laboratory conditions and in a much shorter time frame than in nature. Apart from transfers into the nucleus, transfers from the plastids to the mitochondria also appear to happen quite frequently (Stern and Palmer, 1984; Tsunewaki, 2011; Wang et al., 2007). Transfers from the nucleus to the mitochondria also seem to occur, albeit more rarely (Rodríguez-Moreno et al., 2011). Insertions of DNA fragments into the plastid genome happen very rarely and can be considered negligible (Smith, Crosby, and Lee, 2011; Straub et al., 2013). See Figure 1.2 for a visual representation.

Genes rarely remain functional in the nucleus after transfer. Instead, most coding sequences seem to become non-functional after integration due to deletions, insertions and recombination (Sheppard and Timmis, 2009), with a large part of the NUPTs and NUMTs being eliminated from the nuclear genome, as demonstrated for rice by Matsuo et al. (2005), who calculated that it takes about a million years for the nuclear genome of rice to eliminate 80% of the inserted NUPTs. A process, which appears to be a necessity for the nuclear genome to counteract the frequent insertions of organelle DNA, so rapid expansion is prevented.



Kleine T, et al. 2009.
Annu. Rev. Plant Biol. 60:115–38

Figure 1.2: *A schematic showing the different types of DNA transfer between cellular compartments. 1) transfer from mitochondrion to nucleus (NUMT) 2) transfer from plastid to nucleus (NUPT) 3) transfer from plastid to mitochondrion (MIPT) 4) transfer from nucleus to mitochondrion 5) transfer from mitochondrion to plastid. Adapted from Kleine, Maier, and Leister (2009).*

Since organelle genes are closer in structure to prokaryotic genes than to eukaryotic ones, genes transferred from the plastid to the nucleus need to acquire the promoter sequence of an upstream gene. This can happen through deletion of the region in between due to short repeat sequences or by repair of double-stranded breaks as demonstrated by Stegemann and Bock (2006). They also showed that an AT-rich noncoding region downstream of an inserted gene seems to facilitate RNA cleavage and polyadenylation for the generation of a functional transcript and they hypothesized that targeting sequences for the transfer of the gene products back to the plastid could be aquired the same way as eukaryotic promoters. But an analysis of *Arabidop-*

sis genes with cyanobacterial, plastid and yeast genomes by Martin et al. (2002), showed that most plastid genes in the nucleus aren't targeted back to their original compartment, indicating that they acquired new roles in the cell.

Functional transfers are not limited to the transfer of entire genes: organellar DNA can instead provide already existing nuclear genes with new exons as shown by Noutsos et al. (2007). These studies, among others, show the impact of organellar DNA transfer on plant evolution.
Where organelle DNA fragments insert into the nuclear genome of plants is still open for debate, but Matsuo et al. (2005) showed that in *Oryza sativa subsp. japonica* integration in the pericentromeric regions of the chromosomes is preferred and they hypothesized that centromeric and pericentromeric regions are better able to engulf NUPTs and NUMTs than other parts of the chromsome. Michalovova, Vyskot, and Kejnovsky (2013) confirmed these findings for rice and were also able to show the same pattern for *Arabidopsis thaliana*, but they found a more even distribution of inserts over the chromosomes for the larger genomes of maize and soybean. However, Wang and Timmis (2013) concluded in their study that organelle DNA preferentially inserts into open chromatin regions in *O. sativa subsp. indica*.

More recent and less diverged inserts are apparently also longer, indicating transfer of long sequences at once, and then decay and fracture over time (Richly and Leister, 2004; Huang et al., 2005). Michalovova, Vyskot, and Kejnovsky (2013) provided further evidence for this process and also proposed a model based on the dynamics of transposable elements (TEs), where organellar DNA initially inserts near the centromeres and then becomes fragmented due to insertions of TEs, with ectopic recombination between homologous TEs leading to deletion of inserted sequences (see figure 1.3). NUPTs (and NUMTs) also not only become fragmented but decay over time, with the majority of mutations apparently being the result of deamination of 5-methylcytosine, transforming the cytosine into thymine, which can lead to a faulty repair of the nucleotide mismatch (Huang et al., 2005).

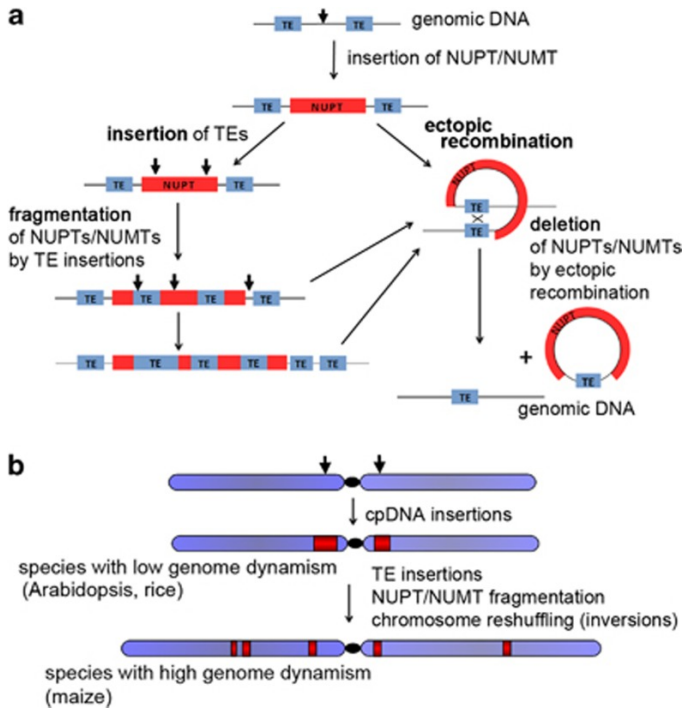
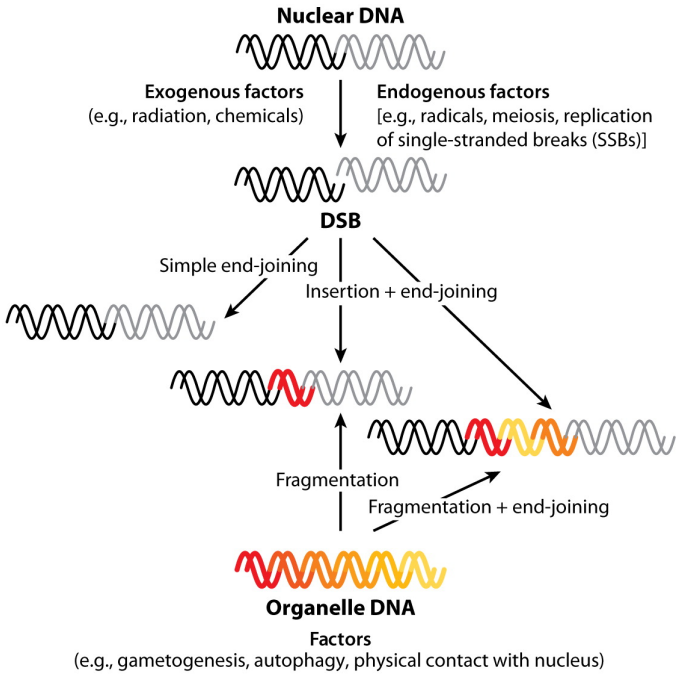


Figure 1.3: *The transposable elements dynamics model proposed by Michalovova, Vyskot, and Kejnovsky (2013). a) Insertion of TEs leading to fragmentation of NUPTs or outright deletion by ectopic recombination b) Different TE dynamics and chromosome reshuffling leads to movement of the fragments away from the centromeres. Adapted with permission of the author from Michalovova, Vyskot, and Kejnovsky (2013).*

The exact underlying mechanisms for transfer of plastid DNA to the nucleus and integration into the same are not completely clear yet, but non-homologous end joining (NHEJ) during the repair of double-stranded breaks (DSBs) seems to be involved (see figure 1.4), as shown, for example, by Lloyd and Timmis (2011), who were able to compare insertion sites before and after the integration event by using a further refined method based on the transplastomic lines used for the tobacco studies (Huang, Ayliffe, and Timmis, 2003; Stegemann et al., 2003). Wang and Timmis (2013) also found microhomologies at NUPT integration sites, thereby providing further evidence for the role of NHEJ during the insertion of organellar DNA into the nucleus genome.



Kleine T, et al. 2009.
Annu. Rev. Plant Biol. 60:115–38

Figure 1.4: *Factors hypothesized to be involved in the insertion of organellar DNA, among them non-homologous end joining. Adapted from Kleine, Maier, and Leister (2009).*

Identification of NUPTs is done bioinformatically by means of sequence similarity searches using the BLAST algorithm (Altschul et al., 1990), by comparing the plastid genome with the nuclear genome of the plant (Matsuo et al., 2005; Michalovova, Vyskot, and Kejnovsky, 2013; Noutsos et al., 2007). Complete plant nuclear genomes are not always readily available, however, since the successful assembly of plant genomes is exceedingly difficult. Plant genomes are often very large, even among eukaryotes, and often possess a higher level of ploidy due to whole genome duplication events (Bowers et al., 2003; Vanneste et al., 2014). Two or more haplotypes can influence the accuracy of an assembly and lead to false segmental duplications, where polymorphic regions are erroneously assembled into separate contigs, giving the impression of a recent duplication event (Kelley and Salzberg, 2010). However, genes are not only duplicated during whole genome duplication events. Chromosomal recombination may lead to the generation of tandemly arrayed genes, as shown for example by Rizzon, Ponger, and Gaut (2006) for *Oryza sativa* and *Arabidopsis thaliana*. The activity of different types of

transposable elements can also cause additional duplications (Jiang et al., 2004; Schnable et al., 2009; Wang et al., 2006). The majority of repetitive DNA in most plant species is formed by transposable elements themselves, as discussed by Lee and Kim (2014), with as much as 85% of the total genome in maize being composed of different TEs (Schnable et al., 2009). Repeats that are longer than the insert size of paired-end reads can lead to incorrect chromosomal rearrangement, or if the assembly program operates conservatively, to a fragmented assembly (Treangen and Salzberg, 2012). Plant genomes are therefore published as drafts, since a complete sequence, free of gaps, is unrealistic with the current technology. Plastid sequences on the other hand are easier to obtain, which is reflected by the number of published plant and plastid sequences on the NCBI platform. As of the writing of this thesis, 230 plant genome drafts have been published. In comparison, 743 entries for chloroplasts and an additional 180 sequences for plastids are available (numbers obtained from: http://www.ncbi.nlm.nih.gov/genome/browse/, January 21, 2016).

The use of Next-generation sequencing data for the detection, reconstruction and analysis of NUPTs, when no nuclear genome is available, is still a sparsely researched field of study. Samaniego Castruita et al. (2015) recently published a pipeline called "odintifier", which aims to provide more reliable reference-based organelle assemblies by simultaneously detecting and reconstructing "organelle-derived inserted sequences" (odins), using a haplotype-phasing algorithm. They note, however, that phasing algorithms are only optimisied for two alleles, whereas organellar inserts can be essentially multiallelic in species with a high transfer rate of organelle DNA, as is the case in many plants. Therefore a reliable reconstruction of NUPT sequences is not possible yet for such species with this approach.
In this project, a different approach was used, based on extracting NUPTs and MIPTs by exact mapping of paired-end reads and subsequent BLAST alignment of unmapped reads (Altschul et al., 1990; Li and Durbin, 2009). The foundation for this method was a workflow originally developed by Thomas Eder in a previous project, at the time PhD student at the Division of Computational Systems Biology of the University of Vienna in collaboration with Gerald Schneeweiss, associate professor at the Department of Botany and Biodiversity Research, University of Vienna.

One plant family particularly suited for the study of NUPTs are the Orobanchaceae (broomrapes). All of its genera, except one, are either holoparasitic or hemiparasitic, which means that they depend on their host plants for nutrients and water. Holoparasites, having lost the ability for photosynthesis,

are fully dependent on their hosts, while hemiparasites may be either facultative or obligate parasitic and have retained the ability for photosynthesis. The relaxed evolutionary pressure on the plastome in these parasitic plants has led to significant chloroplast genome reduction, the degree of which varies from species to species (McNeal et al., 2013; Wicke et al., 2013). This indicates a higher amount of transfers from the plastid and subsequent loss of those sequences, compared to non-parasitic species. For this project, data from four different species of the family Orobanchaceae were used. Two of the species are holoparasites with a strongly reduced plastid genome, namely *Phelipanche ramosa* and *Orobanche crenata*. The third species, *Schwalbea americana* is hemiparasitic, while the fourth species, *Lindenbergia philippensis* is non-parasitic, allowing comparisons between these different lifestyles.
As a first step, the initial workflow created by Thomas Eder had to be adapted to partly different bioinformatic software. Subsequently its robustness was to be tested by comparison of the coverage distribution of NUPTs/MIPTs across the respective plastid chromosome for each species. A highly similar coverage distribution would demonstrate the reproducibility of the workflow. Then, in the next step, the goal was to assemble the extracted NUPT/MIPT reads into the actual, larger fragments and to study those in detail, if the reconstruction of specific single fragments was possible.
Another point of interest was the detection of sequences transferred to the nucleus and subsequently lost from the plastids of the two holoparasitic species *Phelipanche ramosa* and *Orobanche crenata*. Such sequences were to be detected by means of alignment against the plastid chromosomes of the other two species. *Phelipanche ramosa* should then show a higher number of NUPTs/MIPTs that were lost from its more reduced plastid chromosome, compared to *Orobanche crenata*.
In addition, over the course of the project, scaffolds of the mitochondrial genomes for all four species were provided, which allowed, at least partially, the separation of NUPTs and MIPTs. If the separation was successful, the relative amount of MIPTs should increase with the size of the respective mitochondrial genome of the species. The mitochondrial scaffolds also allowed the comparison of the transfer dynamics to the nucleus of plastids and mitochondria for each species, with the expectation that the three parasitic species should show higher transfer dynamics from their plastids, due to an increased instability, relative to their mitochondria.
The final step was the extraction of autochthonous nuclear sequences, so sequences that did not arise from relatively recent insertions from either plastids or mitochondria, for use in further studies. Such nuclear reads would then show lower sequence duplication levels, compared to the total original data set, which still contains sequences from the organelles.

Chapter 2

Materials and Methods

2.1 Initial workflow

By mapping paired-end next-generation sequencing reads of a plant species against its respective plastid genome, one can extract autochthonous plastid sequences, that is, read pairs where both reads successfully mapped in the correct orientation to each other, due to the high stringency and global alignment of the reads that is involved with the mapping procedure (Li and Durbin, 2009). Read pairs where only one read successfully mapped against the plastid genome represent sequences transferred to the nucleus or the mitochondrion relatively recently, since one read was still able to align completely. Subsequently aligning those reads that could not be mapped to the plastid genome, locally by means of the BLAST algorithm (Altschul et al., 1990), allows detection of older NUPTs and MIPTs that have decayed and fragmented over time. This workflow allowed the detection of NUPTs and MIPTs and their separation into different categories, according to their evolutionary distance to the plastid chromosome, based on whether the read pair in question had mapping or BLAST hits. The resulting categories were `m_m`, `m_0`, `b_m`, `b_b` and `b_0`. The term `m_m` stands for read pairs where both successfully mapped against the plastid chromosome and such read pairs are therefore considered autochthonous plastid sequences. The two categories `m_0` and `b_m` represent relatively recent NUPT/MIPT read pairs, since one of the mates still successfully mapped, whereas the other either had a BLAST hit or no hit at all. The last two categories `b_b` and `b_0` are read pairs which are evolutionary most distant to the plastid chromosome, with either a BLAST hit for both mates or only for one. The separate steps of this initial workflow, on which the other workflows used during the project are based, are outlined in the following sections. The unpublished draft plastid chromosome sequences

of the four species can be found in the supplement folder `data`. The published plastid chromosome sequences have the following RefSeq IDs: *Lindenbergia philippensis*: `NC_022859.1`, *Phelipanche ramosa*: `NC_023465.1`, *Orobanche crenata*: `NC_024845.1`, *Schwalbea americana*: `NC_023115.1`.
Both versions of the plastid chromosomes were used as reference in the initial workflow. The custom scripts used for the different steps can be found in supplement folder `initial_analyses`.

2.1.1 Basic statistics

Raw sequence reads in paired-end fastq format had originally been provided to Thomas Eder by Gerald Schneeweiss for all four plant species described in the introduction. The material for the plants had been obtained from the botanical garden of the University of Bonn, except for *Schwalbea americana*, which had been collected in the field (courtesy of K. Kirkman). DNA extraction and purification had been performed following a protocol of McNeal et al. (2006). Preparation of tagged libraries for paired-end Illumina sequencing had been done by GATC Biotech (Konstanz, Germany), using TruSeq DNA Sample Preparation Kit2. Sequencing was then performed with standard protocols on the Illumina HiSeq 2000.

The first step of the workflow was an assessment of the overall quality of the raw reads, which was done by means of the FastQC tool (Andrews, 2015), a software package that consists of several modules, each producing basic statistics about the raw data sets and/or graphical output. The FastQC version used was 0.11.3. An overview of the raw reads is given in Section 3.1.1.

2.1.2 Preprocessing

The second step of the workflow was preprocessing the raw reads, that is, filtering of low quality reads and trimming of low quality bases at the ends of the reads. This was done by means of PRINSEQ-lite, a bioinformatic freeware tool (Schmieder and Edwards, 2011). The PRINSEQ-lite version used was 0.20.4. This was the first deviation from the original workflow, where the FASTX-toolkit (Gordon, 2015) had been used instead. Prinseq-lite was chosen for the reiteration of the workflow because of its more up-to-date state and the increased comfort delivered by a tool having all of its functions in one package. The PRINSEQ-lite commands used were: `-fastq`, `-fastq2`, `-trim_qual_left` 20, `-trim_qual_right` 20,

`-trim_ns_left` 1, `-trim_ns_right` 1, `-min_len` 90, `-min_qual_mean` 20, `-ns_max_n` 1, `-out_format` 3, `-out_good`, `-log`, `-out_bad null`

2.1.3 Mapping

Splitting of the reads into the different categories (`m_m`, `m_0`, `b_m`, `b_b` and `b_0`) is done in two steps, the first being exact mapping against the respective plastid chromosome sequence in FASTA format.
For the mapping process the program BWA, which stands for "Burrows-Wheeler-Aligner" (Li and Durbin, 2009), was used. The 0.6.2 version was chosen, despite there already being newer versions available, because the newest versions of BWA always caused untraceable errors, while the older version, which had also been used by Thomas Eder, ran stably and did not cause any problems.Custom scripts for this part of the workflow can be found in the supplement folder `initial_analyses`, subfolder `03_bwa`.
The BWA settings used, which can also be found in the custom scripts, were:

```
index -a is
bwa aln -k 1 -M 5 -t 8 -q 20
```

2.1.4 BLAST

From the resulting BAM file read pairs with two mapping hits, where forward and reverse read both map, read pairs with one hit and those with no hit at all where extracted using custom scripts and the Samtools software package (Li et al., 2009), version 0.1.19.
The read pairs with one hit and those with no hit were then fed into the next step of the workflow: alignment with BLAST (Altschul et al., 1990).

The reads extracted from the mapping process were split into smaller-sized chunks and then aligned against the respective plastid chromosome sequence. Subsequent extraction of reads with a BLAST hit and those with no hit, again done with custom scripts, resulted in the aforementioned read categories: `m_m`, `m_0`, `b_m`, `b_b` and `b_0`. The e-value chosen was 1E-4, with results subsequently being filtered for minimum coverage of the alignment of 50%, minimum identity of 85% and a minimum query length of 100bp. The version of the blastall software used was 2.2.26. Custom scripts for this part of the workflow can be found in the supplement folder `initial_analyses`, subfolder `04_blast`.

2.1.5 Coverage map

The final step of the original workflow was to extract the coverage at each position of the reference plastid sequence from both the mapping and the BLAST results by means of custom scripts and then to plot those coverage values against the reference plastid sequence with the program DNAPlotter (Carver et al., 2009), which has its own interactive user interface for drawing linear and circular chromosomes with added graphs. Custom scripts for this part of the workflow can be found in the supplement folder `initial_analyses`, subfolder `05_extract_results`.

2.2 Assembly

One particular point of interest in the study of organellar DNA transfers to the nucleus in plants is the complete reconstruction of such fragments in the nuclear genome. Since the assembly of plant nuclear genomes is very difficult, as outlined in the introduction, a method to reconstruct NUPTs without an available nuclear genome of the respective plant would be tremendously useful. This would allow functional analysis, study of the flanking regions in the nuclear genome and other possibilities of research.
Hypothetically, it should be possible to take the reads that were detected as NUPTs/MIPTs in the initial workflow, and assemble them into such larger fragments.
To this end, two different approaches were chosen. The first approach should allow the extraction of only those reads that aligned to a specific region of the respective plastid chromosome and subsequently assemble them, to recreate actual NUPT/MIPT fragments that have a particular locus of the plastid chromosome as evolutionary origin.
The custom scripts for the assembly can be found in supplement folder `initial_analyses`, subfolder `06_region_assembly`.

2.2.1 Targeted assembly approach

The first approach required the implementation of scripts that enabled extraction of only those reads that had a BLAST or mapping hit in a specified region, by simply typing in a base position of the chromosome and a distance upstream and downstream to define the region and then feeding the extracted reads into the assembly software SPADES (Bankevich et al., 2012), version 3.1.0. Afterwards, basic information about the contigs was extracted from the SPADES results.
The SPADES settings used were:

```
spades.py -k 21,33,55,77 --careful --12
```

To allow easier selection of regions of interest for the first approach, circular coverage plots were drawn, but this time separately for each NUPT/MIPT category (`m_0`, `b_m`, `b_0`, `b_b`), with chromosomal repeats and annotation added as additional circles. The repeats were identified by Vmatch (Kurtz, 2015), version 2.2.4, using the following settings:
`vmatch -l 15 -h 3 -d -p -evalue 0.1 -seedlength 6 -v`

These settings were taken from a publication by Wicke et al. (2013), where plastid chromosomes of the orobanchaceae family had been examined. The above average peaks of the coverage plot, combined with the position of repeats should allow the selection of a transfer "hotspot" on the plastid chromosome.
The annotation was performed by means of the CPGAVAS (Liu et al., 2012) automatic annotation pipeline, which is specialized on the annotation of plastid chromosomes. Work up until this point was done with unpublished chromosome sequences, provided to Thomas Eder by Gerald Schneeweiss for development of his workflow. This was done to ensure that the modified workflow produced similar results and just get the assembly approaches working, because it was unclear at that point, if these steps even had to be repeated with the newer plastid chromosomes published by Wicke et al. (2013). Only after the assembly approaches were finished, were the newer plastid sequences used (except for *Phelipanche ramosa*, see section 3.1.4 for the explanation).

2.2.2 Reverse assembly approach

The second assembly approach was basically the reversal of the first, assembling all reads with a BLAST or mapping hit, regardless of the region on the chromosome, and then aligning the contigs back to the plastid chromosome to identify loci which had a high similarity to a contig. Realignment was performed with the progressiveMauve algorithm (Darling, Mau, and Perna, 2010) with standard settings, version 2.3.1. The same custom scripts were used for the second approach as for the first approach, but instead of a specific region, the entire plastid chromosome was specified as the target region in the command line input.

2.3 Detection of MIPTs

2.3.1 Alignment of plastid and mitochondrion

Since mitochondrial scaffolds were obtained over the course of the project, detection of MIPTs and their (at least partial) separation from NUPTs became possible. The first step was the direct alignment of the published plastid GenBank files of each species with the corresponding mitochondrial contigs via the aforementioned progressiveMauve algorithm (Darling, Mau, and Perna, 2010). For the species *Phelipanche ramosa* the published GenBank file had also been used, because the comparison of the initial workflow results for the published plastid chromosome sequences had not yet been done at the time.

Because transfer from the plastid to the mitochondrion has been observed to occur frequently (Stern and Palmer, 1984; Tsunewaki, 2011; Wang et al., 2007), the expectation was that relatively recent transfers should be detectable by this approach. The visualization by Mauve would then conveniently allow documentation of plastid genes with high sequence similarity to the mitochondrial contigs. Initially only the mitochondrial scaffolds for three of the four species were available (for a basic overview see Table 2.1).

Table 2.1: *Basic overview of the three initial mitochondrial scaffolds.*

Species	Combined length	Number of contigs
Lindenbergia philippensis	344278	7
Phelipanche ramosa	1216025	7
Schwalbea americana	1906325	11

2.3.2 Separation of NUPTs and MIPTs and evaluation of transfer dynamics

Once both plastid chromosome sequences and mitochondrial scaffolds were available, separation of NUPTs and MIPTs was possible. To this end an additional workflow was created, in principle based on the initial workflow discussed earlier, but for the specific purpose of separating NUPTs and MIPTs from each other. For this analysis, newer mitochondrial scaffolds were available, this time for all four species (see Table 2.2). These four mitochondrial

draft sequences can be found in the supplemental folder `data`.

Table 2.2: *Basic overview of the four mitochondrial scaffolds.*

Species	Combined length	Number of contigs
Lindenbergia philippensis	344278	7
Phelipanche ramosa	963472	6
Orobanche crenata	1114444	7
Schwalbea americana	1840673	13

The first step of the workflow involved the identification of "lost NUPTs and MIPTs" for the two holoparasitic species *Phelipanche ramosa* and *Orobanche crenata*, by mapping those reads of both species, that did not have a mapping hit against the respective plastid chromosome, against the plastid chromosome of *Lindenbergia philippensis* with BWA (Li and Durbin, 2009). Those reads that also didn't have a mapping hit against this chromosome sequence were also aligned with BLAST (Altschul et al., 1990) against the plastid chromosomes of *Lindenbergia philippensis* and *Schwalbea americana*. Then the NUPT/MIPT reads detected this way were compared with the NUPT/MIPT reads detected by mapping and BLAST alignment against the own respective plastid chromosome of the species. Only those reads that were detected with the alignment against the plastid chromosomes of *Lindenbergia philippensis* and *Schwalbea americana* were added to the initial NUPT/MIPT read set. The initial NUPT/MIPT set for each species was taken from the initial workflow, but with all categories combined.

This was done to detect NUPTs/MIPTs for the two holoparasites that were not detectable by alignment against their own, strongly reduced plastid chromosomes, to which we refer as "lost NUPTs/MIPTs". For consistency, the BWA and BLAST settings were the same as used in the initial workflow. The NUPT/MIPT reads were then mapped with BWA and afterwards aligned with BLAST against the respective mitochondrial scaffold of each species, and successfully aligned reads were filtered out. Such filtered reads represent MIPTs or "secondary NUPTs", sequences that were transferred first from the plastid to the mitochondrion, and then from the mitochondrion to the nucleus. A comparison of the read numbers before and after alignment against the mitochondrial sequences would then allow an estimation of the proportion of MIPTs among the NUPT/MIPT data sets of each species and shed light on the different dynamics between the plastids and the mitochon-

dria of the different species. The scripts necessary for the workflow can be found in the supplemental folder `nupt_extraction`.

Now, similar to the just discussed workflow to extract NUPTs and filter out MIPTs and secondary NUPTS, one can also extract NUMTs, transfers from the mitochondrion to the nucleus. A comparison of NUPTs and NUMTs proportions would then give insight into the organelle transfer dynamics to the nucleus for each species. This is possible, because sequences, for which it isn't clear whether they transferred to the nucleus or are the result of transfer between plastid and mitochondrion, are filtered out. The scripts for this workflow are in the supplemental folder `numt_extraction`.

2.3.3 Extraction of nuclear reads

The workflows so far allowed the extraction of NUPTs and NUMTs. However, extraction of autochthonous nuclear sequences, that is, nuclear sequences not derived from organelle insertions, could be used for comparative repeat analysis between such autochthonous sequences and NUPTs and NUMTs.

The process to extract autochthonous nuclear reads involved the following steps: reads that did not have a mapping or a BLAST hit against the respective plastid chromosome were mapped with BWA and aligned with BLAST against the respective mitochondrial scaffold. Those reads that did not have either a mapping or a BLAST hit were extracted. Again, for the two holoparasites *Phelipanche ramosa* and *Orobanche crenata*, reads were aligned with BLAST against the plastid chromosomes of *Lindenbergia philippensis* and *Schwalbea americana*, to filter out "lost NUPTs/MIPTs".

The resulting read sets were then analysed with FastQC to extract the sequence duplication levels, with the expectation that the duplication levels should be lower in the extracted nuclear reads than in the plastid reads (read pairs where both mates successfully mapped against the plastid chromosome), but also lower than in the intial preprocessed read sets. The scripts of this workflow are in supplement folder `nuclear_read_extraction`.

Chapter 3

Results and Discussion

3.1 Initial workflow

Since the first step was to reproduce the workflow originally created by Thomas Eder and adapt it to the newer versions of the used bioinformatic software, the results had to be compared to ensure that the workflow was still working and producing consistent results.

The repetition of the workflow also allowed to become accustomed to the used data sets and to typical bioinformatic tools and workflow structures. In addition, using the adapted workflow on a never version of the plastid chromosomes should demonstrate the robustness of the workflow, that is, its ability to work with different data. Thomas Eder had used unpublished draft versions of the plastid chromosomes during the creation of his workflow, while for the adapted version of the workflow both the old plastid chromosomes had been used, as well as the newer versions, published by Wicke et al. (2013). Section 3.1.3 compares the results of the initial and the adapted workflow, using the unpublished draft plastid chromosomes as reference. Section 3.1.4 then discusses the differences observed when using the published plastid chromosome sequences.

3.1.1 Basic statistics

The initial basic statistics of the raw sequences were created with FastQC, showing overall relatively good quality of the reads with a peak in low quality at the end positions of the sequences, often seen in Illumina reads (see figure 3.1 and table 3.1).

Since not all of the paired-end reads actually had a mate, filtering of the

Table 3.1: *Basic overview of the raw reads.*

Data sets	Number of reads	Average length	GC
NG-5721_Or64_6998_read1	127736441	101	39
NG-5721_Or64_6998_read2	127736441	101	39
NG-5721_Lp79_6999_read1	81637856	101	36
NG-5721_Lp79_6999_read2	81637856	101	37
NG-6192_QC43_lib14213_1_ID1099	182804687	101	41
NG-6192_QC43_lib14213_2_ID1099	182804687	101	42
NG-6192_SA58_lib14214_1_ID1099	205834621	101	40
NG-6192_SA58_lib14214_2_ID1099	205834621	101	41

singletons was necessary. In addition, since the quality of the reads dropped towards right-end base positions, quality-based trimming of the reads from the right-end side was in order, with a subsequent filtering of reads with an average quality score of less than 20. Because the average length of the reads was determined with 101bp and only very few below 99bp, the cutoff length was set to 90bp. All of these steps and others could be performed simultaneously with PRINSEQ-lite. For the complete list of settings used see Section 2.1.2.

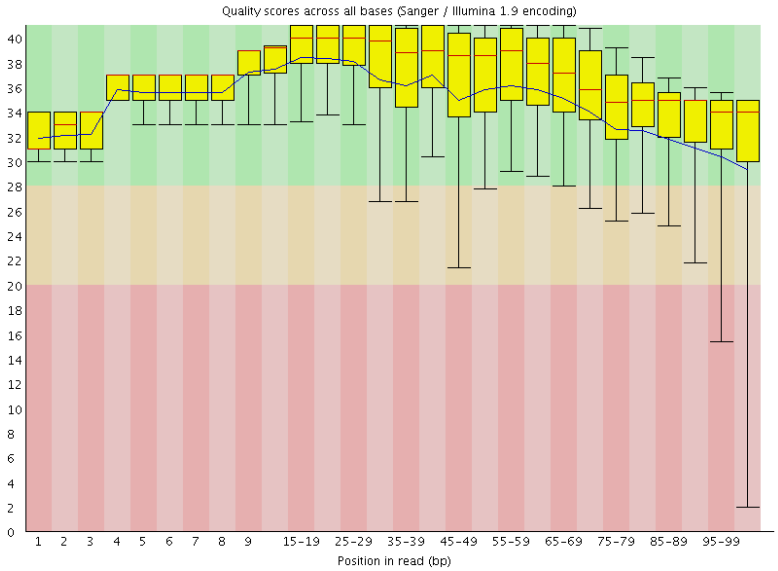


Figure 3.1: *Phred quality scores against base position of Lindenbergia philippensis raw data. Quality scores are represented in the form of box-whisker plots, with the red line being the median and the blue curve being the mean.*

3.1.2 Preprocessing

Subsequent preprocessing showed slight deviations to the original workflow, which was to be expected, as well as a higher stringency of the FASTX-Toolkit. This is because it wasn't possible to exactly recreate the FASTX settings and not because PRINSEQ-lite is inherently less stringent. The preprocessing results with PRINSEQ-lite and the FASTX-Toolkit can be seen in Tables 3.2 and 3.3).

Table 3.2: *Preprocessing results with PRINSEQ-lite for all four species.*

Data sets	Reads in	Reads out	Ratio out/in
Lindenbergia philippensis forward reads	81637856	66279818	81.19%
Lindenbergia philippensis reverse reads	81637856	66279818	
Phelipanche ramosa forward reads	127736441	101189093	79.22%
Phelipanche ramosa reverse reads	127736441	101189093	
Orobanche crenata forward reads	182804687	140162925	76.67%
Orobanche crenata reverse reads	182804687	140162925	
Schwalbea americana forward reads	205834621	167640562	81.44%
Schwalbea americana reverse reads	205834621	167640562	

Table 3.3: *Preprocessing results with the FASTX-Toolkit for all four species.*

Data sets	Reads in	Reads out	Ratio out/in
Lindenbergia philippensis forward reads	81637856	63908996	78.28%
Lindenbergia philippensis reverse reads	81637856	63908996	
Phelipanche ramosa forward reads	127736441	96721928	75.72%
Phelipanche ramosa reverse reads	127736441	96721928	
Orobanche crenata forward reads	182804687	124918780	68.33%
Orobanche crenata reverse reads	182804687	124918780	
Schwalbea americana forward reads	205834621	145889215	70.88%
Schwalbea americana reverse reads	205834621	145889215	

3.1.3 Comparison of workflow results

A comparison of the output of the original workflow and the newer version can be seen in figures 3.2 to 3.5. Each chart shows, for one of the four species, a side-by-side comparison of the read pair numbers of the different NUPT/MIPT categories between the original workflow and the newer version. The numbers are normalized against the total input read pairs in millions (see Tables 3.2 and 3.3).

A very similar overall amount of NUPTs/MIPTs is shown in Figures 3.2 to 3.5 for each species. The distribution of the different NUPT/MIPT categories is also similar for all species, except for *Phelipanche ramosa*. While the distribution pattern for *Phelipanche ramosa* is different, the overall amount is similar. This difference is most likely caused by a different composition of the preprocessed reads, thereby resulting in a different composition of the different NUPT/MIPT categories for this species. On the whole these results show the reproducability of the workflow, but also demonstrate the importance of consistency in the preprocessing step.

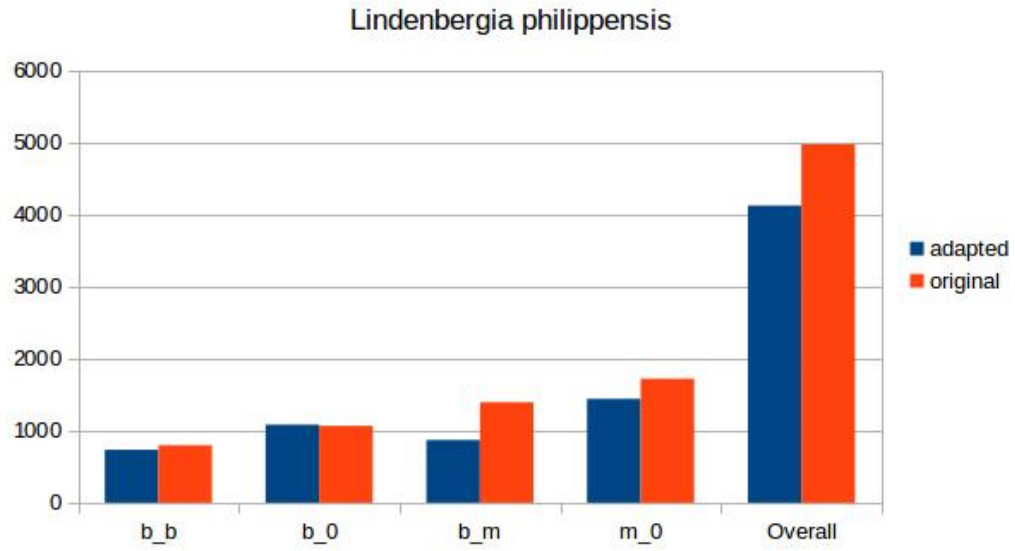


Figure 3.2: *Side-by-side comparison of the read pair numbers of the different NUPT/MIPT categories normalized against the total input read pairs in million. Here shown for the species Lindenbergia philippensis.*

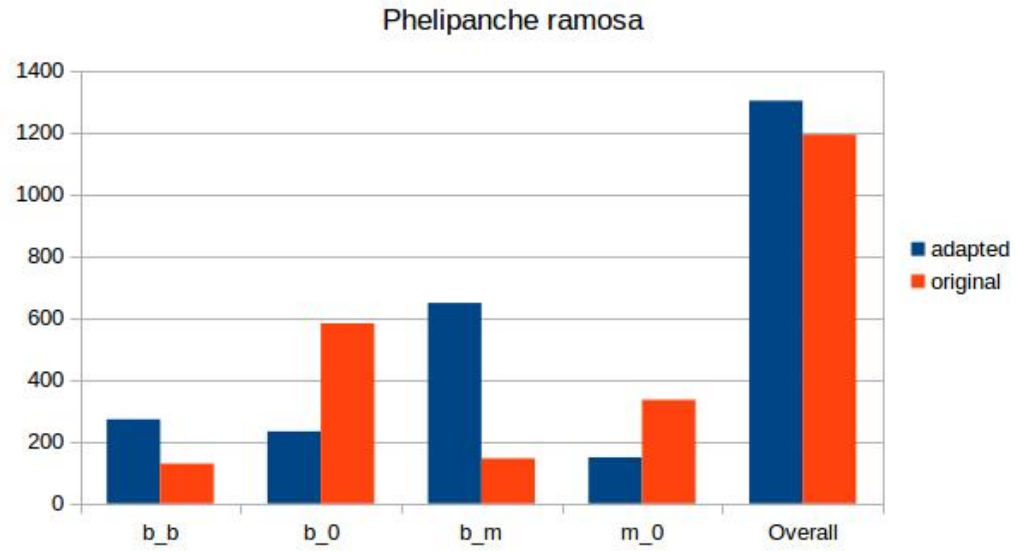


Figure 3.3: *Side-by-side comparison of the read pair numbers of the different NUPT/MIPT categories normalized against the total input read pairs in million. Here shown for the species Phelipanche ramosa.*

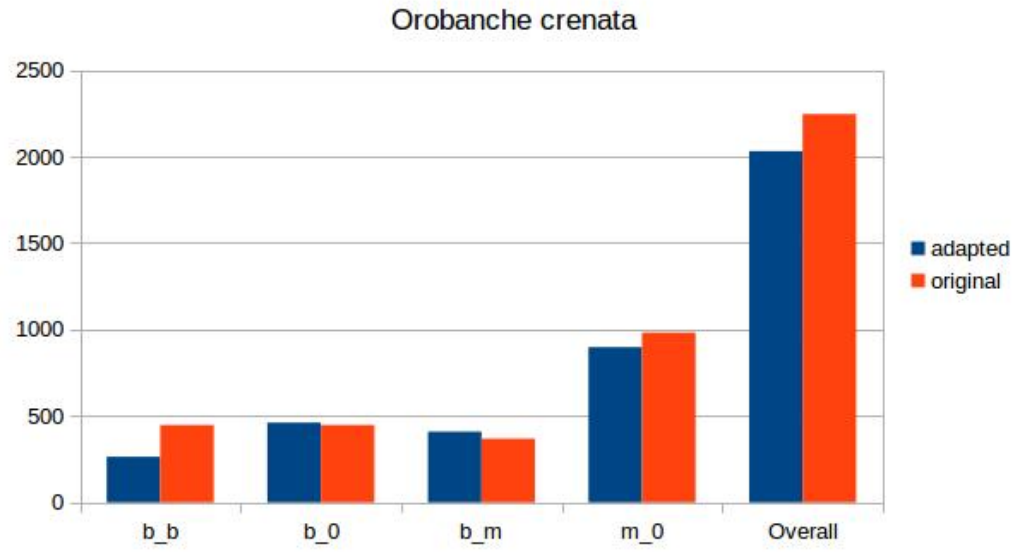


Figure 3.4: *Side-by-side comparison of the read pair numbers of the different NUPT/MIPT categories normalized against the total input read pairs in million. Here shown for the species Orobanche crenata.*

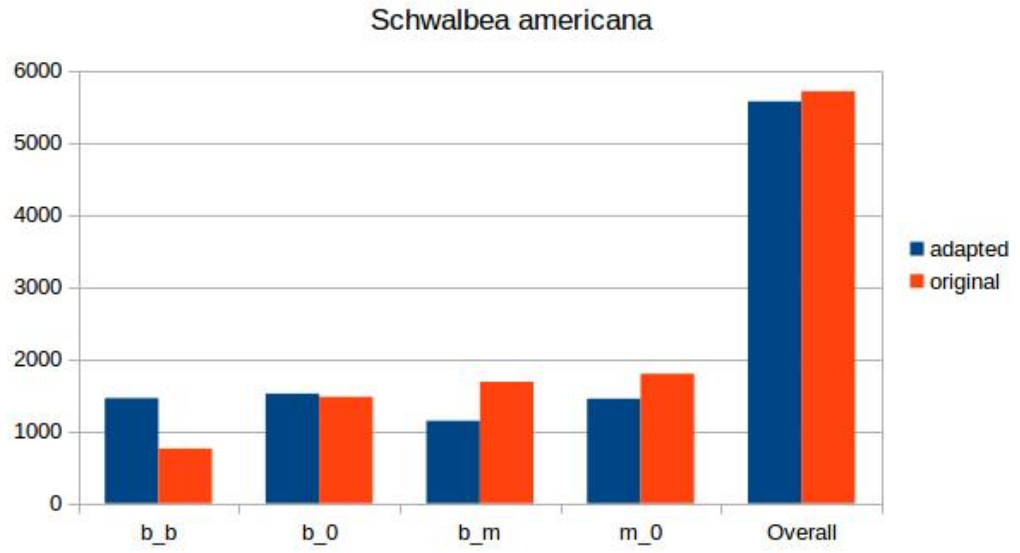


Figure 3.5: *Side-by-side comparison of the read pair numbers of the different NUPT/MIPT categories normalized against the total input read pairs in million. Here shown for the species Schwalbea americana.*

To further investigate the results of the two workflow versions, circular coverage maps were drawn. By comparing the patterns of coverage distribution between the redundant, original workflow and the new version, one can quickly assess the consistency of the output if used bioinformatic software tools are updated or exchanged for different tools. However, the actual purpose of these coverage maps, when they were created by Thomas Eder, was to evaluate whether specific coverage patterns were visible for each NUPT/MIPT category for each species. The hypothesis was that the areas on the plastid chromosome with increased transfer to the nucleus or the mitochondria would show coverage spikes, at least at the flanking regions of NUPT/MIPT fragments.

The coverage plot from the original workflow in Figure 3.6 shows the coverage on each position of the plastid chromosome of *Lindenbergia philippensis*. What can be seen is that the coverage of the `m_m` reads, that is, the reads that are considered autothochtonous for the plastid chromosome, is distributed much more evenly than the coverage of the different NUPT/MIPT categories, suggesting that certain areas along the chromosome are more prone to fragmentation than others, which was expected. Coverage spikes, visible for example in category `b_0`, might be flanking regions of larger transferred

fragments.

Also clearly visible are the two inverted repeats of the plastid chromosome, where most reads map to one repeat, leading to a coverage spike, with the other repeat having below average coverage as a result (best visible in the two innermost rings). The coverage plot for *Lindenbergia philippensis* in Figure 3.7 was created with the new workflow version, using the same unpublished draft sequence for the plastid chromosome as reference. The coverage distribution patterns of the two maps are almost identical, despite partly different tools having been used. The missing segments in the inverted repeats of categories `b_0 and b_b` (the two innermost rings) are a peculiarity of the DNAPlotter software, where coverage values of "0" simply are not shown in the graph, despite them being below the average. A comparison with the coverage map created by Thomas Eder with the original workflow (see Figure 3.6), shows an almost identical coverage distribution.

Figures 3.8 and 3.9 show the coverage maps for *Phelipanche ramosa* for the original and the new workflow, respectively. Despite showing a different distribution of read pair numbers for the different NUPT/MIPT categories (see Table 3.3), the coverage maps for this species show virtually identical patterns. This clearly demonstrates the robustness of the workflow. While the preprocessing step resulted in different read pair numbers, the coverage distribution for each category is still the same, allowing identification of regions with higher transfer activity.

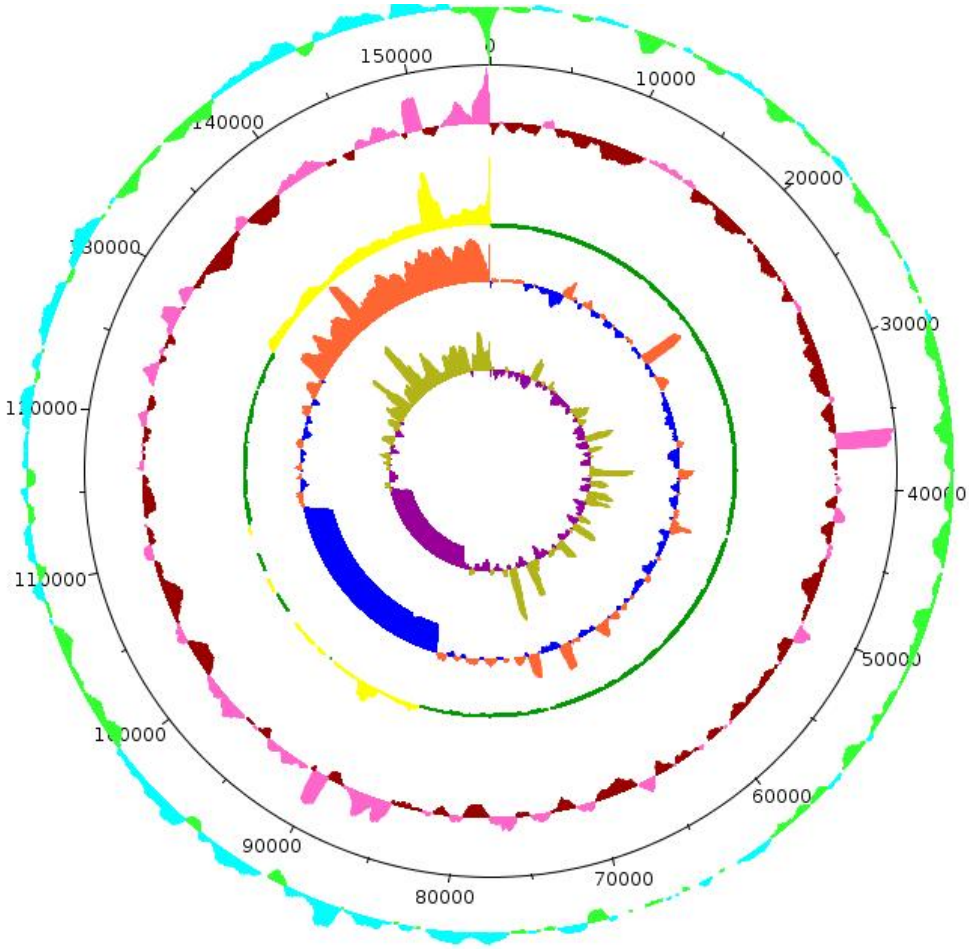


Figure 3.6: *Coverage plot from the original workflow for Lindenbergia philippensis. Each circular plot represents a NUPT/MIPT category and above average coverage points outwards and is coloured differently than below average coverage, which points inwards. The order of the categories from inner to outer is:* `b_0`, `b_b`, `b_m`, `m_0` *and* `m_m`.

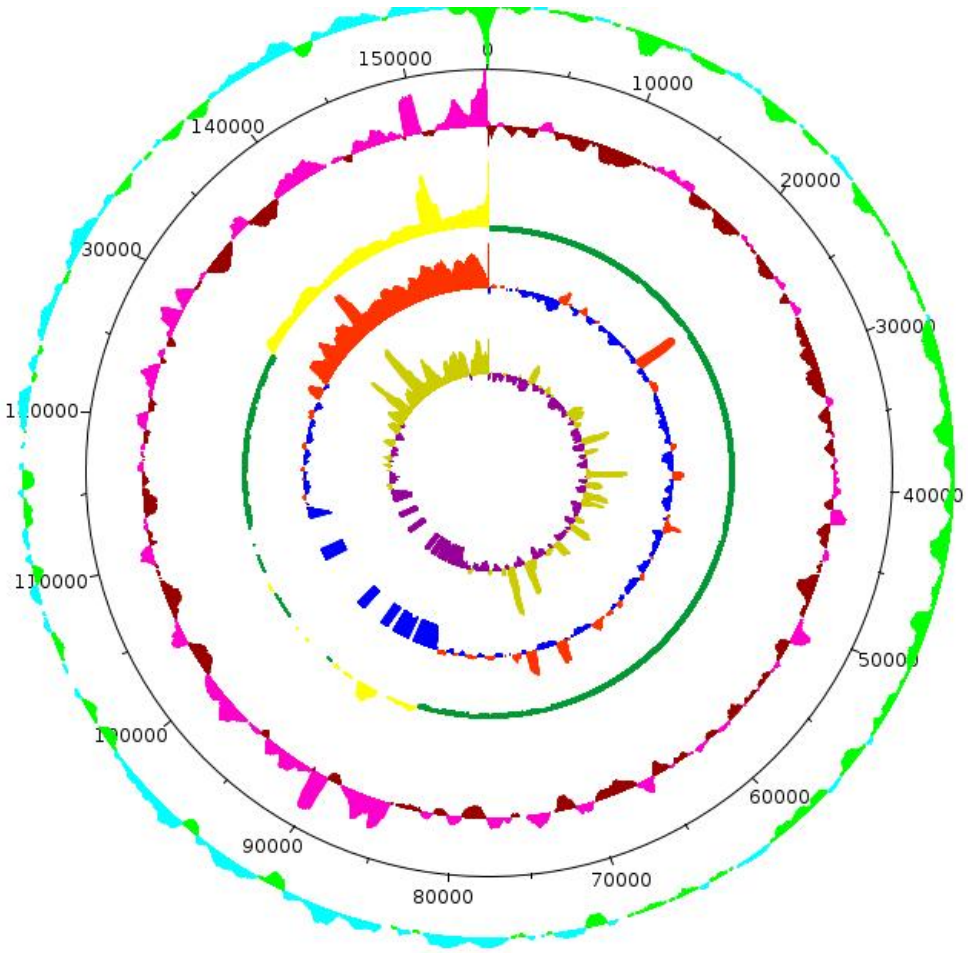


Figure 3.7: *Coverage plot created by the adapted workflow for Lindenbergia philippensis. Each circular plot represents a NUPT/MIPT category and above average coverage points outwards and is coloured differently than below average coverage, which points inwards. The order of the categories from inner to outer is:* `b_0`, `b_b`, `b_m`, `m_0` *and* `m_m`.

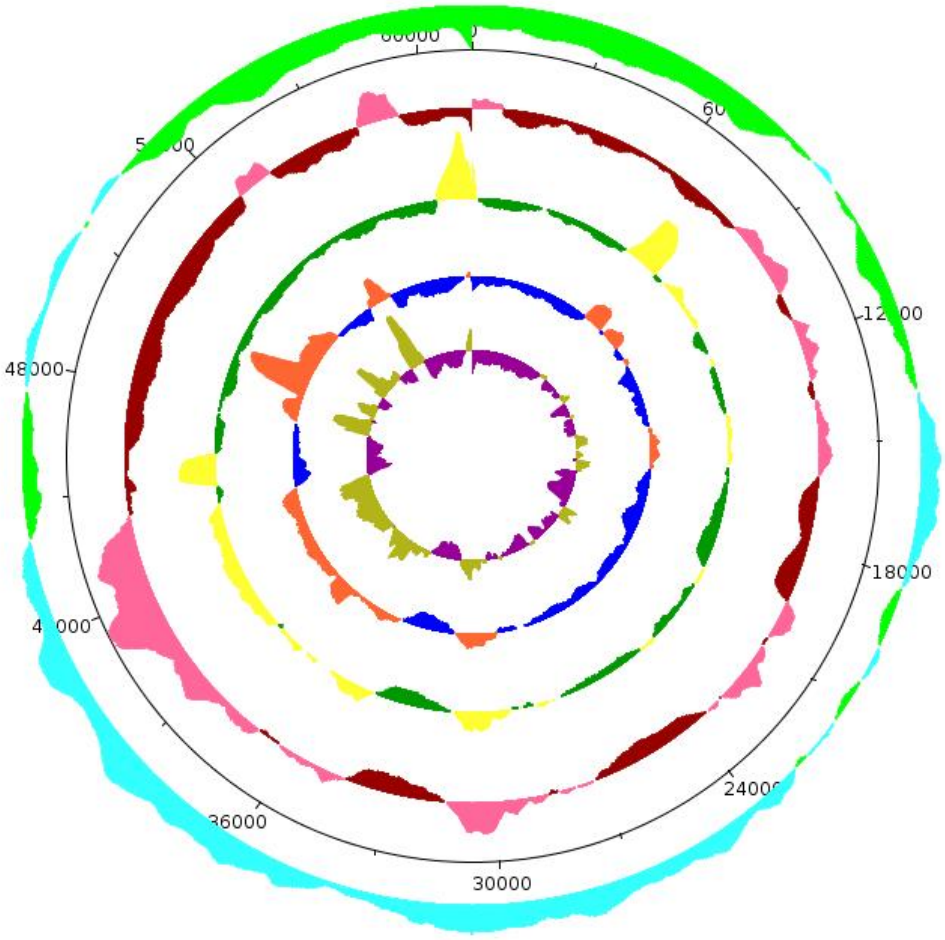


Figure 3.8: *Coverage plot from the original workflow for Phelipanche ramosa. Each circular plot represents a NUPT/MIPT category and above average coverage points outwards and is coloured differently than below average coverage, which points inwards. The order of the categories from inner to outer is:* `b_0`, `b_b`, `b_m`, `m_0` *and* `m_m`.

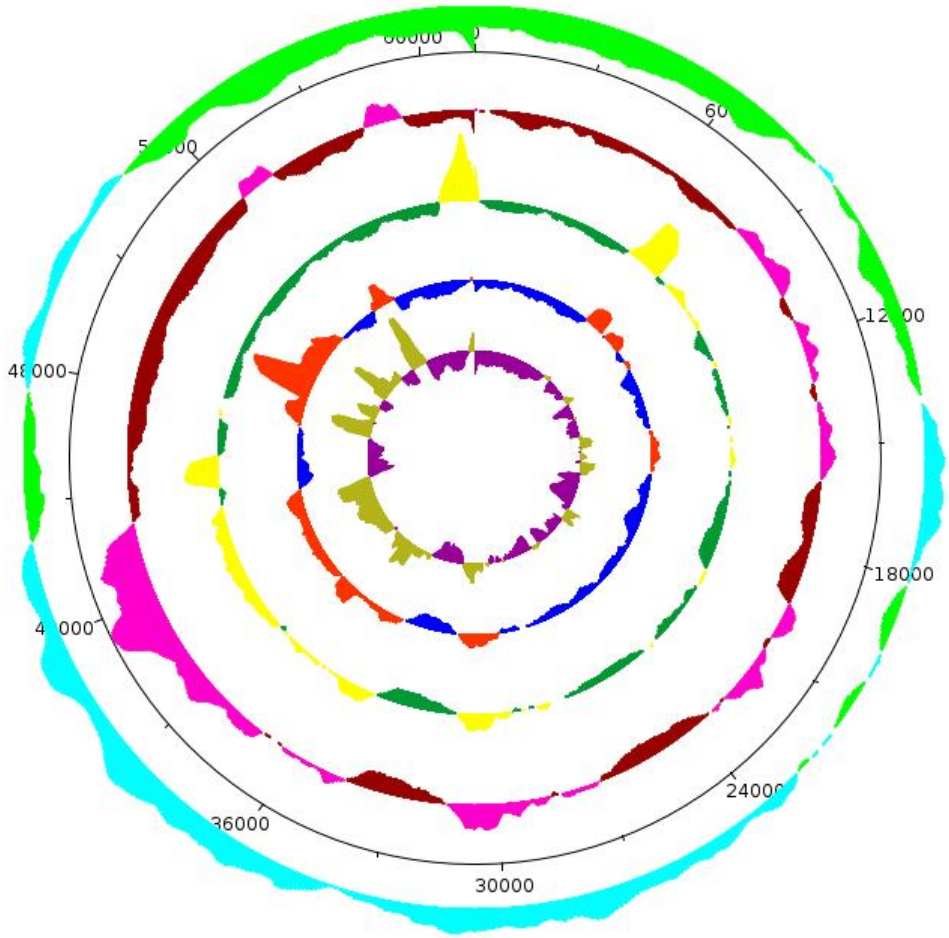


Figure 3.9: *Coverage plot created by the adapted workflow for Phelipanche ramosa. Each circular plot represents a NUPT/MIPT category and above average coverage points outwards and is coloured differently than below average coverage, which points inwards. The order of the categories from inner to outer is:* `b_0`, `b_b`, `b_m`, `m_0` *and* `m_m`.

3.1.4 Comparison of draft and published sequences

The new, adapted workflow was also used with the published plastid chromosome sequences. This was done for consistency and to compare the results between the draft sequences and the published sequences. A comparison of the resulting read pair counts for each NUPT/MIPT category showed only slight deviations for all species, except for *Phelipanche ramosa*.
Figure 3.10 shows a vast increase in overall NUPT/MIPT read pair numbers, when using the published plastid chromosome sequence as reference for the workflow. This increase is caused by the much larger number of reads of the categories `b_0` and `b_b`, that is, evolutionary more distant NUPTs/MIPTs. However, the vast majority of these additional NUPTs/MIPTs align to the last 600 nucleotides of the plastid chromosome. Since precisely the same workflow was used for the published plastid chromosome sequences as for the draft versions and since the input preprocessed fastq files were also the same, the different plastid chromosome sequence appears to be the cause. The coverage distribution of the different categories can be seen in Figure 3.11. The coverage spikes at the end of the chromosome are so large that for categories `b_0` and `b_b` the coverage visualization on the remaining regions of the chromosome is only visible when zoomed in.

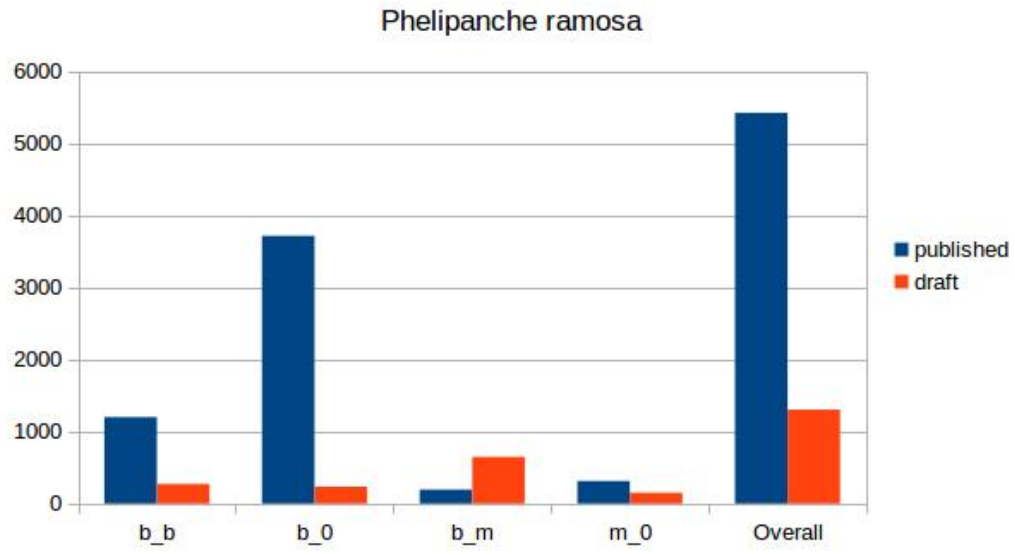


Figure 3.10: *Side-by-side comparison of the read pair numbers of the different NUPT/MIPT categories normalized against the total input read pairs in million. Here shown for the species Phelipanche ramosa, with the draft version and the published version of the plastid chromosome sequence as reference for the workflow, respectively*

To test whether the coverage spikes at the end of the chromosome were merely an artefact caused by the linearization of the actually circular plastid chromosome, the chromosome sequence was split in half and the two parts were switched in position, so that the region in question was now positioned in the middle of the sequence. The workflow was performed again, with the altered plastid chromosome sequence as reference. Figure 3.12 shows that the coverage spike changed position to the middle of the chromosome sequence. Whether this particular region on the plastid chromosome is actually an extreme hotspot of NUPT/MIPT transfers is doubtful.

A comparison of the total amount of NUPTs/MIPTs for each species with the respective amount of read pairs where both mates successfully mapped against the plastid chromosome (see Figure 3.13), shows that it is highly unlikely that such a high amount of NUPTs/MIPTs should be detected for *Phelipanche ramosa*. This species would then have the lowest amount of plastid sequences, while simultaneously having by far the highest amount of total NUPTs/MIPTs. Even when considering that the plastids of holoparasitic species might have an increased incidence of transfers, this seems implausible, when looking at the results of *Orobanche crenata*, the other holoparasitic species. A possible explanation for the coverage spike at the last 600 nucleotides of the published plastid chromosome of *Phelipanche ramosa* is that this region shows sequence similarity to repetitive elements in the nuclear genome.

For further analyses the draft version of the plastid chromosome *Phelipanche ramosa* was used, because of the questionable results created with the published plastid chromosome sequence (except for the direct alignment of plastid chromosome sequence and mitochondrial scaffolds, see Section 3.3.1).

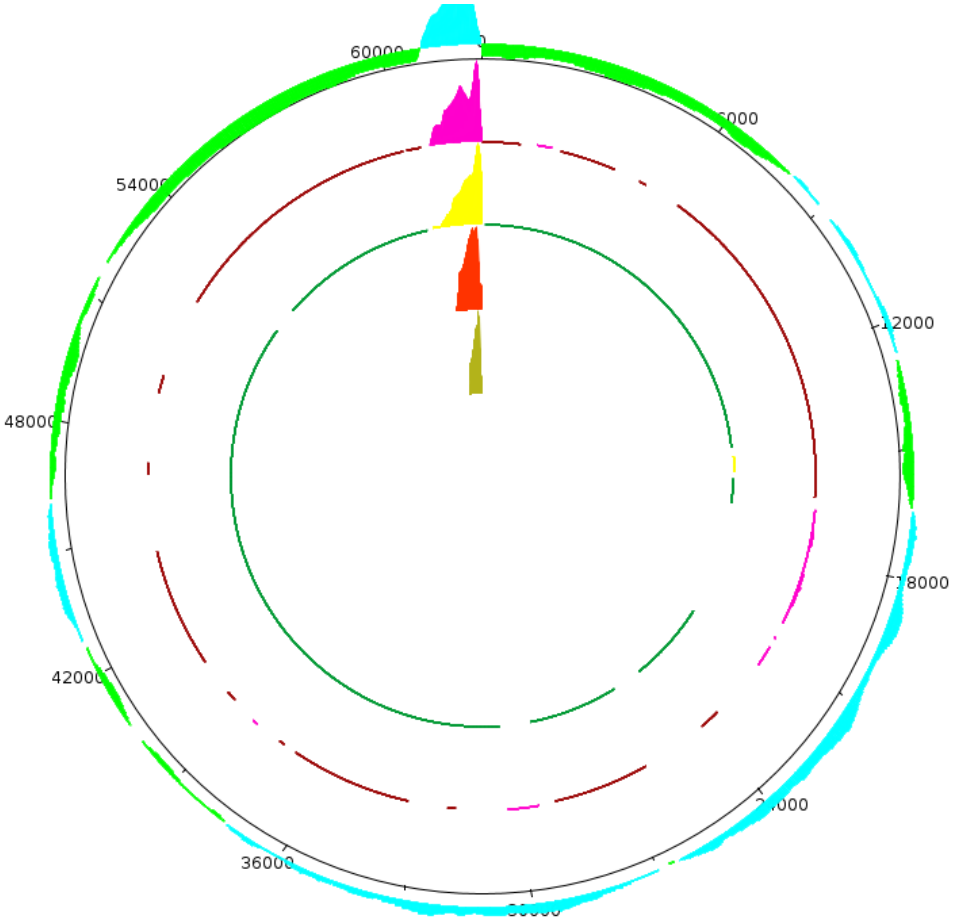


Figure 3.11: *Coverage plot for Phelipanche ramosa with the published plastid chromosome sequence as reference. Each circular plot represents a NUPT/MIPT category and above average coverage points outwards and is coloured differently than below average coverage, which points inwards. The order of the categories from inner to outer is:* `b_0`, `b_b`, `b_m`, `m_0` *and* `m_m`.

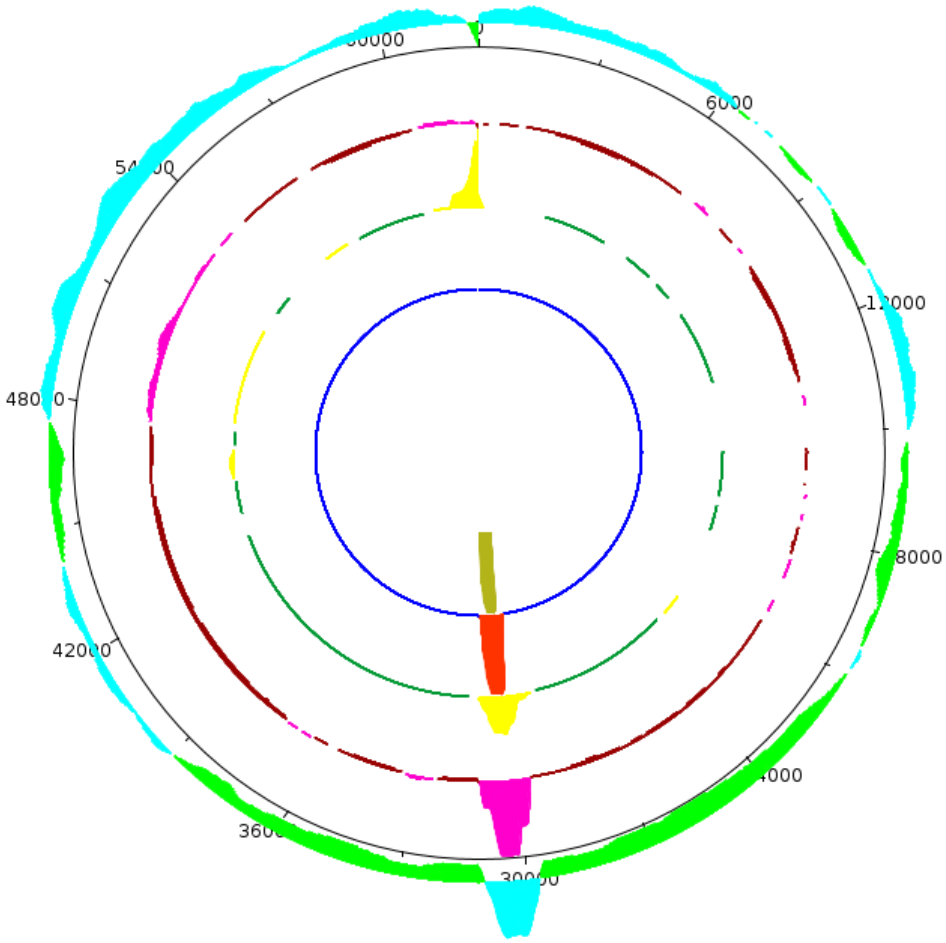


Figure 3.12: *Coverage plot for Phelipanche ramosa with the altered version of the published plastid chromosome sequence as reference. Each circular plot represents a NUPT/MIPT category and above average coverage points outwards and is coloured differently than below average coverage, which points inwards. The order of the categories from inner to outer is:* `b_0`, `b_b`, `b_m`, `m_0` *and* `m_m`.

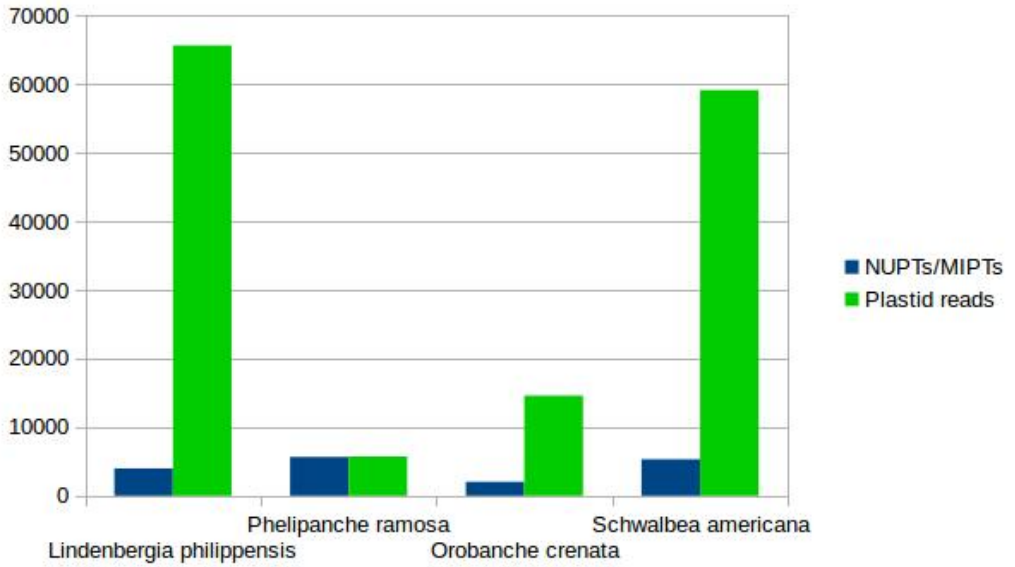


Figure 3.13: *A side-by-side comparison of the total amount of NUPTs/MIPTs with the number of read pairs of category* `m_m` *for all four species. The read pair numbers are normalized against total input read pairs in millions.*

3.2 Assembly

3.2.1 Targeted assembly approach

Since the first assembly approach should allow targeting of a specific region on the plastid, a more detailed coverage plot was drawn with DNAplotter, to facilitate the selection of the region on the plastid (see Figure 3.14). The method was tested only with data from the species *Phelipanche ramosa*, since it was unclear whether this approach would be further developed over the course of the project.

Figure 3.14: *Coverage plot for NUPT/MIPT category* `b_0` *of species Phelipanche ramosa. From inner to outer: annotation, repeats, coverage plot (with green being above average and purple being below average).*

The original motivation was, that the coverage plot in combination with the annotation and the repeats on the plastid might allow a more specific targeting of regions to assemble, using above average coverage spikes as indicator. For the test run, the region encompassing the gene ycf2 was chosen, the longest plastid gene, whose function is yet unknown.
The developed custom scripts allowed extraction of only those reads that had a BLAST or mapping hit in the specified region, which in this case was from base pairs 37000-44500. To increase the coverage, read pairs from categories `b_0` and `b_b` were merged into category `b`, while read pairs from categories `b_m` and `m_0` were merged into category `m`. Subsequent assembly of the extracted reads with Spades showed a pattern of resulting contig lengths for category `b` (see Figure 3.15 and Table 3.4), with a large number of small contigs and very few longer ones. The initial expectation was that the targeted assembly approach should allow the assembly of a specific, longer fragment, which could then be analysed. However, while this initial expectation was disappointed, the observed pattern makes sense biologically. It fits the model that once a fragment from a plastid chromosome has been inserted into a nuclear chromosome, it is subject to recombination within the nuclear chromosome, driven by repeats and transposable elements, which leads to further fragmentation of the insert into smaller pieces over the course of evolution (Huang et al., 2005; Michalovova, Vyskot, and Kejnovsky, 2013; Richly and Leister, 2004).

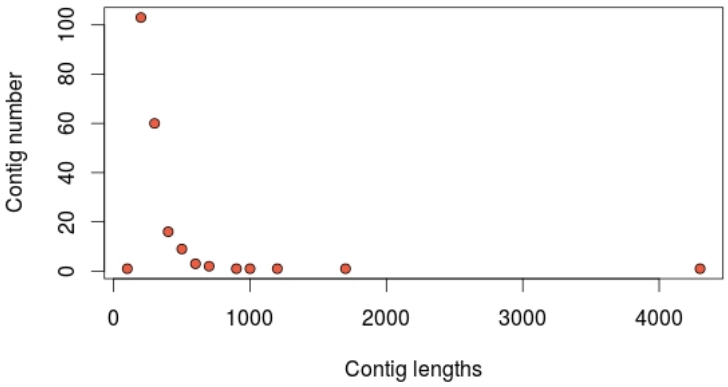


Figure 3.15: *Distribution of contig lengths created from read pairs of category* `b` *of P. ramosa, which had a BLAST hit between 37000bp and 44500bp on the plastid chromosome of P. ramosa.*

Table 3.4: *Basic statistics for the assembly of read pairs of category* `b` *of P. ramosa, which had a BLAST hit between* 37000*bp and* 44500*bp on the plastid chromosome of P. ramosa.*

Total length of sequences	71249
Total number of sequences	199
N50	327
Shortest contig	131
Longest contig	4336
GC	35.99%

Since reads that only had a BLAST hit represent older NUPTS/MIPTS that have undergone more recombinations over time, the longest contigs are then expected to be shorter than for read pairs that also had a mapping hit, and the number of short contigs is expected to be relatively higher.

In contrast, read pairs that also had a mapping hit are now expected to form longer contigs on the whole and fewer very short contigs, which is exactly what can be observed with the assembly of read pair category `m` for *Phelipanche ramosa* for plastid region 37000bp-44500bp (see Figure 3.16 and Table 3.5).

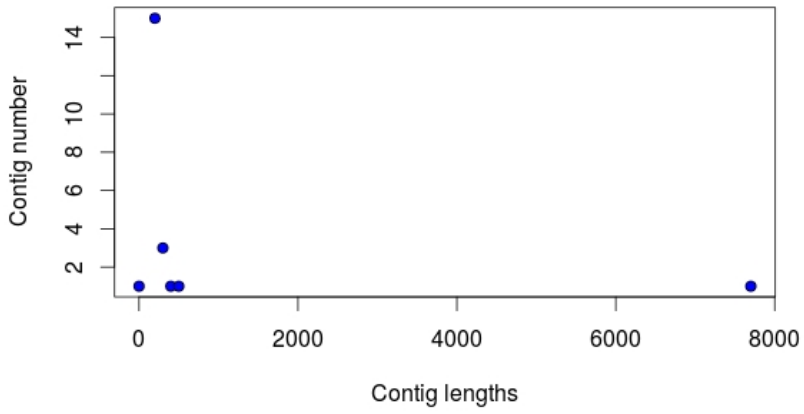


Figure 3.16: *Distribution of contig lengths created from read pairs of category* `m` *of P. ramosa, which had a mapping hit between 37000bp and 44500bp on the plastid chromosome of P. ramosa.*

Table 3.5: *Basic statistics for the assembly of read pairs of category* `m` *of P. ramosa, which had a mapping hit between 37000bp and 44500bp on the plastid chromosome of P. ramosa.*

Total length of sequences	13689
Total number of sequences	22
N50	7753
Shortest contig	80
Longest contig	7753
GC	36.65%

While the initial expectation of the targeted assembly of distinct fragments was disappointed, the increasing fragmentation and shortening of transferred fragments over the course of evolution could be shown.

Without a nuclear genome for the respective species as reference there is no way to verify if contigs actually represent nuclear fragments and are therefore eligible for further analysis to obtain reliable results. Especially shorter contigs might actually be misassemblies rather than true NUPTS/MIPTS. To further develop this approach a different experiment would have to be performed, with a different species for which a nuclear genome draft, the plastid chromosome, and the mitochondrial chromosome are available. Only then could the reliability of the fragment assembly be assessed and improved, for example by comparing the performance of different assembly programs and settings. After sufficient reliability has been demonstrated, the method could be applied to species without an available nuclear genome.

3.2.2 Reverse assembly approach

The second approach to the assembly was to not only assemble those reads that had a BLAST or mapping hit in a specific region of the plastid chromosome, but to assemble all reads of a category across the entire plastid chromosome and then to realign the resulting contigs to the plastid chromosome, to enable the identification of loci that show high similarity to a contig.

Table 3.6 shows the results of the assembly with reads from *Phelipanche ramosa* of category `b`. The high number of small contigs and relatively few longer contigs were expected and are in accordance with the results of the

first approach for category b (see Figure 3.15 and Table 3.4). Realignment of the contigs to the plastid chromosome was performed with the progressiveMauve alogrithm , which calculates so called „Locally Collinear Blocks“ (LCB), regions that are free of genome rearrangements (Darling, Mau, and Perna, 2010).

Figure 3.17 shows the Mauve UI, with the *Phelipanche ramosa* plastid chromosome on the top and the assembled contigs of category b for *Phelipanche ramosa* on the bottom. Contig borders are shown as vertical red lines and LCBs are shown as coloured rectangles, with the LCBs of the contigs and their counterparts of the chromosome being connected by lines of the same colour. The colouring inside the LCBs shows the similarity between the connected LCBs at the respective base pair positions. A completely coloured area denotes high similarity, whereas a white area denotes the opposite.

Mauve was chosen for the realignment step for its graphical demonstration capabilities. One of the problems of the Mauve algorithm, however, is that duplicate sequences are ignored during the calculation of the LCBs which leads to a loss of information. So should this approach be continued in future research, selection of a different alignment method is advised.

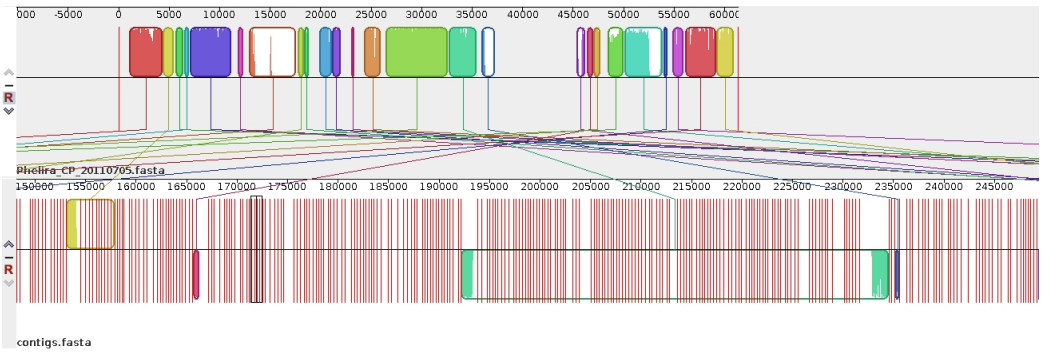


Figure 3.17: *Alignment of contigs of category* b *from P. ramosa against the plastid chromosome.*

Table 3.6: *Assembly results of category* `b` *for P. ramosa, using all reads that had a BLAST hit anywhere on the plastid chromosome.*

Total length of sequences	405022
Total number of sequences	1027
N50	359
Shortest contig	121
Longest contig	6885
GC	36.17%

In contrast to the assembly of category `b`, the assembly of category `m` produced fewer and longer contigs and also fewer and longer LCBs (see Figure 3.18 and Table 3.7), which is again in accordance with the showcased results of the first approach (see Figures 3.15 and 3.16 and Tables 3.4 and 3.5), with the evolutionary more distant category `b` reads showing a higher fragmentation than the evolutionary more recent category `m` reads.

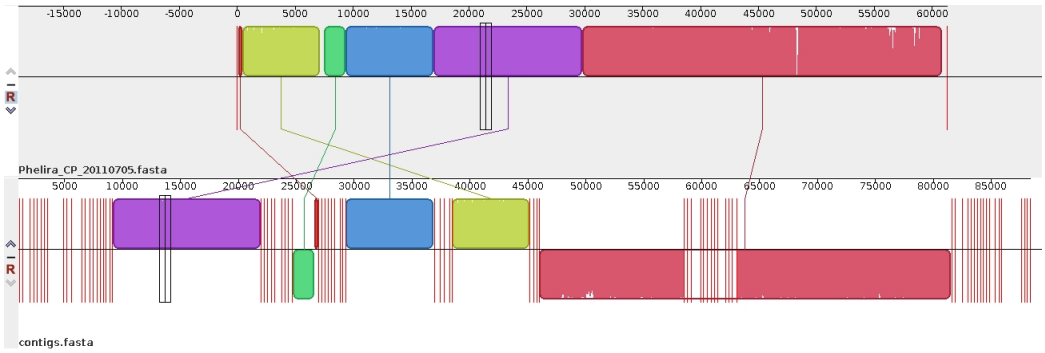


Figure 3.18: *Alignment of contigs of category* `m` *from P. ramosa against the plastid chromosome.*

Table 3.7: *Assembly results of category* `m` *for P. ramosa, using all reads that had a mapping hit anywhere on the plastid chromosome.*

Total length of sequences	88385
Total number of sequences	81
N50	7650
Shortest contig	228
Longest contig	18572
GC	34.04%

Unfortunately, since nuclear genomes for the Orobanchaceae species in question are not available, the assembly of contigs that represent larger NUPTs/MIPTs is currently of limited as discussed above, since there is no way to assess the actual accuracy of the assembly, making a reliable quantification and analysis of NUPTs/MIPTs fragments impossible. However, the results of both assembly methods strengthen the model that NUPTs/MIPTs fragment over time.

3.3 Detection of MIPTs

3.3.1 Alignment of plastid and mitochondrion

The results of the alignments of the plastid sequences and the respective mitochondrial scaffolds were summarized in a table for each species (see Tables 3.8 to 3.10), with each region of sequence similarity denoted by its start and end base pair position on the plastid. For a comprehensive list of the functions of the detected genes see the review by Wicke et al. (2011).

Table 3.8: *Results of the direct alignment of the Lindenbergia philippensis plastid and the mitochondrial contigs.*

Start position (bp)	End position (bp)	Complete ORF	Partial ORF
9114	11994		atpA
25431	25570		rpoB
31550	33095	tRNA-D	
39732	39908		psaB
42355	42533		psaA
44838	46040		Ycf3
46199	47108	tRNA-S	rps4
53625	53721	tRNA-M	
67923	69408	tRNA-W	rpl33
		tRNA-P	
72480	72570		ClpP
86058	86596		rpl2
101679	103130		16SrRNA
104468	104720		tRNA-I
			tRNA-A
105381	105626		tRNA-A
			23SrRNA
105772	108373		23 SrRNA
109883	110034	tRNA-N	

Table 3.9: *Results of the direct alignment of the Phelipanche ramosa plastid and the mitochondrial contigs.*

Start position (bp)	End position (bp)	Complete ORF	Partial ORF
7288	8476	TRNA-N	
		4.5SrRNA	
8660	10611		23SrRNA
10628	11174		16SrRNA
11890	13377		16SrRNA
14646	14799		rps12
17642	17949		Ycf2
29981	24058	tRNA-H	

Continued on next page

Table 3.9 – continued from previous page

Start position (bp)	End position (bp)	Complete ORF	Partial ORF
39824	40117	tRNA-W	tRNA-P

Table 3.10: *Results of the direct alignment of the Schwalbea americana plastid and the mitochondrial contigs.*

Start position (bp)	End position (bp)	Complete ORF	Partial ORF
1	80	tRNA-H	
1114	1351		psbA
6558	7558	tRNA-Q	
10685	11270		atpA
11582	12602		atpA
			atpF
15037	16344		atpI
			rps2
19782	20026	rpoC2	
20158	20232	rpoC2	
22130	22520	rpoC1	
28466	28557		Intergenic
30623	30831	tRNA-D	Intergenic
31506	31759		tRNA-T
			Intergenic
33403	33557		psbD
33926	33545	psbC	psbD
35634	36730	tRNA-S	
		tRNA-G	
37076	38369		rps14
			psaB
41645	42904	ycf3 (exon 3)	psaB
44864	45063		Intergenic
45364	46313	tRNA-S	rps4
47282	47786	tRNA-L	
55797	55949		atpB
56521	56853		atpB
57784	57868		tRNA-M
64170	64233		petA
64654	64746		Intergenic
66110	66196		psbE
67542	67977	tRNA-W	
		tRNA-P	
70620	71489	rps12	clpP
72006	72679		clpP
73348	74442		psbB
75073	75823	psbN	
		psbH	
		petB	
78469	78538		Intergenic

Continued on next page

Table 3.10 – continued from previous page

Start position (bp)	End position (bp)	Complete ORF	Partial ORF
83982	85733		rpl22
			rps19
			rpl2
85738	85850		rpl2
86428	86916	tRNA-I	rpl23
87404	87537		ycf2
87942	88134		ycf2
89586	89627		ycf2
89692	90107		ycf2
91102	91461		ycf2
92087	92241		ycf2
92272	92356		ycf2
92491	92560		ycf2
93653	93737		tRNA-L
95552	99222		ndhB
			Intergenic
99617	100055		16SrRNA
100068	100390		16SrRNA
100404	100847		16SrRNA
100848	101702		16SrRNA
101946	102646		tRNA-I
102819	102878		tRNA-A
102954	103244		tRNA-A
104049	104438		23SrRNA
104488	104795		23SrRNA
106054	106918	4.5SrRNA	23SrRNA
106979	107176		5SrRNA
109749	110288		ndhH

The initial expectation here was that in all three species, genes or their fragments with plastid origin should be detectable, albeit highly mutated. Also the two species with the larger mitochondrial scaffolds, *Phelipanche ramosa* (over 1,21 Mbp) and *Schwalbea americana* (over 1.91 Mbp), should show a larger amount of regions with plastid origin than *Lindenbergia* with its smaller scaffold (over 0.344 Mbp).

While all species show locally collinear blocks, the results of *Phelipanche ramosa* and *Schwalbea americana* differ dramatically. As expected, *Schwalbea americana* shows a high amount of similar regions, with some genes of the plastid being fragmented over several mitochondrial contigs, for example ycf2 (see figure 3.19), the largest gene in the plastid chromosome with, as of yet, undetermined function.

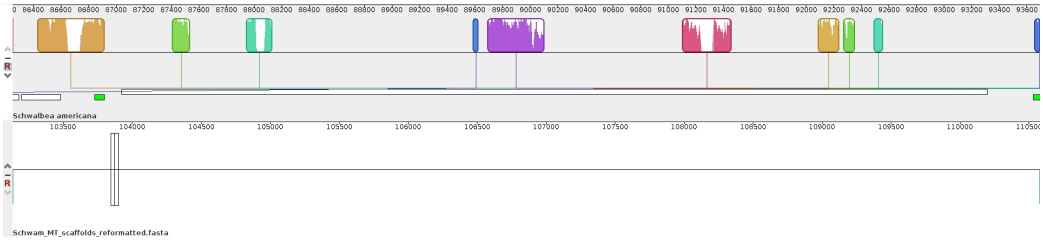


Figure 3.19: *Depiction of the fragments of the ycf2 gene, which are found in different mitochondrial contigs. The coloured boxes are the locally colinear blocks and the long white bar directly below the blocks is the ycf2 gene.*

In another instance, several genes of the *Schwalbea americana* plastid have been transferred to the mitochondrion as one fragment, namely psbN, psbH and part of petB (see figure 3.20). The genes psbN and psbH are part of photosystem II and petB is part of the cytochrome $b6/f$ complex. Compared to other locally collinear blocks, this fragment also shows relatively high sequence similarity, indicating that it is younger than other inserts. This would be consistent with the model that inserts are initially large and then decay and fracture over evolutionary time (Huang et al., 2005; Michalovova, Vyskot, and Kejnovsky, 2013; Richly and Leister, 2004).

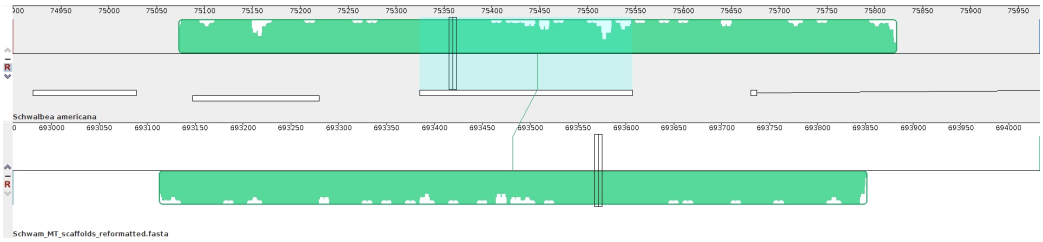


Figure 3.20: *The locally colinear block of Schwalbea americana encompassing the genes psbN, psbH and part of petB.*

Phelipanche ramosa on the other hand shows the fewest locally collinear blocks, despite also having a rather large mitochondrial genome, as indicated by the combined length of the contigs. The reasons for this discrepancy in comparison to the results of *Schwalbea americana* is unclear. In this case, the published plastid chromosome sequence of *Phelipanche ramosa* had been used for the alignment. Because the Mauve algorithm ignores duplicates, it is possible that the shown number of LCBs for this species is an underrepresentation. It is also possible that the mitochondria of *Phelipanche ramosa* show relatively low transfer dynamics, despite the estimated large size of the

mitochondrial genome. To obtain a better quantification of the amount of MIPTs, a different approach was used, as explained in the following section.

3.3.2 Separation of NUPTs and MIPTs and evaluation of transfer dynamics

During the workflow for the separation of NUPTs and MIPTs, as described in the respective section of Materials and Methods, "lost NUPTs/MIPTs" are also detected for the species *Phelipanche ramosa* and *Orobonchа crenata*. These are NUPTs/MIPTs that are relatively recent, but still cannot be detected with the initial workflow, due to the reduction of the plastid chromosomes in the two holoparasites. Aligning those reads of the two species, that did not even have a BLAST hit on the respective plastid chromosome, against the plastid chromosomes of *Lindenbergia philippensis* and *Schwalbea americana* should detect NUPTs/MIPTs that had been transferred to the nucleus or the mitochondrion and were then lost from the plastid chromosome.

As can be seen in Table 3.11 and Figure 3.21, a certain additional amount of NUPTs/MIPTs was detected by alignment against the plastid chromosomes of *Lindenbergia philippensis* and *Schwalbea americana*. When comparing the ratio of the plastid chromosome sizes with the ratio of lost NUPTs/MIPTs between *Phelipanche ramosa* and *Orobanche crenata* (0.70 and 0.60, respectively), it becomes apparent that the size difference of the chromosomes roughly corresponds to the difference in the amount of lost NUPTs/MIPTs detected. This is what one would expect biologically.

Table 3.11: *This table compares the number of NUPT/MIPT read pairs that were detectable with the initial workflow (initial NUPTs/MIPTs) with the number of NUPTs/MIPTs read pairs that were additionally detected by aligning against the plastid chromosomes of Lindenbergia and Schwalbea (lost NUPTs/MIPTs). Read pair numbers are normalized against the respective number of total input read pairs in millions.*

Species	Initial NUPTs/MIPTs	Lost NUPTs/MIPTs
Phelipanche ramosa	1169	302
Orobanche crenata	1826	188

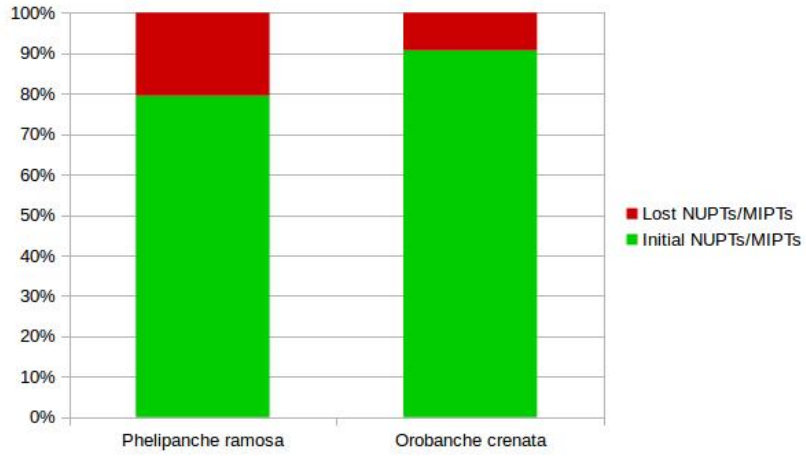


Figure 3.21: *The relative amount of lost NUPTs/MIPTs among the total amount of NUPTs/MIPTs for the species Phelipanche ramosa and Schwalbea americana.*

When looking at the results for the same procedure with the published plastid chromosome sequence as reference for *Phelipanche ramosa* (see figure 3.22), a different pattern is visible. In this case, *Phelipanche ramosa*, with its more reduced genome compared to *Orobanche crenata*, shows a lower relative amount of NUPTs/MIPTs that were lost from the plastid chromosome. This seems rather unlikely and shows again that the results of the initial workflow for the published plastid chromosome sequence appear not to reflect the biological truth, while the results created with the unpublished plastid chromosome are plausible.

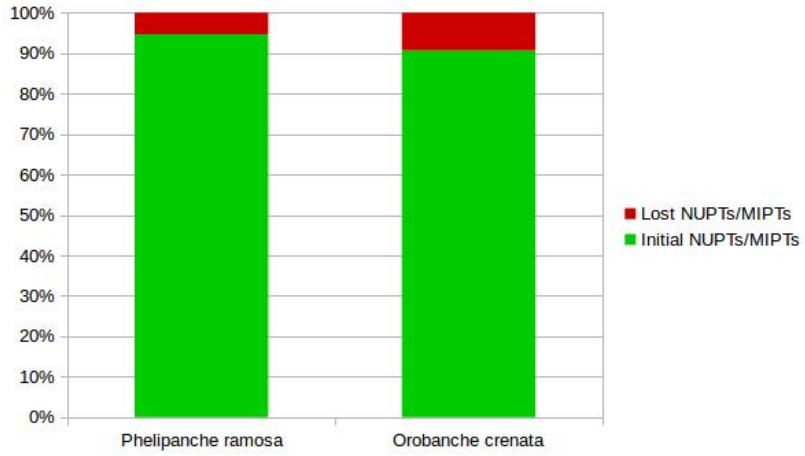


Figure 3.22: *The relative amount of lost NUPTs/MIPTs among the total amount of NUPTs/MIPTs for the species Phelipanche ramosa and Schwalbea americana. Here, the published plastid chromosome sequence had been used as a reference for Phelipanche ramosa.*

The subsequent filtering of putative MIPTs and secondary NUPTs, as described in Materials and Methods, allows the comparison of NUPT and MIPT proportions for each species (see table 3.12 and figure 3.23). The numbers shown are normalized against the number of input read pairs in millions.

Table 3.12: *The number of read pairs for each species that are strictly NUPTs and the number of read pairs that also show sequence similarity to the mitochondrion, and are therefore putative MIPTs (or secondary NUPTs). Read pair numbers are normalized against the respective number of total input read pairs in millions.*

Species	NUPTs	MIPTs
Lindenbergia philippensis	3526	434
Phelipanche ramosa	1055	416
Orobanche crenata	1577	437
Schwalbea americana	2911	2382

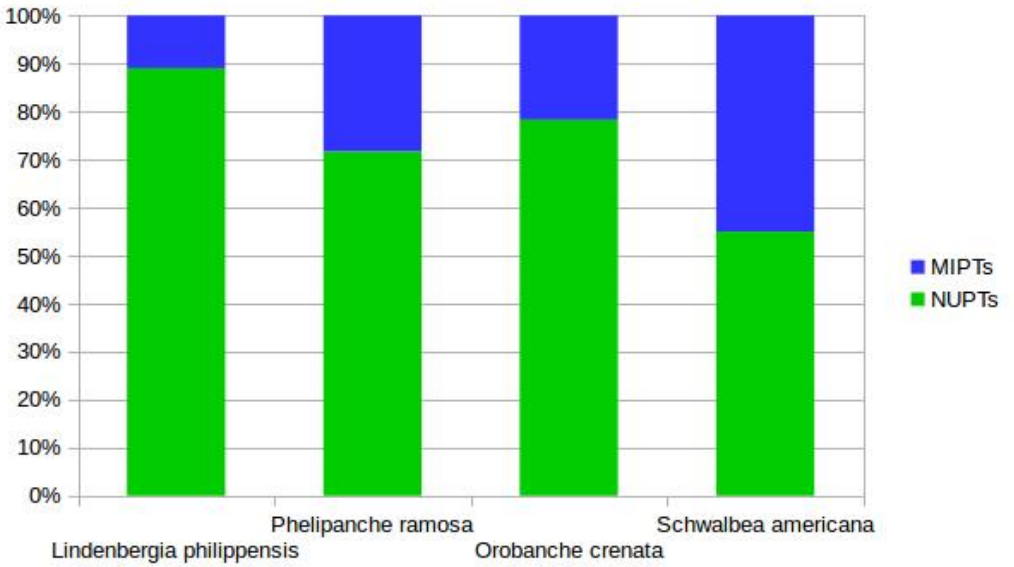


Figure 3.23: *The relative amount of NUPTs and MIPTs (with secondary NUPTs) among the total amount of transfers from the plastids for all four species.*

Schwalbea americana shows, as expected, the highest proportion of putative MIPTs. The differences of the relative amount of MIPTs among the species roughly correspond to the differences in size of the total mitochondrial genome (the amount of autochthonous mitochondrial reads, see Figure 3.24. This is again what one would expect, since larger mitochondrial genomes would also offer more opportunities of insertion of plastid sequences and the size increase of the mitochondrial genome is partly caused by influx of sequences from the plastid and the nucleus (Rodríguez-Moreno et al., 2011; Stern and Palmer, 1984; Tsunewaki, 2011; Wang et al., 2007).

Filtering of strict NUMTs (that is, without MIPTs), mitochondrial transfers to the nucleus, can be performed with the workflow described in the Section 2.3.2. The numbers are normalized against total input read pairs in millions. When the numbers of strict NUMTs are compared to the respective number of strict NUPT read pairs for each species, a first look at the organelle transfer dynamics to the nucleus for each species becomes possible (see Table 3.13 and Figure 3.25).

Looking only at these data, one could come to the conclusion that the transfer dynamics of the mitochondria of *Phelipanche ramosa*, *Orobanche crenata* and *Schwalbea americana* are much higher than the transfer dynamics of their

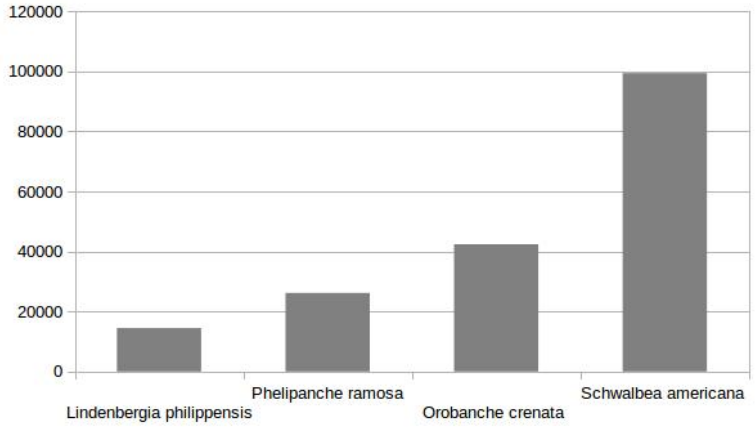


Figure 3.24: *The number of autochthonous mitochondrial read pairs for each species (read pair where both mates successfully mapped against the mitochondrial scaffold). The numbers are read pair counts normalized against the respective total input read pairs in millions.*

Table 3.13: *Strict NUPT and NUMT numbers (so without MIPTs) for each Species. The numbers are read pair counts normalized against the respective total input read pairs in millions.*

Species	NUPTs	NUMTs
Lindenbergia philippensis	3526	1215
Phelipanche ramosa	1055	2520
Orobanche crenata	1577	3721
Schwalbea americana	2911	10812

plastids. However, this does not take into account the number and size of the plastids and mitochondria in the cell. A greater number and size of mitochondria could by itself lead to an increased transfer rate. One way of taking this into account is to look at the ratio of autochthonous read pairs of plastids and mitochondria, so read pairs where both mates were successfully mapped to either the plastid chromosomes or the mitochondrial scaffolds respectively (see Table 3.14 and Figure 3.26).

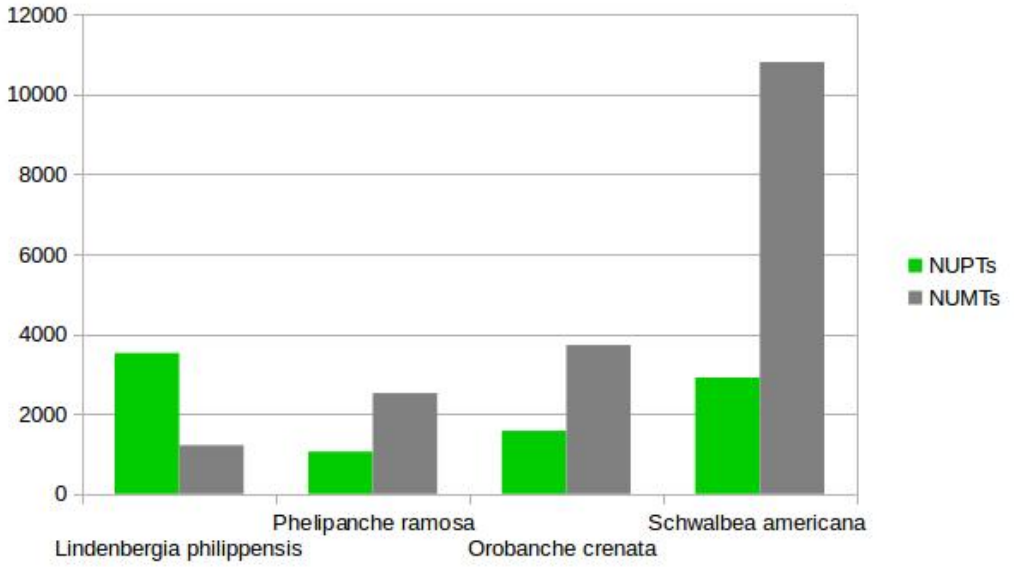


Figure 3.25: *Strict NUPT and NUMT numbers (so without MIPTs) for each Species. The numbers are read pair counts normalized against the respective total input read pairs in millions.*

Table 3.14: *Autochthonous read pair numbers for each species, so read pairs where both mates successfully mapped against the plastid chromosome or the mitochondrial scaffold respectively. The numbers are read pair counts normalized against the respective total input read pairs in millions.*

Species	Plastid reads	Mitochondrion reads
Lindenbergia philippensis	65637	14467
Phelipanche ramosa	5399	26187
Orobanche crenata	14592	42384
Schwalbea americana	59142	99433

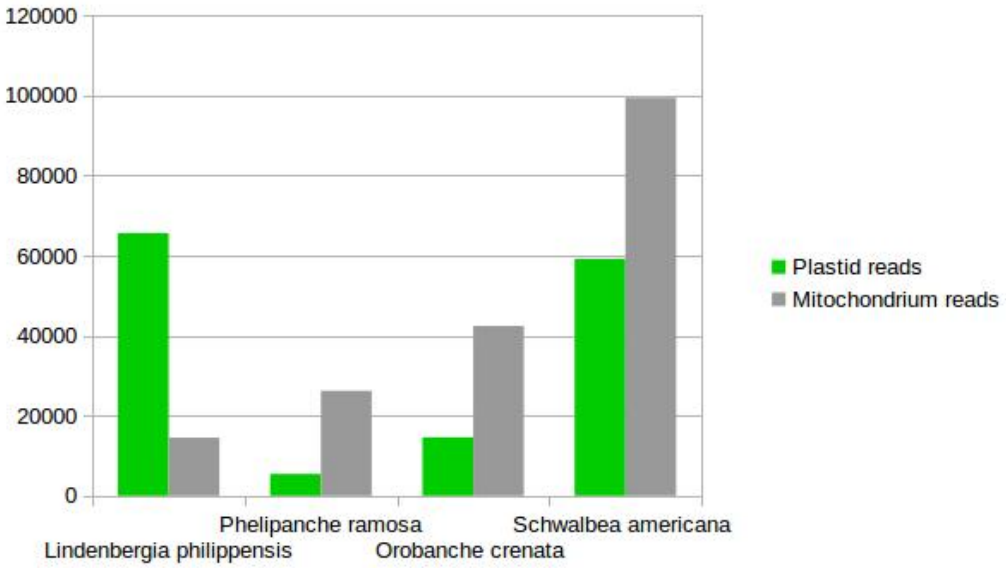


Figure 3.26: *Autochthonous read pair numbers for each species, so read pairs where both mates successfully mapped against the plastid chromosome or the mitochondrial scaffold respectively. The numbers are read pair counts normalized against the respective total input read pairs in millions.*

What is immediately apparent in figure 3.26 is that the ratio of plastid reads to mitochondrial reads strongly differs between *Lindenbergia philippensis* and the other three species. Now one can compare the ratio of strict NUPTs to the combined number of organelle transfers to the nucleus (strict NUPTs + strict NUMTs), calculated from table 3.13, with the ratio of plastid reads to total organelle reads (plastid reads + mitochondrion reads), calculated from table 3.14, in a chi-squared "goodness of fit" test. If the result is not statistically significant, this would mean that the transfer dynamics of the organelles of the respective species are equal between plastids and mitochondria, so both are equally stable. Then the number of transfer events for each organelle would only depend on the amount of organelles in the cell, the size of the organelle genome and the average genome copy number in each organelle. However, the expectation was that the parasitic species, *Phelipanche ramosa*, *Orobanche crenata* and *Schwalbea americana*, should show higher transfer dynamics and therefore instability of the plastid, due to reduced evolutionary pressure on the plastome, even when compared to the mitochondria. The non-parasitic species *Lindenbergia philippensis* on the other hand should show higher mitochondrial transfer dynamics, since in plants plastid genomes are usually much more stable than mitochondrial genomes.

The results of the chi-squared test for *Lindenbergia philippensis* were statistically significant (χ-squared = 183.4532, df = 1, $p - value < 2.2e - 16$), indicating a relatively higher transfer dynamic of the mitochondria, compared to the plastids. The same is true for *Schwalbea americana* (χ-squared = 1517.907, df = 1, $p - value < 2.2e - 16$), which also shows the higher mitochondrial transfer dynamics.
The holoparasitic species, *Phelipanche ramosa* (χ-squared = 172.63, df = 1, $p - value < 2.2e - 16$) and *Orobanche crenata* (χ-squared = 48.0127, df = 1, $p - value = 4.235e - 12$), also show statistically significant results, but in their case this means a higher transfer from the plastids, compared to the mitochondria.
The results are as expected, with the exception of *Schwalbea americana*. This means that the plastids of *Schwalbea americana* are still more stable than the mitochondria, in contrast to the two holoparasitic species. The results of *Phelipanche ramosa* and *Orobanche crenata* demonstrate the instability of their plastids. However, there are two uncertainties that must be considered. First, the mitochondrial scaffolds are unfinished and for *Phelipanche ramosa* the unpublished plastid chromosome sequence was used. Second, reads that showed sequence similarity between plastid and mitochondrion

were excluded, to look only at the transfers directly to the nucleus from both plastid and mitochondrion. This also excludes reads that transferred both from the plastid to the mitochondrion and to the nucleus. But because the proportion of such reads can't be estimated without an available nuclear genome draft, such reads had to be excluded for this comparison.

3.3.3 Extraction of nuclear reads

Extraction of autochthonous nuclear reads, that is, sequences not derived by transfers from the organelles, can be performed with the workflow described in the respective section of Materials and Methods. If the extraction is successful, we would expect much lower sequence duplication levels for the nuclear reads compared to organelle reads (read pairs where both mates successfully mapped to the plastid chromosome or the mitochondrial scaffold). Furthermore, we would also expect slightly lower sequence duplication levels of the nuclear reads compared to the preprocessed reads, which still contain the organelle reads and all the still detectable sequences transferred from the organelles to the nucleus (NUPTs and NUMTs). The different sequence duplication levels of the forward reads of *Phelipanche ramosa* can be seen in Figures 3.27 to 3.29.

As can be seen in Figures 3.27 to 3.29, the nuclear read sample of *Phelipanche ramosa* shows the lowest overall duplication levels (the highest percentage of sequences remaining after deduplication). The plastid read sample shows by far the highest duplication levels and the preprocessed sample only slightly higher ones than the nuclear reads, which was expected. The difference between the forward and reverse read samples was marginal for both *Phelipanche ramosa* and *Orobanche crenata*.
Since FastQC always uses the first 100 000 reads as sample for the duplication levels analysis, the approximate frequencies of reads remaining after deduplication can be calculated from the percentages, using the average between forward and reverse reads, to test if the difference in duplication levels between the preprocessed and the nuclear reads is statistically significant (see Tables 3.15 and 3.16).

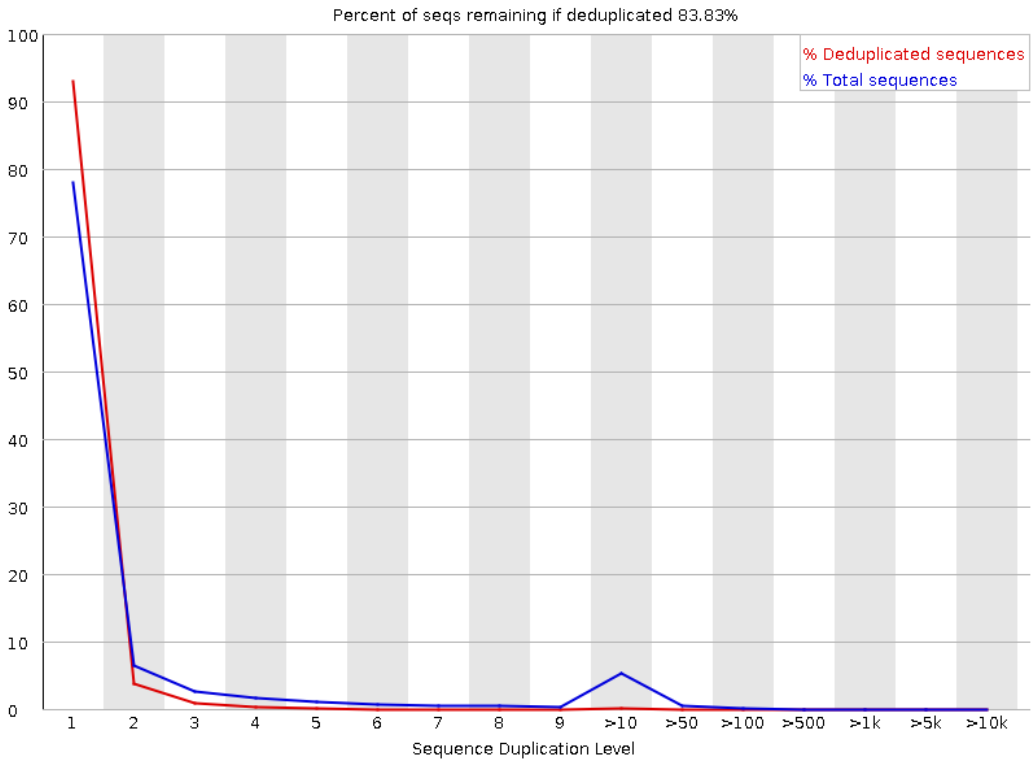


Figure 3.27: *Sequence duplication levels of the preprocessed reads of Phelipanche ramosa (forward reads). The blue line shows the percentage of all sequences that have a certain duplication level. The red line shows only the percentage of deduplicated sequences that had a certain duplication level. Above the figure the percentage of sequences remaining from the total set after deduplication is shown. The higher this percentage, the lower the overall duplication levels in the given set.*

Table 3.15: *Calculated frequencies of remaining deduplicated sequences and duplicates for both the preprocessed read sample and the nuclear read sample of Phelipanche ramosa.*

	Preprocessed set	**Nuclear set**
Deduplicated sequences	83885	84940
Duplicates	16115	15060

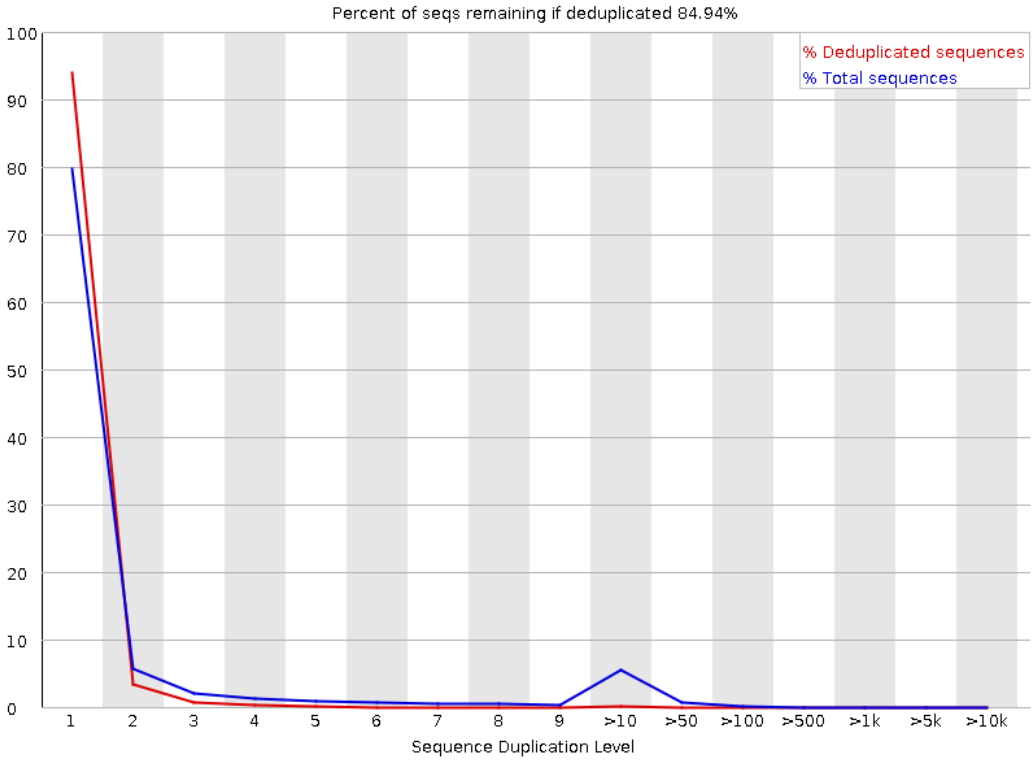


Figure 3.28: *Sequence duplication levels of the nuclear reads of Phelipanche ramosa (forward reads). The blue line shows the percentage of all sequences that have a certain duplication level. The red line shows only the percentage of deduplicated sequences that had a certain duplication level. Above the figure the percentage of sequences remaining from the total set after deduplication is shown. The higher this percentage, the lower the overall duplication levels in the given set.*

Table 3.16: *Calculated frequencies of remaining deduplicated sequences and duplicates for both the preprocessed read sample and the nuclear read sample of Orobanche crenata.*

	Preprocessed set	**Nuclear set**
Deduplicated sequences	75815	78820
Duplicates	24185	21180

A chi-squared test showed a statistically significant result for both *Phelipanche ramosa* (χ-squared = 42.215, df = 1, $p-value = 8.177e-11$) and for *Orobanche crenata* (χ-squared = 257.2771, df = 1, $p - value < 2.2e - 16$). These extracted nuclear reads are autochthonous, meaning that these are sequences that haven't been derived by inserted fragments from the organelles.

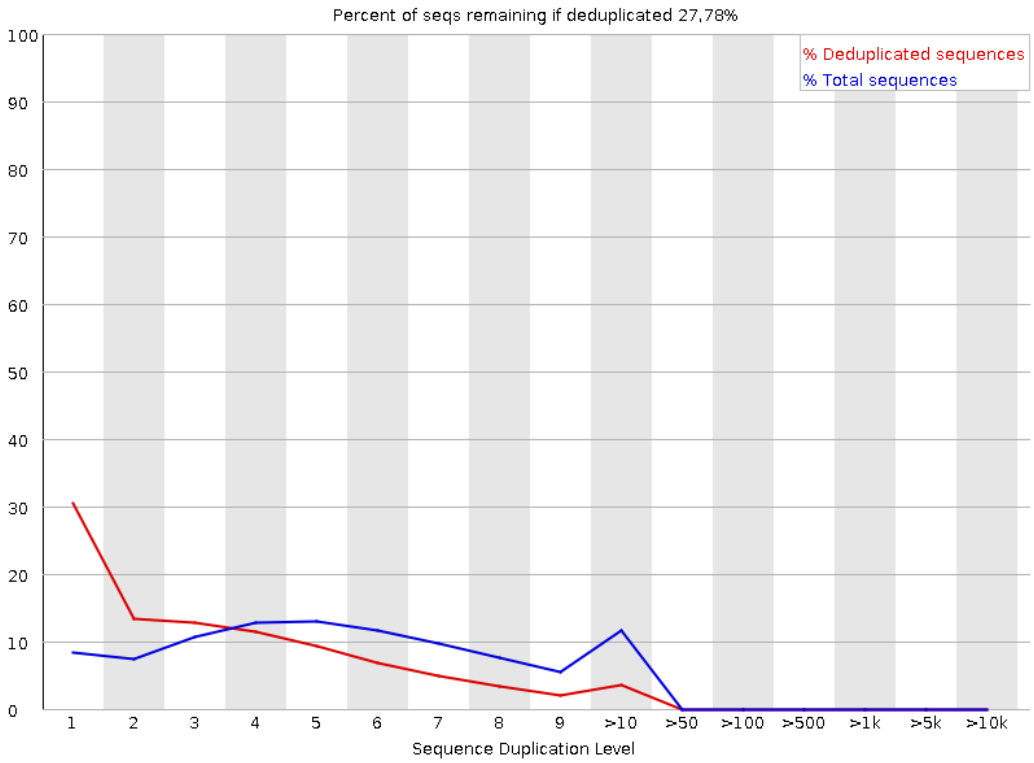


Figure 3.29: *Sequence duplication levels of the plastid reads of Phelipanche ramosa (forward reads). The blue line shows the percentage of all sequences that have a certain duplication level. The red line shows only the percentage of deduplicated sequences that had a certain duplication level. Above the figure the percentage of sequences remaining from the total set after deduplication is shown. The higher this percentage, the lower the overall duplication levels in the given set.*

Of course only relatively recent insertions can be detected by sequence similarity searches, so older insertions cannot be excluded with this method. Still, random subsampling from the NUPT, NUMT and nuclear read sets would allow a comparative repeat analysis between the organelle transfer sequences and the nuclear sequences, using a program like RepeatExplorer (Novák et al., 2013). However, such an approach goes beyond the scope of this project.

3.4 Conclusions and Outlook

The study of promiscuous DNA in plants without an available nuclear genome draft, using only whole genome next-generation data and organellar chromosomes, shows potential, but is also subject to technical limitations. So far, it is possible to extract different kinds of promiscuous DNA and to observe biological patterns. This includes the increase in the relative amount of MIPTs, with increasing size of the mitochondrial genome, or the larger amount of "lost NUPTs/MIPTs" for *Phelipanche ramosa*, with its more reduced plastome, compared to *Orobanche crenata*. The higher transfer dynamics to the nucleus of the plastids of the two holoparasitic species, compared to their mitochondria, could also be shown, as well as the reproducibility of the initial workflow, that formed the foundation for all subsequent analyses.
As long as the reliable reconstruction of complete NUPT fragments without a nuclear genome draft is not possible, however, more detailed studies of promiscuous DNA are restriced to those species with sequenced nuclear genomes. While the assembly of reads with different evolutionary distance to the plastid chromosome showed the difference in fragmentation between older and younger fragments, the results cannot be verified at this point. Further analysis of the reconstructed fragments would in turn only lead to unreliable results.

The development of a reconstruction method would demand a different experimental setup, using next-generation sequencing data of a species with an available nuclear genome draft, to evaluate the accuracy of the resulting fragments. Only then would also a meaningful comparison between different assembly software and settings be possible. The assembly approaches demonstrated in this thesis could serve as a foundation for a more advanced method, with the optimal solution probably being a haplotype-phasing algorithm specifically optimised for the multi-allelic situation presented by promiscuous DNA in plants. The establishment of such a method would then allow the examination and characterization of complete NUPT fragments for every plant species with sequenced organelle chromosomes.

Bibliography

Allen, J.F. (1993a). "Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes". *J. Theor. Biol.* 165: 609–631.

Allen, J.F. (1993b). "Redox control of gene expression and the function of chloroplast genomes: a hypothesis". *Photosynth. Res.* 36: 95–102.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). "Basic local alignment search tool". *J. Mol. Biol.* 215: 403–410.

Andrews, S. (2015). *FastQC: a quality control tool for high throughput sequence data.* Available online at: URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc` (visited on 12/11/2015).

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". *J. Comput. Biol.* 19: 455–477.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). "Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events". *Nature* 422: 433–438.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., and Parkhill, J. (2009). "DNAPlotter: circular and linear interactive genome visualization". *Bioinformatics* 25: 119–120.

Darling, A.E., Mau, B., and Perna, N.T. (2010). "progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement". *PLoS ONE* 5: e11147. DOI: `10.1371/journal.pone.0011147`.

Deutsch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K.V., Allen, J.F., Martin, W., and Dagan, T. (2008). "Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor". *Mol. Biol. Evol.* 25: 748–761.

Ellis, J. (1982). "Promiscuous DNA-chloroplast genes inside plant mitochondria". *Nature* 299: 678–679.

Gordon, A. (2015). *FASTX-toolkit*. Available online at: URL: http://hannonlab.cshl.edu/fastx_toolkit (visited on 07/18/2015).

Huang, C.Y., Ayliffe, M.A., and Timmis, J.N. (2003). "Direct measurement of the transfer rate of chloroplast DNA into the nucleus". *Nature* 422: 72–76.

Huang, C.Y., Grünheit, N., Ahmadinejad, N., Timmis, J.N., and Martin, W. (2005). "Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes". *Plant Physiol.* 138: 1723–1733.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. (2004). "Pack-MULE transposable elements mediate gene evolution in plants". *Nature* 431: 569–573.

Kanevski, I. and Maliga, P. (1994). "Relocation of the plastid rbcL gene to the nucleus yields functional ribulose-1,5-bisphosphate carboxylase in tobacco chloroplasts". *Proc. Natl. Acad. Sci. USA* 91: 1969–1973.

Kelley, D.R. and Salzberg, S.L. (2010). "Detection and correction of false segmental duplications caused by genome mis-assembly". *Genome Biology* 11: R28. DOI: 10.1186/gb-2010-11-3-r28.

Kleine, T., Maier, U.G., and Leister, D. (2009). "DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis". *Annu. Rev. Plant Biol.* 60: 115–138. DOI: 10.1146/annurev.arplant.043008.092119.

Kurtz, S. (2015). *The Vmatch large scale sequence analysis software.* Available online at: URL: http://www.vmatch.de (visited on 01/20/2015).

Lee, S.I. and Kim, N.S. (2014). "Transposable elements and genome size variations in plants". *Genomics Inform.* 12: 87–97. DOI: 10.5808/GI.2014.12.3.87.

Li, H. and 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., and 1000 Genome Project Data Processing Subgroup (2009). "The Sequence alignment/map (SAM) format and SAMtools". *Bioinformatics* 25: 2078–2079.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., and Guan, X. (2012). "CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences". *BMC Genomics* 13: 715. DOI: 10.1186/1471-2164-13-715.

Lloyd, A.H. and Timmis, J.N. (2011). "The Origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome". *Mol. Biol. Evol.* 28: 2019–2028.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). "Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus". *Proc. Natl. Acad. Sci. USA* 99: 12246–12251.

Matsuo, M., Ito, Y., Yamauchi, R., and Obokata, J. (2005). "The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux". *Plant Cell* 17: 665–675.

McNeal, J.R., Bennett, J.R., Wolfe, A.D., and Mathews, S. (2013). "Phylogeny and origins of holoparasitism in Orobanchaceae". *Am. J. Bot.* 100: 971–983.

McNeal, J.R., Leebens-Mack, J.H., Arumuganathan, K., Kuehl, J.V., Boore, J.L., and dePhamphilis, C.W. (2006). "Using partial genomic fosmid libraries for sequencing complete organellar genomes". *Biotechniques* 41: 69–73.

Michalovova, M., Vyskot, B., and Kejnovsky, E. (2013). "Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization". *Heredity* 111: 314–320.

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., and Leister, D. (2007). "Nuclear insertions of organellar DNA can create novel patches of functional exon sequences". *Trends Genet.* 23: 597–601.

Novák, P., Neumann, P., Pech, J., and Steinhaisl J. Macas, J. (2013). "RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads". *Bioinformatics* 29: 792–793. DOI: 10.1093/bioinformatics/btt054.

Park, S., Grewe, F., Zhu, A., Ruhlman, T.A., Sabir, J., Mower, J.P., and Jansen, R.K. (2015). "Dynamic evolution of Geranium mitochondrial genomes through multiple horizontal and intracellular gene transfers". *New Phytol.* 208: 570–583.

Richly, E. and Leister, D. (2004). "NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs". *Mol. Biol. Evol.* 21: 1972–1980.

Rizzon, C., Ponger, L., and Gaut, B.S. (2006). "Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice". *PloS Comput. Biol.* 2: e115. DOI: 10.1371/journal.pcbi.0020115.

Rodríguez-Moreno, L., Gonzalez, V.M., Benjak, A., Martí, M.C., Puigdomènech, P., Aranda, M.A., and Garcia-Mas, J. (2011). "Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a sig-

nificant amount of DNA having a nuclear origin". *BMC Genomics* 12: 424.

Samaniego Castruita, J.A., Zepeda Mendoza, M.L., Barnett, R., Wales, N., and Gilbert, M.T.P. (2015). "Odintifier – A computational method for identifying insertions of organellar origin from modern and ancient high throughput sequencing data based on haplotype phasing". *BMC Bioinformatics* 16: 232. DOI: 10.1186/s12859-015-0682-1.

Schmieder, R. and Edwards, R. (2011). "Quality control and preprocessing of metagenomic datasets". *Bioinformatics* 27: 863–864. DOI: 10.1093/bioinformatics/btr026.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., et al. (2009). "The B73 maize genome: complexity, diversity and dynamics". *Science* 326: 1112–1115.

Shahmuradov, I.A., Akbarova, Y.Y., Solovyev, V.V., and Aliyev, J.A. (2003). "Abundance of plastid DNA insertions in nuclear genomes of rice and Arabidopsis". *Plant Mol. Biol.* 52: 923–934.

Sheppard, A.E. and Timmis, J.N. (2009). "Instability of Plastid DNA in the Nuclear Genome". *PLoS Genet.* 5: e1000323. DOI: 10.1371/journal.pgen.1000323.

Smith, D.R., Crosby, K., and Lee, R.W. (2011). "Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis". *Genome Biol. Evol.* 3: 365–371. DOI: 10.1093/gbe/evr001.

Stegemann, S. and Bock, R. (2006). "Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus". *Plant Cell* 18: 2869–2878.

Stegemann, S., Hartmann, S., Ruf, S., and Bock, R. (2003). "High-frequency gene transfer from the chloroplast genome to the nucleus". *Proc. Natl. Acad. Sci. USA* 100: 8828–8833.

Stern, D.B. and Lonsdale, D.M. (1982). "Mitochondrial and chloroplast genomes of maize have a 12-kb DNA sequence in common". *Nature* 299: 698–702.

Stern, D.B. and Palmer, J.D. (1984). "Extensive and widespread homologies between mitochondrial DNA and chloroplast DNA in plants". *Proc. Natl. Acad. Sci. USA* 81: 1946–1950.

Straub, S.C., Cronn, R.C., Edwards, C., Fishbein, M., and Liston, A. (2013). "Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (apocynaceae)". *Genome Biol. Evol.* 5: 1872–1885. DOI: 10.1093/gbe/evt140.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). "Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes". *Nat. Rev. Genet.* 5: 123–135.

Treangen, T.J. and Salzberg, S.L. (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions". *Nat. Rev. Genet.* 13: 36–46. DOI: 10.1038/nrg3117.

Tsunewaki, K. (2011). "Interorganellar DNA transfer in wheat: dynamics and phylogenetic origin". *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* 87: 529–549.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). "Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary". *Genome Res.* 24: 1334–1347. DOI: 10.1101/gr.168997.113.

Von Heijne, G. (1986). "Mitochondrial targeting sequences may form amphiphilic helices". *EMBO J.* 5: 1335–1342.

Wang, D. and Timmis, J.N. (2013). "Cytoplasmic organelle DNA preferentially inserts into open chromatin". *Genome Biol. Evol.* 5: 1060–1064.

Wang, D., Wu, Y.W., Shih, A.C., Wu, C.S., Wang, Y.N., and Chaw, S.M. (2007). "Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA". *Mol. Biol. Evol.* 24: 2040–2048.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., et al. (2006). "High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes". *Plant Cell* 18: 1791–1802.

Wicke, S., Müller, K.F., dePamphilis, C.W., Quandt, D., Wickett, N.J., Zhang, Y, Renner, S.S., and Schneeweiss, G.M. (2013). "Mechanisms of functional and physical genome reduction in photosynthetic and non-photosynthetic parasitic plants of the broomrape family". *Plant Cell* 25: 3711–3725.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., and Quandt, D. (2011). "The evolution of the plastid chromosome in land plants: gene content, gene order, gene function". *Plant Mol. Biol.* 76: 273–297.

Abstract

The development of plastids was accompanied by a significant reduction of the genome of the cyanobacteria-like ancestor. A large part of the genome of this ancestor has been transferred to the nucleus early on in plastid evolution. Transfers from the plastid have not stopped, however, and appear to play an important role in the evolution of plants, providing functionally transferred organelle genes or novel exons for nuclear genes. Fragments in the nuclear genome that originated from the plastid have been termed NUPTs, while fragments in the mitochondrial genome have been termed MIPTs. NUPTs have been studied so far by performing sequence similarity searches between the plastid chromosome and the nuclear genome of the respective plant. Nuclear genome drafts are only available for relatively few species, however, because the assembly of a plant genome is highly challenging. In this thesis, next-generation sequencing data of three parasitic and one autotrophic species of the plant family Orobanchaceae is used to demonstrate the robustness and reproducibility of a workflow for the detection and extraction of NUPT and MIPT sequences without an available nuclear genome sequence. Then, based on the results of this initial workflow, a novel approach for the reconstruction of NUPTs/MIPTs is introduced, showing the increasing fragmentation of NUPTs/MIPTs over the course of evolution. In addition, a bioinformatic workflow for the detection of sequences that have been transferred to the nucleus and lost from the reduced plastomes of the two holoparasitic species *Phelipanche ramosa* and *Orobanche crenata* is demonstrated. The difference in the amount of detected lost NUPTs/MIPTs between the two holoparasites roughly corresponds to the size difference of their reduced plastid genomes. In a second step, this workflow separates NUPTs and MIPTs. The relative amount of MIPTs among all transfers from the plastid appears to increase with the size of the mitochondrial genome of the respective species. Another workflow allows the extraction of NUMTs, transfers from the mitochondrion to the nucleus, and subsequent comparison of the transfer dynamics of the plastids and mitochondria of each species, showing the increased instability of the plastids for the holoparasitic species. In a final step, the autochthonous nuclear reads are extracted and verified by detection of lowered sequence duplication levels. The source code for the workflows is provided in a supplement folder.

Abstract

Die Entwicklung der Plastiden war von einer signifikanten Verkleinerung des Genoms des Cyanobakterien-ähnlichen Vorfahren begleitet. Ein großer Teil des Genoms dieses Vorfahrens war dabei bereits in einer frühen Phase der Plastidenevolution in den Nukleus transferiert worden. Transfers von den Plastiden geschehen jedoch nach wie vor und scheinen eine wichtige Rolle in der Pflanzenevolution zu spielen, durch die Übertragung funktionaler Gene von den Organellen oder durch die Erweiterung von Kern-Genen um neue Exons. Fragmente, die von den Plastiden ins Kerngenom übertragen wurden, werden als NUPTs bezeichnet, Fragmente die in die Mitochondrien gewandert sind, als MIPTs. NUPTs wurden bisher mit Hilfe von Sequenzähnlichkeit zwischen Plastidgenom und Kerngenom untersucht. Sequenzierte Kerngenome sind jedoch nur für relativ wenige Pflanzenspezies verfügbar, da die Assemblierung von Pflanzengenomen äußerst herausfordernd ist. In dieser Arbeit werden Next-Generation Sequencing Daten von drei parasitären und einer autotrophen Spezies der Pflanzenfamilie Orobanchaceae benutzt, um die Reproduzierbarkeit eines Workflows zu zeigen, der die Extraktion von NUPT- und MIPT-Sequenzen ohne verfügbares Kerngenom ermöglicht. Basierend auf diesem Workflow wird ein Ansatz für die Rekonstruktion von NUPTs/MIPTs eingeführt, dessen Ergebnisse die Fragmentierung von NUPTs/MIPTs im Laufe der Evolution zeigen. Ein weiterer Workflow, für die Detektion von Sequenzen die in den Nukleus transferiert und daraufhin aus den Plastomen der holoparasitären Spezies *Phelipanche ramosa* und *Orobanche crenata* verschwunden sind, wird demonstriert. Der Unterschied in der Menge der verlorenen NUPTs/MIPTs zwischen den beiden Holoparasiten entspricht annähernd dem Größenunterschied der reduzierten Plastidsequenzen. In einem zweiten Schritt trennt dieser Workflow NUPTs und MIPTs, wobei der relative Anteil an MIPTs mit der Größe des mitochondrialen Genoms zu steigen scheint. Überdies wird durch einen dritten Workflow die Extraktion von NUMTs, Transfers vom Mitochondrium zum Kerngenom, ermöglicht. Ein darauffolgender Vergleich der Transferdynamik der Plastiden und der Mitochondrien zeigt die erhöhte Instabilität der Plastiden der beiden holoparasitären Spezies. In einem letzten Schritt werden autochthone Sequenzen des Kerngenoms extrahiert und über die Detektion eines niedrigereren Grads an Sequenzduplikationen verifiziert. Der Quellcode für die Workflows steht in einem Supplement Ordner zur Verfügung.